BMC Proceedings

BioMed Central

Proceedings

Open Access

# Growth mixture modelling in families of the Framingham Heart Study

Berit Kerner*[1] and Bengt O Muthén[2]

Addresses: [1]Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, University of California, 695 Charles E Young Drive South, Room 4357B, Box 951761, Los Angeles, California 90095-1761, USA and [2]Professor Emeritus, University of California, 2005 East Moore Hall, Los Angeles, California 90095, USA

E-mail: Berit Kerner* - bkerner@mednet.ucla.edu; Bengt O Muthén - bmuthen@ucla.edu
*Corresponding author

*from* Genetic Analysis Workshop 16
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009

*BMC Proceedings* 2009, **3**(Suppl 7):S114 doi: 10.1186/1753-6561-3-S7-S114

This article is available from: http://www.biomedcentral.com/1753-6561/3/S7/S114



## Abstract

Growth mixture modelling, a less explored method in genetic research, addresses unobserved heterogeneity in population samples. We applied this technique to longitudinal data of the Framingham Heart Study. We examined systolic blood pressure (BP) measures in 1060 males from 692 families and detected three subclasses, which varied significantly in their developmental trajectories over time. The first class consisted of 60 high-risk individuals with elevated BP early in life and a steep increase over time. The second group of 131 individuals displayed first normal BP, but showed a significant increase over time and reached high BP values late in their life time. The largest group of 869 individuals could be considered a normative group with normal BP on all exams. To identify genetic modulators for this phenotype, we tested 2,340 single-nucleotide polymorphisms on chromosome 8 for association with the class membership probabilities of our model. The probability of being in Class 1 was significantly associated with a very rare variant (rs1445404) present in only four individuals from four different families located in the coding region of the gene *EYA* (eyes absent homolog 1 in Drosophila) ($p = 1.39 \times 10^{-13}$). Mutations in *EYA* are known to cause brachio-oto-renal syndrome, as well as isolated renal malformations. Renal malformations could cause high BP early in life. This result awaits replication; however, it suggests that analyzing genetic data stratified for high-risk subgroups defined by a unique development over time could be useful for the detection of rare mutations in common multi-factorial diseases.

## Background

Longitudinal data analysis in genetic research is a new and emerging field with great potential. Genetic analysis of cross-sectional data generally assumes homogeneity in a sample with regard to the observed phenotype. However, longitudinal follow-up on the outcome variables often suggests that, nevertheless, individuals may differ in their development over time. These individual differences may even cluster into distinct subgroups with diverse environmental and genetic risk factors. A method

that can be used to further explore this unobserved heterogeneity is growth mixture modelling (GMM) [1-4]. In GMM, the assumption of a single average growth curve is relaxed and different unobserved groups of individuals or latent subclasses are allowed to vary in their growth parameters, such as estimates of means, variances, and covariate influences. This flexible modelling framework allows for growth curves that differ in shape. Multi-level data, such as individuals nested in families, can easily be integrated into this model.

The Framingham Heart Study is one of the largest longitudinal clinical studies for which genetic material is available [5]. This study followed families over more than 50 years for common cardiac and metabolic disorders such as hypertension, diabetes, obesity, and heart disease. Current analyses have studied genetic and environmental risk factors in this sample by assuming a homogeneous population. In our study, we explore possible heterogeneity in this sample by relaxing the single-population assumption and allowing for parameter differences across unobserved subpopulations. Using GMM on systolic blood pressure (SBP) measures in 1060 males of the Original Cohort and the Offspring Cohort in 692 families, we detected three subclasses that varied significantly in their developmental trajectories. The growth curve of the first class ($n$ = 60 individuals) was characterized by a high mean SBP early in life and an early, steep slope. The second class ($n$ = 131 individuals) had a low mean SBP at a young age followed by a steep increase in SBP over time. The third class ($n$ = 869 individuals) could be conceptualized as a normative class. Members of this subclass had low SBP at Exam 1 and the SBP remained low throughout the follow-up exams. Because previous studies had suggested a risk locus for high SBP in males on chromosome 8 in these data [6], we tested 2340 single-nucleotide polymorphisms (SNPs) on this chromosome for association with the class membership probabilities of our model. The probability of being in Class 1 was significantly associated with the coding SNP rs1445404 in the gene *EYA* (eyes absent homolog 1 in Drosophila) ($p = 3.07 \times 10^{-13}$). This very rare variant represents a miss-sense mutation in exon 3 of the gene. The minor allele of this SNP was present in only four individuals from four different families. Mutations in *EYA* were found in patients with brachio-oto-renal syndrome, as well as in individuals with isolated renal malformations [7,8]. Renal malformations can cause high blood pressure early in life. This result needs to be replicated, but it suggests that analyzing genetic data stratified for high-risk subgroups defined by a unique development over time could give an advantage for the detection of high-risk and rare mutations in common multi-factorial diseases.

## Methods

We used 1060 male individuals in 692 families from the Original Cohort and the Offspring Cohort of the Framingham Heart Study. SBP measured at four time points and spanning about 30 years of follow-up was used as outcome variable, whereas body mass index (BMI) and treatment for hypertension (HTNRX) were included in the model as time-varying covariates. Individuals with missing values on the covariates were excluded from the analysis, but missing values on the outcome variable at any time point were estimated with maximum-likelihood estimation under the assumption of missing at random (MAR) [9]. We fitted a three-level GMM by relaxing the assumption of identical parameter values across all mixture groups. The model was estimated by maximum likelihood using estimation-maximization (EM) algorithm. The first level described the variation over time, the second level described the variation over individuals, and the third level described the variation over families. Age was allowed to vary across the cohorts at each time point. We tested the fit of the model to the data by comparing the Bayesian information criteria (BIC) of the different class solutions for the non-nested models, with smaller BIC values indicating a better model fit [10]. The entropy of the classification and the posterior probability of belonging to a single class were taken into consideration as well. The analysis was performed with the computer software program Mplus [11].

The three-level growth mixture model for systolic blood pressure $y_{tij}$ for time point $t$, individual $i$, and pedigree $j$ is described as follows using the individual-level latent class variable $c_{ij}$ with $K$ classes. The Level 1 model is

$$y_{tij}\,|_{c_{ij}=k} = b_{0ij} + b_{1ij}(\text{age}_{tij} - 50) + b_2(\text{age}_{tij} - 50)^2 + b_3(\text{age}_{tij} - 50)^3 + b_4 bmi_{tij} + b_5 htx_{tij} + b_6 htx_{tij}(\text{age}_{tij} - 50) + e_{tij}, \quad (1)$$

where $e_{tij}$ follows a first-order autocorrelation structure. The Level 2 model is

$$b_{01j}\,|_{c_{ij}=k} = b_{0j} + u_{0ij} \quad (2)$$

$$b_{1ij}\,|_{c_{ij}=k} = b_{1j} + u_{1ij}, \quad (3)$$

where the $u$ values are bivariate normal within latent class. The Level 3 model is

$$b_{0j}\,|_{c_{ij}=k} = b_{00k} + v_{0j} \quad (4)$$

$$b_{1j}\,|_{c_{ij}=k} = b_{10k} + v_{1j}, \quad (5)$$

where the $v$ values are bivariate normal within class and uncorrelated with the $u$ values. The latent class probabilities follow a multinomial logistic regression with random intercepts,

$$P(c_{ij} = k) = \frac{e^{\alpha kj}}{\sum_{s=1}^{K} e^{\alpha sj}} \quad (6)$$

$$\alpha_{kj} = \alpha + \zeta_j, \quad (7)$$

where $\zeta$ is normally distributed.

The estimated class membership probability was then used as phenotype in a quantitative trait (QT) association analysis by testing one class against the other two classes performed with the software program GOLDENHELIX [12]. We tested 2340 SNPs on chromosome 8 for association with this phenotype. The significance of association was tested with the correlation/trend regression test under the basic, allelic model. A cut-off value of $p < 10^{-8}$ was used for genome-wide significance. We then performed 100,000 permutations of the data to evaluate the significance of our finding.

## Results

We identified three distinct subgroups in this data set with regard to SBP development over time. The conventional 1-class random effect (multilevel) growth model was outperformed by a 2-class GMM in terms of BIC (BIC = 35527 for the 1-class model versus BIC = 33847 for the 2-class model). A 3-class model gave the lowest (best) BIC (BIC = 33845). In the 4-class model, BIC increased (BIC = 33865). The entropy for the 3-class model was not very high (entropy = 0.66). The classes varied in mean, slope, and shape of the growth curves (Figure 1). Class 1 consisted of 60 individuals with high SBP values early in life and a steep growth curve. Class 2 included 131 individuals. Members of this class started out with a low mean SBP, but developed a steep rise in SBP later in life. Class 3 contained a normative group of 869 individuals. The SBP measures in this group were normal and remained normal throughout follow-up.

Association analysis with SNPs on chromosome 8 revealed two signals with genome-wide significance. The first one was an association between Class 1 membership probability and SNP rs1445404 located in the third exon of the gene *EYA* (eyes absent homolog 1 in Drosophila) ($p = 1.39 \times 10^{-13}$, OR>8.1). A total of four individuals, one homozygote and three heterozygotes from four different families, had a rare C allele instead of the wild-type G allele (6.6% of the individuals in Class 1). This miss-sense mutation in exon 3 changes an

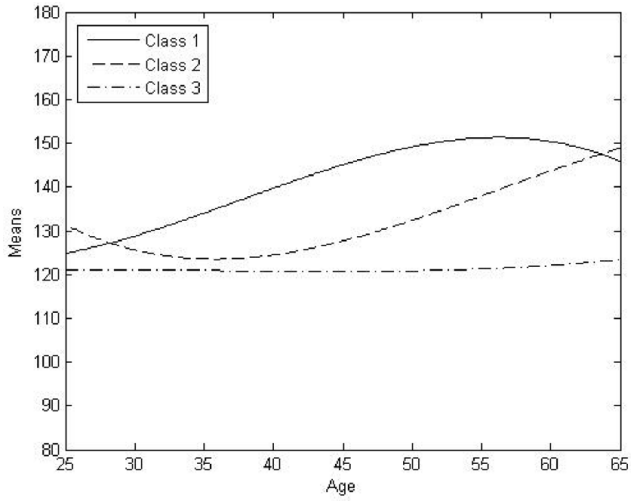


**Figure 1**
**Growth curves of the three latent classes over time.** Age is on the x-axis. The mean systolic blood pressure (SBP) is indicated on the y-axis.

alanine to a proline at amino acid position 20 of the protein, with likely consequences for the protein structure and folding of the protein. The homozygous individual and two of the heterozygous individuals belonged to Class 1. Interestingly, one heterozygous individual was assigned to Class 3; however, this individual was the youngest in this group of mutation carriers and the only one who was treated for hypertension as early as Exam 2. He also developed cardiac disease at age 49. This longitudinal course could indicate a misclassification due to very aggressive treatment of SBP at a very early stage.

The second signal was an association between Class 1 membership and the SNP rs6601495 located in the second exon of the gene *RP1L1* (retinitis pigmentosa 1-like) ($p = 3.8 \times 10^{-13}$). This miss-sense mutation changes serine to threonine at amino acid position 112. Two individuals in this data set, one homozygote and one heterozygote, carried the rare C allele and both were members of Class 1. The homozygous individual was also homozygous for the SNP rs1445404. Whereas the C allele of this SNP is absent in the European and Asian population, homozygotes and heterozygotes for this variant are common among the Sub-Saharan African population (allele frequency of the C allele: 0.86). Whether this genomic variant indicates an ethnic admixture in this data set remains to be explored. Permutation testing with 100,000 permutations revealed a permutation $p$-value of 0.0001 for the single-marker permutation for marker rs1445404 and 0.0003 for marker rs6601495, indicating that of 100,000

permutations of the phenotype, only 10 and 30 permutations, respectively, reached the same or better results than indicated here. Permutation of the full model taking all the markers on chromosome 8 into account revealed a permutation *p*-value = 0.0044 for marker rs1445404, and 0.0085 for marker rs6601495 after Bonferroni correction for multiple testing.

## Discussion

We demonstrated here that GMM is a powerful tool to address unobserved population substructure in longitudinal data sets. By assigning individuals into different risk groups based on phenotype development over time, we were able to identify rare genomic variants that were present only in one group and absent in the others. Our approach used a two-stage design, in which we first defined class membership probabilities, and in a second step performed a quantitative trait association analysis. This approach may be biased. The low number of individuals who carried the identified rare mutations prohibited incorporating the genotype information of the SNPs directly into the model. The fact that carriers of these rare mutations were found only in one class and not in the other classes, however, justified our approach. Our study was limited by the very small size of the high-risk group, which might lead to spurious associations. In order to correctly interpret our finding, it would be necessary to replicate the results in a larger sample. Because statistical replication might require very large data sets, given the very rare nature of the mutation, an alternative approach would be a biological validation. Functional consequences of the mutation could be tested for by renal ultrasound or renal function tests in the affected individuals. A further limitation of our approach is the sensitivity to population stratification and admixture.

## Conclusion

GMM is a useful tool to detect subgroups in heterogeneous populations. We demonstrate here that family structure can easily be incorporated into the model. We successfully identified a high-risk group with steep growth over time. Members of this latent class had high blood pressure early in life with continuous steep increase. The class membership probability showed significant association with a rare mendelian variant in a gene that is involved in renal development. However, for correct interpretation of this result, replication in a larger sample or biological validation would be essential. Our approach may be a useful complement to the commonly used case/control association design because it provides more power to identify rare variants associated with a severe phenotype.

## List of abbreviations used

BIC: Bayesian information criterion; BMI: Body mass index; EM: Estimation maximization; GMM: Growth mixture model; HTNRX: Treatment for hypertension; QT: Quantitative trait; MAR: Missing at random; SBP: Systolic blood pressure; SNP: Single-nucleotide polymorphism.

## Competing interests

Dr. B. Kerner declares to have no financial interest or potential conflicts of interest. Dr. B. Muthen is a co-developer of the commercial computer program Mplus used in the analyses.

## Authors' contributions

BK carried out some of the growth mixture modelling and the association testing of the SNPs. She drafted the manuscript. BOM designed the growth mixture model and performed some of the analyses, wrote the input for Mplus, supervised the modelling, and edited the paper. BK and BOM participated equally in the design of the study and the interpretation of the results. Both authors read and approved the final manuscript.

## Acknowledgements

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. This work was supported by NIMH grant K08 MH74057-01 to BK. We are grateful to the researchers who have collected the data over the years and have made them available for analysis to the participants of GAW16. We also thank the families who participated in the study.

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/3?issue=S7.

## References

1. Muthén B and Shedden K: **Finite mixture modeling with mixture outcomes using the EM algorithm.** *Biometrics* 1999, **55:**463–469.
2. Muthén B, Brown CH, Masyn K, Jo B, Khoo ST, Yang CC, Wang CP, Kellam S, Carlin J and Liao J: **General growth mixture modeling for randomized preventive interventions.** *Biostatistics* 2002, **3:**459–475.
3. Muthén B: **Latent variable analysis: growth mixture modeling and related techniques for longitudinal data.** *Handbook of Quantitative Methodology for the Social Sciences* Newbury Park, Sage Publications: Kaplan D 2004, 345–368.
4. Muthén B and Asparouhov T: **Growth mixture modeling: analysis with non-Gaussian random effects.** *Longitudinal Data Analysis* Boca Raton, Chapman & Hall, CRC Press: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G 2008, 143–165.
5. Cupples LA, Arruda HT, Benjamin EJ, D'Agostino RB Sr, Demissie S, DeStefano AL, Dupuis J, Falls KM, Fox CS, Gottlieb DJ, Govindaraju DR, Guo CY, Heard-Costa NL, Hwang SJ, Kathiresan S, Kiel DP, Laramie JM, Larson MG, Levy D, Liu CY, Lunetta KL, Mailman MD, Manning AK, Meigs JB, Murabito JM, Newton-Cheh C, O'Connor GT, O'Donnell CJ, Pandey M, Seshadri S, Vasan RS, Wang ZY, Wilk JB, Wolf PA, Yang Q and Atwood LD: **The Framingham Heart Study 100 K SNP genome-wide association study resource: overview of 17 phenotype working group reports.** *BMC Med Genet* 2007, **8(Suppl 1):**S1.

6. Briollais L, Tzontcheva A and Bull S: **Framingham Heart Study. Multilevel modeling for the analysis of longitudinal blood pressure data in the Framingham Heart Study pedigrees.** *BMC Genet* 2003, **4(Suppl 1):**S19.
7. Hoskins BE, Cramer CH II, Tasic V, Kehinde EO, Ashraf S, Bogdanovic R, Hoefele J, Pohl M and Hildebrandt F: **Missense mutations in EYA1 and TCF2 are a rare cause of urinary tract malformations.** *Nephrol Dial Transplant* 2008, **23:**777–779.
8. Orten DJ, Fischer SM, Sorensen JL, Radhakrishna U, Cremers CW, Marres HA, Van Camp G, Welch KO, Smith RJ and Kimberling WJ: **Branchio-oto-renal syndrome (BOR): novel mutations in the EYA1 gene, and a review of the mutational genetics of BOR.** *Hum Mutat* 2008, **29:**537–544.
9. Little RJ and Rubin DB: **Statistical Analysis with Missing Data.** New York, John Wiley & Sons; Second2002.
10. McLachlan GJ and Peel D: **Finite Mixture Models.**New York, Wiley & Sons; 2000.
11. Muthen LK and Muthen BO: **Mplus User's Guide.**Los Angeles, Muthen & Muthen; Third1998 http://www.statmodel.com/.
12. **GOLDENHELIX.** http://www.goldenhelix.com.