PopPAnTe: A Tool for Population and Pedigree Association Testing

Alessia Visconti1\*, Mashael Al-Shafai2,3,4, Wadha A Al Muftah2,3,4, Shaza B Zaghlool3, Massimo Mangino1, Karsten Suhre3† and Mario Falchi1†

**Abstract**

**Background:** Family-based designs, from twin studies to isolated populations with their complex genealogical data, are a valuable resource for genetic studies of heritable molecular biomarkers. Existing software for family-based studies have mainly focused on facilitating association between response phenotypes and genetic markers, and no user-friendly tools are at the present available to straightforwardly extend association studies in related samples to large datasets of generic quantitative data, as those generated by current *-omics* technologies.

**Results:** We developed PopPAnTe, a user-friendly Java program, which evaluates the association of quantitative data in related samples. Additionally, PopPAnTe implements data pre and post processing, region based testing, and empirical assessment of associations.

**Conclusions:** PopPAnTe is an integrated and flexible framework for pairwise association testing in related sample with large number of predictors and response variables. It works either with family data of any size and complexity, or, when the family structure is unknown, it uses genetic similarity information between individuals as those inferred from genome-wide genetic data. Therefore it facilitates the usage of biobank data collections from population isolates when extensive genealogical information is missing.

**Keywords:** Association Studies; Heritability; *-omics* data; Family Data; Isolated Population; Population Genetics

**Background**

Family-based designs, from complex genealogical structure to twin studies, are a valuable resource for genetic studies. The primary aim of currently available software accounting for population stratification and/or relatedness in the statistical model (*e.g.*, EMMA [1], Merlin [2], GenABEL [3], QTDT [4]), is to evaluate the association between genetic SNP markers and response phenotypes and, to date, very few tools are available to test association of large quantitative datasets generated by high throughput *-omics* technologies (*e.g.,* epigenomic versus metabolomic data, or transcriptomic versus metagenomic data) in familial samples , For instance, although modeling of genealogical data can be performed by *coxme* [5] and *kinship2* [6] R packages, R is not a particularly efficient environment to carry out hundred thousand or million tests. Therefore, we implemented a user-friendly Java program, PopPAnTe, to perform exact association tests of between large quantitative datasets in family-based studies. Relationships between individuals can be either described by known family structures of any size and complexity, or by genetic similarity matrices (GSM) inferred from genome-wide genetic data. This approach is particularly useful where some degree of hidden relatedness (including population stratification is expected, but extensive genealogical information is missing or incomplete. For instance, genealogical information going back more than three or four generations may be difficult to be retrieved for individuals recruited in large-scale biobank started in genetic isolates as those from the Middle East.

**Implementation**

PopPAnTe assesses the relationship between quantitative dependent variables (*responses*) and quantitative independent variables (*predictors*) within a variance component framework in order to model the resemblance among relatives. The association of a single predictor with a single response variable is described as

ri=μ+βpi+􏰂 ψijCij+gi+ei (1)

where ri represents the response value for the i-th individual, μ the response mean, β the estimate of the predictor value pi, ψj the estimate of the j-th covariate C, and gi and ei the polygenic and environment effect, respectively.

The total response variance is partitioned into polygenic and environmental variances (the latter including also measurement errors), and the variance covariance matrix is calculated as

ω = 2Φσa2 + Iσe2

where Φ is the matrix of the relatedness matrix between each pair of individuals, I is the identity matrix, and σa2 and σe2 are the additive genetic and environmental variance, respectively.

Within the same framework, PopPAnTe allows the evaluation of the narrow heritability of any quantitative response variance included in the analysis.

The significance of the association is calculated using a formal likelihood-ratio test comparing the likelihood of the alternative model described in equation (1) to the likelihood of a null model where the effect of the predictor is constrained to zero. To speed-up the evaluation, PopPAnTe clusters variables having the same pattern of missingness (*i.e.*, the same missing values in a subset of individuals), then evaluates once the likelihood of the null model, and reuses the value to assess every variable included in the cluster. PopPAnTe also allows the evaluation of empirical p-values by randomly permuting the predictor values among subjects and re-assessing the association under the null hypothesis. When family structures are provided as input, predictor values are randomly permuted among family members. To speed-up the performance PopPAnTe implements an adaptive permutation approach [7], stopping the generation of randomly permuted samples earlier when there is little or no evidence of significance.

**Pedigree versus Population analysis**

When genealogical information is available PopPAnTe evaluates the relatedness matrix from the known pedigree relationships using a recursive procedure and assuming pedigree founders as unrelated [8]. This results in a variance-covariance matrix that is usually both symmetric and semi-positive definite. Therefore, the maximum likelihood estimates of the variance components can be assessed through efficient Cholesky decomposition.

When the family structure is not available, a GSM can be estimated from genome-wide genetic data with any of several well-established tools, such as PLINK [9], GCTA [10], or LDAK [11], and given as input to PopPAnTe. The property of positive-definiteness, however, often does not hold for GSMs and in these cases the maximum likelihood estimates is evaluated using the LU decomposition. Additionally, PopPAnTe implements the QR decomposition to solve the rare cases where the variance-covariance matrix is not invertible and the LU decomposition cannot be used. To speed-up the evaluation of the variance components, PopPAnTe allows the user to set a minimum threshold below which individuals can be considered as unrelated.

**Region-based testing**

When predictors can be ordered in space, as it is for instance for gene expression or epigenetic markers, PopPAnTe allows the computation of region-based association tests by gathering information from flanking predictors included in a sliding window of user-defined size, whose values are replaced with their first principal component. By definition the first principal component accounts for as much of the variability in the data as possible, and can thus be used to summarise the joint distribution of all variables included in a given region for gene- or region-based association studies (e.g., [12] and [13]).

**Data pre- and post-processing**

Quantile normalisation [14] can be automatically applied to each variable to improve normality of the response variables and of the predictors. Moreover, PopPAnTe implements two approaches to correct the association test for unwanted biological and technical variability (*e.g*., batch effects). When the source of the confounders is known, it can be directly included in the association model. To deal with unknown source of biological and technical co-variation, PopPAnTe can integrate in the association model the principal components that are required to explain a user-specified percentage of variation.

PopPAnTe implements the Benjamini-Hochberg procedure (BH step-up procedure) to control the false discovery rate [15], and, to aid in results interpretation and further analyses, it generates basic Quantile-Quantile and Manhattan plots – the latter only when genomic data that can be ordered in space (*e.g.,* CpG loci) are used as predictors.

Finally, when the family structure is available, to determine whether an association has been generated by a uniform contribution of all the families within the sample, or by a strong contribution of a small number of families, PopPAnTe reports, for each test, the percentage of families showing a positive contribution and the Gini coefficient [16] assessed on family contribution to the χ2 statistics.

**Results and Discussion**

We assessed PopPAnTe’s performance in two case studies. First, we performed an epigenome-wide association study of body mass index (BMI) using data from a family study from the Qatari isolated population. Then we performed a trascriptome-wide association study of BMI in a cohort of healthy female Caucasians twins. We aimed at testing whether using a pedigree-based kinship matrix (both in an inbred and in an outbred population) or a GSM yields to the same results. Moreover, we used GCTA [10] in the first case study, and LDAK [11] in the second to show that the results are independent from the tool used to estimate the GSM.

**Epigenome-wide Association Study in a Qatari Family Study**

The Qatari population is an isolated population characterised by a large number of consanguineous families. A detailed description of the subjects and methylation data included in this study has been reported previously described in Zaghlool *et al.* [17]. Briefly, we used genome-wide methylation and SNP data generated from whole blood on the Infinium HumanMethylation450 Bead-Chip (Illumina Inc, San Diego, CA) and the Illumina HumanOmni2.5-8M BeadChip, respectively. DNA methylation Beta-values were measured for 123 individuals, 88 of whom had both genotype data and BMI information. Genealogical data spanning three generations and connecting these subjects in 13 families were also available. We used GCTA to calculate a GSM between pairs of individuals using all autosomal SNP markers with minor allele frequency > 0.01. We compared heritability estimates of the methylation values at CpG loci and their association with body mass index (BMI) in the Qatari family study using either the family information or the inferred GSM. Age, sex, and cell-type proportions as estimated using the Houseman method [18] were included in the model as fixed effects.

We observed a very high concordance correlation coefficient [19] of the effect size estimates for the association between CpG methylation states and BMI (rβ = 0.99; Figure 1, left), as well as of the CpG-specific component of genetic and environmental variances (rσa2 = 0.99 and rσe2 = 0.90, respectively; Figure 1, right).

**Trascriptome-wide Association Study in UK Twins**

The TwinsUK adult twin registry includes about 12,000 subjects, predominately females [20]. Genotyping of the TwinsUK dataset was performed with a combination of Illumina HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M chips and imputation was performed using the IMPUTE software package (v2), as previously described [21]. Expression profiling in subcutaneous adipose tissue was measured using Illumina Human HT-12 V3 BeadChips for 825 female individuals [22], 778 of whom had both genotype data and BMI information. We used LDAK to calculate a GSM based on allelic correlation across autosomes. Before calculation, we excluded SNPs with minor allele frequency < 0.01. We compared effect size estimates of the gene expression profiles versus BMI, using either the family information or the inferred GSM. Age was included in the model as fixed effect.

We observed a very high concordance correlation coefficient of the effect size estimates for the association between gene expression levels and BMI (rβ = 0.99; Figure 2, left)), as well as of the gene-specific genetic component of genetic and environmental variances (rσa2 = 0.99 and rσe2 = 0.99, respectively; Figure 2, right).

**Conclusions**

PopPAnTe is a user-friendly platform-independent Java program that enables pairwise association testing of large number of predictors and response variables in related sample. We showed that PopPAnTe achieves highly comparable results both when the genealogical information is available and when the relatedness between subjects is inferred from genome-wide SNP data using two case studies. PopPAnTe can thus facilitate the usage of biobank collections from population isolates when extensive genealogical information is missing.

**Availability and requirements**

Project name: PopPAnTe

Project home page: FIXME

Operating system(s): Platform independent

Programming language: Java 1.7 or higher

Other requirements: None

License: GNU GPL 3 or higher

Any restrictions to use by non-academics: None

**List of abbreviations**

BMI: Body mass index

GSM: Genetic similarity matrix

**Competing interests**

The authors declare that they have no competing interests.

**Author’s contributions**

AV, KS, and MF designed the software. AV developed PopPAnTe, and performed the computational experiments. MAS, WAAM, SBZ, and KS generated the Qatari family study data. MM provided the data for the TwinsUK cohort and tested the software. AV and MF wrote the manuscript. All authors read and approved the final manuscript.

**Acknowledgements**

We thank the Qatar Diabetes Association (QDA) and Cindy McKeon for their help in sample recruitment. We are grateful to all study participants for their contribution to this research study. AV and MF thank S. Burbidge and M. Harvey of the Imperial College High-Performance Computing service for their assistance. The two studies were conducted in concordance with the Helsinki declaration of ethical principles for medical research involving human subjects. The studies were approved by the relevant institutional review boards in Qatar (Institutional Review Board of Weill Cornell Medical College in Qatar, ethical approval numbers 2012–003 and 2012–0025), and in the UK (Guy’s and St. Thomas’ Hospital Ethics Committee). Written informed consent was obtained from every participant in each study.

Funding. This work was supported by the Qatar Science Leadership Program at the Research Division, a program funded by the Qatar Foundation, and by Bioinformatics and Virtual Metabolomics Core of Weill Cornell Medical College in Qatar. MF and AV are supported by the British Skin Foundation (grant 5044i). The TwinsUK study was funded by the Wellcome Trust; European Community’s Seventh Framework Programme (FP7/2007-2013). The study also receives support from the National Institute for Health Research (NIHR)- funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy’s and St Thomas’ NHS Foundation Trust in partnership with King’s College London. SNP Genotyping was performed by The Wellcome Trust Sanger Institute and National Eye Institute via NIH/CIDR.

**Author details**

1Department of Twin Research and Genetic Epidemiology, King’s College London, London, UK.

2Department of Genomics of Common Diseases, Imperial College London, London, UK.

3Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Doha, Qatar, Doha, Qatar.

4Research Division, Qatar Science Leadership Program, Qatar Foundation, Doha, Qatar.

**References**

1. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E.: Efficient control of population structure in model organism association mapping. Genetics 178(3), 1709–1723 (2008)

2. Abecasis, G.R., Cherny, S.S., Cookson, W.O., Cardon, L.R.: Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics 30(1), 97–101 (2002)

3. Aulchenko, Y.S., Ripke, S., Isaacs, A., Van Duijn, C.M.: Genabel: an r library for genome-wide association analysis. Bioinformatics 23(10), 1294–1296 (2007)

4. Abecasis, G., Cardon, L., Cookson, W.: A general test of association for quantitative traits in nuclear families. The American Journal of Human Genetics 66(1), 279–292 (2000)

5. Therneau, T.: coxme: Mixed effects Cox models. R package version 2(3) (2012)

6. Therneau, T., Atkinson, E., Sinnwell, J., Matsumoto, M., Schaid, D., McDonnell, S.: kinship2: Pedigree

Functions, R Package Version 1.5. 0, 2013

7. Che, R., Jack, J.R., Motsinger-Reif, A.A., Brown, C.C.: An adaptive permutation approach for genome-wide

association study: evaluation and recommendations for use. BioData Min 7(9) (2014)

8. Thompson, E.A.: Pedigree Analysis in Human Genetics. Johns Hopkins University Press Baltimore, ??? (1986)

9. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P.,

De Bakker, P.I., Daly, M.J., et al.: PLINK: a tool set for whole-genome association and population-based

linkage analyses. The American Journal of Human Genetics 81(3), 559–575 (2007)

10. Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M.: GCTA: a tool for genome-wide complex trait analysis. The

American Journal of Human Genetics 88(1), 76–82 (2011)

11. Speed, D., Hemani, G., Johnson, M.R., Balding, D.J.: Improved heritability estimation from genome-wide

SNPs. The American Journal of Human Genetics 91(6), 1011–1021 (2012)

12. Gauderman, W.J., Murcray, C., Gilliland, F., Conti, D.V.: Testing association between disease and multiple

SNPs in a candidate gene. Genetic epidemiology 31(5), 383–395 (2007)

13. Agerbo, E., Mortensen, P.B., Wiuf, C., Pedersen, M.S., McGrath, J., Hollegaard, M.V., Nørgaard-Pedersen, B.,

Hougaard, D.M., Mors, O., Pedersen, C.B.: Modelling the contribution of family history and variation in single nucleotide polymorphisms to risk of schizophrenia: a Danish national birth cohort-based study. Schizophrenia research 134(2), 246–252 (2012)

14. Wichura, M.J.: Algorithm AS 241: The percentage points of the normal distribution. Applied Statistics, 477–484 (1988)

15. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289–300 (1995)

16. Gini, C.: Variabilit`a e mutabilit`a. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi 1 (1912)

17. Zaghlool, S.B., Al-Shafai, M., Al Muftah, W.A., Kumar, P., Falchi, M., Suhre, K.: Association of DNA methylation with age, gender, and smoking in an Arab population. Clinical epigenetics 7(1), 6 (2015)

18. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., Kelsey, K.T.: DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics 13(1), 86 (2012)

19. Lin, L.: A note on concordance correlation coefficient. Biometrics 56, 324–325 (2000)

20. Moayyeri, A., Hammond, C.J., Valdes, A.M., Spector, T.D.: Cohort profile: TwinsUK and healthy ageing twin

study. International journal of epidemiology 42(1), 76–85 (2013)

21. Mangino, M., Hwang, S.-J., Spector, T.D., Hunt, S.C., Kimura, M., Fitzpatrick, A.L., Christiansen, L., Petersen, I., Elbers, C.C., Harris, T., et al.: Genome-wide meta-analysis points to CTC1 and ZNF676 as genes regulating telomere homeostasis in humans. Human molecular genetics 21(24), 5385–5394 (2012)

22. Grundberg, E., Small, K.S., Hedman, ̊A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri,

E., Barrett, A., et al.: Mapping cis-and trans-regulatory effects across multiple tissues in twins. Nature genetics

44(10), 1084–1089 (2012)

**Figures**



**Figure 1 Epigenome-wide Association Study in a Qatari Family Study.** Comparison of the results obtained in the Epigenome-wide association studies when the relatedness between subjects was evaluated using the family structures and when it was inferred from genome-wide SNPs by means of GCTA [10]. Left panel: effect size estimates for the association between CpG methylation states and BMI. Right panel: estimated genetic (σa2, in blue) and environmental (σe2, in red) variances.



**Figure 2 Trascriptome-wide Association Study in UK Twins.** Comparison of the results obtained in the Trascriptome-wide Association when the relatedness between subjects was evaluated using the family structures and when it was inferred from genome-wide SNPs by means of LDAK [11]. Left panel: effect size estimates for the association between gene expression levels and BMI. Right panel: estimated genetic (σa2, in blue) and environmental (σe2, in red) variances.