HHS Public Access

Author manuscript

Ann Gerontol Geriatr Res. Author manuscript; available in PMC 2015 April 15.

Published in final edited form as:

Ann Gerontol Geriatr Res. 2014 ; 1(4): .

Genetic Structures of Population Cohorts Change with Increasing Age: Implications for Genetic Analyses of Human aging and Life Span

Anatoliy I. Yashin[1,2], **Deqing Wu**[1], **Konstantin G. Arbeev**[1], **Liubov S. Arbeeva**[1], **Igor Akushevich**[1], **Alexander Kulminski**[1], **Irina Culminskaya**[1], **Eric Stallard**[1], and **Svetlana V. Ukraintseva**[1,2,*]

[1]Biodemography of Aging Research Unit, Duke University, USA

[2]Duke Cancer Institute, Duke University, USA

Abstract

Background—Correcting for the potential effects of population stratification is an important issue in genome wide association studies (GWAS) of complex traits. Principal component analysis (PCA) of the genetic structure of the population under study with subsequent incorporation of the first several principal components (PCs) in the GWAS regression model is often used for this purpose.

Problem—For longevity related traits such a correction may negatively affect the accuracy of genetic analyses. This is because PCs may capture genetic structure induced by mortality selection processes in genetically heterogeneous populations.

Data and Methods—We used the Framingham Heart Study data on life span and on individual genetic background to construct two sets of PCs. One was constructed to separate population stratification due to differences in ancestry from that induced by mortality selection. The other was constructed using genetic data on individuals of different ages without attempting to separate the ancestry effects from the mortality selection effects. The GWASs of human life span were performed using the first 20 PCs from each of the selected sets to control for possible population stratification.

Results—The results indicated that the GWAS that used the PC set separating population stratification induced by mortality selection from differences in ancestry produced stronger genetic signals than the GWAS that used PCs without such separation.

Conclusion—The quality of genetic estimates in GWAS can be improved when changes in genetic structure caused by mortality selection are taken into account in controlling for possible effects of population stratification.

[*]**Corresponding author** Svetlana V. Ukraintseva, Biodemography of Aging Research Unit, 2024 West Main Street, Box 90420, Durham, NC 27705, USA, Tel: 1-919-668-2712; Fax: 1-919-684-3861; svo@duke.edu.

Keywords

Genetics of aging; Longevity; Genetic associations; Principal component analysis; Genetic structure; Mortality selection; Heterogeneous population

INTRODUCTION

Despite evident progress in understanding systemic biological mechanisms involved in the regulation of aging related changes and life span in humans, many problems remain unsolved and continue challenging researchers. One such problem deals with the low effectiveness of the genome wide association studies (GWAS) of human aging and longevity: most estimated associations do not reach the genome wide level of statistical significance and suffer from the lack of replication [1–5]. Several reasons are likely to be responsible for this situation. Their better understanding will create a background for obtaining more accurate estimates of genetic parameters.

One reason is fundamental: it is due to the fact that complicated relationships between genotypes and phenotypes are not properly described by the simple statistical models used in GWAS. For example, the effects of some genetic variants on mortality risks differ at distinct age intervals (e.g., harmful effects of genetic markers at one age interval can become neutral or even beneficial at the next [6–8]. These effects may depend on other genetic variants carried by individuals, as well as non-genetic (e.g., environmental, behavioral, social-economic) factors that change over age and time. Such relationships between genetic factors and mortality risks are difficult to study efficiently in the framework of Cox's regression model. This is because Cox's model assumes the proportionality of hazards. This assumption does not allow us to distinguish between intersecting hazard rates for carriers and non-carriers of genetic variants. The genetic factors with such effects were detected in [6, 7] and later confirmed in [9–11]. More sophisticated approaches are needed to address the complexity of the dynamic relationships between genotypes and phenotypes [12].

Another reason is less fundamental but not less important. It may be due to the underutilized potential of data available for analyses. Using the fact that human populations are genetically heterogeneous, and that the absence of genetic measurements is a special kind of missing data problem, one can develop efficient methods of analyses that take such incompleteness of data into account. Approaches that benefit from joint analyses of genetic and non-genetic data in genetic studies of centenarians are described in [6, 7]. More sophisticated elaborations of these problems are described in [13–15].

A third reason is the lack of attention to the bio demographic nature of the data on human life span. These data reflect the process of mortality selection in a genetically heterogeneous population. Individuals in such populations have genetic differences in individual susceptibility to death. The presence of such differences is assumed in any genetic study of human longevity. In [6, 7] we showed how the effectiveness of statistical methods can be substantially improved when the genetic heterogeneity of the population under study is taken into account in statistical analyses of data.

One more opportunity is to make the procedure of controlling for population stratification in GWAS of human longevity more efficient. This opportunity is investigated below. We show that methods of correcting for the potential effects of population stratification traditionally used in GWAS have to be used with care when the data are subject to mortality selection. Improper use of these methods may substantially reduce the accuracy of genetic analyses in GWAS of human longevity. We present an approach that avoids such problems and improves p-values of associations estimated in GWAS of human longevity.

MATERIALS AND METHODS

Traditional methods of controlling for population stratification using PCA do not pay attention to the age structure of the population at the time of bio-specimen (e.g., blood) collection [16]. Such an approach may result in inefficient genetic analyses of human longevity because in addition to genetic differences in ancestry the selected PCs may capture genetic structure induced by mortality selection processes. The development of efficient methods of GWAS requires separating the effects of population stratification due to differences in ancestry from those induced by mortality selection. For this purpose we used data on genotyped individuals collected in the Framingham Heart Study (FHS) [17]. The genetic data were represented by 550,000 SNPs. Genotyping was conducted using Affymetrix 500K and 50K (non-overlapping) arrays. The life span data were available for 1529 participants from the Original FHS cohort. The quality control (QC) procedure included 95% call rate for the sample and 95% call rate for SNPs, HWE p-value >1E-7. After applying the QC procedure, the data on 1111 individuals with data on life span and 429,783 SNPs were available for analyses. Following Price [16] we constructed a set of principal components (PC)s using 1009 unrelated individuals out of 1111 members of the Original Framingham cohort. Note that this group included individuals of different ages. Then we applied a mixed effect model realized in the EMMAX computer program [18] to perform GWAS of human life span separately for 432 males and 679 females using the first 20 PCs, smoking habit (ever or never), and birth cohorts as observed covariates.

Then we constructed the second set of PCs using data on 1625 unrelated members of the Off spring FHS cohort whose ages did not exceed 60 years at the time of bio specimen collection and repeated GWAS of life span data on the same individuals with the new set of PCs keeping other covariates the same. Note that almost all individuals from this cohort have data on life span. For 204 study subjects with censored life spans the life span data were imputed by adding mean residual life span to each age at censoring. We expected that the p-values of genetic associations obtained in GWAS of human life span for the same genetic variants will be smaller for analyses based on the second set of PCs (constructed from the data on younger individuals from the Offspring FHS cohort) than in those based on the corresponding analyses of the first set of PCs (constructed from the data on individuals of all age categories from the Original FHS cohort). The QQ-plots of observed p-values versus p-values corresponding to the null-hypotheses are used to compare the results of analyses.

RESULTS AND DISCUSSION

The results of the analyses are shown in the QQ-plots in the four panels of Figure 1 for males and females. The association signals were substantially stronger in the case where the PCs were constructed using data on individuals whose ages did not exceed age 60. These results indicate that the effects of mortality selection may influence the genetic structure of the population cohort when its members are getting older. If at the time of bio-specimen collection the study population comprises individuals of different ages (e.g., young adults; old and oldest-old individuals) then, in addition to genetic structure due to differences in ancestry, the PCs constructed from genetic data on such a population will capture genetic structure due to mortality selection in the subgroup of individuals surviving to the old and oldest-old ages. In this case the correction for population stratification performed by including PCs as observed covariates in genetic models will tend to reduce or nullify the estimates of the associations of genetic variants with the targeted longevity traits (i.e., genetic variants increasing or reducing mortality risks).

Controlling for Population Stratification in GWAS

Population stratification is caused by nonrandom mating between groups of individuals. This is often due to earlier physical separation of subpopulations followed by different patterns of genetic drift of allele frequencies in each group. It may take many generations of random mating to eliminate this type of stratification. Population stratification can be a problem for genetic studies of complex traits where the association found could be due to the underlying structure of the population. A widely used approach for correcting for population stratification is based on principal component analysis (PCA) [16]. The method identifies several top principal components (PCs) and uses them as observed covariates in the association analyses. However, the PCA approach may not properly adjust for population stratification in genetic studies of human longevity if the population under study at the time of bio-specimen (e.g., blood) collection is represented by individuals of different age groups including young adults and old and oldest-old individuals. This is because mortality selection may generate additional genetic structure in the study population. This additional structure which involves genetic variants affecting life span may be inadvertently captured by the selected principal components. Hence, controlling for the potential effects of population stratification in GWAS of human aging and longevity may reduce or completely eliminate the association of genetic variants with longevity traits, i.e., it may weaken the signals from the genetic variants one is trying to detect.

Mortality selection modifies the genetic structure of the population

Genetic analyses of human longevity are based on the assumption that some genetic variants or combinations of such variants have positive or negative effects on human life span. In other words these genetic factors may reduce or increase mortality risk. Many other variants and genetic combinations have nothing to do with longevity and they do not affect mortality risk. This means that the population under study is genetically heterogeneous. It is well known from demography that heterogeneous population cohorts experience a process of mortality selection in which genetically frail or vulnerable individuals tend to die first leaving more genetically robust individuals in the cohort [19]. This process changes the

genetic structure of the population cohort, which means that the frequencies of genetic variants as well as combinations of variants that affect life span (mortality risks) change as the age of the cohort members increases. An important insight from this observation is that the genetic structure of a population comprising sub-populations of individuals of different ages may differ from that of a population of individuals of the same age, e.g., a given birth cohort, for which the only source of genetic structure at birth, or at any age prior to significant selective mortality, is due to differences in ancestry.

Although the need for correcting for the potential effects of population stratification in GWAS is clear and is recognized in many publications, most researchers do not recognize that the genetic structure generated by mortality selection should be distinguished from that due to differences in ancestry. Ignoring the difference between these two sources of genetic structure may result in the loss of power in genetic analyses, weak genetic signals; and failure to replicate research findings. Hence, the straight forward use of PCA may be an inefficient method for correcting for the potential effects of population stratification in GWAS of human longevity.

CONCLUSION

These analyses underscore the importance of taking the bio demographic nature of the data into account in genetic association studies. Specifically, if genes are involved in life span regulation then the genetic structure of the population cohorts must change with age. The presence of individuals of different ages (including old and oldest-old individuals) in the study population induces additional genetic structure due to mortality selection. This structure may be captured by the PCs used to correct for the potential effects of population stratification. The use of such PCs in GWAS regression models may compromise the results of the analyses: the estimates of the effects of genetic variants affecting mortality risk may be reduced. A practical way of addressing this problem is to construct PCs using the younger part of the study population (e.g., the subpopulation of offspring that have not reached age 60 years) that did not yet experience significant mortality selection. Such PCs are likely to capture the effects of population stratification due to differences in ancestry. At the same time, they will allow an evaluation of the effects of genetic variants involved in the mortality selection process. Detecting such variants is the primary goal of the GWAS of human life span.

The evaluation of how much the initial genetic structure of the cohort will change in transition from young to the old and from old to the oldest old ages requires additional study. The results will depend on our understanding of the details of genetic mechanisms involved in regulation of aging and longevity, the number of genes involved in mortality selection process in humans, the effects of their mutual interactions, as well as their interactions with external forces and conditions on life span. These analyses will require new models of mortality selection in genetically heterogeneous populations. Such models have to allow for incorporating available knowledge about genetics of aging and longevity in statistical analyses of data, testing hypotheses about mechanisms of genetic influence on health and longevity related traits, roles of separate genes, signaling and metabolic pathways, as well as genetic networks in the process of aging and mortality.

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Numbers R01AG046860, P01AG043352, and P30AG034424. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. 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.

ABBREVIATIONS

GWAS	Genome Wide Association Studies
PCA	Principal Component Analysis
PC	Principal Component

REFERENCES

1. Deelen J, Beekman M, Uh HW, Helmer Q, Kuningas M, Christiansen L, et al. Genome-wide association study identifies a single major locus contributing to survival into old age the APOE locus revisited. Aging Cell. 2011; 10:686–698. [PubMed: 21418511]
2. Nebel A, Kleindorp R, Caliebe A, Nothnagel M, Blanché H, Junge O, et al. A genome-wide association study confirms APOE as the major gene influencing survival in long-lived individuals. Mech Ageing Dev. 2011; 132:324–330. [PubMed: 21740922]
3. Walter S, Atzmon G, Demerath EW, Garcia ME, Kaplan RC, Kumari M, et al. A genome-wide association study of aging. Neurobiol Aging. 2011; 32:2109. [PubMed: 21782286]
4. Anne BN, Stefan W, Kathryn LL, Melissa EG, Eline PS, Kaare C, et al. A Meta-analysis of Four Genome-Wide Association Studies of Survival to Age 90 Years or Older: The Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. J Gerontol A Biol Sci Med Sci. 2010; 65A:478–487.
5. Kathryn LL, Ralph BDA, David K, Emelia JB, Chao YG, Raju G. Genetic correlates of longevity and selected age-related phenotypes: a genome-wide association study in the Framingham Study. BMC Med Genet. 2007; 8:S13. [PubMed: 17903295]
6. Yashin AI, De BG, Vaupel JW, Tan Q, Andreev KF, Iachine IA, et al. Genes, demography, and life span: The contribution of demographic data in genetic studies on aging and longevity. Am J Hum Genet. 1999; 65:1178–1193. [PubMed: 10486337]
7. Yashin AI, De BG, Vaupel JW, Tan Q, Andreev KF, Iachine IA, et al. Genes and longevity: lessons from studies of centenarians. J Gerontol A Biol Sci Med Sci. 2000; 55:319–328.
8. Kulminski AM, Culminskaya I, Arbeev KG, Ukraintseva SV, Stallard E, Arbeeva L, et al. The role of lipid-related genes, aging-related processes, and environment in healthspan. Aging Cell. 2013; 12:237–246. [PubMed: 23320904]
9. Atzmon G, Rincon M, Schechter CB, Shuldiner AR, Lipton RB, Bergman A, et al. Lipoprotein genotype and conserved pathway for exceptional longevity in humans. PLoS Biol. 2006; 4:e113. [PubMed: 16602826]
10. Bergman A, Atzmon G, Ye K, MacCarthy T, Barzilai N. Buffering mechanisms in aging: a systems approach toward uncovering the genetic component of aging. PLoS Comput Biol. 2007; 3:e170. [PubMed: 17784782]
11. Huffman DM, Deelen J, Ye K, Bergman A, Slagboom EP, Barzilai N, et al. Distinguishing between longevity and buffered-deleterious genotypes for exceptional human longevity: the case of the MTP gene. J Gerontol A Biol Sci Med Sci. 2012; 67:1153–1160. [PubMed: 22496539]

12. Arbeev KG, Akushevich I, Kulminski AM, Arbeeva LS, Akushevich L, Ukraintseva SV, et al. Genetic model for longitudinal studies of aging, health, and longevity and its potential application to incomplete data. J Theor Biol. 2009; 258:103–111. [PubMed: 19490866]
13. Yashin AI, Arbeev KG, Ukraintseva SV. The accuracy of statistical estimates in genetic studies of aging can be significantly improved. Biogerontology. 2007; 8:243–255. [PubMed: 17160500]
14. Yashin AI, Arbeev KG, Wu D, Arbeeva LS, Kulminski AM, Akushevich I, et al. How the quality of GWAS of human lifespan and health span can be improved. Front Genet. 2013; 4:125. [PubMed: 23825477]
15. Arbeev KG, Ukraintseva SV, Arbeeva LS, Akushevich I, Kulminski AM, Yashin AI. Evaluation of genotype-specific survival using joint analysis of genetic and non-genetic subsamples of longitudinal data. Biogerontology. 2011; 12:157–166. [PubMed: 21193960]
16. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]
17. Giroux E. The Framingham Study and the Constitution of a Restrictive Concept of Risk Factor. Social History of Medicine. 2013; 26:94–112.
18. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–354. [PubMed: 20208533]
19. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography. 1979; 16:439–454. [PubMed: 510638]

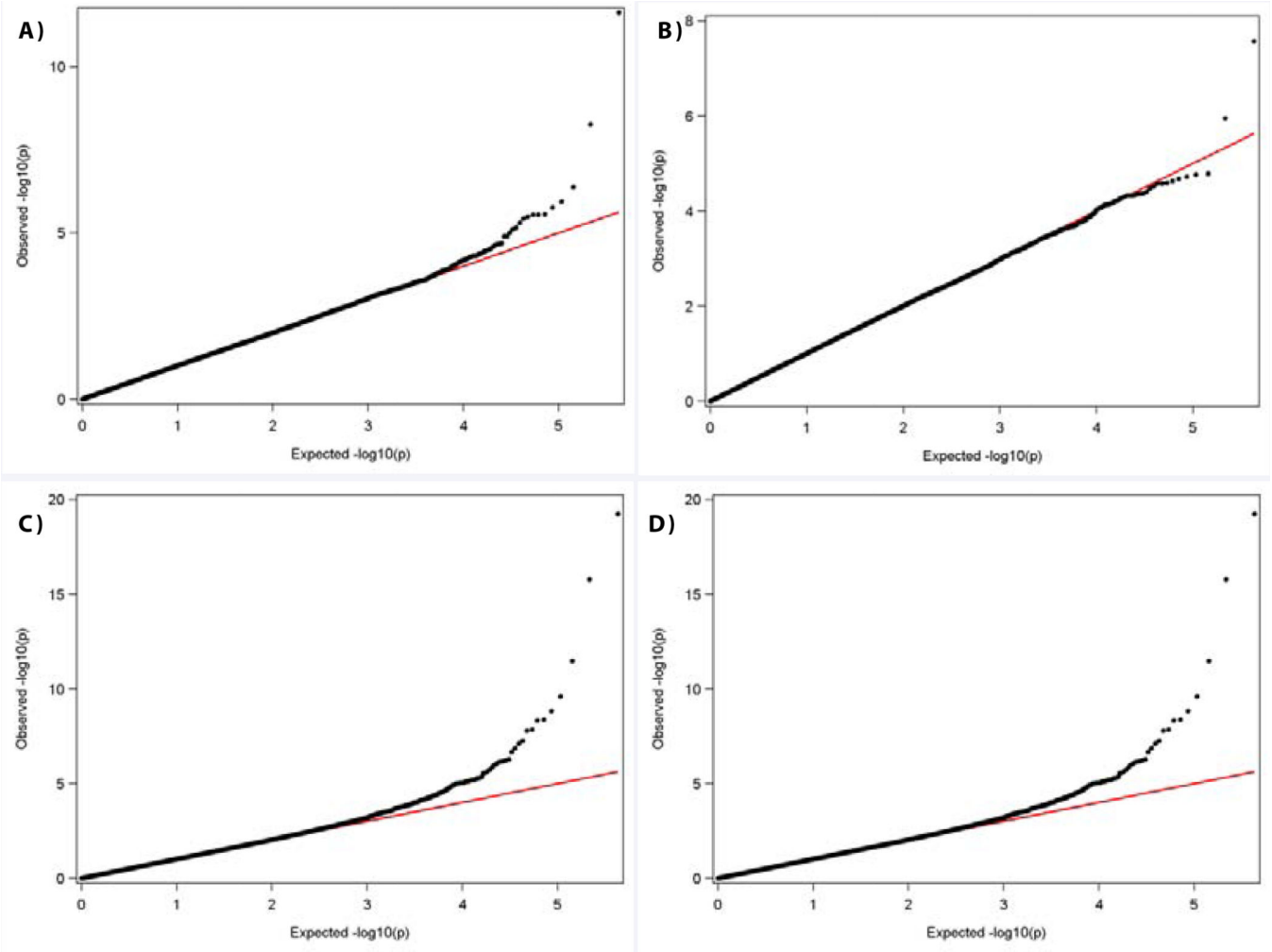

Figure 1.

QQ plots for the results of GWAS of human life span for the members of the Original Framingham cohort obtained using the EMMAX program controlling for birth cohorts, smoking, and 20 PCs. Top panels represent the results for 679 females (**A**) and 432 males (**B**) obtained using 20 PCs constructed using genetic data on 1009 unrelated subjects from the Original Framingham cohort (without separation effects of PS due to differences in ancestry from that due to mortality selection in a genetically heterogeneous population). Bottom panels represent the results for 679 females (**C**) and 432 males (**D**) obtained using 20 PCs constructed using genetic data on 1625 unrelated subjects from the Framingham cohort. The PCs were constructed using genetic data on 1625 unrelated subjects from the offspring FHS cohort among individuals whose age was less than or equal to 60 years at the time of bio specimen collection.