FEMA-Long: Modeling unstructured covariances for discovery of time-dependent effects in large-scale longitudinal datasets

List of authors:

Pravesh Parekh[1,2,*], Nadine Parker[1], Diliana Pecheva[2], Evgeniia Frei[1], Marc Vaudel[4,5,6], Diana M. Smith[7,8], Alison Rigby[9], Piotr Jahołkowski[1], Ida Elken Sønderby[1,10], Viktoria Birkenæs[1,11], Nora Refsum Bakken[1,11], Chun Chieh Fan[12], Carolina Makowski[13], Jakub Kopal[1], Robert Loughnan[2,14], Donald J. Hagler Jr.[2], Dennis van der Meer[1], Stefan Johansson[4,15], Pål Rasmus Njølstad[4,15], Terry L. Jernigan[3,7,13,16], Wesley K. Thompson[12], Oleksandr Frei[1,17], Alexey A. Shadrin[1], Thomas E. Nichols[18,19], Ole A. Andreassen[1], Anders M. Dale[2,20,21*]

Affiliations:

[1]Centre for Precision Psychiatry, Division of Mental Health and Addiction, University of Oslo and Oslo University Hospital, Oslo, Norway.

[2]Center for Multimodal Imaging and Genetics, J. Craig Venter Institute, La Jolla, CA, USA.

[3]Department of Radiology, School of Medicine, University of California San Diego, La Jolla, CA, USA.

[4]Mohn Center for Diabetes Precision Medicine, Department of Clinical Science, University of Bergen, Bergen, Norway.

[5]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway.

[6]Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Bergen, Norway.

[7]Center for Human Development, University of California San Diego, La Jolla, CA, USA.

[8]Medical Scientist Training Program, University of California San Diego, La Jolla, CA, USA.

[9]Neuroscience Graduate Program, University of California San Diego, La Jolla, CA, USA.

[10]Department of Medical Genetics, Oslo University Hospital & University of Oslo, Oslo, Norway.

[11]PsychGen Center for Genetic Epidemiology and Mental Health, Norwegian Institute of Public Health, Oslo, Norway.

[12]Center for Population Neuroscience and Genetics, Laureate Institute for Brain Research, Tulsa, OK, USA.

[13]Department of Psychiatry, School of Medicine, University of California San Diego, La Jolla, CA, USA.

[14]Population Neuroscience and Genetics Lab, University of California San Diego, La Jolla, CA, USA.

[15]Department of Pediatrics and Adolescents, Haukeland University Hospital, Bergen, Norway.

[16]Department of Cognitive Science, University of California San Diego, La Jolla, CA, USA.

[17]Department of Pharmacy, Section for Pharmacology and Pharmaceutical Biosciences, University of Oslo, Oslo, Norway.

[18]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK.

[19]Oxford Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK.

[20]University of California San Diego, La Jolla, CA, USA.

[21]University of Oslo, Oslo, Norway.

*Correspondence should be addressed to:

Dr. Pravesh Parekh: praveshparekh@gmail.com (PP)

Prof. Anders M. Dale: andersmdale@gmail.com (AMD)

Abstract

While linear mixed-effects (LME) models are common for analyzing longitudinal data, most users rely on random intercepts or simple stationary covariance, due to unavailability of computationally tractable solutions. Here, we extend the Fast and Efficient Mixed-Effects Algorithm (FEMA) and present FEMA-Long, a computationally tractable approach to flexibly modeling longitudinal covariance suitable for high-dimensional data. FEMA-Long can: i) model unstructured covariance, ii) model covariates as smooth functions using splines, iii) discover time-dependent effects of covariates with spline interactions, and iv) use these flexible longitudinal modeling strategies to perform longitudinal genome-wide association studies and discover time-dependent genetic effects, in a computationally scalable manner, suitable for high-dimensional data. Through extensive simulations, we show that estimates from FEMA-Long are accurate, while being up to several thousand times faster and with minimal carbon footprint. To show the utility of FEMA-Long for discovering novel biological signal, using data from the Norwegian Mother, Father and Child Cohort Study (MoBa), we performed a longitudinal genome-wide association study with non-linear SNP-by-time interaction on length, weight, and BMI of 68,273 infants with up to six measurements in the first year of life. We found dynamic patterns of random effects including time-varying heritability and genetic correlations, as well as several genetic variants showing time-dependent effects, highlighting the applicability of FEMA-Long to enable novel discoveries. FEMA-Long is available at: https://github.com/cmig-research-group/cmig_tools.

Keywords: mixed-effects modeling; GWAS; longitudinal analysis; green algorithm; MoBa; MBRN

Author summary

Most large-scale datasets have complexities such as repeated measures, related individuals, or other dependencies across samples, preventing the use of standard regression approaches for analysis. In such circumstances, linear mixed-effects modeling is often employed. However, for high-dimensional datasets, fitting these models is quite challenging. Further, most standard uses of linear mixed-effects modeling focus on simpler covariance models, which may not hold. Here, we introduce FEMA-Long, a novel computationally efficient analytical framework for fitting linear mixed-effects models with time-varying random effects, as well as allowing the effect of the covariates to change smoothly over time by using splines. This is particularly relevant when, for example, studying the effect of genetic variants on phenotypes, where the effects could be non-linear over time. The FEMA-Long framework allows time-varying heritability as well as discovery of genetic variants that show time-dependent effects. By performing a genome-wide association study on data from the Norwegian Mother, Father and Child Cohort Study (MoBa) using FEMA-Long, we show the discovery of genetic variants with time-dependent effects on infant length, weight, and BMI during the first year of life. Our results highlight the potential of using

FEMA-Long to make novel discoveries that can lead to biological insights on the genetics of complex traits as well as improve the potential of using genetics for personalized prediction.

Introduction

Large-scale longitudinal datasets are critical for novel discoveries in health and medicine and can transform healthcare (1). Longitudinal data allows studying the dynamic processes underlying health and disease development/progression by modeling trajectories, examining time-dependent effects of variables, investigating association with health indicators and outcomes, and making predictions of outcomes. Examples of large-scale densely-sampled datasets include the UK Biobank (https://www.ukbiobank.ac.uk), All of Us (https://allofus.nih.gov), the Adolescent Brain Cognitive Development℠ Study (https://abcdstudy.org), and many others (2).

To leverage the full potential of these studies, we need computationally efficient modeling strategies, optimized for large-scale high-dimensional datasets. These strategies need to account for non-independence of observations (e.g., repeated measurements and relationships between individuals), other sources of variance (e.g., shared environment), covariance in the data (e.g., correlations across time), flexible modeling of longitudinal trajectories, and enabling discovery of time-dependent effects. Linear mixed-effects (LME) models are a powerful class of analytical strategies that can account for these issues in longitudinal data such as non-independence of the samples and other known sources of variance modeled as random effects, in addition to including variables with non-linear effects on the outcome and non-linear interactions. Examples of these random effects include the similarity of measurements for the same subject, variance explained by genetics (heritability), variance attributable to shared environment, etc.

When using LME models on longitudinal datasets, random effects should be specified as time-varying. For example, heritability can vary significantly over time (3–8). Similarly, correlations of measurements within subjects can change over time (e.g., autocorrelation). However, time-varying random effects are seldom modeled, given the significant computational burden of fitting such models. Thus, researchers use simplified LME models like compound symmetry (constant covariance over time) or making assumptions about the form of covariance (e.g., autoregressive models) which may not fit the data well. In addition, there is a lack of LME modeling software that scales to the needs of high-dimensional data (9) such as when estimating the marginal effects of several million single nucleotide polymorphisms (SNPs) in genome-wide association studies (GWAS), or when fitting mass-univariate models to millions of outcome variables, as is common in neuroimaging. This has severely limited our ability to maximize the use of high-dimensional longitudinal datasets.

Previously, we developed the Fast and Efficient Mixed-Effects Algorithm (FEMA) (10) for fitting a large number of LME models in a matter of seconds to minutes. FEMA massively accelerates LME fitting by using a covariance-regression approach to estimate random effects parameters and then quantizes these for vectorized fitting of many phenotypes simultaneously. For example, FEMA can fit ~169,000 LME models with ~10,000 observations in ~54 seconds, a speedup of more than 3000 times, compared to standard LME

solvers. In the present work, we introduce FEMA-Long, a framework adapted and optimized for high-dimensional longitudinal datasets. The features of FEMA-Long include i) modeling time-varying random effects using unstructured covariance, ii) modeling non-linear effects using splines, iii) examining time-dependent effects using spline interactions, and iv) performing GWAS, supporting the discovery of time-dependent effects. Through extensive simulations we show that FEMA-Long accurately models known sources of variance in a computationally fast manner while maintaining false positive rates. We demonstrate the use of FEMA-Long by performing a longitudinal GWAS on length, weight, and body mass index (BMI) of 68,273 infants with up to six repeated measurements (299,447 observations) during the first year of life, a highly dynamic period of development, and discovering SNPs showing time-dependent effects.

Description of the method

Nomenclature

We use the term *phenotype* to refer to a continuous outcome variable (or dependent variable) with the error in the phenotype approximately following a Gaussian distribution; we use the term covariates to refer to independent variables, inclusive of covariates of interest and covariates of no interest; and the term *random effects* refers to various known sources of nesting and similarity between observations in the data. For longitudinal data, the term *subject* or *participant* refers to a unique individual followed over time while the term *observation* refers to one measurement for a given subject. *Visit* or *timepoint* corresponds to an ordered set of repeated measurements for the same subject; for example, *visit 1* means the first measurement (first observation) on the subject, with increasing visit numbers indicating subsequent measurements. For mathematical notation, we use boldface for matrices or vectors, and regular font weight for scalar quantities.

FEMA-Long algorithm overview

Given a set of continuous phenotype(s), covariates, and random effects, the goal is to estimate the fixed effects (beta coefficients for the covariates) while accounting for random effects. In the original FEMA framework (10), we first calculate the ordinary least squares (OLS) residuals of the phenotype; then, we use them to estimate the variance components of the random effects using method of moments (MoM) or a regression estimator with a non-negativity constraint (11). Then, the fixed effects are estimated using a generalized least squares (GLS) solution which incorporates a covariance matrix using the estimated variance components of the random effects. For scalability to large number of variables (such as for voxel-wise or vertex-wise data derived from brain imaging data), we use a binning approach – for all outcome variables that have similar variance components for the random effects, the GLS estimation is performed using a common covariance matrix that relies on the average of the variance components of all the outcome variables within that bin.

Here, we present FEMA-Long, an extension of FEMA that is adapted and optimized for large longitudinal datasets. The FEMA-Long framework includes i) time-varying covariances in the random effects using unstructured covariances; ii) modeling non-linear effects using splines; iii) examining time-dependent effects using spline interactions; and iv) performing GWAS and other similar omics analyses, where we test the association of a large number of variables of interest (e.g., genotyping matrix with millions of SNPs) with outcome(s), independently (i.e., marginal rather than joint effect), after controlling for covariates and random effects.

In FEMA-Long, we use unstructured covariance to model time-varying covariance of random effects. The elements of the unstructured covariance matrix are estimated by using the MoM estimation for every pair of visits that are nested within subjects. This results in a matrix of covariance terms for each random effect, which is then used for GLS estimation of the fixed effects. Importantly, this method does not assume a balanced design of all subjects having data for all the visits.

FEMA-Long also includes the option to model continuous outcome variables as a smooth function (for example, smooth function of age) using splines, represented by a set of basis functions. We also allow interaction terms between covariates, including those covariates that are modeled via splines. While the estimation of the unstructured covariance matrix uses pre-defined set of timepoints (visits), the spline modeling uses a continuous variable (such as age). This, therefore, brings the framework of generalized additive mixed models to FEMA, allowing flexible modeling of outcome variables as well as performing non-linear interaction analyses.

Finally, while the original FEMA was designed for many outcome variables and a common model (mass-univariate analyses), FEMA-Long includes a variant of FEMA that is intended for GWAS, where we fit the outcome(s) repeatedly on many models, each differing only in their genetic covariate. For this, we leverage the Frisch–Waugh–Lovell theorem (12,13): first, use stage-1 FEMA to estimate the random effects variance components and fixed effects (omitting genetic variants); then, we GLS residualize genotype vector for the covariates using the stage-1 variance components. Finally, we perform a GLS estimation using the GLS residuals of the phenotype and the GLS residuals of the genotype vector. This framework also allows modeling the interaction of genetic variants with spline basis functions, thereby allowing discovery of genetic variants that may show a nonlinear, time-dependent effect on the phenotype.

Further details on each of these steps is described below.

General FEMA model set up (stage-1)

Let a dataset have n observations indexed with $i \in 1 \dots n$. Let $\boldsymbol{X}$ be a full-rank matrix of p covariates , with each column of $\boldsymbol{X}$ being a different covariate. $\boldsymbol{X}_i$ are the values of the covariates for observation i (X_i is a scalar if $p = 1$ and a vector if $p > 1$). The phenotype is indicated by $\boldsymbol{y}$, thus y_i indicates the value of the

phenotype for observation i. For brevity, we present the framework for a single phenotype (i.e., $\boldsymbol{y}$ is a vector), although FEMA fits the same model for each phenotype.

The linear model setup is:

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of p weights associated with each covariate, and ε_i is the unmodeled measurement error for each observation i. In a mixed modeling framework, we assume that the error term, ε, is normally distributed with zero mean and variance-covariance $\boldsymbol{V}$:

$$\varepsilon \sim \mathcal{N}(0, \boldsymbol{V}) \tag{2}$$

Then, the total residual variance is $\sigma^2 = \frac{\mathrm{tr}(\boldsymbol{V})}{n}$. This normalized residual variance can be decomposed as a combination of random effects:

$$\frac{\boldsymbol{V}}{\sigma^2} = \sigma_F^2\boldsymbol{Z}_F\boldsymbol{Z}_F^\top + \sigma_A^2\boldsymbol{Z}_A\boldsymbol{Z}_A^\top + \sigma_S^2\boldsymbol{Z}_S\boldsymbol{Z}_S^\top + \sigma_\epsilon^2\boldsymbol{I} \tag{3}$$

where, generically, $\boldsymbol{Z}$ is a random effects indicator matrix with entries of Z_i indicating membership of observation i; for example, in equation (3), $\boldsymbol{Z}_F$ represents family with a value of 1 indicating observations that belong to the same family, entries of $\boldsymbol{Z}_A$ represent the genetic relationship between a pair of observations, $\boldsymbol{Z}_S$ represents subject effect with a value of 1 indicating repeated measurements for the same subject, and $\boldsymbol{I}$ represents the independent unmodeled variance across each observation; each random effect is associated with a corresponding variance component σ^2. The choice of which random effects should be included depends on the data.

Within FEMA, we first use OLS to get an estimate of the fixed effects:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \tag{4}$$

Then, the OLS residuals are:

$$\hat{y}_i^{\mathrm{res,OLS}} = y_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} \tag{5}$$

and the total residual variance $\hat{\sigma}^2$ estimated as:

$$\hat{\sigma}^2 = \frac{1}{(n-p)}\sum_{i=1}^{n}\left(\hat{y}_i^{\mathrm{res,OLS}}\right)^2 \tag{6}$$

Following (3), for each pair of observations $\{i, i'\}$, the expected value of the product of corresponding residuals $\hat{y}_i^{\mathrm{res,OLS}}$ and $\hat{y}_{i'}^{\mathrm{res,OLS}}$ is:

$$E\,\frac{\left(\hat{y}_i^{\mathrm{res,OLS}} \times \hat{y}_{i'}^{\mathrm{res,OLS}}\right)}{\hat{\sigma}^2} \approx \sigma_F^2\left((Z_FZ_F^\top)_{\{i,i'\}}\right) + \sigma_A^2\left((Z_AZ_A^\top)_{\{i,i'\}}\right) + \sigma_S^2\left((Z_SZ_S^\top)_{\{i,i'\}}\right) + \sigma_\epsilon^2 \tag{7}$$

where the E operator is the expectation with respect to the random effects and generically $(ZZ^\top)_{\{i,i'\}}$ denotes the $\{i, i'\}$ element of the random effects design matrix. Solving (7) results in the estimated values of random effects variance components, which can be used to construct a covariance matrix $\widehat{\boldsymbol{V}}$:

$$\widehat{\boldsymbol{V}} = \hat{\sigma}_F^2 \boldsymbol{Z}_F \boldsymbol{Z}_F^\top + \hat{\sigma}_A^2 \boldsymbol{Z}_A \boldsymbol{Z}_A^\top + \hat{\sigma}_S^2 \boldsymbol{Z}_S \boldsymbol{Z}_S^\top + \hat{\sigma}_\varepsilon^2 \boldsymbol{I} \tag{8}$$

Finally, the stage-1 fixed effects can be estimated using the GLS solution:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} y \tag{9}$$

with variance

$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X}\right)^{-1} \tag{10}$$

More details of these steps can be found in (10). We refer to this overall procedure as stage-1, which includes estimation of the random effects parameters as well as GLS estimation of the fixed effects coefficients.

Estimating unstructured covariance

The original framework in FEMA assumed compound symmetry (i.e., constant covariance over time; sometimes also referred as diagonal covariance patterns since only the variance components are being modeled), an assumption which may not hold for longitudinal studies. In FEMA-Long, we allow modeling arbitrary unstructured covariance, where the parameters of the unstructured covariance matrix are estimated using the available data. Specifically, using the OLS residuals from equation (5), we solve equation (11) for every pair of visits (v, v'):

$$E\,\frac{\left(\hat{y}_i^{\mathrm{res,OLS}} \times \hat{y}_{i'}^{\mathrm{res,OLS}}\right)_{vv'}}{\hat{\sigma}^2} \approx \Omega_{F,vv'}\left(\left(Z_F Z_F^\top\right)_{\{i,i'\}}\right)_{vv'} + \Omega_{A,vv'}\left(\left(Z_A Z_A^\top\right)_{\{i,i'\}}\right)_{vv'} + \Omega_{S,vv'}\left(\left(Z_S Z_S^\top\right)_{\{i,i'\}}\right)_{vv'} \tag{11}$$

When solving (11), if $v = v'$, we use non-negativity constrained solution to estimate the variance component for that visit number v; if $v \neq v'$, we drop the non-negativity constraint and estimate the covariance of the random effects between pairs of visits v and v'. Then, the estimated variance-covariance components can be put together to get $v \times v$ variance-covariance matrix for each random effect. Note that when estimating the variance components (i.e., the diagonal entries of the variance-covariance matrices), the subject and error terms are colinear, and therefore these variance parameters ($\widehat{\Omega}_{S,vv}^2$ and $\widehat{\Omega}_{\epsilon,vv}^2$) cannot be uniquely identified, and thus are estimated as one. Similarly, when estimating the covariance components (i.e., the off-diagonal entries of the variance-covariance matrices), the system of equations constructed from 11 makes the error term unidentifiable (see **Supplementary Figures S1** and **S2** for a demonstration of why these terms cannot be identified). The implication of this is that the estimated subject covariance matrix should be interpreted as describing the longitudinal stability of the phenotype within individuals, rather than as separate subject and noise variance components: the off-diagonal covariances between visits reflect

the degree to which measurements remain stable within individuals over time, while diagonal elements capture the total variance at each visit, including both stable subject differences and visit-specific noise. If the estimated subject covariance matrix were purely diagonal, this would indicate little longitudinal stability and variability dominated by visit-specific noise; conversely, substantial off-diagonal covariance would indicate that measurements are correlated across visits, consistent with persistent subject-level differences over time.. This overall formulation does not require a balanced dataset – i.e., subjects with differing number of visits are allowed and the covariances are estimated according to the available data, with no observations being removed.

After estimating the elements of the unstructured covariance matrix, we construct the covariance matrix $\widehat{\boldsymbol{V}}$ following equation (8)), with appropriate visit-wise reordering (such that the correct variance or covariance term for each random effect is used). Then, we use this covariance matrix $\widehat{\boldsymbol{V}}$ to perform GLS estimation for the fixed effects (equation (9)). When computing the standard errors using equation (10), we use a nearest symmetric positive semidefinite algorithm (based on (14) and implemented in (15)) to ensure that the covariance matrix is positive semidefinite; if this process fails, we generate a convergence warning message.

Modeling spline basis functions and omnibus test

FEMA-Long allows modeling the outcome as a set of smooth basis functions $f(t)$ (typically splines; indexed by $s \in \{1 \dots S\}$) of continuous variable $\boldsymbol{t}$:

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sum_{s=1}^{S} \alpha_s f_s(t) + \varepsilon_i \tag{12}$$

This leads to representing the variable $\boldsymbol{t}$ by $\boldsymbol{S}$ derived columns in the design matrix $\boldsymbol{X}$. Then, the set of coefficients, α_s, define the non-linear function $\mathcal{F}(\boldsymbol{t}) = \sum_{s=1}^{S} \alpha_s f_s(t)$, which represents the trajectory of the outcome over, for example, time. These basis functions can also be used to model interactions with another covariate.

In FEMA, the default option is to create natural cubic splines with unit height at knots, similar to the *nsk* option in the *splines2* package (16,17) in *R* (with intercept set as TRUE). Users can specify the knot placement based on domain knowledge, experiment design, or choose to place the knots at the quartiles of the variable. We also support creating other spline basis functions, namely natural cubic splines or B-splines which internally rely on a call to *ns* or *bs* functions in the *splines* package (18) in *R*.

By default, we modify the created basis functions such that the linear (or main) effect of $\boldsymbol{t}$ can be retained in the model. First, we create a linearly spaced vector of values $\boldsymbol{t}_{\text{lin}}$ over the range of values in $\boldsymbol{t}$. Then, we create basis functions using these values. At this point, the user has an option of regressing out powers of $\boldsymbol{t}_{\text{lin}}$ – for example, regressing out the zeroth and first powers of $\boldsymbol{t}_{\text{lin}}$ would result in demeaned basis functions

that are orthogonal to the linear effect of $\boldsymbol{t}_{\text{lin}}$. Next, we perform a singular value decomposition on the demeaned basis functions. We then rescale the orthonormal basis functions such that each basis function has a bounded value between $[-1, 1]$. Finally, we linearly interpolate the modified basis functions to get corresponding basis function values for $\boldsymbol{t}$. These modified orthogonal basis functions are added as columns of $\boldsymbol{X}$. If the design matrix of the covariates already includes an intercept, then, the intercept taken together with these orthogonal basis functions covers the span of the variable.

Once the coefficients have been estimated, we can perform an omnibus Wald test to test the null hypothesis that the linear combination of the estimated coefficients is zero. Generically, given a vector or matrix of weights $\boldsymbol{L}$ with rank r equal to number of linearly independent rows, for a vector of estimated coefficients $\widehat{\boldsymbol{\beta}}$ the multivariate Wald statistics W is (19):

$$W = \left(\boldsymbol{L}\widehat{\boldsymbol{\beta}}\right)^{\top}\left\{\boldsymbol{L}\widehat{\mathrm{Cov}}\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{L}^{\top}\right\}^{-1}\left(\boldsymbol{L}\widehat{\boldsymbol{\beta}}\right) \tag{13}$$

where $\widehat{\mathrm{Cov}}(\boldsymbol{\beta})$ is the estimated variance-covariance matrix of the fixed effects. The Wald statistic W follows a χ^2 distribution with r degrees of freedom. Note that equation (13) can be used to perform Wald tests on any combination of coefficients estimated in (12); i.e., not just limited to the estimated spline coefficients. Therefore, for an omnibus test of the null hypothesis that the linear combination of estimated coefficients of the basis functions is zero, $\boldsymbol{L} = \boldsymbol{I}(S)$, i.e., an identity matrix of size S (the number of basis functions). Alternatively, it is possible to use an F distribution to calculate the p value with two degrees of freedom: the numerator degree of freedom is the rank of $\boldsymbol{L}$ and the denominator degree of freedom is $n - p - S$ (see (20) for a discussion on the similarity and differences in these approaches).

Scaling FEMA-Long for GWAS-like analyses

Within GWAS, the goal is to estimate the effect of genetic variants on the phenotype(s), considering one variant at a time. Let $\boldsymbol{G}$ represent the time-invariant genotyping matrix, in which entry $g_i \in \{0, 1, 2\}$ denotes the best-guess genotype (imputed genotypes which have a posterior probability above a threshold, followed by conversion to discrete count indicating the most probable genotype (21)) for the single nucleotide polymorphism (SNP) for observation i; 0 represents homozygous for the reference allele, 1 represents heterozygous, and 2 represents homozygous for the other allele. Let γ_g be the weight associated with this SNP $\boldsymbol{g}$. Since we are only interested in the marginal effect of each SNP $\boldsymbol{g}$ in $\boldsymbol{G}$, instead of fitting the full model (with covariates) for every SNP, we employ a two-stage procedure.

The Frisch-Waugh-Lovell theorem (12,13) states that the estimate of coefficients from a full regression model are equivalent to regression done in parts (or partial regressions) if both the X and y variables are residualized (22). This implies that we can estimate γ_g by first fitting a reduced model without genetic effect (stage-1), then residualizing each genetic variant $\boldsymbol{g}$ for the covariates, and finally fitting a second

model using the residuals from stage 1 as the dependent variable and the residualized genotype as the independent variable (see (23) for a proof of the Frisch-Waugh-Lovell theorem in the GLS case).

Concretely, we first fit a reduced model excluding the genotyping vector. We estimate the variance components (compound symmetry or unstructured covariance) and the GLS estimates for the covariates using the stage-1 model as described previously. Then, the GLS residuals can be calculated as:

$$\hat{y}_i^{\text{res,GLS}} = y_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} \tag{14}$$

Next, we residualize $\boldsymbol{g}$ for the covariates. The coefficients of $\boldsymbol{X}$ on $\boldsymbol{g}$, $\boldsymbol{\zeta}$, can be calculated as:

$$\hat{\boldsymbol{\zeta}} = \left(\boldsymbol{X}^\top\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{V}}^{-1}\boldsymbol{g} \tag{15}$$

resulting in the GLS residuals:

$$\hat{g}_i^{\text{res,GLS}} = g_i - \boldsymbol{X}_i\hat{\boldsymbol{\zeta}} \tag{16}$$

It is important to note that this residualized genotype vector $\widehat{\boldsymbol{g}}^{\text{res,GLS}}$ is phenotype specific. This follows from $\widehat{\boldsymbol{V}}$ being composed from phenotype specific variance components (equation (8)). Now, using the stage-1 residuals $\widehat{\boldsymbol{y}}^{\text{res,GLS}}$ and the residualized genotype vector $\widehat{\boldsymbol{g}}^{\text{res,GLS}}$, we can estimate $\hat{\gamma}_g$ as:

$$\hat{\gamma}_g = \left(\widehat{\boldsymbol{g}}^{\text{res,GLS}\top}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{g}}^{\text{res,GLS}}\right)^{-1}\widehat{\boldsymbol{g}}^{\text{res,GLS}\top}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{y}}^{\text{res,GLS}} \tag{17}$$

Then, for each genetic variant (and for each phenotype), the GLS residuals can be calculated as:

$$\tilde{y}_{i,g}^{\text{res,GLS}} = \hat{y}_i^{\text{res,GLS}} - \hat{g}_i^{\text{res,GLS}}\hat{\gamma}_g \tag{18}$$

We use the notation $\tilde{y}_{i,g}^{\text{res,GLS}}$ to distinguish these residuals (which are specific to each genetic variant) from $\hat{y}_i^{\text{res,GLS}}$ which are the stage-1 GLS residuals obtained from FEMA after accounting for the covariates (without the genetic variant). Therefore, the residuals $\widehat{\boldsymbol{y}}^{\text{res,GLS}}$ from (14) are the residualized phenotype for (17) which are estimated using the residualized genotype $\widehat{\boldsymbol{g}}^{\text{res,GLS}}$ (from (16)).

The mean squared error is given by:

$$\tilde{\sigma}_g^2 = \frac{1}{(n-p-1)}\sum_{i=1}^{n}\left(\tilde{y}_{i,g}^{\text{res,GLS}}\right)^2 \tag{19}$$

where the denominator is normalized by $(n-p-1)$ to account for the genotyping vector $\boldsymbol{g}$ as an additional independent variable (13,24). The variance for each $\hat{\gamma}_g$ can be estimated as:

$$\text{Var}(\hat{\gamma}_g) = \left(\widehat{\boldsymbol{g}}^{\text{res,GLS}\top}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{g}}^{\text{res,GLS}}\right)^{-1} \tag{20}$$

Since the $\widehat{\boldsymbol{V}}$ term has been normalized by $\hat{\sigma}^2$, the standard error is scaled by $\tilde{\sigma}_g^2$.

The above framework can be applied to other GWAS-like analyses where one is interested in examining the effect of many predictors, taken one at a time (for example, given protein assays, the effect of each protein on a phenotype). Note, that in this framework, we assume that the individual contribution of each SNP is small and unlikely to impact the estimation of the covariance components in stage-1 (see Supplementary Material and Supplementary Figures S46-S51 for a comparison of the estimates between the two approaches).

Modeling time-dependent genetic effects using spline interactions

This GWAS framework can also include interaction terms – linear interaction or smooth interaction. In case of linear interaction, if the main effect of the interacting variable has already been included as an X variable in stage-1 regression, then, $\boldsymbol{g}$ is a matrix with the first column containing the main effect of the genetic variant and the second column containing the elementwise (Hadamard) product of the genetic variant and the interacting variable. In this case, we residualize both columns of $\boldsymbol{g}$ for $\boldsymbol{X}$ (equation (15)), followed by estimation using equation (17), resulting in two coefficients: the coefficient for the main effect and the coefficient for the interaction term. The denominator of the mean squared error (equation (19)) is adjusted for two parameters, i.e., $(n - p - 2)$.

For modeling spline interactions, $\boldsymbol{g}$ is a matrix having the elementwise product of the genetic variant and the smooth function $f(t)$. The estimation follows the same procedure, with the estimation of as many coefficients as the columns of $\boldsymbol{g}$. If the basis functions have been appropriately modified and a constant term added (see the section on Modeling spline basis functions), then both the main effect of the genetic variant as well as the interaction of the genetic variant with smooth functions can be estimated. Generically, if there are S basis functions (with or without the constant term), the denominator in equation (19) is penalized by $(n - p - S)$. Having estimated the coefficients, we can use equation (13) to either perform an omnibus test across the main and the time-dependent genetic effect, or perform an omnibus test for just the time-dependent effect and a separate test for just the main effect of the SNP (if desired). Depending on the number of tests performed, multiple comparison correction may be necessary.

Concretely, the GWAS interaction can be framed as modeling the effects of $\boldsymbol{g}$ on the outcome by altering regression coefficients α_s in (12):

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sum_{s=0}^{S}(\alpha_s + \gamma_s g_i)f_s(t) + \varepsilon_i \tag{21}$$

where γ_s coefficients are equivalent to the interaction term between variable $\boldsymbol{g}$ and the basis functions $f(t)$. Note that $s = 0$ corresponds to the main effect of the SNP. This formulation is equivalent to:

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sum_{s=1}^{S} \alpha_s f_s(\boldsymbol{t}) + \gamma_g g_i + \sum_{s=1}^{S} \gamma_s g_i f_s(t) + \varepsilon_i \tag{22}$$

For example, if the time $\boldsymbol{t}$ was modeled as smooth functions $f(\boldsymbol{t})$, then after estimating the regression coefficients, FEMA-Long can be used to calculate the following quantities:

The estimated cross-sectional effect of the genotype $\boldsymbol{g}$ on the outcome, at a specific time $\boldsymbol{t}_0$:

$$\hat{\eta}_{\boldsymbol{t}_0} = \frac{dy}{dg} = \sum_{s=0}^{S} \hat{\gamma}_s f_s(\boldsymbol{t}_0) \tag{23}$$

and the instantaneous effect of genotype $\boldsymbol{g}$ on the rate of change of the outcome at time $\boldsymbol{t}_0$:

$$\hat{\eta}'_{\boldsymbol{t}_0} = \frac{d^2y}{dgd\boldsymbol{t}} = \sum_{s=1}^{S} \hat{\gamma}_s f'_s(\boldsymbol{t}_0) \tag{24}$$

Both these quantities can be plotted as a function of time, $\boldsymbol{t}$, to assess the effect of $\boldsymbol{g}$ on the outcome over time.

Simulations

Verification and Comparison

Methods

To validate the FEMA-Long approach for fitting LMEs with unstructured covariances, we performed a series of simulations. All analyses were performed using MATLAB R2023a (25).

Simulation 1: parameter recovery

In the first simulation, we created phenotypes with known variance-covariance matrices and known effect sizes. Then, using FEMA, we estimated the parameters and compared them with the ground truth. We created a grid of known sources of variances: fixed effects $\mathrm{Var}_{\mathrm{FFX}} \in \{0.1, 0.2, \ldots, 0.9\}$, family effect (i.e., variance attributed to common environment) $\mathrm{Var}_{\mathrm{Fam}} \in \{0.1, 0.2, \ldots, 0.9\}$, subject effect (i.e., variance attributed to repeated measurements) $\mathrm{Var}_{\mathrm{Sub}} \in \{0.1, 0.2, \ldots, 0.9\}$, and noise $\mathrm{Var}_{\mathrm{Noise}} \in \{0.1, 0.2, \ldots, 0.9\}$. The total phenotypic variance was the sum of these four variances with the total variance being 1; i.e., $\mathrm{Var}_y = \mathrm{Var}_{\mathrm{FFX}} + \mathrm{Var}_{\mathrm{Fam}} + \mathrm{Var}_{\mathrm{Sub}} + \mathrm{Var}_{\mathrm{Noise}} = 1$. This resulted in 84 combinations of variances.

For each of these combinations, we simulated data for 5,000 unique individuals with five repeated measurements; these individuals were grouped in families such that a family had no more than five members; then, we randomly subset the data to result in a total of 20,000 observations, reflecting repeated and missing measurements, as well as differing number of individuals across families. We simulated the

age at each measurement using the mean and standard deviation of age as $\{0 \pm 0;\ 60 \pm 10;\ 180 \pm 12;\ 240 \pm 14;\ 365 \pm 15\}$ days – this scenario represents repeated measures data in infants with measurements conducted at birth, two months, six months, eight months, and one year.

For each of the 84 combinations of variances, we created an arbitrary 5×5 covariance matrix for family effect ($\mathrm{V_{Fam}}$) and subject effect ($\mathrm{V_{Sub}}$). We simulated 15 covariates– intercept, spline basis functions of the simulated age, and other random X variables drawn from a normal distribution. The ground truth beta coefficients were drawn from a uniform distribution $\beta^{\mathrm{sim}} \sim \mathcal{U}[-0.2, 0.2]$ such that $\mathrm{y_{FFX}} = X\beta^{\mathrm{sim}}$, having mean $\overline{y_{\mathrm{FFX}}}$ and standard deviation σ_{FFX}; we then mean centered and scaled y_{FFX} to have $\mathrm{Var_{FFX}}$ variance. Then, using V_{Fam} and V_{Sub} covariance matrices, we drew samples from a multivariate normal (MVN) distribution with zero means and scaled them to have $\sqrt{\frac{\mathrm{Var_{Fam}}}{\mathrm{Var_{RFX}}}}$ and $\sqrt{\frac{\mathrm{Var_{Sub}}}{\mathrm{Var_{RFX}}}}$ variances each, where $\mathrm{Var_{RFX}} = \mathrm{Var_{Fam}} + \mathrm{Var_{Sub}}$. In other words, $y_{\mathrm{Fam}} = \sqrt{\frac{\mathrm{Var_{Fam}}}{\mathrm{Var_{RFX}}}} \times \mathrm{MVN}(0, V_{\mathrm{Fam}})$ and $y_{\mathrm{Sub}} = \sqrt{\frac{\mathrm{Var_{Sub}}}{\mathrm{Var_{RFX}}}} \times \mathrm{MVN}(0, V_{\mathrm{Sub}})$. Therefore, $y_{\mathrm{RFX}} = \sqrt{\mathrm{Var_{RFX}}} \times (y_{\mathrm{Fam}} + y_{\mathrm{Sub}})$. Finally, we simulated noise drawn from a uniform distribution and scaled it to have zero mean and $\mathrm{Var_{Noise}}$ variance. Thus, the simulated phenotype was $y = y_{\mathrm{FFX}} + y_{\mathrm{RFX}} + y_{\mathrm{noise}}$.

For this simulated phenotype, we fitted the model using FEMA, specifying the covariance type to be unstructured and the random effects to be family and subject effects. As an additional comparator, we fit the same model using the *glmmTMB* toolbox version 1.1.10 (26) in R version 4.2.1 (18). The model in glmmTMB was specified as: $y \sim 1 + X + \mathrm{us}(\mathrm{timepoint} + 0|\mathrm{FID}) + \mathrm{us}(\mathrm{timepoint} + 0|\mathrm{IID}), \mathrm{dispformula} = \sim 0$, where 1 indicates the intercept, X indicates the remaining covariates (i.e., excluding the intercept), us specifics an unstructured covariance matrix for family (FID) and subject (IID) terms, the timepoint denotes a factor variable indicating which visit the observation belonged to, and the dispformula component disables the addition of an additional error term as that would be an overparameterized model (27). After preliminary testing, we noticed convergence issues with glmmTMB; therefore, we changed the default settings to have higher number of iterations, specifically setting iter.max and eval.max to 5,000.

We compared the beta coefficients for the covariates and the estimated variance-covariance matrices for the family and subject effects from both FEMA-Long and glmmTMB with the ground truth and with each other. For the beta coefficients, $\beta^{\mathrm{true}}_{\mathrm{intercept}} = \beta^{\mathrm{sim}}_{\mathrm{intercept}} - \overline{y_{\mathrm{FFX}}}/\sigma_{\mathrm{FFX}}$, and for the remaining X variables, $\beta^{\mathrm{true}}_{X} = \beta^{\mathrm{sim}}_{X}/\sigma_{\mathrm{FFX}}$; for the family effect variance-covariance matrix, $V^{\mathrm{true}}_{\mathrm{Fam}} = V_{\mathrm{Fam}} \times \mathrm{Var_{Fam}}$; and for the subject effect variance-covariance matrix, $V^{\mathrm{true}}_{\mathrm{Sub}} = V_{\mathrm{Sub}} \times \mathrm{Var_{Sub}} + I \times \mathrm{Var_{noise}}$, where I was an identity matrix of size five (corresponding to number of visits). Then, we calculated the root mean squared error (RMSE) for the fixed effects and the family and subject variance-covariance matrices to assess how far the estimated

values were from the true simulated parameters, and additionally to quantify the difference between the estimated parameters from FEMA-Long and glmmTMB. For each of the 84 combinations of variances explained by covariates and random effects, we repeated this entire process 50 times.

Simulation 2: false positive rate

To evaluate the false positive rate, we repeated the strategy from simulation one. The covariates included the intercept, age, and other random 98 X variables drawn from a random normal distribution, resulting in 100 X variables. However, the covariates had zero contribution to the simulated phenotype. The total phenotypic variance was 1, i.e., $\mathrm{Var}_y = \mathrm{Var}_{\mathrm{Fam}} + \mathrm{Var}_{\mathrm{Sub}} + \mathrm{Var}_{\mathrm{Noise}} = 1$, resulting in 36 different combinations. For each simulation setting, we created 10 y variables that have different amount of contribution from family, subject, and noise terms and repeated these simulations 1,000 times. Using these simulated data, we fitted the models using the unstructured covariance in FEMA-Long as well as repeated the model fitting using compound symmetry covariance type. Then, we examined the distribution of the p-values from both the models.

Simulation 3: computational time and carbon footprint

To assess the computational performance of FEMA-Long with unstructured covariance, we simulated data with varying number of observations and varying number of outcome variables. We fixed $\mathrm{Var}_{\mathrm{Fam}} = 0.20$, $\mathrm{Var}_{\mathrm{Sub}} = 0.30$, $\mathrm{Var}_{\mathrm{FFX}} = 0.35$, and $\mathrm{Var}_{\mathrm{noise}} = 0.15$. We set the number of covariates to 50. We created 10 log-spaced values for the number of observations (rounded off), varying between 10^3 and 10^5. Similarly, we created 10 log-spaced values for the number of outcome variables (rounded off), varying between 10^0 and 10^4. For each of these combinations of number of observations and number of outcome variables, we estimated the time taken by FEMA. We repeated the time measurement five time and then took the median to get a robust estimate of the time taken to run the analysis. We also performed the time measurement using parallel computing with 32 parallel jobs and 2 threads per job. As comparators, we used the glmmTMB toolbox and the *lme4* (28) toolbox (version 1.1-28); since these toolboxes are not designed for multiple outcome variables, we only timed it for one outcome variable (across the 10 values of number of observations) and then scaled the timing for larger number of outcome variables. For fitting the models with *lme4*, we used the *lmer* function, specifying the random effects as $\mathrm{timepoint} + 0|\mathrm{FID}$ and $\mathrm{timepoint} + 0|\mathrm{IID}$; in addition, we specified a *lmer* control object that disabled the check for number of observations being greater than or equal to the number of random effects levels (i.e., we set check.nobs.vs.nRE to ignore). We ran the timing measurements on a compute cluster with two AMD EPYC 7702 64-core processors, with the job specification of 64 CPUs per task and 5GB memory per CPU. In addition, to estimate the carbon footprint of FEMA-Long, we used the formula from (29) and the data from the green-algorithms.org v3.0. We specified the number of cores as 64 and the memory usage as 320

GB; we set the CPU model to "Any" and the location to "World". For comparison, we also estimated the carbon footprint for the glmmTMB and the lme4 toolboxes.

Simulation 4: impact of sample size and missingness on parameter recovery

Since FEMA-Long is designed for large sample sizes and the estimation of the covariance parameters uses pairs of timepoints, we performed an additional analysis to examine the impact of missing observations on parameter recovery. Specifically, we simulated data for 5,000 unique individuals with five repeated measurements; these individuals were assigned family numbers such that a family had no more than five members. The probability of a missing visit was between 0.1 and 0.8, while ensuring that subsequent visits had a higher missingness probability; since there are five repeated measurements, there are 15 possible visit combinations (five variances and ten covariances) – we varied the minimum number of observations between any pairs of visits $\min_{\text{numObs}}$ between 500 to 800, in increments of 100; the resulting total number of observations n_{Obs} were 12,000, 15,000, 18,000, and 20,000. To clarify, there were 16 scenarios: four $\min_{\text{numObs}}$ and four n_{Obs}. For simulating arbitrary covariance matrices (family and subject covariances), we allowed the within-visit variance to vary between 0.2 and 0.8, and the correlation between visits (scaled to covariances) to vary between -0.7 and 0.7; after ensuring that the simulated covariance matrix was positive semidefinite, we additionally ensured that the condition number of the matrices did not exceed 1000. Then, for these 16 scenarios, the rest of the simulation strategy was the same as simulation 1. We compared the estimated parameters with the ground truth and with glmmTMB as a comparator for 16 scenarios, each having 84 combinations of variance values, each of them having 15 covariates, one outcome variable, and the process repeated 50 times.

Simulation 5: impact of sample size and missingness on false positives

Similar to simulation 4, to examine the impact of missing observations on the number of false positives, we followed the simulation strategy from simulation 4 to create 16 different scenarios with differing $\min_{\text{numObs}}$ and n_{Obs}. We examined the distribution of the p-values (using unstructured covariance and compound symmetry) for the 16 scenarios, each having 36 combination of variance values (similar to simulation 2), each of them having 100 covariates, 10 outcome variable, and the entire process repeated 1000 times.

Results

Simulation 1: FEMA-Long accurately models unstructured covariances

For the fixed effects, the average (over 50 iterations) RMSE across 84 simulations was $0.0121 \pm 0.0018\,[0.0073 - 0.0150], r_{\text{true,FEMA}} = 0.996$ (**Figure 1a**, **Supplementary Figure S3**). For the random effects parameters, the average RMSE for the family effect was $0.0160 \pm 0.0058\,[0.0064 - 0.0312], r_{\text{true,FEMA}} = 0.997$ (**Figure 1b**, **Supplementary Figure S4**), and for the subject effect was

$0.0122 \pm 0.0036\,[0.0051 - 0.0218], r_{\text{true,FEMA}} = 0.999$ (**Figure 1c**, **Supplementary Figure S5**), indicating that the estimates were close to the ground truth.

The average RMSE between the estimates from FEMA-Long and from glmmTMB toolbox for the fixed effects was $0.0004 \pm 0.0001\,[0.0002 - 0.0007], r_{\text{glmmTMB,FEMA}} = 1.00$ (**Supplementary Figure S6**), for the family effect was $0.0076 \pm 0.0028\,[0.0028 - 0.0151], r_{\text{glmmTMB,FEMA}} = 0.9996$ (**Supplementary Figure S7**), and for the subject effect was $0.0055 \pm 0.0016\,[0.0020 - 0.0104], r_{\text{glmmTMB,FEMA}} = 0.9998$ (**Supplementary Figure S8**), showing that the FEMA-Long estimates were almost identical to glmmTMB.

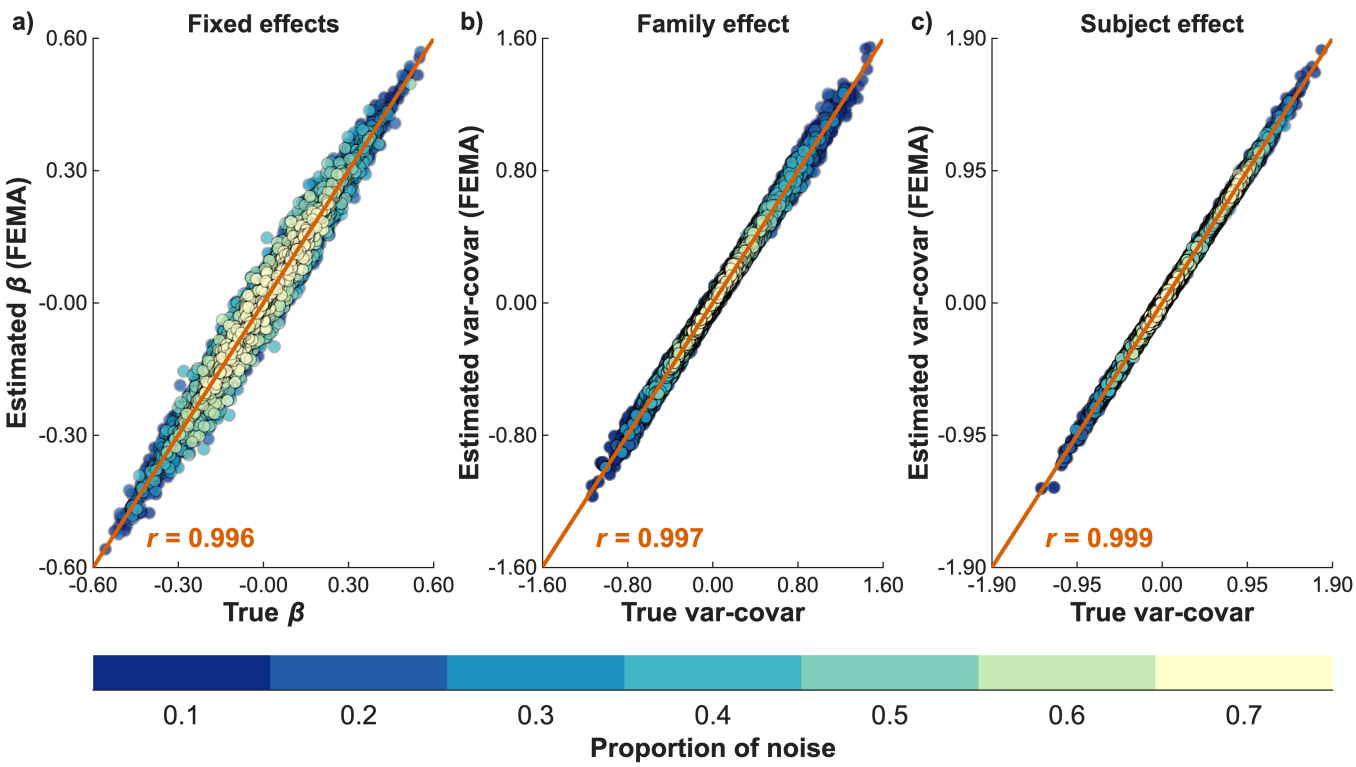

Figure 1: Comparison of estimated parameters from FEMA-Long with ground truth. Each panel shows the scatterplot of the simulated ground truth and the estimated parameters from FEMA-Long across 50 iterations of 84 simulation settings with different amounts of variances explained by fixed, family, and subject effects, as well as noise; the amount of noise in the simulated phenotype is indicated by the color of the points; the orange line indicates the least square fit and the correlation between ground truth and estimated parameters are indicated in orange text; **a)** beta coefficients for the covariates; **b)** variance-covariance components for the family random effect; and **c)** variance-covariance components for the subject effect.

Simulation 2: FEMA-Long controls for false positives

Across the 36 simulations, when using unstructured covariance, the p-values were uniformly distributed under the null ($-\log_{10} p$-values following the diagonal; green points in **Figure 2** and **Supplementary Figure S9**). However, for the compound symmetry covariance pattern, in almost all scenarios, the p-values had a non-uniform distribution with several of the p-values being smaller than expected (leftward deflection of the $-\log_{10} p$-values from the diagonal; orange points in **Figure 2** and **Supplementary Figure S9**). These results show that the unstructured covariances p-values are well-calibrated.

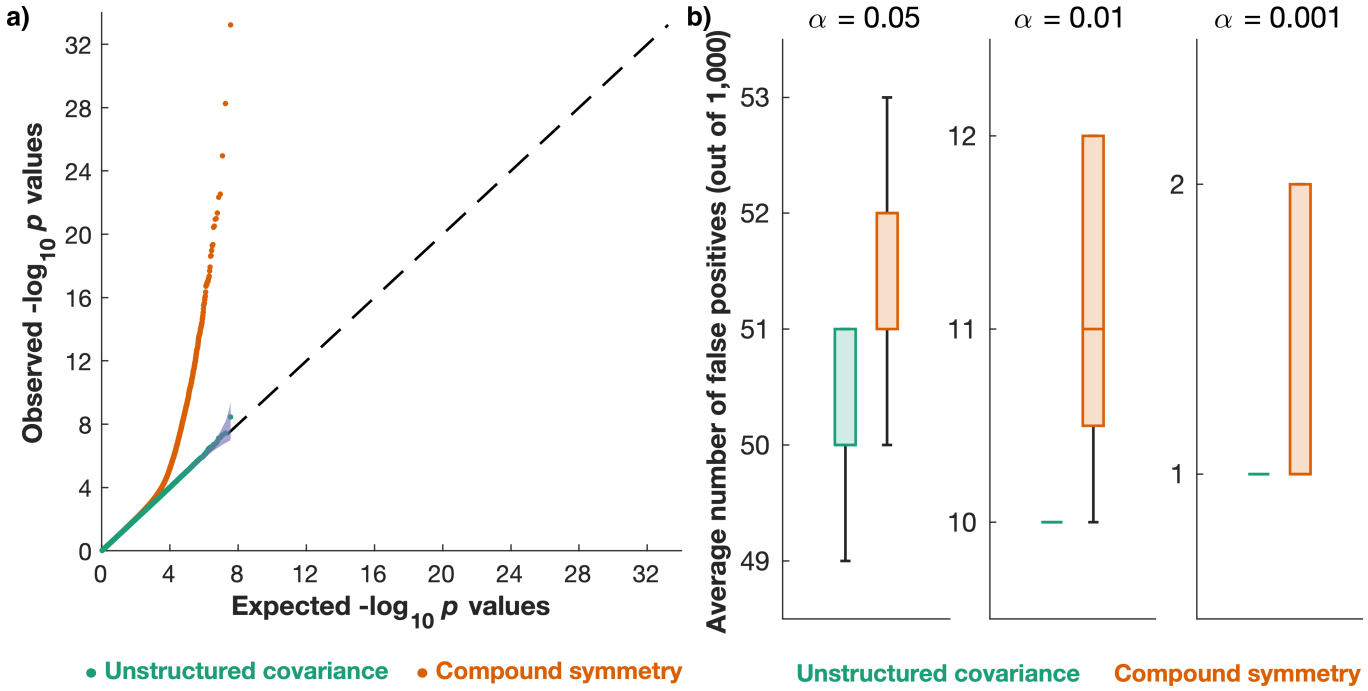

Figure 2: Examining false positives under the null. a) Distribution of $-\log_{10} p$-values across 1000 iterations of 36 simulation settings with differing amounts of variances explained by family effect, subject effect, and noise for unstructured covariance (green points) and compound symmetry (orange points); each simulation iteration consisted of 100 X variables and 10 outcome variables; the purple filled area indicates the 95% confidence interval based on inverse beta distribution; **b)** box-plots showing the average number of false positives (rounded off) at different alpha values across 100 X variables and 10 y variables for unstructured covariance (green bars) and compound symmetry (orange bars); the whiskers mark the values 1.5 times the interquartile range away from 75th and 25th percentile; under the null, the expected number of false positives are 50 ($\alpha = 0.05$), 10 ($\alpha = 0.01$), and 1 ($\alpha = 0.001$).

Simulation 3: FEMA-Long is faster and greener

FEMA-Long was fast across a range of observations and outcomes (**Figure 3**). For a single outcome, for 1,000 observations, FEMA-Long took ~0.14 seconds; in comparison, glmmTMB took ~9.76 seconds and lmer took ~85.66 seconds (i.e., FEMA-Long was ~71 times faster than glmmTMB and ~622 times faster than lmer); for 100,000 observations, this time went up to ~34.8 seconds (FEMA-Long; **Figure 3a**), ~1013 seconds (glmmTMB; **Supplementary Table S1**), and ~3105 seconds (lmer; **Supplementary Table S5**); therefore, for a single phenotype, FEMA-Long was 29-106 times faster than glmmTMB and 89-622 times faster than lmer. The carbon footprint for glmmTMB for a single outcome varied between 1.78 gCO_2e and 185.01 gCO_2e; for lmer the carbon footprint varied between 8.52 gCO_2e and 567 gCO_2e; in comparison, for FEMA-Long, the carbon footprint was between 0.03 gCO_2e and 6.35 gCO_2e; therefore, similar to computational timing, FEMA-Long was between 29 and 106 times greener than glmmTMB (**Supplementary Table S7**) and between 89 and 622 times greener than lmer (**Supplementary Table S9**).

For 100,000 observations and 10,000 outcome variables, FEMA-Long took ~47.5 minutes and had a carbon footprint of 520.51 gCO_2e (serial, **Figure 3b** and **Supplementary Tables S2** and **S7**) and ~4.2 minutes with a carbon footprint of 46.46 gCO_2e (parallel, **Figure 3c** and **Supplementary Tables S2** and **S7**); in contrast, extrapolating the timing for glmmTMB, the same analysis would have taken ~117.25 days and would have had a carbon footprint of 1.85 TCO_2e (**Supplementary Tables S2** and **S7**); similarly, extrapolating the

timing for lmer, the same analysis would have taken ~359.36 days and would have had a carbon footprint of ~5.67 TCO_2e (**Supplementary Tables S6** and **S9**). Overall, for more than one phenotype, across a range of observations, FEMA-Long was between ~86-3,554 times faster and greener than glmmTMB (serial) and between ~84-42,096 times faster and greener than glmmTMB (parallel; **Supplementary Table S2** and **Supplementary Table S7**); compared to lmer, FEMA-Long was between ~263-15,785 times faster and greener (serial) and between ~259-287,190 times faster and greener (parallel; **Supplementary Tables S6** and **S9**). We encountered convergence issues with glmmTMB; increasing the maximum number of iterations and evaluations to 5000 and re-examining the time taken by glmmTMB, we found FEMA-Long to be up to 6,902 (serial) and 93,659 times faster and greener (parallel) than glmmTMB (**Supplementary Table S3, S4,** and **S8**).

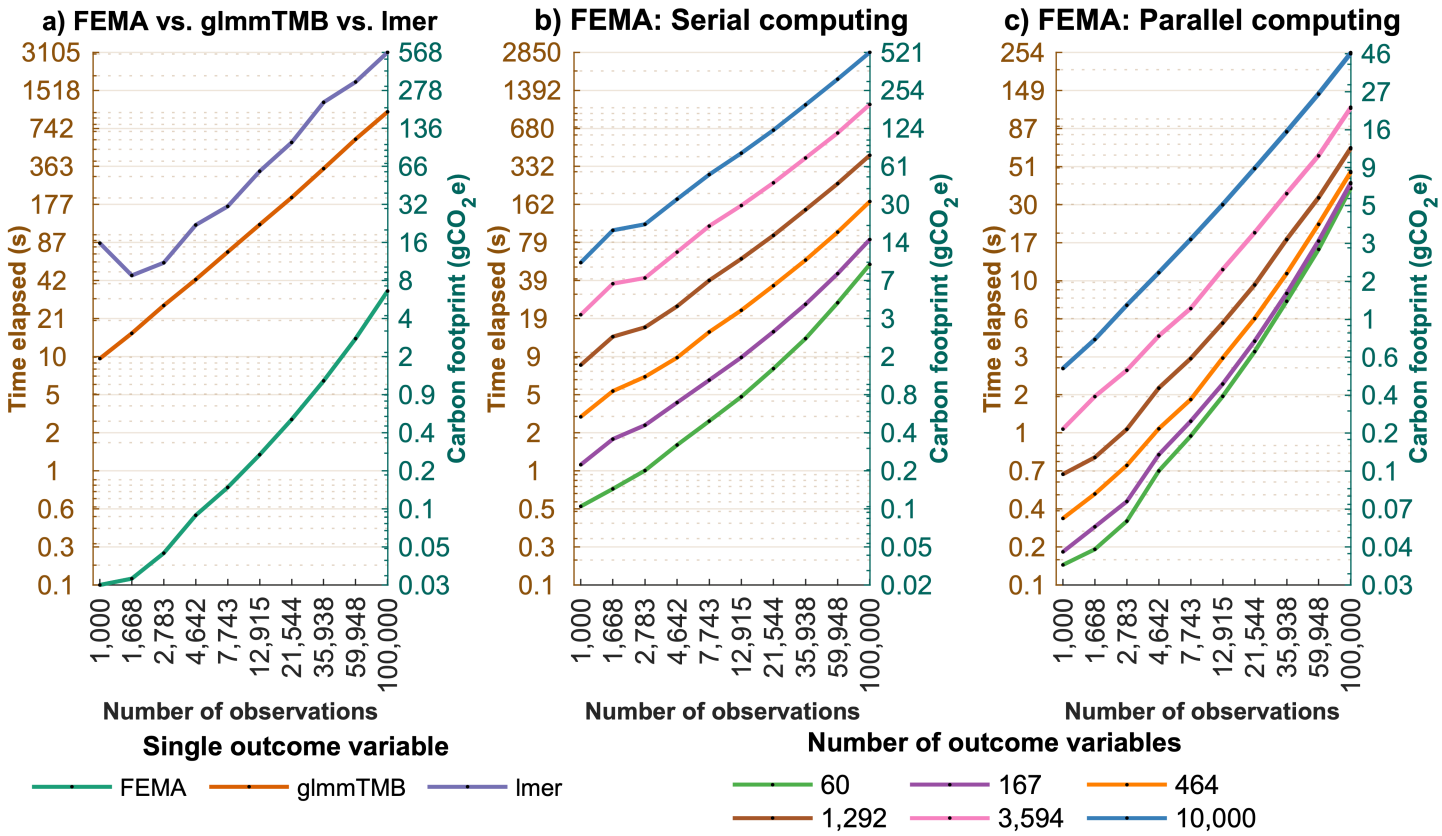

Figure 3: Time taken and carbon footprint for fitting linear mixed-effects models with unstructured covariances. In each panel, the line(s) indicate the time taken (in seconds) on the left *y*-axis and the carbon footprint (in grams of carbon dioxide emission) on the right *y*-axis for fitting linear mixed-effects models with unstructured covariances for family and subject random effects and 50 covariates across a range of number of observations (shown on the *x*-axis). Panel **a)** shows the time taken by and the carbon footprint for FEMA-Long (green), glmmTMB (orange), and lmer (from the lme4 toolbox; purple) for a single outcome variable across a range of number of observations. Panel **b)** shows the time taken by and the carbon footprint of FEMA-Long with serial computing, and panel **c)** shows the time taken by and the carbon footprint of FEMA-Long with parallel computing (32 parallel processes) across a range of increasing number of outcome variables, indicated by the color of the line. In all the panels, both axes are log-spaced and the *y*-axes tick labels are rounded off.

Simulation 4: FEMA-Long estimates are stable across sample sizes and missingness

When we compared the FEMA-Long estimates with the ground truth across a range of sample size and differing minimum number of observations between pairs of visits, we found the FEMA-Long estimates to be highly correlated to the ground truth ($r_{\text{true,FEMA}} = 0.987 - 0.999$; **Supplementary Figures S10 – S25**).

The average (over 50 iterations) RMSE across 84 simulations for the fixed effects ranged between 0.0150 ± 0.0025 ($n_{\text{Obs}} = 12{,}000$, $\text{min}_{\text{numObs}} = 500$) and 0.0111 ± 0.0018 ($n_{\text{Obs}} = 15{,}000$, $\text{min}_{\text{numObs}} = 800$); for the family effect covariance parameters, the average RMSE ranged between 0.0128 ± 0.0033 ($n_{\text{Obs}} = 12{,}000$, $\text{min}_{\text{numObs}} = 500$) and 0.0085 ± 0.0025 ($n_{\text{Obs}} = 15{,}000$, $\text{min}_{\text{numObs}} = 800$); for the subject effect covariance parameters, the average RMSE ranged between 0.0152 ± 0.0042 ($n_{\text{Obs}} = 12{,}000$, $\text{min}_{\text{numObs}} = 500$) and 0.0082 ± 0.0027 ($n_{\text{Obs}} = 15{,}000$, $\text{min}_{\text{numObs}} = 800$); see **Supplementary Table S10** for details. For smaller n_{Obs} such as 12,000 observations, increasing $\text{min}_{\text{numObs}}$ resulted in reduced RMSE with the ground truth; at larger n_{Obs} such as 20,000 observations, we did not see any differences when increasing $\text{min}_{\text{numObs}}$, likely indicating a stable performance for large samples.

On comparing the FEMA-Long estimates with glmmTMB across a range of sample sizes and minimum number of observations between pairs of visits, the FEMA-Long parameters were comparable to the glmmTMB estimated parameters (**Supplementary Table S11**). The average (over 50 iterations) RMSE across 84 simulations for the fixed effects ranged between 0.0006 ± 0.0002 ($n_{\text{Obs}} = 12{,}000$, $\text{min}_{\text{numObs}} = 500$) and 0.0002 ± 0.0001 ($n_{\text{Obs}} = 15{,}000$, $\text{min}_{\text{numObs}} = 800$); for the family effect covariance parameters, the average RMSE ranged between 0.0053 ± 0.0021 ($n_{\text{Obs}} = 12{,}000$, $\text{min}_{\text{numObs}} = 500$) and 0.0022 ± 0.0010 ($n_{\text{Obs}} = 15{,}000$, $\text{min}_{\text{numObs}} = 800$); for the subject effect covariance parameters, the average RMSE ranged between 0.0064 ± 0.0019 ($n_{\text{Obs}} = 12{,}000$, $\text{min}_{\text{numObs}} = 500$) and 0.0024 ± 0.0007 ($n_{\text{Obs}} = 15{,}000$, $\text{min}_{\text{numObs}} = 800$). Similar to the comparison with ground truth, we observed that for smaller n_{Obs} such as 12,000 observations, increasing $\text{min}_{\text{numObs}}$ resulted in reduced RMSE, while at larger n_{Obs} such as 20,000 observations, increasing $\text{min}_{\text{numObs}}$ did not make a difference to the average RMSE.

Simulation 5: FEMA-Long false positive calibration improves with sample size

We examined the distribution of p-values under a range of n_{Obs} and $\text{min}_{\text{numObs}}$; in each of these 16 scenarios, across the 36 simulations, unstructured covariance generally resulted in uniformly distributed p-values (**Supplementary Figures S26**, **S29**, **S32**, and **S35**), except for two settings: $\text{Var}_{\text{Fam}} = 0.7, \text{Var}_{\text{Sub}} = 0.2, \text{Var}_{\text{Noise}} = 0.1$ and $\text{Var}_{\text{Fam}} = 0.8, \text{Var}_{\text{Sub}} = 0.1, \text{Var}_{\text{Noise}} = 0.1$; in these two settings, we observed that there were a few combinations of n_{Obs} and $\text{min}_{\text{numObs}}$, where the distribution of the unstructured covariance p-values closely followed the null distribution but deviated away from the null towards the tail of the distribution: $n_{\text{Obs}} = 12{,}000$, $\text{min}_{\text{numObs}} = 500$ (**Supplementary Figure S27**), $n_{\text{Obs}} = 12{,}000$, $\text{min}_{\text{numObs}} = 600$ (**Supplementary Figure S28**), $n_{\text{Obs}} = 15{,}000$, $\text{min}_{\text{numObs}} = 500$ (**Supplementary Figure S30**), $n_{\text{Obs}} = 15{,}000$, $\text{min}_{\text{numObs}} = 600$ (**Supplementary Figure S31**), $n_{\text{Obs}} = 18{,}000$, $\text{min}_{\text{numObs}} = 600$ (**Supplementary Figure S33**), and $n_{\text{Obs}} = 18{,}000$, $\text{min}_{\text{numObs}} = 700$ (**Supplementary Figure S34**). On further examination of these cases, we observed that the miscalibration of the p-values

was caused by one or few iterations (out of the 1000) where the resulting distribution of visits was extremely imbalanced (**Supplementary Figure S36**), leading to a small number of common observations between the last pairs of visits, possibly leading to an unstable estimation of the covariance components. As the sample size increased, the false positive calibration for the unstructured covariance improved; in contrast, the compound symmetry covariance pattern resulted in inflated Q-Q plots.

Application: GWAS on anthropometric features in infants

Study sample

To demonstrate the features of FEMA-Long (unstructured covariance, non-linear modeling, scalability to large datasets, performing GWAS, and discovery of time-dependent genetic effects), we performed a GWAS on the length, weight, and BMI of infants in the first year of life using the large-scale Norwegian Mother, Father, and Child Cohort Study (MoBa) (30–32). The MoBa Study is a population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health. Participants were recruited from all over Norway from 1999-2008. The women consented to participation in 41% of the pregnancies. The cohort includes approximately 114,500 children, 95,200 mothers and 75,200 fathers. The current study is based on version 12 of the quality-assured data files released for research in 2019. The establishment of MoBa and initial data collection was based on a license from the Norwegian Data Protection Agency and approval from The Regional Committees for Medical and Health Research Ethics. The MoBa cohort is currently regulated by the Norwegian Health Registry Act. The current study was approved by The Regional Committees for Medical and Health Research Ethics (2016/1226/REK Sør-Øst C).

Genetic data

Blood samples were obtained from both parents during pregnancy and from mothers and children (umbilical cord) at birth (33), which were used for genotyping. We used the quality controlled genetic data from the MoBaPsychGen pipeline v1 (34). The release consists of data from 76,577 children, 77,634 mothers, and 53,358 fathers; we restricted ourselves to the data on children who had not withdrawn their consent from MoBa as of November 1, 2024.

Phenotypes

The phenotypes of interest were the length, weight, and BMI of infants at six timepoints during the first year of life: birth, six weeks, three months, six months, eight months, and twelve months. Phenotypic information was retrieved from questionnaire data, and from the Medical Birth Registry of Norway (MBRN); the MBRN is a national health registry containing information about all births in Norway (see Supplementary Materials for details on data curation and quality checks prior to GWAS). The final sample size for GWAS was 68,273 infants with 299,447 observations having complete data on length, weight, and

BMI (i.e., each of the 68,273 subjects had all three measurements); see **Supplementary Figure S10** for distribution of age at each timepoint, and **Supplementary Figures S11-S13** for plots showing the trajectories of length, weight, and BMI.

Genome-wide association study

For the 68,273 infants, with up to six repeated measurements totaling 299,447 observations on length, weight, and BMI during the first year of life, we performed a longitudinal GWAS using FEMA-Long on the autosomes. We standardized the phenotypes to have zero mean and unit standard deviation. The covariates included the intercept, six spline basis functions of age (after the transformations described in the Methods section), dummy-coded sex, twenty genetic principal components (PC), and 23 dummy-coded genotyping batch variables, resulting in a total of 51 covariates. We created the spline basis functions by placing a knot at the median age for each timepoint: 44, 94, 181, 246, and 369 days, and boundary knots at the extremes (0 and 425 days; see, **Supplementary Figure S14** and **Supplementary Figure S15**).

The random effects included the effect of family or shared environment, the genetic relationship matrix (GRM), and the subject effect (or repeated measurements). We defined families using the mother ID of the participants; in addition, for a subset of participants where the mother ID was different, but the father ID was the same, we assigned those individuals to the same family; this resulted in 46,453 unique families. To calculate GRM, we used a subset of 500,243 SNPs which were directly genotyped in at least one imputation batch or had an imputation INFO score > 0.99785 in all imputation batches; we mean imputed any missing genotyping data and standardized the SNPs (zero mean, unit standard deviation); then, the GRM was calculated as the correlation coefficient between the standardized SNPs. Therefore, the stage-1 model (without GWAS) for each phenotype was:

$$\text{Phenotype} \sim 1 + s(\text{age}) + \text{sex} + \text{PC}_{1-20} + \text{Batch} + \text{us}(1|\text{Family}) + \text{us}(1|\text{GRM}) + \text{us}(1|\text{Subject}) \quad (25)$$

We performed GWAS on 6,981,748 SNPs (34) after mean imputing missing genotyping data, and examined both the main effect of the SNPs as well as the interaction of each SNP with the six basis functions, resulting in seven coefficients per SNP per phenotype. Specifically, after fitting the stage-1 model and residualizing each genotype for the covariates in (25), the model for each residual SNP, G_{res}, was:

$$\text{Phenotype}_{\text{residuals,GLS}} \sim G_{\text{res}} + G_{\text{res}} \odot s(\text{age}) \quad (26)$$

where $G_{\text{res}} \odot s(\text{age})$ is the element-wise multiplication of G_{res} with every column of the basis function matrix. Then, for each phenotype, and each SNP, we performed a Wald test to determine if any of these coefficients had a non-zero effect. To accelerate computing, we split the genetic data into chunks of 5,000 SNPs, processing eighteen chunks in parallel, with each parallel worker having seven threads, run on a compute cluster with two AMD EPYC 7702 64-core processers; the job specification specified 128 CPUs per task and 30 GB RAM per task. Additionally, we also ran a standard longitudinal GWAS, without

including $s(\text{age})$ and $G_{\text{res}} \odot s(\text{age})$ terms (run on the same hardware with the same specification, but twenty parallel chunks with six threads each).

To quantify genomic inflation factor (λ_{GC}), we used the Wald test p values, converted them to chi-squared statistics (using $1 - p$ values as input to the inverse of the chi-square cumulative distribution function with 1 degree of freedom); then, λ_{GC} was calculated as:

$$\lambda_{\text{GC}} = \frac{\text{median}(\chi^2_{\text{obs}})}{\text{median}(\chi^2_{\text{exp}})} \tag{27}$$

where χ^2_{obs} were the observed chi-squared statistics, while χ^2_{exp} were the expected chi-squared statistics under the null (this simplifies to using the chi-squared statistics for a value of 0.5, since the under the null, the p values are uniformly distributed). We quantified λ_{GC} for both the scenarios: when allowing SNP effects to vary as smooth function of age, and when performing a standard longitudinal GWAS without any interactions or smooth terms.

Results: examining unstructured covariances and discovery of age-dependent genetic effects

The stage-1 analysis was completed in ~3 minutes; examining the random effects unstructured covariance revealed dynamic patterns over time (**Figure 4**). Specifically, the proportion of within-visit variance attributed to genetic effects (within-visit heritability) for each phenotype showed differences across timepoints, ranging between 0.42–0.91 for length, 0.45–0.84 for weight, and 0.55–0.73 for BMI (diagonal entries, middle column, **Figure 4**). Interestingly, genetic correlations between timepoints (off-diagonal entries, middle column, **Figure 4)** also showed fluctuations over time with different patterns across phenotypes. For example, the correlations between birth and other timepoints for length and weight were fairly stable, while BMI showed a decaying pattern. The heritability was relatively lower for the first three timepoints for length and weight compared to other timepoints, while the heritability for BMI was relatively stable from 3m to 12m period. The family effect explained a small amount of variance for the first three timepoints for length and weight, but not as much for BMI. The variances explained by subject effect (and noise) also changed over time: a stronger effect for the first three timepoints for length but not as much for later timepoints, and a similar trend for weight and BMI but with greater variance explained for later timepoints.

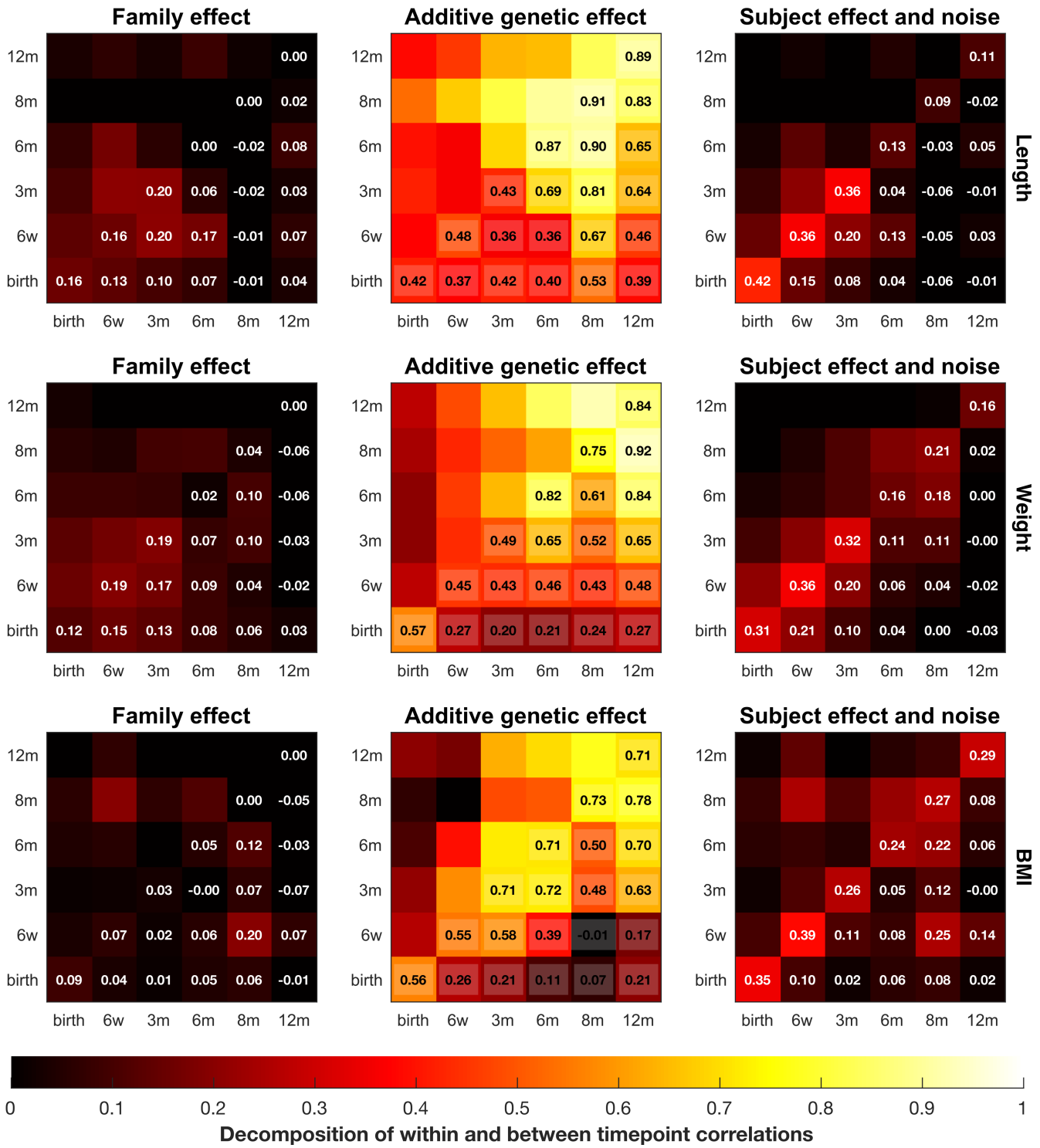

Figure 4: Decomposition of within and between timepoint correlations into family, additive genetic, and subject effects. The panels are the normalized unstructured covariance matrices for length, weight, and BMI in infants in the first year of life; the diagonal values are the variances explained by shared family environment, the additive genetic effect (heritability), and the subject effect (which also includes noise) for each timepoint while the off-diagonals reflect the correlations between time points.

The GWAS analysis, completed in ~16.88 hours, revealed 2,836, 4,367, and 6,707 SNPs showing a main effect or a time-dependent effect on length ($\lambda_{GC} = 1.10$), weight ($\lambda_{GC} = 0.91$), and BMI ($\lambda_{GC} = 1.36$) respectively at $\alpha = 5 \times 10^{-8}$; in comparison, longitudinal GWAS (without including splines or interactions; analysis completed in ~3.13 hours) revealed 2,126, 660, and 2,797 significant SNPs for length ($\lambda_{GC} = 1.25$), weight ($\lambda_{GC} = 1.13$), and BMI ($\lambda_{GC} = 1.21$) respectively (**Figure 5a**, **Figure 6a**, and **Figure**

7a). Here, the $\lambda_{GC} > 1$ is in line with other recent large-scale GWAS analyses and likely reflects the high polygenicity of the traits (see **Supplementary Figures S43-S45** for Q-Q plots). We used the SNP2GENE module of FUMA (35) to examine the number of independently significant SNPs and the number of identified genetic loci. For length, FUMA identified 134 independent SNPs and 54 loci when allowing the SNP effect to vary over time, as compared to 126 independent SNPs and 49 loci when modeling only the longitudinal main effect of SNPs; for weight, FUMA identified 201 independent SNPs and 46 loci when allowing the SNP effect to vary over time, as compared to 45 independent SNPs and 23 loci when modeling only the longitudinal main effect of SNPs; for BMI, FUMA identified 242 independent SNPs and 59 loci when allowing the SNP effect to vary over time, as compared to 124 independent SNPs and 44 loci when modeling only the longitudinal main effect of SNPs. We examined the fitted spline trajectories for the top 10 SNPs for each trait, one SNP per chromosome (**Figure 5b**, **Figure 6b**, and **Figure 7b**. For length, most of the effects were linear; however, the selected SNPs for weight and BMI showed more dynamic patterns with non-linear trajectories. Notably, even the same selected SNPs for weight and BMI showed differing patterns in the direction and timing of the effect changes across the first year of life (e.g., rs2767486 and rs13322435 (3)). These trajectories highlight the importance of modeling the time-dependent effects and reveal how the SNP effects change over time.

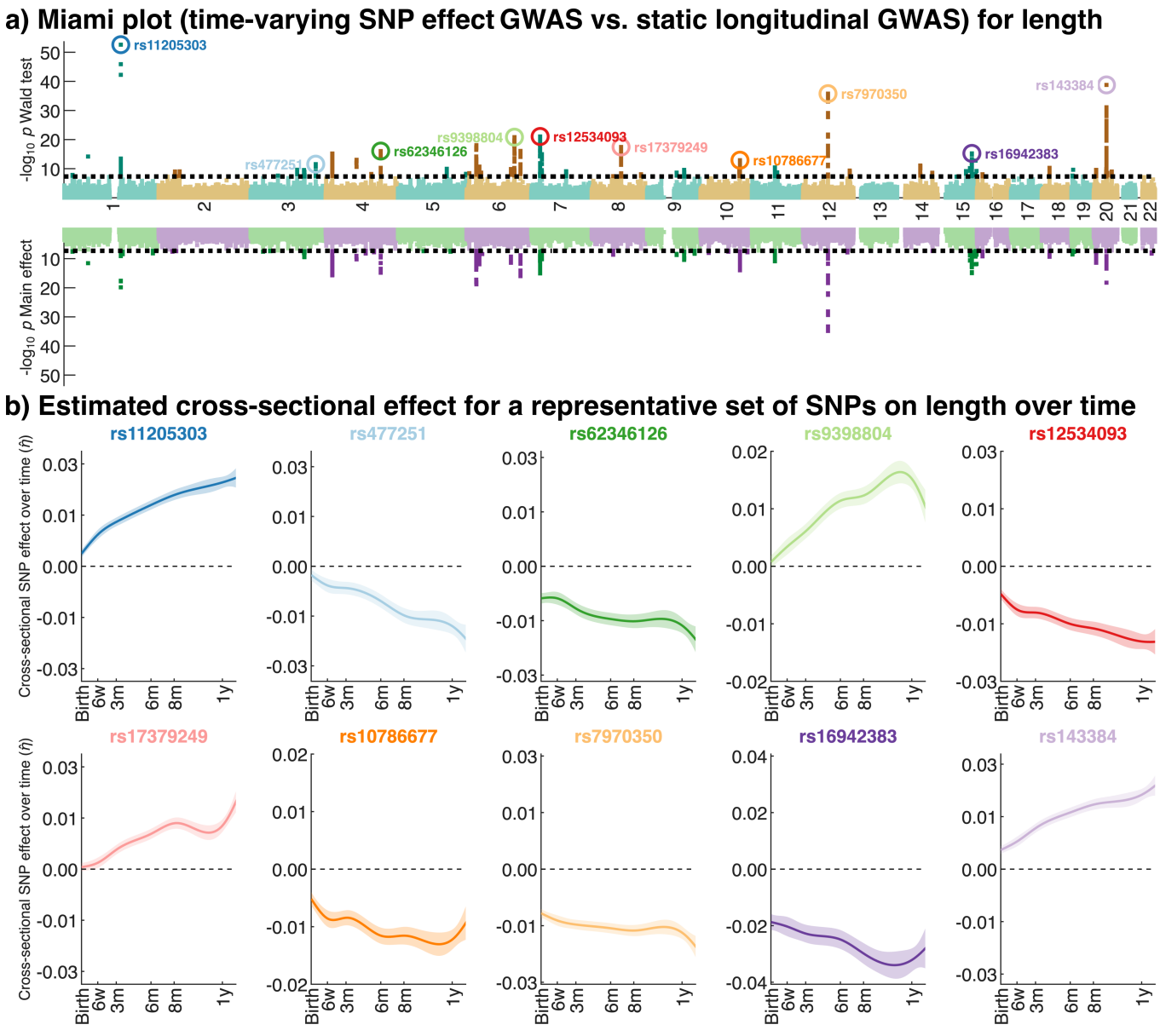

Figure 5: Miami plots comparing the omnibus effect of main and time-dependent effect of SNPs with the longitudinal main effect of SNPs on length of infants in the first year of life, and the estimated SNP effect over time for selected SNPs. a) the top part shows the $-\log_{10}(p)$-values for the Wald test while the lower part shows the $-\log_{10}(p)$-values for the longitudinal main effect; **b)** the solid lines show the cross-sectional effect of selected SNPs over time (SNPs marked in panel **a)**); the corresponding standard errors are shown by the band surrounding the solid lines.

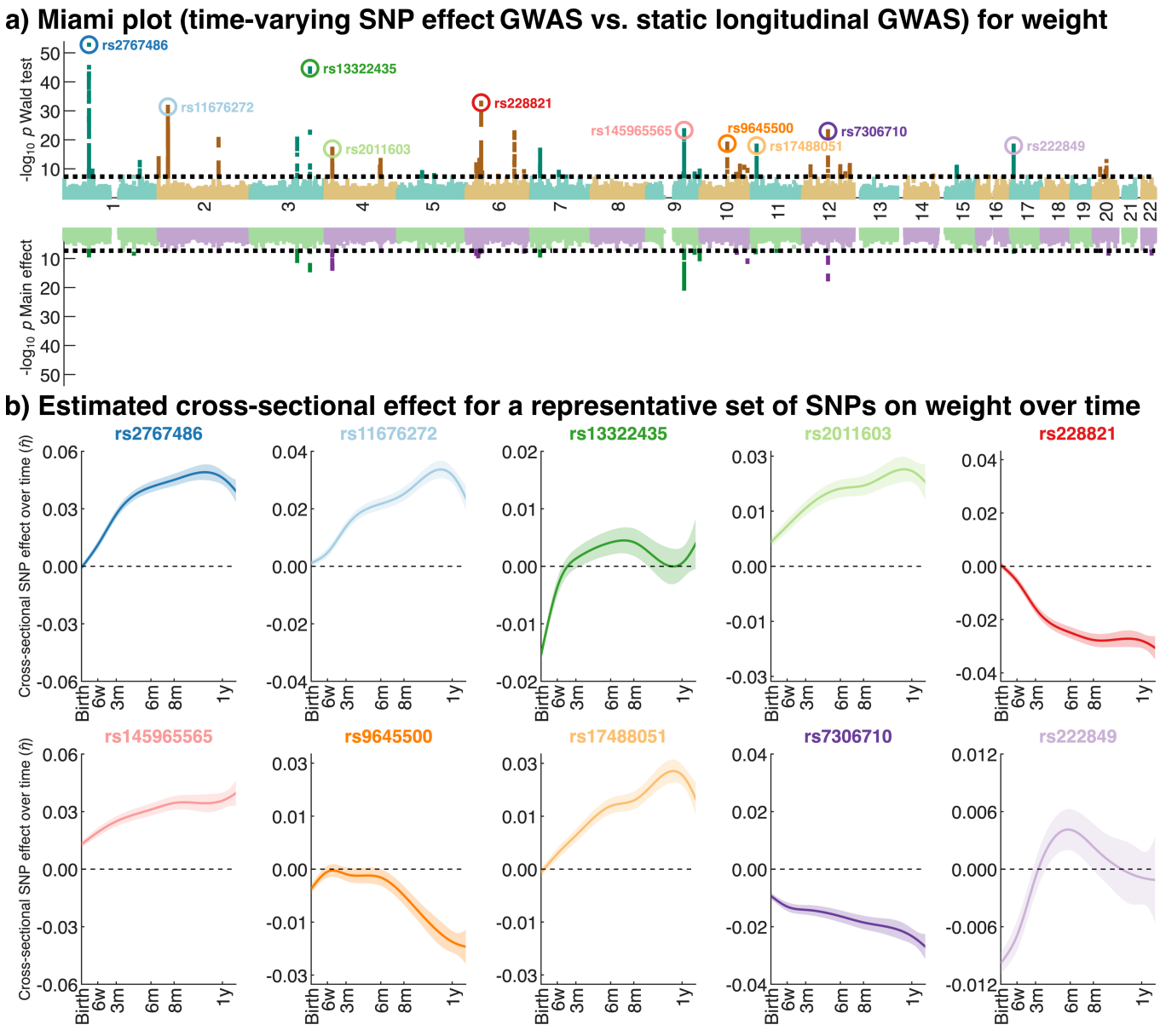

Figure 6: Miami plots comparing the omnibus effect of main and time-dependent effect of SNPs with the longitudinal main effect of SNPs on weight of infants in the first year of life, and the estimated SNP effect over time for selected SNPs. a) the top part shows the $-\log_{10}(p)$-values for the Wald test while the lower part shows the $-\log_{10}(p)$-values for the longitudinal main effect; **b)** the solid lines show the cross-sectional effect of selected SNPs over time (SNPs marked in panel **a)**); the corresponding standard errors are shown by the band surrounding the solid lines.

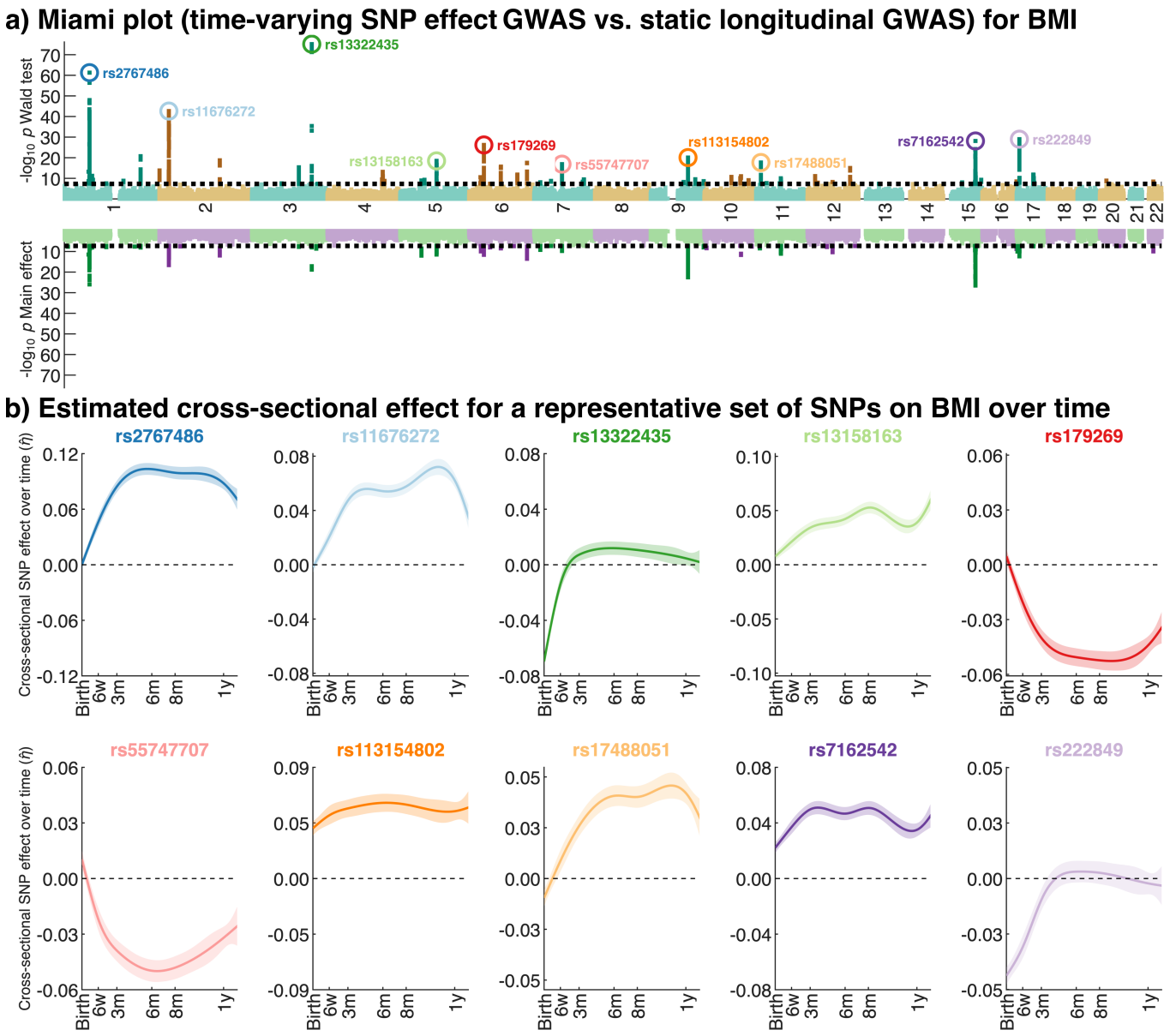

Figure 7: Miami plots comparing the omnibus effect of main and time-dependent effect of SNPs with the longitudinal main effect of SNPs on BMI of infants in the first year of life, and the estimated SNP effect over time for selected SNPs. a) the top part shows the $-\log_{10}(p)$-values for the Wald test while the lower part shows the $-\log_{10}(p)$-values for the longitudinal main effect; **b)** the solid lines show the cross-sectional effect of selected SNPs over time (SNPs marked in panel **a)**); the corresponding standard errors are shown by the band surrounding the solid lines.

Discussion

We have adapted and extended our computationally fast linear mixed-effects modeling method FEMA for large-scale high-dimensional longitudinal datasets. The features of FEMA-Long include: i) modeling time-varying random effects using unstructured covariance, ii) incorporating splines for modeling non-linear effects of variables, iii) examining time-dependent effects using spline interactions, and iv) performing GWAS-like analyses by scaling FEMA to estimate marginal effect of millions of variables, including time-dependent effects. Our simulation results show that FEMA-Long can accurately model these parameters, while controlling for false positives, and provides exceptional computational performance, tens to thousands of times faster than a standard LME implementation, and at the same time having a much smaller

carbon footprint. We have also demonstrated that FEMA-Long is scalable to scenarios with a large number of observations and outcome variables such as GWAS or whole-brain neuroimaging analyses.

As previously mentioned, random effects like heritability can vary significantly over time (3–8). Similarly, longitudinal data usually shows some form of time-varying correlation. Our application of FEMA-Long to length, weight, and BMI in infants during the first year of life revealed time-varying random effect covariances including heritability. These patterns also differed between similar phenotypes. Generally, increasing heritability for later timepoints (compared to birth) potentially indicates that the phenotypes at earlier timepoints might be more influenced from other sources like the maternal genotype (36,37). Importantly, it underscores the need to model these using unstructured covariance as these covariances cannot be well-approximated with simpler covariance patterns like compound symmetry or autoregressive models.

Additionally, when studying dynamic phenomenon such as development, it is essential to model time-dependent effects. Our time-dependent GWAS results revealed several SNPs showing time-dependent genetic effects. In contrast, the discovery of these SNPs was stunted when only examining the longitudinal main effect of SNPs, most notably for weight and BMI. Charting the effect of these SNPs over time revealed several interesting patterns including non-monotonic and curvilinear patterns. These results highlight the two-fold importance of modeling spline interactions for every SNP: first, allowing discovery of SNPs showing non-linear time-dependent effects; and second, ensuring more accurate predictions when using these estimated effect sizes (e.g., polygenic risk scores). Both these aspects cannot be captured by just the main effect of the SNPs and will likely be missed even with linear interaction.

The use of unstructured covariance has received relatively little attention in fields with high dimensional data. For example, within neuroimaging, there are no software packages for modeling unstructured covariances, besides AFNI's *3dLME* and *3dLMEr* (38) which supports compound symmetry and order 1 autoregressive process (9). In genetics, prior work on longitudinal GWAS include: i) computing a rate of change between timepoints (39–41); ii) using standard LME packages (42,43); iii) fitting a growth curve-like model to the phenotype and performing GWAS on the model parameters (44,45); and iv) novel tools for longitudinal GWAS like fGWAS (46), GALLOP (47), trajGWAS (48), L-BRAT and RGMMAT (49), RVMMAT (50), and SPA_{GRM} (51). Despite the availability of these tools, they have one or more limitations including inability to: i) handle related individuals, ii) model additional sources of variances, iii) model linear and/or non-linear SNP interactions, iv) estimating effect sizes, and v) scale to a large samples and/or large-number of phenotypes. Importantly, it is rare to see a longitudinal GWAS study specifying a time-varying covariance pattern. The closest such recent approach (52) used a range of LMEs with linear or cubic smoothing splines, or cubic slopes for age, with and without a continuous AR(1) covariance pattern. However, this work still relied on collapsing repeated observations into a single measurement (followed by

a meta-analysis) as well as removing related observations. In addition, the authors highlighted the significant computational burden when scaling to large cohorts.

FEMA-Long addresses all the above-mentioned limitations, providing a unified framework for performing large-scale analyses, modeling unstructured covariances, non-linear effects, and discovery of time-dependent effects, including time-dependent genetic effects. There are, however, some limitations to our work. First, the MoM estimator is not as efficient as maximum likelihood estimator. Second, currently FEMA-Long only supports continuous outcome variables, assuming approximately normally distributed errors. Third, for GWAS, we assume that the individual contribution of SNPs is small and unlikely to affect the covariance estimation of the random effect. Fourth, the computational time, especially for GWAS, could be further improved. Fifth, when performing the visit-wise estimation of the covariance pattern, for certain scenarios where the number of shared observations between the visits are too low, the method of moments estimation could be unstable. While FEMA and FEMA-Long are designed for large samples, we envision improving the stability of the covariance estimation by updating our method to perform a joint estimation of all parameters instead of visit-wise estimation; other possibilities include introducing a shrinkage or regularization on the estimation of the covariance parameters and/or enforcing smoothness during the estimation; additionally, we could interpolate the visit-wise estimated covariances, resulting in smoothly changing covariances, and further extending the scope of time-varying random effects. Future work will explore these and incorporate additional features like binning the phenotypes, computation of age-specific risk scores, extension to non-continuous phenotypes such as Bernoulli distribution, allowing statistical inference on the random effects estimates, using longitudinal trajectories for prediction, and the application of FEMA-Long to longitudinal complex trait phenotypes such as mental health.

To conclude, FEMA-Long is a powerful novel method that will enable researchers to apply large-scale LME models with unstructured covariances as well as discover time-dependent effects. The unified framework of FEMA-Long not only allows inclusion of subjects with a differing number of timepoints but also handles scenarios like sample relatedness and other sources of variance while providing a computationally fast and environmentally green solution for application to genomics, neuroimaging, and other fields. We expect FEMA-Long to facilitate discoveries in the growing number of large longitudinal datasets becoming available.

Acknowledgements

The Norwegian Mother, Father and Child Cohort Study is supported by the Norwegian Ministry of Health and Care Services and the Ministry of Education and Research. We are grateful to all the participating families in Norway who take part in this on-going cohort study. We thank the Norwegian Institute of Public Health (NIPH) for generating high-quality genomic data. This research is part of the HARVEST collaboration, supported by the Research Council of Norway (#229624). We also thank the NORMENT

Centre for providing genotype data, funded by the Research Council of Norway (#223273), South East Norway Health Authorities and Stiftelsen Kristian Gerhard Jebsen. We further thank the Center for Diabetes Research, the University of Bergen for providing genotype data and performing quality control and imputation of the data funded by the ERC AdG project SELECTionPREDISPOSED, Stiftelsen Kristian Gerhard Jebsen, Trond Mohn Foundation, the Research Council of Norway, the Novo Nordisk Foundation, the University of Bergen, and the Western Norway Health Authorities.

We acknowledge helpful suggestions from Drs. Gleda Kutrolli and Dominic Holland. We thank Dr. Elizabeth Corfield for leading the MoBaPsychGen quality control and imputation pipeline.

Some of the color schemes used for various figures were taken from ColorBrewer (https://colorbrewer2.org/). Some of the figures use the tight_subplot function (Pekka Kumpulainen (2023). tight_subplot(Nh, Nw, gap, marg_h, marg_w) (https://www.mathworks.com/matlabcentral/fileexchange/27991-tight_subplot-nh-nw-gap-marg_h-marg_w), MATLAB Central File Exchange. Retrieved April 5, 2023).

Competing interests

Dr. Anders M. Dale is a Founding Director and holds equity in CorTechs Labs, Inc. (DBA Cortechs.ai), Precision Pro, Inc., and Precision Health and Wellness, Inc. Dr. Dale is the President and a Board of Trustees member of the J. Craig Venter Institute (JCVI) and holds an appointment as Professor II at University of Oslo in Norway.. Dr. Andreassen has received speaker fees from Lundbeck, Janssen, Otsuka, Lilly, and Sunovion and is a consultant to Cortechs.ai. and Precision Health. Dr. O. Frei is a consultant to Precision Health. The other authors declare no competing interests.

Author contribution

Conceptualization: PP, NP, OF, AAS, TEN, OAA, AMD

Data curation: PP, EF, MV, PJ, VB, NRB, OF

Formal analysis: PP

Funding acquisition: PP, OAA, AMD

Investigation: PP, EF

Methodology: PP, NP, DP, WKT, OF, AAS, TEN, AMD

Project administration: OAA, AMD

Resources: IES, TLJ, OAA, AMD

Software: PP, DP, EF, DMS, AR, CCF, JK, OF, AAS, TEN, AMD

Supervision: OF, AAS, TEN, OAA, AMD

Visualization: PP, NP, PJ, TEN, AMD

Writing – original draft: PP

Writing – review & editing: PP, NP, DP, EF, MV, DMS, AR, PJ, IES, VB, NRB, CCF, CM, JK, RL, DJH, DM, SJ, PRN, TLJ, WKT, OF, AAS, TEN, OAA, AMD

Data and code availability

FEMA-Long is available on GitHub at: https://github.com/cmig-research-group/cmig_tools. Data from the Norwegian Mother, Father and Child Cohort Study is managed by the Norwegian Institute of Public Health. Access requires approval from the Regional Committees for Medical and Health Research Ethics (REC), compliance with GDPR, and data owner approval. Participant consent does not allow individual-level data storage in repositories or journals. Researchers seeking access for replication must apply via www.helsedata.no. The synthetic data used for simulations can be generated using the code provided here: https://github.com/parekhpravesh/FEMA-Long.

Funding

PP acknowledges funding and salary support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant 801133. **PP** additionally acknowledges salary support from the Research Council of Norway grant 324252, the National Institutes of Health grants U24DA041123 and U24DA055330, and the Wellcome Leap, CARE Program ("FEMA-AD"). **NP** is supported by the Research Council of Norway grant 324252. **EF** was supported by the Research Council of Norway (RCN) (#324499). **MV** is supported by the Research Council of Norway (grant #301178), the European Research Council (grant #101171420), and the University of Bergen. **DS** was supported by the National Institute on Drug Abuse with a Ruth L. Kirschstein Individual Predoctoral NRSA award (5F30DA057078-02) and the National Institute of General Medical Sciences of the National Institutes of Health under award T32GM154642 (UC San Diego Medical Scientist Training Program). **PJ** is supported by the European Union's Horizon 2020 Research and Innovation Programme (#964874; RealMent). **IES** is supported by Southern and Eastern Norway Regional Health Authority (#2020060, #2025037) and NIH R01MH129858. **VB** is supported by European Union's Horizon 2020 Research and Innovation Programme (RealMent, Grant No. 964874). **NRB** was supported by the Research Council of Norway (RCN)(#271555/F21) and European Union's Horizon 2020 Research and Innovation Programme (RealMent, Grant No. 964874). **CM** is supported by the National Institute of Mental Health (R00MH132886) and the Brain Behavior Research Foundation (#31876). **JK** was supported by a Marie Skłodowska-Curie Postdoctoral Fellowship under the European Union's Horizon Europe research and

innovation programme (Grant Agreement No. 101150746). **DM** is supported by the Research Council of Norway (RCN) 324252. **SJ** is supported by Helse Vest's Open Research Grant (grants #912250 and F-12144), the Novo Nordisk Foundation (grant NNF20OC0063872) and the Research Council of Norway (grant #315599). **PRN** was supported by grants from the European Research Council (AdG #293574), Stiftelsen Trond Mohn Foundation (Mohn Center of Diabetes Precision Medicine), the University of Bergen, Haukeland University Hospital, the Research Council of Norway (FRIPRO grant #240413), and the Novo Nordisk Foundation (grant #54741). **OF** acknowledges funding support from the South-Eastern Norway Regional Health Authority (#2022073) and the Research Council of Norway (#324499). **AAS** is supported by the Research Council of Norway (grant #326813). **TEN** acknowledges funding from the National Institutes of Health grant 5U24DA041123. **AMD** acknowledges funding from the National Institutes of Health grants U24DA041123, R01AG076838, U24DA055330, and OT2HL161847. Parts of this work used the TSD (Tjeneste for Sensitive Data) facilities, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo, IT-Department (USIT, tsd-drift@usit.uio.no), using resources provided by UNINETT Sigma2 – the National Infrastructure for High Performance Computing and Data Storage in Norway (NS9703S). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Denny JC, Collins FS. Precision medicine in 2030—seven ways to transform healthcare. Cell. 2021 Mar 18;184(6):1415–9. doi:10.1016/j.cell.2021.01.015 PubMed PMID: 33740447.

2. Atlas of Longitudinal Datasets [Internet]. 2024 [cited 2025 Feb 19]. Available from: https://atlaslongitudinaldatasets.ac.uk/

3. Helgeland Ø, Vaudel M, Sole-Navais P, Flatley C, Juodakis J, Bacelis J, et al. Characterization of the genetic architecture of infant and early childhood body mass index. Nat Metab. 2022 Mar;4(3):344–58. doi:10.1038/s42255-022-00549-1

4. Jr TJB. The Wilson Effect: The Increase in Heritability of IQ With Age. Twin Research and Human Genetics. 2013 Oct;16(5):923–30. doi:10.1017/thg.2013.54

5. Brown WM, Beck SR, Lange EM, Davis CC, Kay CM, Langefeld CD, et al. Age-stratified heritability estimation in the Framingham Heart Study families. BMC Genetics. 2003 Dec 31;4(1):S32. doi:10.1186/1471-2156-4-S1-S32

6. Akingbuwa WA, Hammerschlag AR, Bartels M, Middeldorp CM. Systematic Review: Molecular Studies of Common Genetic Variation in Child and Adolescent Psychiatric Disorders. Journal of the American Academy of Child & Adolescent Psychiatry. 2022 Feb 1;61(2):227–42. doi:10.1016/j.jaac.2021.03.020 PubMed PMID: 33932494.

7. Jelenkovic A, Sund R, Hur YM, Yokoyama Y, Hjelmborg J v B, Möller S, et al. Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. Sci Rep. 2016 Jun 23;6(1):28496. doi:10.1038/srep28496

8. Silventoinen K, Pietiläinen KH, Tynelius P, Sørensen TIA, Kaprio J, Rasmussen F. Genetic regulation of growth from birth to 18 years of age: The Swedish young male twins study. American Journal of Human Biology. 2008;20(3):292–8. doi:10.1002/ajhb.20717

9. van der Horn HJ, Erhardt EB, Dodd AB, Nathaniel U, Wick TV, McQuaid JR, et al. A cautionary tale on the effects of different covariance structures in linear mixed effects modeling of fMRI data. Human Brain Mapping. 2024;45(7):e26699. doi:10.1002/hbm.26699

10. Parekh P, Fan CC, Frei O, Palmer CE, Smith DM, Makowski C, et al. FEMA: Fast and efficient mixed-effects algorithm for large sample whole-brain imaging data. Human Brain Mapping. 2024;45(2):e26579. doi:10.1002/hbm.26579

11. Zou T, Wei L, Hansheng W, Tsai CL. Covariance Regression Analysis. Journal of the American Statistical Association. 2017 Jan 2;112(517):266–81. doi:10.1080/01621459.2015.1131699

12. Frisch R, Waugh FV. Partial Time Regressions as Compared with Individual Trends. Econometrica. 1933;1(4):387–401. doi:10.2307/1907330

13. Lovell MC. Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis. Journal of the American Statistical Association. 1963 Dec 1;58(304):993–1010. doi:10.1080/01621459.1963.10480682

14. Higham NJ. Computing a nearest symmetric positive semidefinite matrix. Linear Algebra and its Applications. 1988;103:103–18. doi:https://doi.org/10.1016/0024-3795(88)90223-6

15. D'Errico J. nearestSPD [MATLAB] [Internet]. MATLAB Central File Exchange; [cited 2026 Mar 16]. Located at: MATLAB Central File Exchange. Available from: https://www.mathworks.com/matlabcentral/fileexchange/42885-nearestspd

16. Wang W, Yan J. splines2: Regression Spline Functions and Classes [Internet]. 2024. Available from: https://CRAN.R-project.org/package=splines2

17. Wang W, Yan J. Shape-Restricted Regression Splines with R Package splines2. Journal of Data Science. 2021;19(3):498–517. doi:10.6339/21-JDS1020

18. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2024. Available from: https://www.R-project.org/

19. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis [Internet]. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2011 [cited 2023 Mar 23]. (Wiley Series in Probability and Statistics). Available from: https://www.wiley.com/en-ae/Applied+Longitudinal+Analysis%2C+2nd+Edition-p-9780470380277

20. Gałecki A, Burzykowski T. Linear Mixed-Effects Models Using R: A Step-by-Step Approach [Internet]. New York, NY: Springer; 2013 [cited 2025 Apr 24]. (Springer Texts in Statistics). Available from: https://link.springer.com/10.1007/978-1-4614-3900-4 doi:10.1007/978-1-4614-3900-4

21. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience. 2019 Jul 1;8(7):giz082. doi:10.1093/gigascience/giz082

22. Basu D. The Yule-Frisch-Waugh-Lovell Theorem for Linear Instrumental Variables Estimation [SSRN Scholarly Paper] [Internet]. Rochester, NY; 2023 [cited 2024 Jul 16]. Available from: https://papers.ssrn.com/abstract=4514656 doi:10.2139/ssrn.4514656

23. Krishnakumar J. Chapter 5 Time Invariant Variables and Panel Data Models: A Generalised Frisch–Waugh Theorem and its Implications. In: Baltagi BH, editor. Contributions to Economic Analysis [Internet]. Elsevier; 2006 [cited 2024 Jul 16]. p. 119–32. (Panel Data Econometrics). Available from: https://www.sciencedirect.com/science/article/pii/S0573855506740056 doi:10.1016/S0573-8555(06)74005-6

24. Ding P. The Frisch–Waugh–Lovell theorem for standard errors. Statistics & Probability Letters. 2021 Jan 1;168:108945. doi:10.1016/j.spl.2020.108945

25. The MathWorks Inc. MATLAB version: 9.14.0 (R2023a) [Internet]. Natick, Massachusetts, United States: The MathWorks Inc.; 2023. Available from: https://www.mathworks.com

26. Brooks ME, Kristensen K, Benthem KJ van, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. The R Journal. 2017;9(2):378–400.

27. Kristensen K, McGillycuddy M. Covariance structures with glmmTMB [Internet]. 2024 [cited 2025 Feb 6]. The AR(1) covariance structure. Available from: https://cran.r-project.org/web/packages/glmmTMB/vignettes/covstruct.html#the-unstructured-covariance

28. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015 Oct 7;67:1–48. doi:10.18637/jss.v067.i01

29. Lannelongue L, Grealey J, Inouye M. Green Algorithms: Quantifying the Carbon Footprint of Computation. Advanced Science. 2021 Jun 1;8(12):2100707. doi:10.1002/advs.202100707

30. Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). International Journal of Epidemiology. 2016 Apr 1;45(2):382–8. doi:10.1093/ije/dyw029

31. Magnus P, Irgens LM, Haug K, Nystad W, Skjærven R, Stoltenberg C, et al. Cohort profile: The Norwegian Mother and Child Cohort Study (MoBa). International Journal of Epidemiology. 2006 Oct 1;35(5):1146–50. doi:10.1093/ije/dyl170

32. Brandlistuen RE, Kristjansson D, Alsaker E, Valen R, Birkeland E, Røyrvik EC, et al. Cohort Profile Update: The Norwegian Mother, Father and Child Cohort (MoBa). International Journal of Epidemiology. 2025 Aug;54(5):dyaf139. doi:10.1093/ije/dyaf139

33. Paltiel L, Haugan Anita, Skjerden T, Harbak K, Bækken S, Kristin SN, et al. The biobank of the Norwegian Mother and Child Cohort Study – present status. Norsk Epidemiologi. 2014 Dec;24(1–2):29–35. doi:10.5324/nje.v24i1-2.1755

34. Corfield EC, Shadrin AA, Frei O, Rahman Z, Lin A, Athanasiu L, et al. The Norwegian Mother, Father, and Child cohort study (MoBa) genotyping data resource: MoBaPsychGen pipeline v.1 [Internet]. bioRxiv; 2024 [cited 2024 Jul 17]. p. 2022.06.23.496289. Available from: https://www.biorxiv.org/content/10.1101/2022.06.23.496289v4 doi:10.1101/2022.06.23.496289

35. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017 Nov 28;8(1):1826. doi:10.1038/s41467-017-01261-5

36. Warrington NM, Beaumont RN, Horikoshi M, Day FR, Helgeland Ø, Laurin C, et al. Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. Nat Genet. 2019 May;51(5):804–14. doi:10.1038/s41588-019-0403-1

37. Beaumont RN, Warrington NM, Cavadino A, Tyrrell J, Nodzenski M, Horikoshi M, et al. Genome-wide association study of offspring birth weight in 86 577 women identifies five novel loci and highlights maternal genetic effects that are independent of fetal genetics. Human Molecular Genetics. 2018 Feb 15;27(4):742–56. doi:10.1093/hmg/ddx429

38. Chen G, Saad ZS, Britton JC, Pine DS, Cox RW. Linear mixed-effects modeling approach to FMRI group analysis. NeuroImage. 2013 Jun 1;73:176–90. doi:10.1016/j.neuroimage.2013.01.047

39. Brouwer RM, Klein M, Grasby KL, Schnack HG, Jahanshad N, Teeuw J, et al. Genetic variants associated with longitudinal changes in brain structure across the lifespan. Nat Neurosci. 2022 Apr;25(4):421–32. doi:10.1038/s41593-022-01042-4

40. Gorski M, Rasheed H, Teumer A, Thomas LF, Graham SE, Sveinbjornsson G, et al. Genetic loci and prioritization of genes for kidney function decline derived from a meta-analysis of 62 longitudinal genome-wide association studies. Kidney International. 2022 Sep 1;102(3):624–39. doi:10.1016/j.kint.2022.05.021 PubMed PMID: 35716955.

41. Loomis SJ, Sadhu N, Fisher E, Gafson AR, Huang Y, Yang C, et al. Genome-wide study of longitudinal brain imaging measures of multiple sclerosis progression across six clinical trials. Sci Rep. 2023 Aug 31;13(1):14313. doi:10.1038/s41598-023-41099-0

42. Jin H, Kwak SH, Yoon JW, Lee S, Park KS, Won S, et al. Genome-Wide Association Study on Longitudinal Change in Fasting Plasma Glucose in Korean Population. Diabetes Metab J. 2023 Jan 19;47(2):255–66. doi:10.4093/dmj.2021.0375

43. Wendel B, Papiol S, Andlauer TFM, Zimmermann J, Wiltfang J, Spitzer C, et al. A genome-wide association study of the longitudinal course of executive functions. Transl Psychiatry. 2021 Jul 10;11(1):1–8. doi:10.1038/s41398-021-01510-8

44. Bradfield JP, Kember RL, Ulrich A, Balkhiyarova Z, Alyass A, Aris IM, et al. Trans-ancestral genome-wide association study of longitudinal pubertal height growth and shared heritability with adult health outcomes. Genome Biology. 2024 Jan 16;25(1):22. doi:10.1186/s13059-023-03136-z

45. Couto Alves A, De Silva NMG, Karhunen V, Sovio U, Das S, Taal HR, et al. GWAS on longitudinal growth traits reveals different genetic factors influencing infant, child, and adult BMI. Science Advances. 2019 Sep 4;5(9):eaaw3095. doi:10.1126/sciadv.aaw3095

46. Das K, Li J, Wang Z, Tong C, Fu G, Li Y, et al. A dynamic model for genome-wide association studies. Hum Genet. 2011 Jun 1;129(6):629–39. doi:10.1007/s00439-011-0960-6

47. Sikorska K, Lesaffre E, Groenen PJF, Rivadeneira F, Eilers PHC. Genome-wide Analysis of Large-scale Longitudinal Outcomes using Penalization —GALLOP algorithm. Sci Rep. 2018 May 1;8(1):6815. doi:10.1038/s41598-018-24578-7

48. Ko S, German CA, Jensen A, Shen J, Wang A, Mehrotra DV, et al. GWAS of longitudinal trajectories at biobank scale. The American Journal of Human Genetics. 2022 Mar 3;109(3):433–45. doi:10.1016/j.ajhg.2022.01.018 PubMed PMID: 35196515.

49. Wu W, Wang Z, Xu K, Zhang X, Amei A, Gelernter J, et al. Retrospective Association Analysis of Longitudinal Binary Traits Identifies Important Loci and Pathways in Cocaine Use. Genetics. 2019 Dec 1;213(4):1225–36. doi:10.1534/genetics.119.302598

50. Xu G, Amei A, Wu W, Liu Y, Shen L, Oh EC, et al. Retrospective varying coefficient association analysis of longitudinal binary traits: Application to the identification of genetic loci associated with hypertension. The Annals of Applied Statistics. 2024 Mar;18(1):487–505. doi:10.1214/23-AOAS1798

51. Xu H, Ma Y, Xu L lin, Li Y, Liu Y, Li Y, et al. SPAGRM: effectively controlling for sample relatedness in large-scale genome-wide association studies of longitudinal traits. Nat Commun. 2025 Feb 6;16(1):1413. doi:10.1038/s41467-025-56669-1

52. Burrows K, Heiskala A, Bradfield JP, Balkhiyarova Z, Ning L, Boissel M, et al. A framework for conducting GWAS using repeated measures data with an application to childhood BMI. Nat Commun. 2024 Nov 20;15(1):10067. doi:10.1038/s41467-024-53687-3