Supplementary Materials for Testing Cross-Phenotype Effects of Rare Variants in Longitudinal Studies of Complex Traits

Pratyaydipta Rudra*[1], K. Alaine Broadaway[2], Erin B. Ware[3], Min A. Jhun[3], Lawrence F. Bielak[3], Wei Zhao[3], Jennifer A. Smith[3], Patricia A. Peyser[3], Sharon L.R. Kardia[3], Michael P. Epstein[2], and Debashis Ghosh[1]

[1]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO

[2]Department of Human Genetics, Emory University, Atlanta, GA

[3]Department of Epidemiology, University of Michigan, Ann Arbor, MI

*Correspondence: 13001 E. 17th Place, Aurora, CO 80045, Email: pratyaydipta.rudra@ucdenver.edu, Phone: 919-699-4043

1 Details of data simulation for single time point

The genetic data for our simulations were generated as 20,000 haplotypes of 30 kb in size using COSI, a coalescent model that mimics LD pattern, local recombination rate, and population history for individuals of European descent (Schaffner et al., 2005). We assume ten phenotypes for each subject generated from a multivariate normal distribution with mean vector 0 and 10×10 residual correlation matrix Σ. To model the residual correlation matrix, we considered scenarios of low residual correlation among phenotypes (pairwise correlation among phenotypes selected from a uniform (0,0.3). The low correlation structure was chosen to increase the likelihood of the eventual correlation matrix of longitudinal data being positive definite (if not positive definite, we simulate again). For the high correlation scenario, the pariwise correlations among phenotypes were selected from a uniform (0.3,0.4).

For power models, we simulated data sets in which 5% of the rare variants in our haplotypes were modeled as causal. We set effect size $\beta_{rl} = (0.4 + N(0,0.1))|\log_{10}(MAF_r)|$ for each causal variant r and phenotype l. MAF_r denotes the minor allele frequency of causal variant r. This formulation sets mean effect size of causal variant r as inversely proportional to its MAF, such that very rare variants have on average a larger effect size than less rare variants. Allowing β_{rl} to vary around a normal distribution maintains the relationship between MAF and effect size, while allowing the variant to have a slightly different effect size for each phenotype.

We varied the number of phenotypes associated with the rare variants, such that not all of the tested phenotypes will be dependent on the gene of interest. The number was varied as 0 (null case), 2, 4, 6, and 8. We control correlation among phenotypes through consideration of the relative variance of phenotype explained by the R causal variants. We define this relative variance for phenotype l as $h_l = 2\sum_{r=1}^{R} \beta_{rl}^2 MAF_r(1 - MAF_r)$. As in (Galesloot, Van Steen, Kiemeney, Janss, & Vermeulen, 2014) we define the overall correlation between phenotypes l and l' as $E_{ll'} = \sqrt{1-h_l}\sqrt{1-h_l'}\Sigma_{ll'}$, where $\Sigma_{ll'}$ is the (l,l')th element of the $L \times L$ residual phenotypic correlation matrix. This allows the residual correlation structure among phenotypes to stay at the defined values.

2 Details of GAMuT (Meta)

The method computes the GAMuT p-values for each time point using the linear kernel and uses meta analysis to combine them into a single p-value. The meta-analysis is done as follows. A Gaussian copula approach was used to obtain a standard normal variate from each p-value by applying the inverse of the standard normal distribution function. These values were computed for each simulation to construct a matrix of dimension $T \times nsim$, where $nsim$ is the number of simulations. This matrix can be used to estimate the correlation of the normal variates across time which can then be used for the meta-analysis using Kost's method (Kost & McDermott, 2002). Note that this method requires the matrix from the simulations, and cannot be applied for real data without further assumptions since the correlation structure cannot be estimated from a single replicate.

3 Supplementary figures and tables

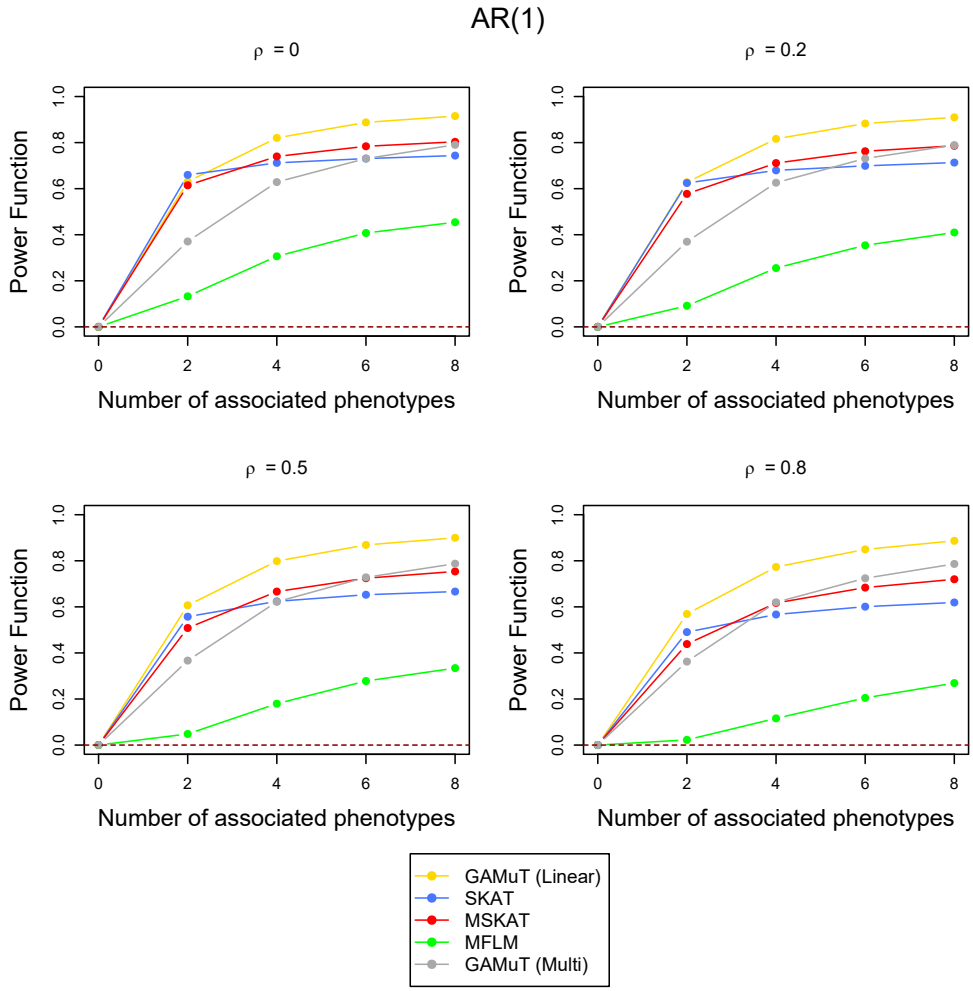

Figure 1: Comparison of power curves for different methods in the quadratic time effect set up. The correlation structure across time is considered to be AR(1) with parameter ρ and the tests are done at the p-value threshold 5×10^{-6}. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test.

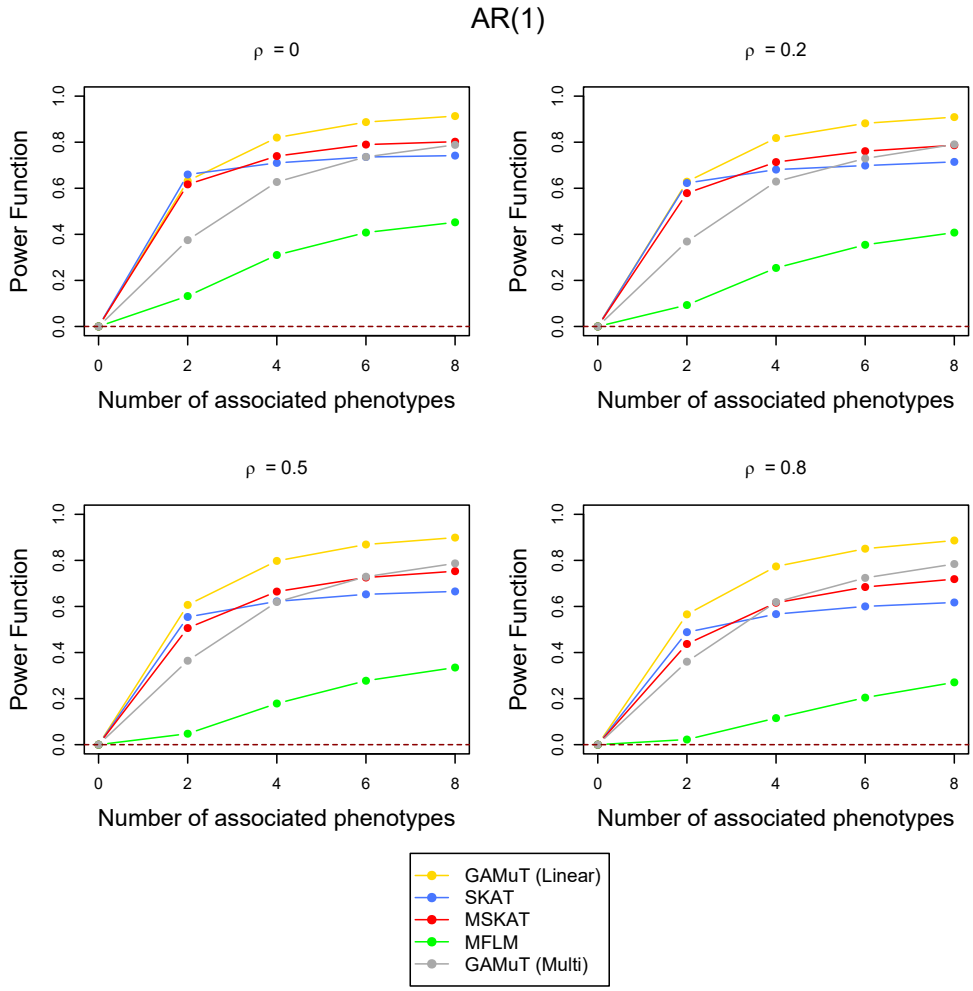

Figure 2: Comparison of power curves for different methods in the sinusoidal time effect set up. The correlation structure across time is considered to be AR(1) with parameter ρ and the tests are done at the p-value threshold 5×10^{-6}. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test.

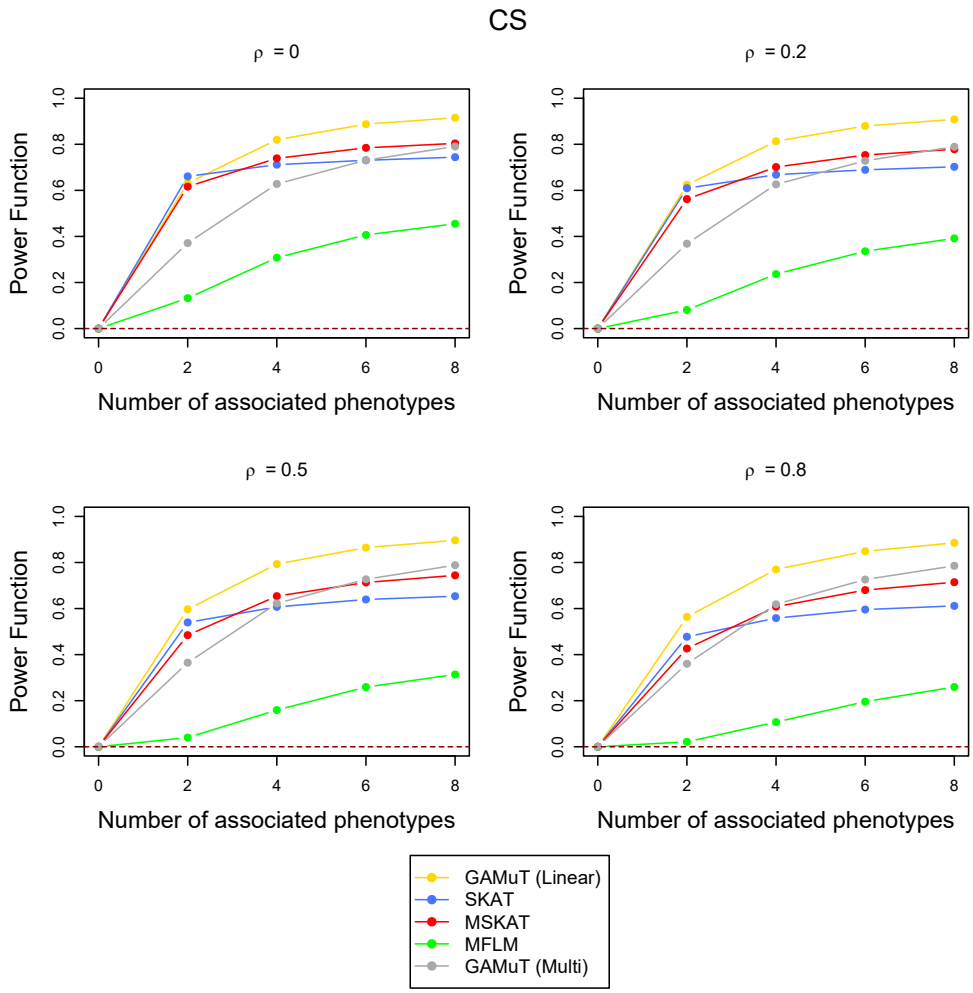

Figure 3: Comparison of power curves for different methods in the linear time effect set up. The correlation structure across time is considered to be Compound Symmetry with parameter ρ and the tests are done at the p-value threshold 5×10^{-6}. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test. The rest of the compound symmetry results are not presented since they are very similar to the AR(1) results

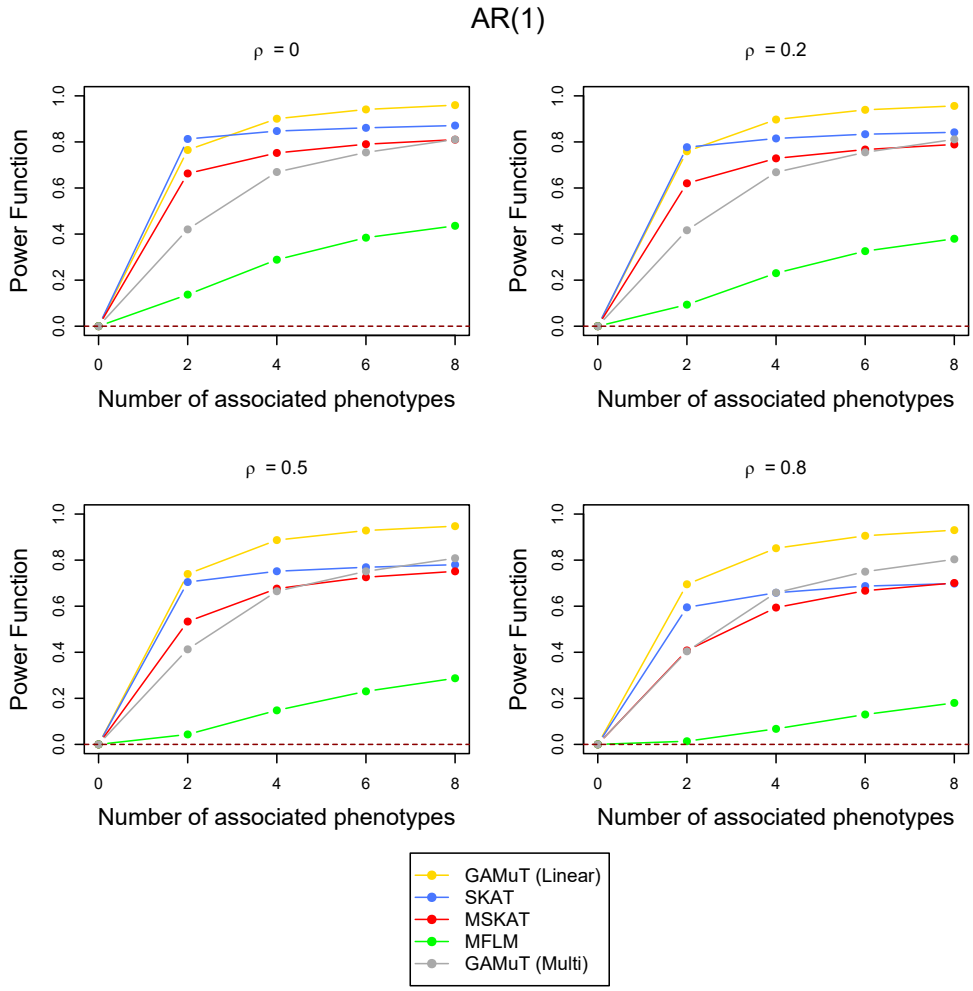

Figure 4: Comparison of power curves for different methods in the linear time effect set up for 6 time points. The correlation structure across time is considered to be AR(1) with parameter ρ and the tests are done at the p-value threshold 5×10^{-6}. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test. Other results with 6 time points are omitted since they are very similar.

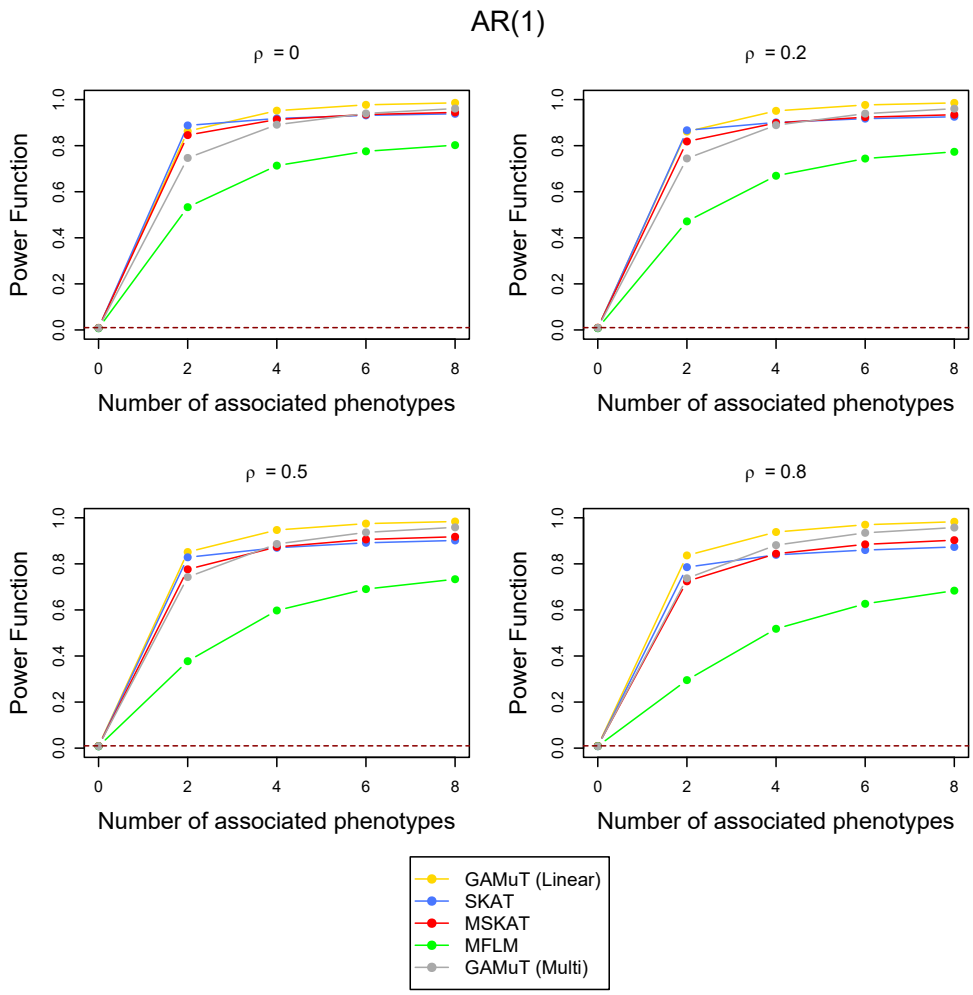

Figure 5: Comparison of power curves for different methods in the linear time effect set up. The correlation structure across time is considered to be AR(1) with parameter ρ and the tests are done at the p-value threshold 0.01. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test. As expected, the power of all methods are much higher at this less stringent p-value threshold. Other results with less stringent p-value threshold are omitted since they are very similar.

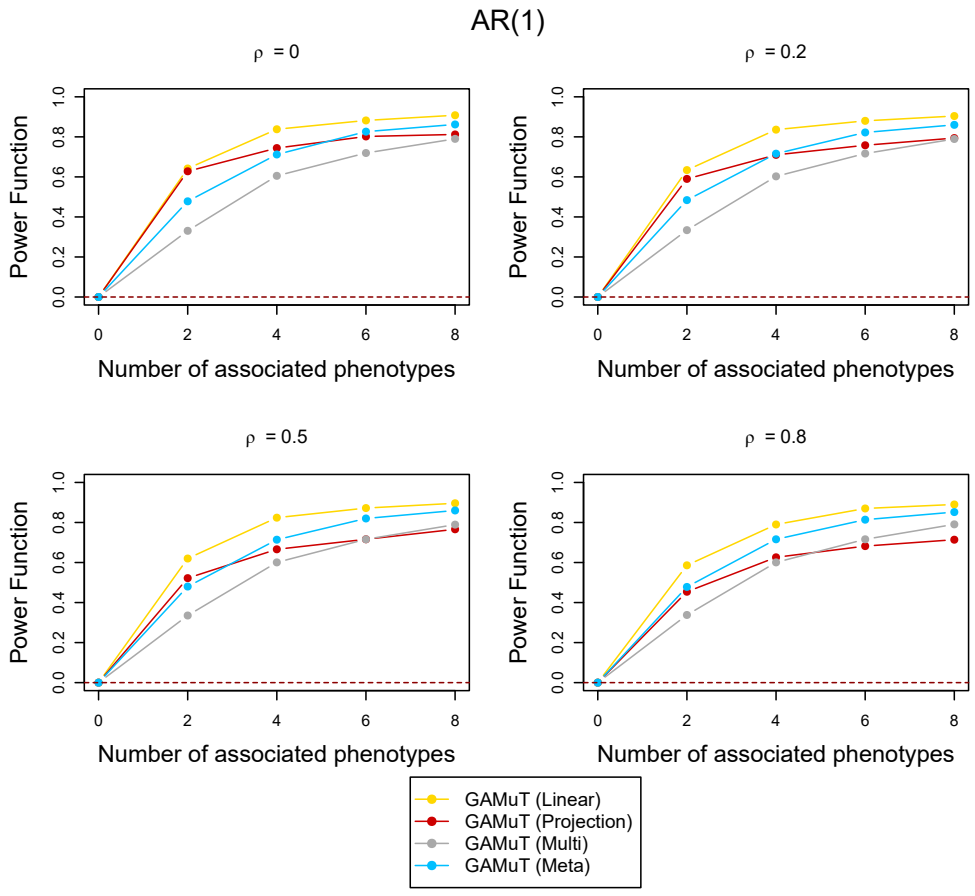

Figure 6: Comparison of power curves for joint analysis using GAMuT with meta-analysis of single time point GAMuT in the linear time effect set up. The correlation structure across time is considered to be AR(1) with parameter ρ and the tests are done at the p-value threshold 5×10^{-6}. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test.

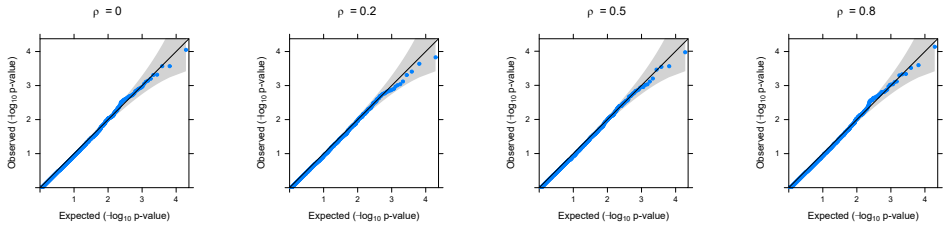

Figure 7: QQ-plots based on application of GAMuT (Linear) on simulated null datasets for the linear time effect set up with high correlation between the phenotypes. The correlation structure across time is considered to be AR(1) with parameter ρ.

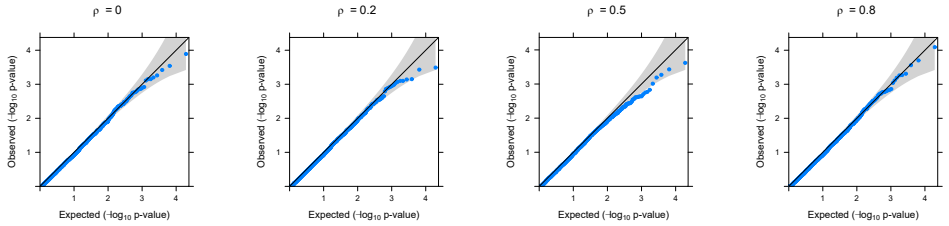

Figure 8: QQ-plots based on application of GAMuT (Projection) on simulated null datasets for the linear time effect set up with high correlation between the phenotypes. The correlation structure across time is considered to be AR(1) with parameter ρ.

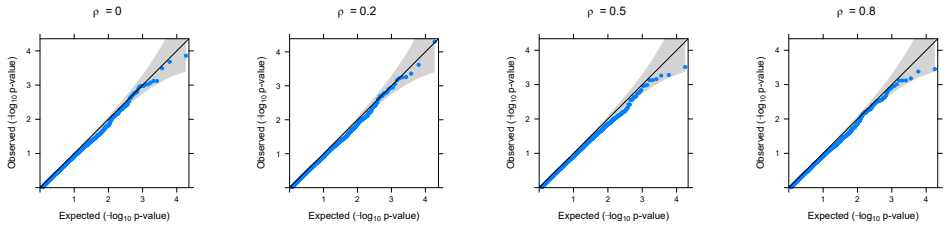

Figure 9: QQ-plots based on application of GAMuT (Linear) on simulated null datasets for the missing data scenario. The correlation structure across time is considered to be AR(1) with parameter ρ.

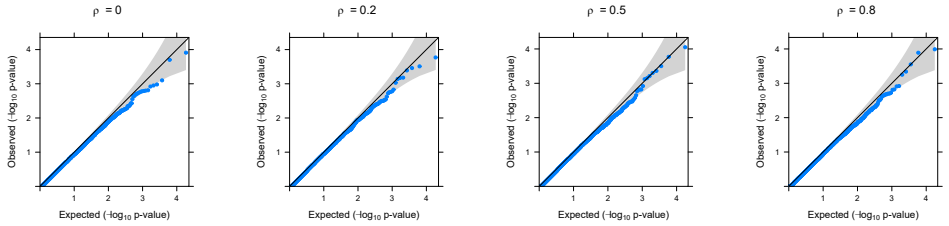

Figure 10: QQ-plots based on application of GAMuT (Projection) on simulated null datasets for the missing data scenario. The correlation structure across time is considered to be AR(1) with parameter ρ.

	GAMuT (Projection)				GAMuT (Linear)				GAMuT (Multi)			
	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
Linear	5	1	3	6	5	11	9	8	8	3	3	3
Linear(high correlation)	2	2	2	0	1	2	0	1	0	0	0	0
Linear (6 time points)	2	1	1	1	0	0	1	2	4	3	3	3
Quadratic	4	2	4	8	6	11	9	8	7	3	3	3
Sine	6	1	2	1	6	11	8	8	6	3	3	3
Binary	3	1	3	3	2	2	3	3	42	29	29	21
Missing	0	0	1	1	2	6	2	2	0	0	1	0
	SKAT				MSKAT				MFLM			
	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
Linear	8	7	3	2	6	2	4	8	0	1	0	23
Linear (high correlation)	4	3	10	8	2	2	2	0	3	2	0	14
Linear (6 time points)	0	0	0	0	2	1	1	1	15	18	8	2
Quadratic	8	7	3	2	4	2	4	8	0	1	0	23
Sine	8	6	3	3	7	2	3	2	0	1	0	23
Binary	0	0	2	2	3	1	3	3	8	2	3	4
Missing	14	0	3	0	0	0	1	2	2	0	0	1

Table 1: **Type-I error (in the scale of 10^{-6}, i.e. a value q in the table should be read as $q \times 10^{-6}$) of different methods under various simulations, target $\alpha = 10^{-5}$. The correlation structure across time is considered to be AR(1) with parameter ρ.**

Chromosome	Gene Name	Number of Variants	GAMuT (Projection)	GAMuT (Linear)	GAMuT (Projection) 95% CI	GAMuT (Linear) 95% CI
1	EFCAB7	6	1.980E-05	0.0104	(0,0.0005)	(0,0.0263)
6	ENPP3	6	2.84E-05	0.0095	(0,0.0001)	(0,0.0171)
6	NQO2	5	0.0007	0.0071	(0,0.0050)	(0,0.0112)
6	FNDC1	10	0.0061	0.0031	(0,0.0012)	(0,0.0035)
7	ZNF655	5	0.0092	0.0063	(0,0.0900)	(0,0.1092)
19	CD33	5	0.0002	2.608E-05	(0,0.0003)	(0,0.0003)
19	ZNF551	5	0.0009	0.0012	(0,0.0030)	(0,0.0104)
19	ZNF667	5	0.0003	0.0079	(0,0.0011)	(0,0.0144)
20	COL9A3	5	0.0000	0.0002	(0,0.0069)	(0,0.0154)

Table 2: **Bootstrap confidence intervals for the p-values based on the application of longitudinal GAMuT on GENOA data. 1000 bootstraps were used.**

References

Galesloot, T. E., Van Steen, K., Kiemeney, L. A., Janss, L. L., & Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PloS one*, *9*(4), e95923.

Kost, J. T., & McDermott, M. P. (2002). Combining dependent p-values. *Statistics & Probability Letters*, *60*(2), 183–190.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., & Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome research*, *15*(11), 1576–1583.