# KMgene: a unified R package for gene-based association analysis for complex traits

Qi Yan[1], Zhou Fang[2] and Wei Chen[1,2]

[1]Division of Pulmonary Medicine, Allergy and Immunology, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh,
[2]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

## Background

- One widely used gene-based test is the sequence kernel association test (SKAT, Wu et al., 2011) based on a KM regression framework.
- After SKAT was introduced for testing independent samples with continuous and binary traits, a number of methods and corresponding tools have been developed to extend the approach to complex traits.
- We introduce KMgene, which combines SKAT-type methods for complex traits and extends them to include their corresponding optimal tests. KMgene can perform association tests between a set of genetic variants and familial, multivariate, longitudinal or survival traits (Table 1, Yan et al., 2018).

**Table 1.** A summary of functions in KMgene package

| | Regular (KM) | Optimal (KM-O) | Interaction (KM-Int) |
|---|---|---|---|
| **Continuous family (F-KM)** | Chen et al. 2013 | Extended | NA |
| **Binary family (Fb-KM)** | Yan et al. 2015 | Extended | NA |
| **Continuous multivariate (M-KM)** | Maity et al. 2012 | Extended | NA |
| **Continuous multivariate family (MF-KM)** | Yan et al. 2015 | Extended | NA |
| **Continuous longitudinal (L-KM)** | Yan et al. 2015 | Yan et al. 2015 | Extended |
| **Survival (CoxKM)** | Chen et al. 2014 | NA | NA |

## Method

**KMgene works in two steps:**

- The first step with function names, *prefix*_Null_Model, fits the model under the null hypothesis (i.e., the genetic effects are zero). The estimates of covariate parameters and covariance matrix are obtained at this step. The covariance matrix can account for relatedness in families, correlation between multivariate traits or between times for longitudinal data.
- The second step with function names, *prefix*, constructs the test statistic and calculates the *p*-value. We use the parameter estimates from step one to construct the test statistic. Since the parameters are estimated under the null hypothesis and used for all genes, they only need to be calculated once for the whole genome-wide analysis, which greatly reduces the computation time.
- According to our derivation, the test statistic follows a mixture of $\chi^2$ distributions and thus we can compute the *p*-values analytically, also leading to improvement in computation.
- The KM statistics can be extended to the optimal test by combining with burden statistics. Analogously, our optimal tests consist of two steps for fitting null models (*prefixO*_Null_Model) and calculating *p*-values (prefixO).

**Input:**

- Genotypes pre-grouped in genes and coded as 0, 1, 2 for the number of copies of minor allele (i.e., additive genetic model);
- Traits and covariates;
- Family pedigree when analyzing familial data.

**Output:**

- Gene-level *p*-values

## Results (real data example)

- Here, as an illustrative example, we apply MFKM_Null_Model() and MFKM() to carry out a gene-based genome wide association test of the correlated lung function phenotypes FEV1 (Forced Expiratory Volume in One Second) and FEV1/FVC (Forced Vital Capacity) ratio (Yan, et al., 2015). We identified *COL6A6* associated with these two traits (Fig 1) and *COL6A6* is known to be in the chronic obstructive pulmonary dis-ease related regions based on Rat Genome Database (RGD)

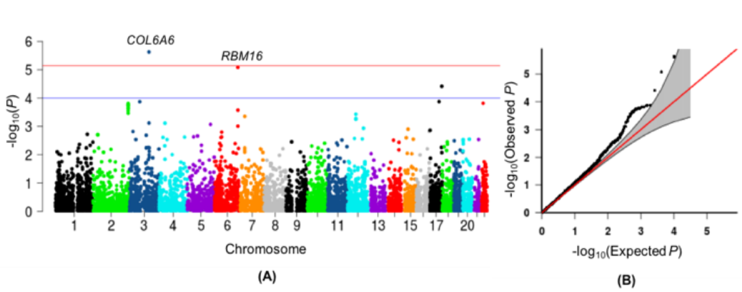

Fig1. (A) Genome wide gene-based results of MFKM on lung function data. Each dot represents p-value of a gene. (B) QQ plot of p-values from the lung function analysis, with 95% pointwise confidence band (gray area).

## Conclusions

- This R package adapts GLMM to conduct gene-based tests for complex traits and uses Cox model for survival trait.
- KMgene can handle genome-wide genotypic datasets with reasonable computational time.
- KMgene currently uses the linear kernel that is the most commonly used kernel in genetic studies.

## Acknowledgement

This work was supported by NIH grants R01HG007358 and R01EY024226.

## References

Wu, M.C., et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;89(1):82-93.

Chen, H., *et al*. Sequence kernel association test for survival traits. *Genet Epidemiol* 2014;38(3):191-197.

Chen, H., Meigs, J.B. and Dupuis, J. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* 2013;37(2):196-204.

Maity, A., Sullivan, P.F. and Tzeng, J.Y. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet Epidemiol* 2012;36(7):686-695.

Yan, Q., et al. A Sequence Kernel Association Test for Dichotomous Traits in Family Samples under a Generalized Linear Mixed Model. Hum Hered 2015;79(2):60-68.

Yan, Q., et al. Associating Multivariate Quantitative Phenotypes with Genetic Variants in Family Samples with a Novel Kernel Machine Regression Method. Genetics 2015;201(4):1329-1339.

Yan, Q., et al. Rare-Variant Kernel Machine Test for Longitudinal Data from Population and Family Samples. Hum Hered 2015;80(3):126-138