Published in final edited form as:

J Genet Genomics. 2018 July 20; 45(7): 411–413. doi:10.1016/j.jgg.2018.06.006.

fGWAS: an R package for genome-wide association analysis with longitudinal phenotypes

Zhong Wang^{a,b,*}, Nating Wang^a, Rongling Wu^{c,d}, and Zuoheng Wang^{e,*}

^aCollege of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, China

^bBaker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14850, USA

^cCenter for Computational Biology, Beijing Forestry University, Beijing 100083, China

^dCenter for Statistical Genetics, Department of Public Health Sciences, Pennsylvania State College of Medicine, Hershey, PA 17033, USA

^eDepartment of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA

The past two decades have witnessed a revolution in identifying genetic risk factors underlying diseases and complex traits using genome-wide association studies (GWAS) (Risch and Merikangas, 1996; Hirschhorn and Daly, 2005; Altshuler et al., 2008). Together with advanced high-throughput technologies for genotyping and sequencing, GWAS have discovered thousands of susceptibility loci for various traits (Welter et al., 2014). The progress has greatly enhanced our knowledge on genetic architecture of complex diseases such as heritability estimation, genetic correlation between diseases, the number of loci and their effect sizes underlying phenotypic variation, as well as gene-gene and gene-environment interactions. Researchers can prioritize genes or genomic regions on the basis of GWAS results for further investigation, and ultimately identify causal functional variants to elucidate disease mechanisms.

Some genetic epidemiological studies have been conducted in cohorts with participants followed over time and repeated measures collected on the trait of interest. Such studies allow for exploration of genetic and environmental risk factors that influence the development of trait over time, and also allow for incorporation of time-varying covariates that are not available in case-control or cross-sectional phenotype data (Wu and Lin, 2006; Fan et al., 2012). However, longitudinal data analysis can be challenging due to the non-independence among repeated measurements from the same subject. Advanced statistical methods have been developed to account for correlation to ensure valid inference, such as random effects models, hierarchical linear models, empirical Bayes models, and growth mixture models. Association analyses that incorporate these methods have been applied to longitudinal GWAS (Das et al., 2011; Furlotte et al., 2012; Wang et al., 2012; Londono et al., 2013; Meirelles et al., 2013; Sikorska et al., 2013). Results from these studies generally

*Corresponding authors. zhongwang@bjfu.edu.cn (Z. Wang), zuoheng.wang@yale.edu (Z. Wang).

confirm that association tests that make full use of repeated measures increase statistical power over the approaches that collapse data into one single value such as average, baseline, or last observation.

For dynamic traits with values varying over time (e.g., human growth), incorporating knowledge of the trait development can boost statistical power in gene mapping (Wu and Lin, 2006). Functional GWAS (*f*GWAS) use mathematical or statistical functions to describe biological principle of trait formation in the GWAS framework (Luo et al., 2010; Das et al., 2011). This method capitalizes on the cumulative phenotypic variation over time to identify association through a dynamic interaction between genes and developmental patterns. It is useful to detect difference in overall growth curves among individuals at different phases of development. In addition, the *f*GWAS model can accommodate sparse longitudinal data collected at subject-specific irregular time points, and thus increases its applicability in clinical genomic studies. Built on the *f*GWAS model, Li et al. (2015) implemented a Bayesian group lasso variable selection method in the setting of high-dimensional single nucleotide polymorphism (SNP) predictors to improve gene detection and genetic effect estimation. More recently, Ning et al. (2017) extended the *f*GWAS model to capture time-varying polygenic effect of complex traits using random regression model through treating SNPs as covariates or factors. Therefore, *f*GWAS is a very flexible modeling framework in association analysis of longitudinal traits. It describes temporal mechanisms of growth curves using parametric, nonparametric, or semiparametric approach. *f*GWAS also employs covariance matrices to capture the correlation structure of trait values measured at different time points on the same subject. However, we note that the choice of mean and covariance structures may impact power in gene identification. In this study, we develop a *f*GWAS package, which performs a genome scan of association between SNPs and a longitudinal trait based on the *f*GWAS model. It is written in the statistical software R and distributed through GitHub that is publicly available under an open source software license GPL version 3.

The *f*GWAS package implements a complete pipeline that includes modules for data import, SNP scanning, result summary, and data visualization. It takes genotype and phenotype files as an input. Users can supply a covariate file if the analysis is adjusted for additional variables such as age, gender, and body mass index (BMI). The *f*GWAS model assumes that the trait values follow a specific mathematic function, thus it requires specification of the trait curve to describe the shape of the phenotypic trajectory. It also requires a covariance structure to describe the correlation of phenotypes measured at different time points. The package currently provides 9 curves to model the mean vector of trait values, including parametric curves such as logistic, bi-logistic, exponential, bi-exponential, power function, and the Emax model, and nonparametric curves based on Legendre orthogonal polynomials up to the 4th order. Users have their choice of curves for specific area of data application. For example, the Emax model is commonly used to describe the pharmacological activity of drugs, and the logistic curve has been used to model human growth. The package also includes a comprehensive set of 13 covariance structures such as the first-order autoregressive [AR(1)] model, the autoregressive moving average (ARMA) model, the first-order structured antedependence [SAD(1)] model (Jaffrezic et al., 2003), compound symmetry, and the Toeplitz matrix. It allows users to pick a specific correlation structure in

data analysis. When there is no prior information available on the shape of the trait curve and/or the covariance structure, the package automatically selects the best combination of the curve type and covariance matrix for a given phenotypic dataset using the maximum likelihood estimation (MLE) method followed by model selection via the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). Then, MLEs of parameters in the mean curve and covariance matrix estimated under the null model of no association are fed to the next SNP scanning step as the initial values. Their MLEs will be obtained via numerical iterations under the alternative model of association. The coefficient of determination R^2 is used to assess if there is any lack of fit in the estimated curves. Users are able to check the R^2 value and visually determine the curve fitting by comparing the fitted curve with the raw phenotypic data. However, users need to be aware of the high cost of computational time when fitting more complex mean and covariance structures in the analysis. The p -value of each SNP is obtained using the likelihood ratio (LR) test.

The *f*GWAS package outputs SNP information and parameter estimates in the mean curves and covariance matrices under both the null and alternative hypotheses. It sorts SNPs according to their p -values. Bonferroni correction and false discovery rate (FDR) are built into the program to adjust for multiple comparisons. A Manhattan plot is generated to display the $-\log_{10}(p\text{-value})$ obtained from the *f*GWAS model along the genomic coordinates. The package also plots the fitted curves in different colors based on the parameter values estimated by the *f*GWAS model in each genotype group, and overlays the curves with the raw phenotypic data. We can visually capture genetic contribution to the developmental patterns over the entire time course at each SNP. Furthermore, the package includes a simulation module to evaluate power of the *f*GWAS model under different parameter settings. We have performed extensive simulations to test the applicability of the package to various datasets obtained from many application areas.

To illustrate the utility of the *f*GWAS package, we apply it to a whole-genome analysis of longitudinal BMI data from the Framingham Heart Study (FHS). FHS is a multigenerational, longitudinal study of risk factors for cardiovascular disease (Splansky et al., 2007). We analyze the BMI data on a set of 2104 unrelated individuals from the first two generation cohorts (original and offspring cohorts). The individuals in the original cohort have measurements at up to 28 time points, and the individuals in the offspring cohort have measurements at up to 8 time points. The time points on an individual can be irregularly spaced, and the number of time points varies across subjects. All samples were genotyped on the Affymetrix 500K array. After quality control, there are 728 males and 950 females who have both genotype and phenotype information. Their ages span from 19 to 86 years old, and the number of BMI measurements on each subject ranges from 3 to 22. We perform a genome-wide longitudinal association analysis of BMI on 349,769 SNPs whose minor allele frequency (MAF) is $\geq 5\%$. *f*GWAS launches the automatic selection procedure and picks the Legendre polynomial of order 2 to fit the BMI trajectory. The AR(1) model is used as the covariance structure. In the Manhattan plot (Fig. 1A), there are 8 SNPs that reach the genome-wide significance level of 0.05. Among them, the SNP rs7806587 on chromosome 7 is in an intron of the E twenty-six Variant 1 (*ETVI*) gene that regulates target genes modulating biological processes like cell growth, angiogenesis, migration, proliferation and differentiation. It was reported that ETV-dependent genes were significantly enriched in

human diabetes and obesity GWAS (Suriben et al., 2015). Fig. 1B demonstrates the fitted BMI trajectories in the three genotype groups *CC*, *CT* and *TT* at rs7806587. We can see that BMI increases with age until around 65 years old and then drops afterwards. Individuals who carry genotype *TT* on average have a lower BMI in their 20–30 years old and a higher BMI in their later life compared to those with genotypes *CC* and *CT*.

In human genome, gene expression profiles change across life span. These changes will exert substantial influence on human traits. In the *f*GWAS framework, trait values are viewed as a developmental curve influenced by SNP genotypes. Once the growth curve of each genotype is estimated, the additive and dominant effects of the SNP in a time course can be obtained. We can further access the developmental landmarks on growth curves such as maximum velocity, maximum acceleration of growth, minimum deceleration of growth, and time lapse between different genotypes. This landmark information is usually obtained from the first and/or second derivatives of the curve function through differential equations. The timing of these developmental landmarks enables us to better understand the shape and process of the trait growth (Wang et al., 2014).

In the *f*GWAS package, we use LR test to assess the association between SNP and trait trajectory. Based on our own experience, we find that MLEs of parameters in the mean curves and covariance matrices can be unstable when testing association on SNPs with low MAF, leading to a very large LR test statistic and potentially a false positive result. This may partly be explained by the small sample size in certain genotype groups. Furthermore, when trait values deviate substantially from a normal distribution, the *f*GWAS model does not perform well because statistical inference is based on the assumption of a normal distribution of trait values. Therefore, we recommend users to conduct quality control to remove low-frequency variants and check trait distribution. If trait values do not resemble a normal distribution, data transformation is needed to achieve normality before performing the *f*GWAS analysis.

Current version of the *f*GWAS package has limitations that can potentially open for future improvement. First, it only accommodates time-varying genetic effects in the *f*GWAS model. We can expand the package to allow for adjustment of other time-varying effects from covariates such as gender and race. We can also incorporate gene-environment interactions that may fluctuate over time to impact trait values. Second, the *f*GWAS model is designed for single marker association analysis. As we have understood that complex traits are influenced by multiple genes, it is important to consider the joint effects of multiple markers. Thus, users can select the top SNPs from the *f*GWAS analysis and conduct additional analyses. For example, the package *gwas.lasso* that we are currently developing based on a Bayesian group lasso variable selection method (Li et al., 2015) can be used to consider the polygenic effects from the top SNPs. The longitudinal sequence kernel association test (LSKAT) we developed can also be used to test multiple variants, rare or common, in a genomic region for association with a longitudinal trait (Wang et al., 2017).

*f*GWAS: <https://github.com/wzhy2000/fGWAS>.

Acknowledgments

This study was supported by the Fundamental Research Funds for the Central Universities (BLX2013026), the National Natural Science Foundation of China (No. 31470675), and the National Institutes of Health (K01AA023321). The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University.

References

- Altshuler D, Daly MJ, Lander ES, 2008 Genetic mapping in human disease. *Science* 322, 881–888. [PubMed: 18988837]
- Das K, Li J, Wang Z, Tong C, Fu G, Li Y, Xu M, Ahn K, Mauger D, Li R, Wu R, 2011 A dynamic model for genome-wide association studies. *Hum. Genet* 129, 629–639. [PubMed: 21293879]
- Fan R, Albert PS, Schisterman EF, 2012 A discussion of gene-gene and gene-environment interactions and longitudinal genetic analysis of complex traits. *Stat. Med* 31, 2565–2568. [PubMed: 22969024]
- Furlotte NA, Eskin E, Eyheramendy S, 2012 Genome-wide association mapping with longitudinal data. *Genet. Epidemiol* 36, 463–471. [PubMed: 22581622]
- Hirschhorn JN, Daly MJ, 2005 Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet* 6, 95–108. [PubMed: 15716906]
- Jaffrezic F, Thompson R, Hill WG, 2003 Structured antedependence models for genetic analysis of repeated measures on multiple quantitative traits. *Genet. Res* 82, 55–65. [PubMed: 14621271]
- Li J, Wang Z, Li R, Wu R, 2015 Bayesian group LASSO for nonparametric varying-coefficient models with application to functional genome-wide association studies. *Ann. Appl. Stat* 9, 640–664. [PubMed: 26478762]
- Londono D, Chen KM, Musolf A, Wang R, Shen T, Brandon J, Herring JA, Wise CA, Zou H, Jin M, Yu L, Finch SJ, Matisse TC, Gordon D, 2013 A novel method for analyzing genetic association with longitudinal phenotypes. *Stat. Appl. Genet. Mol. Biol* 12, 241–261. [PubMed: 23502345]
- Luo J, Berg A, Ahn K, Das K, Li J, Wang Z, Li Y, Wu R, 2010 Functional genome-wide association studies of longitudinal traits In: Pong A, Chow SC (Ed.), *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*. Chapman and Hall/CRC, chp 23, pp. 1–13.
- Meirelles OD, Ding J, Tanaka T, Sanna S, Yang HT, Dudekula DB, Cucca F, Ferrucci L, Abecasis G, Schlessinger D, 2013 SHAVE: shrinkage estimator measured for multiple visits increases power in GWAS of quantitative traits. *Eur. J. Hum. Genet* 21, 673–679. [PubMed: 23092954]
- Ning C, Kang H, Zhou L, Wang D, Wang H, Wang A, Fu J, Zhang S, Liu J, 2017 Performance gains in genome-wide association studies for longitudinal traits via modeling time-varied effects. *Sci. Rep* 7, 590. [PubMed: 28377602]
- Risch N, Merikangas K, 1996 The future of genetic studies of complex human diseases. *Science* 273, 1516–1517. [PubMed: 8801636]
- Sikorska K, Rivadeneira F, Groenen PJ, Hofman A, Uitterlinden AG, Eilers PH, Lesaffre E, 2013 Fast linear mixed model computations for genome-wide association studies with longitudinal data. *Stat. Med* 32, 165–180. [PubMed: 22911890]
- Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, D'Agostino RB Fox CS, Larson MG, Murabito JM, O'Donnell CJ, Vasan RS, Wolf PA, Levy D, 2007 The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am. J. Epidemiol* 165, 1328–1335. [PubMed: 17372189]
- Suriben R, Kaihara KA, Paolino M, Reichelt M, Kummerfeld SK, Modrusan Z, Dugger DL, Newton K, Sagolla M, Webster JD, Liu J, Hebrok M, Dixit VM, 2015 β -Cell insulin secretion requires the ubiquitin ligase COP1. *Cell* 163, 1457–1467. [PubMed: 26627735]
- Wang Y, Huang C, Fang Y, Yang Q, Li R, 2012 Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing. *J. R. Stat. Soc. Ser. C Appl. Stat* 61, 1–24.

- Wang Z, Pang X, Wu W, Wang J, Wang Z, Wu R, 2014 Modeling phenotypic plasticity in growth trajectories: a statistical framework. *Evolution* 68, 81–91. [PubMed: 24111588]
- Wang Z, Xu K, Zhang X, Wu X, Wang Z, 2017 Longitudinal SNP-set association analysis of quantitative phenotypes. *Genet. Epidemiol* 41, 81–93. [PubMed: 27859628]
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H, 2014 The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42, D1001–D1006. [PubMed: 24316577]
- Wu R, Lin M, 2006 Functional mapping - how to map and study the genetic architecture of dynamic complex traits. *Nat. Rev. Genet* 7, 229–237. [PubMed: 16485021]

Fig. 1.

Association results for BMI using *i*GWAS in Framingham Heart Study. **A:** A Manhattan plot. $-\log_{10}(p\text{-value})$ is plotted against chromosome location. **B:** BMI growth curves at the SNP rs7806587. The solid lines, in red, green and blue, are the estimated curves based on the Legendre polynomial of order 2 in the three genotypes *CC*, *CT* and *TT*, respectively. The black solid line is the estimated curve in the whole samples. The dashed lines, in red, green and blue, are the mean curves in the three genotypes *CC*, *CT* and *TT*, respectively. The black dashed line is the mean curve in the whole samples.