Why Demographic Analysis Is Crucial in Studying Genetics of Exceptional Longevity

Anatoliy I. Yashin, Deqing Wu, Konstantin Arbeev, Liubov Arbeeva, Alexander Kulminski, Igor Akushevich, Irina Culminskaya, Eric Stallard, Svetlana Ukraintseva

Center for Population Health and Aging, Duke University

Background and Objective. Population stratification characterized by differences in genetic frequencies in subpopulations comprising population under study is an important problem in genome-wide association studies (GWAS) of human aging and longevity. The use of ten or twenty first principal components (PC) as additional observed covariates in statistical procedure of selecting genetic variants was recommended to cope with this problem. The objective of this paper is to show that such recommendation has to be used with care because stratification may be generated by genes we are looking for in the process of mortality selection.

Data and Methods. We performed GWAS of human lifespan using data on the Original cohort of the Framingham Heart Study (FHS) with and without controlling for population stratification using 20 PC components.

Results. We showed that population stratification does take place in the data and it can be eliminated by adding 20 estimated PCs to the list of observed covariates used in the GWA procedure. Then we also showed that this stratification can be eliminated by controlling for the age at genotyping in the Original FHS cohort.

Conclusion. Mortality selection can produce population stratification in populations with left truncated subsamples of the data. Such stratification contains important information about genetic influence on lifespan and correlated traits.

Introduction

The hypothesis about existence of genetic variants capable of making substantial individual contributions to longevity in a wide range of internal and external conditions ("longevity" genes) was in the focus of many studies of aging and longevity during several last decades. The search for such genes has been motivated by advances in experimental studies of aging and longevity in animal model systems in which single genetic mutations showed substantial influence on lifespan (Kenyon et al., 1993; Bartke et al., 2001; Martin 2005; 2011). In this paper we describe the results of genome wide association study (GWAS) of human lifespan which show that an alternative modulation of human lifespan is possible. We show that proper interpretation of the results of genetic analyses require understanding demographic mechanisms of mortality selection in genetically heterogeneous populations. The use of demographic principles allows for developing efficient methods of genetic analyses of human lifespan. The connections of obtained research findings with evolutionary theories of aging and accumulated evidence about evolutionary conserved genes and molecular biological pathways involved in aging are discussed. .

Data and Methods

Framingham Heart Study Data. The FHS Original cohort was launched at Exam 1 in 1948 and has continued with biennial examinations to the present. The Original FHS cohort consists of 5,209 respondents (55% females) aged 28–62 years residing in Framingham, Massachusetts, between 1948 and 1951. Nearly all subjects were Caucasians. The examination included an interview, physical examination, and laboratory tests. One thousand four hundred and seventy

one participants of the Original FHS cohort were genotyped. Individual information on the SNP genotyping and phenotypic traits collected in the FHS was obtained through the dbGaP website (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v3.p2).

We combined data on 500K and 50K SNPs available in the Framingham data. This procedure resulted in 549,157 SNPs. We applied GWAS quality control procedures to the data on genotyped individuals from the original FHS cohort: sample call rate ≥ 95%; MAF > 5%; HWE > 10^{-4}. Altogether 1,111 individuals (432 males and 679 females) had a sample call rate ≥ 95%, 203 people had censored lifespans.

It is important to note that a substantial part of the Original FHS cohort has information on life spans of the study participants. To be able to use a mixed effect model in the analyses of data on life span using EMMAX, we estimated residual lifespans for individuals censored at a given age by calculating average lifespan of deceased study participants who survived up to this age. Then we used these estimates to impute lifespan for individuals with censored data. In case of 95% sample call rate lifespans of 203 individuals were imputed.

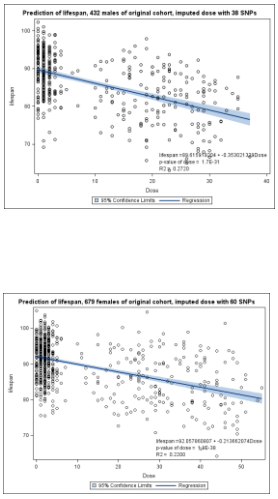


Fig. 1. Lifespan as function of vulnerability alleles for males (top) and females (bottom) panels.

Results. The GWAS of lifespan data (which used an additive genetic model)) resulted in 38 and 60 genome-wide significant genetic variants for males and females, respectively, with 24 common SNPs for the two genders. Note that all these variants have negative associations with lifespan so they will be called "frailty" or "vulnerability" alleles. In these analyses we controlled for observed covariates including smoking and birth cohorts. In order to evaluate how lifespan depends on the number of frailty alleles we calculated the number of such alleles contained in the genomes of each genotyped study participant, and estimated parameters of linear regression models considering individual's life span as dependent and the number of frailty alleles in his/her genome as independent variables. The results are shown in Fig. 1a and Fig 1b for males and females respectively.

One can see from these figures that the cumulative index has a negative effect on lifespan and that this effect is substantial, highly statistically significant and explains about 27 and 22 percent of phenotypic variance for males and females, respectively.

The QQ plots resulting from these analyses are shown in Fig. 2 for males (left) and females (right).

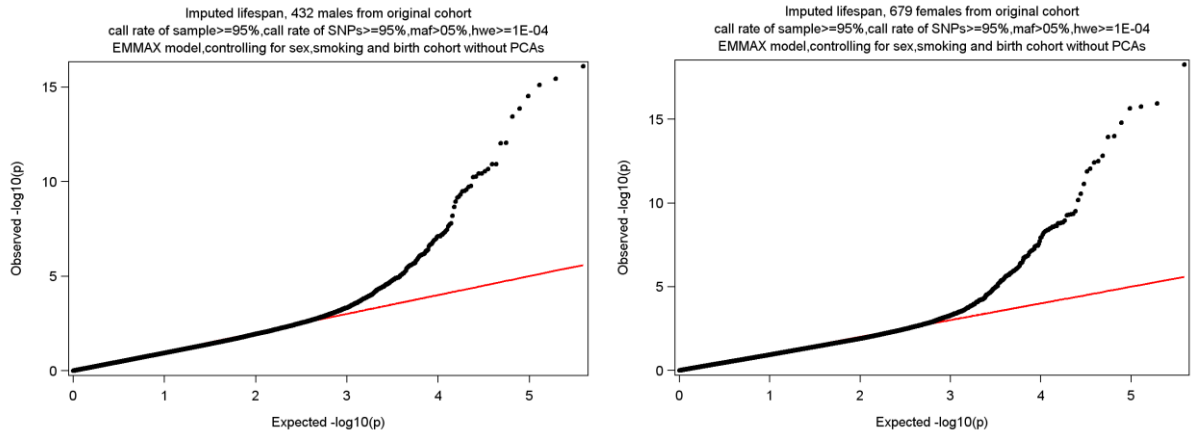


Fig. 2 is about here. The sharp increase in the QQ-plots after $-\log_{10}(p) = 3$ of expected value indicates possible presence of population stratification in genetic data.

One can see from these figures that population stratification in genetic data (i.e., genetic clustering due to differences in ancestry) is likely to take place. One way of controlling for the effects of population stratification is to follow strategy described in Price et al. (2006). These include performing principal component analyses (PCA) of genetic data, identifying PCs for each individual, and use up to 20 of PC components as observed covariates in the allele selection procedure. We performed such analyses of data on lifespans of individuals from the Original FHS cohort. The QQ plot resulting from these analyses are shown in Fig. 3 for males and females, respectively.

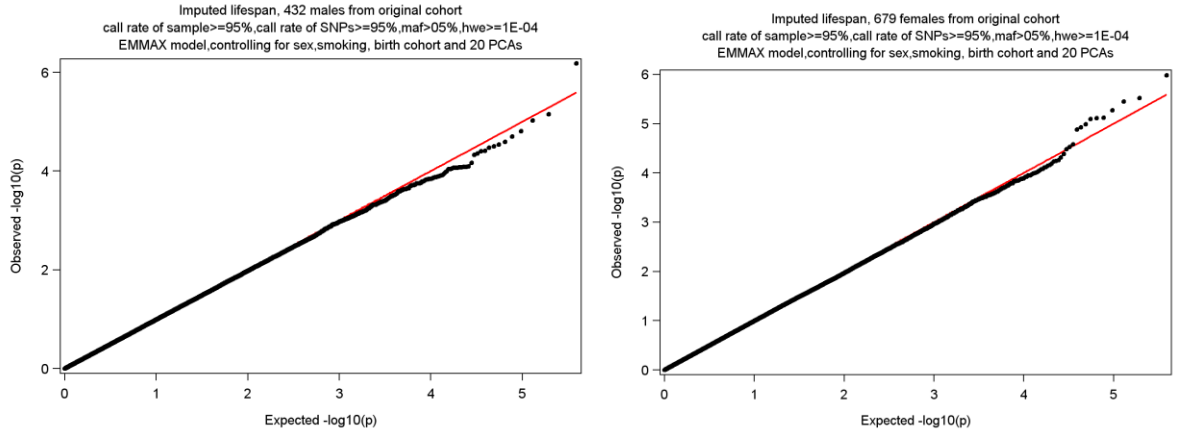


Fig. 3 is about here.
The QQ plots of the results of GWAS of human lifespan corrected for population stratification.

One can see from this figure that the corrected QQ plots look more appropriate: the population stratification effect disappeared. However, all genome-wide significant associations with lifespan disappeared as well. Such radical changes in the QQ plots in response to controlling for population clustering suggest an idea that the use of 20 principal components in the regression model in GWAS could remove part of genetic "structure" induced by mortality selection in genetically heterogeneous population due to the beneficial effects of longevity alleles and deleterious effects of vulnerability alleles on survival.

A glance at the data from the Original FHS cohort indicates that "longevity related structure could result from distribution of the age at blood collection-among study participants

(Fig.4). Assuming that population sub-cohorts comprising the Original FHS cohort have similar genetic background and survival characteristics Fig. 4 illustrates the pattern of left truncation in the cohort data, which in case of genetic influence on lifespan can be responsible for genetic structure in population of the old and oldest old individuals, which will be interpreted as the presence of population stratification in genetic data. Indeed, individuals who survived to old ages are likely to carry a smaller number of vulnerability alleles in their genomes, so the frequencies of such alleles will be smaller in individuals who were genotyped at older ages compared to those genotyped at younger ages. As one can see from Fig. 4 a substantial part of the Original FHS cohort has been brought into the genetic study at age 85 years and older.

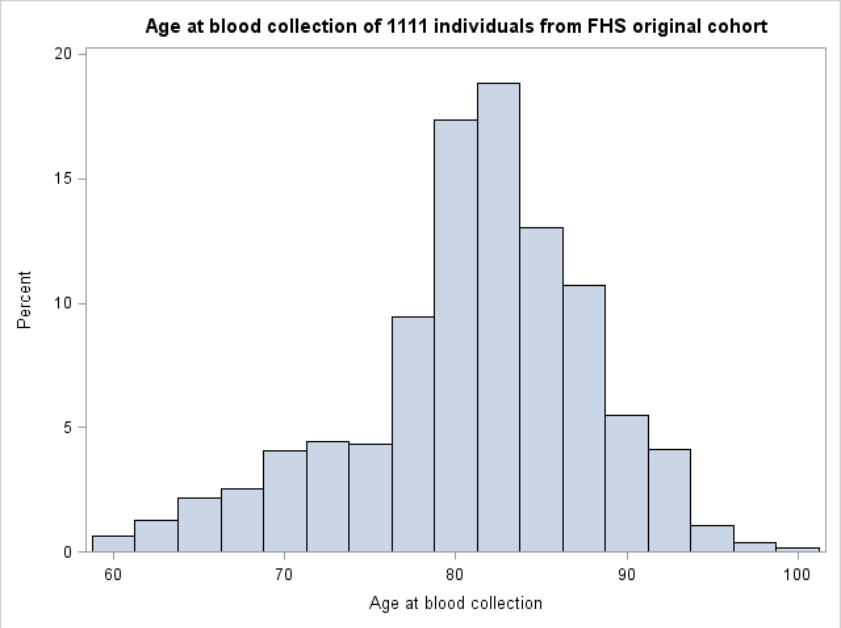


Fig. 4 is about here

Thus, if genetic influence on lifespan does take place and if it is realized through collective effect of large numbers of genetic variants each having a small negative effect on life span than these older groups of individuals are supposed to have smaller frequencies of "vulnerability" alleles. Controlling for population stratification using PCA approach described above will eliminate these effects even if no other sources of population clustering are available.

Thus, if the hypothesis that population stratification is generated by the mortality selection is correct then controlling for the age at genotyping in GWAS procedure should result in the QQ plots similar to those with the PCA correction. Fig. 5 shows that this is exactly the case.

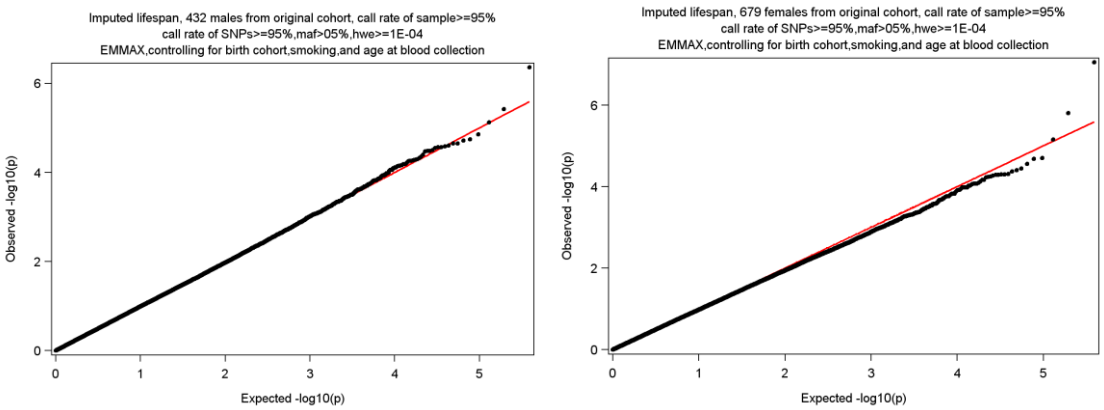


Fig. 5. QQ plots for the results of GWAS of human lifespan obtained using EMMAX program controlling for birth cohort, smoking (ever or never) and age at blood collection.

Table 1

rs num	Chr	# MA	# A	MAF	MAF HP	MAF 1000
rs5491	19	224	2128	**0.10563**		0.075
rs356430	5	218	2020	**0.107921**	0	0.017
rs1399453	12	225	2052	**0.109649**	0	0.024
rs1440483	11	190	2064	**0.092054**	0	0.054
rs1794108	11	159	2104	**0.07557**	0	0
rs2353447	8	230	2092	**0.109943**	0	0.02
rs2586484	17	236	2074	**0.11379**	0.008	0.012
rs2838566	21	252	2098	**0.120114**	0	0.11
rs3738682	1	166	2026	**0.081935**	0.017	0.011
rs4565533	9	447	2088	**0.21408**	0.09	0.06
rs4904670	14	291	2120	**0.137264**	0	0.03
rs5743998	11	198	2044	**0.096869**	0	0.012
rs6007952	22	356	2058	**0.172983**	0.05	0.06
rs6090342	20	266	2060	**0.129126**	0	0.28
rs7894051	10	426	2136	**0.199438**	0.05	0.1
rs8081943	17	148	2176	**0.068015**	0	0.03
rs8135777	22	216	1996	**0.108216**	0	0.023
rs9896996	17	209	2082	**0.100384**	0.035	0.04
rs9925881	16	144	2068	**0.069632**	0	0.05
rs9928967	16	137	2140	**0.064019**	0	0.03
rs9971555	11	232	2092	**0.110899**	0	0.02
rs10845099	12	380	2072	**0.183398**	0.093	0.323
rs11536959	20	155	2132	**0.072702**	0	0.017
rs17067605	5	167	2046	**0.081623**	0	0.004

Table 1 shows MAF of the 24 selected SNPs. The first column shows SNP rs-number; the second -- chromosome number; the third and the fourth columns show the number of minor alleles and the total number of alleles for the corresponding SNP, respectively; the fifth, sixth, and seventh columns show the SNPs' minor allele frequencies in our study, HAP/MAP, and "1000 genome" databases. One can see that chromosomes 2, 3, 4,6,7,9, 13, 15, 18, and 21 are not represented by the selected SNPs. Chromosome 11 is represented by 4 SNPs: rs1440483, rs1794108, rs5743998, and rs9971555. Chromosome 12 is represented by the two SNPs rs1399453 and rs1084509; three SNPs are in chromosome 17; rs2586484, rs8081943, and rs9896996; two others are on chromosome 20: rs6090342 and rs1153695; two are on chromosome 5 rs356430 and rs1706760; two are on chromosome 16 rs9925881 and rs9928967; one SNP is on chromosome 1: rs3738682. Similarly, chromosomes 8,9,19, 21 and 22 have only one SNP: rs2353447, rs4565533, rs2838566, rs6007952, and rs8135777, respectively.

Table 2, females

SNP		ln_a_1	b_1	ln_a_0	b_0	f	p-value
rs6007952	#*	-5.91	0.046	-11.28	0.102	0.7	2.33E-15
rs9971555	#*	-5.22	0.040	-10.78	0.097	0.8	1.44E-15
rs9896996	*	-4.48	0.031	-10.85	0.098	0.9	1.11E-15
rs2353447	#*	-5.00	0.038	-11.41	0.103	0.8	0.00E+00
rs4904670	#*	-5.86	0.048	-11.52	0.104	0.8	0.00E+00
rs6090342	#*	-5.62	0.044	-11.08	0.100	0.7	0.00E+00
rs17067605		-4.79	0.036	-10.55	0.094	0.8	3.33E-16
rs1440483		-4.81	0.036	-10.76	0.096	0.8	0.00E+00
rs2838566	#*	-4.88	0.036	-11.56	0.105	0.9	0.00E+00
rs9925881		-5.12	0.040	-10.03	0.089	0.7	3.22E-12
rs4565533	#*	-5.97	0.047	-12.45	0.114	0.8	0.00E+00
rs5743998		-4.74	0.034	-10.82	0.097	0.8	3.33E-16
rs3738682		-4.95	0.038	-10.65	0.095	0.8	0.00E+00
rs7894051	#*	-5.90	0.047	-12.37	0.113	0.8	0.00E+00
rs2586484	*	-5.19	0.039	-10.81	0.097	0.8	3.89E-15
rs356430	#*	-5.45	0.043	-10.61	0.095	0.7	7.77E-16
rs10845099	*	-5.36	0.039	-11.26	0.102	0.8	2.33E-12
rs8135777	#*	-5.61	0.044	-10.56	0.094	0.7	6.00E-15
rs11536959		-5.19	0.040	-10.21	0.090	0.7	5.02E-13
rs1399453	#*	-5.55	0.044	-10.61	0.095	0.7	1.44E-15
rs9928967		-4.21	0.029	-10.37	0.092	0.9	1.82E-14
rs5491	#*	-6.34	0.054	-10.56	0.094	0.6	0.00E+00
rs1794108		-4.81	0.036	-10.38	0.092	0.8	1.38E-14
rs8081943		-4.19	0.029	-10.52	0.094	0.9	0.00E+00

Table 3, males

SNP		ln_a_1	b_1	ln_a_0	b_0	f	p-value
rs6007952	#*	-5.55	0.045	-8.79	0.078	0.8	4.06E-09
rs9971555	#*	-6.12	0.054	-8.73	0.077	0.7	9.33E-14
rs9896996	*	-5.27	0.044	-8.57	0.075	0.8	2.49E-12
rs2353447	#*	-6.20	0.055	-8.38	0.073	0.7	2.12E-12
rs4904670	#*	-6.24	0.056	-9.81	0.088	0.8	0.00E+00
rs6090342	#*	-6.13	0.054	-8.74	0.077	0.7	1.12E-13
rs17067605		-5.25	0.044	-8.75	0.078	0.8	1.55E-14
rs1440483		-7.25	0.070	-8.65	0.076	0.6	0.00E+00
rs2838566	#*	-6.06	0.054	-9.47	0.085	0.8	0.00E+00
rs9925881		-5.65	0.049	-8.21	0.071	0.7	2.75E-11
rs4565533	#*	-5.798	0.048	-9.56	0.086	0.8	2.66E-13
rs5743998		-6.03	0.053	-8.23	0.072	0.6	6.10E-11
rs3738682		-6.15	0.056	-8.38	0.075	0.6	4.63E-14
rs7894051	#*	-6.00	0.051	-9.76	0.088	0.8	1.79E-14
rs2586484*		-6.40	0.057	-8.55	0.075	0.7	9.06E-14
rs356430	#*	-7.10	0.067	-8.26	0.072	0.5	2.80E-13
rs10845099*		-6.30	0.054	-8.84	0.078	0.7	7.22E-12
rs8135777	#*	-5.74	0.050	-8.54	0.075	0.7	8.03E-13
rs11536959		-5.30	0.045	-8.32	0.073	0.7	2.54E-12
rs1399453	#*	-5.85	0.051	-8.80	0.078	0.7	1.99E-14
rs9928967		-5.12	0.043	-8.45	0.074	0.8	1.41E-13
rs5491	#*	-6.37	0.056	-8.00	0.069	0.6	4.18E-09
rs1794108		-5.50	0.048	-8.55	0.075	0.7	3.50E-14
rs8081943		-5.74	0.050	-8.54	0.075	0.7	8.03E-13

Tables 2 and 3 show the results of association study performed with 24 selected genetic variants using the advanced method of genetic analyses. This method is a particular case of a more general approach described in Arbeev et al. (2011). The idea of the method is in use with genetic information contained in both population age distribution at the time of genotyping and in the follou up data. In this analyses we assumed that mortality rates for carriers and non-carriers of detected genetic variants are described by the Gompertz's curves. We derived likelihood functions for each of two subsets of data containing genetic information. Then we constructed the likelihood function for combined data as a product of the two likelihoods and use maximum likelihood method for estimating parameters of the two mortality rates (for carriers and non-carriers of corresponding variants). The null-hypothesis about the absence of genetic association (the two mortality rates are the same) was tested using the likelihood ratio test. The results are shown in Table 1 for females and Table 2 for males. One can see from these tables that the difference between mortality rates is highly statistically significant.

In the analyses of genetic connection with lifespan we imputed the data for individuals with censored lifespans by adding average residual lifespan to the age at censoring. Strictly

speaking the multiple imputation procedure has to be used in this case. Instead of doing this we performed alternative analyses of selected 24 genetic variants using a method that combines the two datasets containing genetic information. The results of this alternative analyses indicate that all selected SNPs are associated with lifespan, and this association is statistically significant.

The polygenic risk score ("genetic dose") constructed from 24 SNPs showed significant influence on lifespan with and without controlling for observed covariates.

Discussion

In this paper we provided evidence that human longevity can be modulated by a large number of genetic variants having negative influence on lifespan. In this context the exceptional longevity is likely to be the result of the absence of large number of "vulnerability" alleles rather than the presence of one or more "longevity" alleles in genomes of long lived individuals. Genetic analyses of complex traits involves population data which could involve sub-groups of individuals having different ancestry. The possibility of such situation, known as "population stratification", has to be taken into account in genetic analyses of complex traits. We showed that in genetic studies of human longevity population stratification may result from mortality selection in heterogeneous populations. We performed alternative association analyses of detected 24 genetic variants common for males and females and found that their associations with the trait are highly statistically significant. We described functions of genes linked to these variants and their roles in metabolic and signaling pathways. We also showed, that controlling for population stratification in genetic studies of aging and longevity using methods of principal component analyses has to be used with care. Mortality selection can produce effects of population stratification in populations with left truncated sub-cohorts. Such stratification contains important information about genetic influence on lifespan and correlated traits and has to be properly controlled.

Researchers studying the genetics of human aging and longevity recognize the complexity and multifactorial nature of the related traits. An important scientific question which has huge practical implications is about genetic contribution into human lifespan. A substantial role of single genes in lifespan is demonstrated by Mendelian inheritance disorders which effects can be: dominant (e.g., achondroplasia -- imperfect bone development causing dwarfism; Marfan syndrome - a connective tissue disorder causing long limbs and heart defects), recessive (cystic fibrosis - a disorder of the glands causing excess mucus in the lungs and problems with pancreas function and food absorption; sickle cells disease - a condition causing abnormal red blood cells; Tay Sachs disease - an inherited autosomal recessive condition that causes a progressive degeneration of the central nervous system which is fatal (usually by age 5), X-linked e.g, Duchenne's muscular dystrophy - a disease of muscle wasting; hemophilia - a bleeding disorder caused by low levels, or absence of, a blood protein that is essential for clotting). Mutations showing recessive effects may improve fitness under special conditions (e.g., sickle cells protect against malaria). Mutations with dominant effects may be results of mutation-selection bias.

The possibility that exceptional longevity can result from the absence of a large number of harmful genetic factors ("vulnerability" alleles), each contributing to a small increase in the chances of premature death has not been carefully investigated in GWAS. The existence of many such "vulnerability" genes is in accord with the mutation accumulation (MA) antagonistic pleiotropy (AP) hypotheses of aging (Medawar, 1952 Edney and /Gill, 1968). Both hypotheses have been used for explaining an increase in age trajectory of mortality rates with age. Although

studies testing these hypotheses keep producing controversial results, they still remain on the research agenda in many studies (Moorad and Hahn, 2008).

The following arguments support the view that genetic variants detected in this study have high chances of being truly associated with lifespan. First, all selected genetic variants show genome wide significant negative associations with lifespan. The fact that 24 out of 60 genetic variants selected for their genome wide significant genetic associations with female lifespan also show highly significant associations with male lifespan increases the chances that selected 24 variant are true positives. (Only 3 variants out of 60 would be expected as false-positive). The call rate for each detected variant exceeds 95% among all genotyped FHS study participants, and exceeds 90% among individuals from the Original FHS cohort.

To test how sensitive the results of GWAS are to the results of lifespan imputation we treated ages at censoring as ages at death, and performed GWAS of such data. These analyses resulted in 51 and 60 genome wide significant SNPs for males and females, respectively. Intersection of these sets resulted in 21 SNPs common for males and females. 15 out of these 21 SNPs also belong to the set of 24 SNPs detected earlier. These 15 SNPs are marked by "*" in Tables 2 and 3.

The empirical distribution of imputed lifespan is shown in Fig. 1. To make it more "normalized" we used Box-Cox transformation of this distribution. The GWAS of transformed data resulted in 43 SNPs negatively associated SNPs for females, and 33 such SNPs for males. The intersection of these sets with the set of 24 SNPs resulted in 12 SNPs. These SNPs are marked by "#" sign in Tables 2 and 3.

Strictly speaking the GWAS of lifespan data in the presence of censoring should follow multiple imputation procedure (Rubin, 2008). This procedure, however, is time consuming and may increase the number of false negative findings. Therefore we use an alternative method based on combining the follow up data on lifespan in the original FHS cohort with data on genetic frequencies from age distribution of the age at blood collection. The method is a particular case of more general procedure described in Arbeev et al. (2011). In these analyses the mortality rates for carriers and non-carriers of each of 24 preselected alleles were described by the Gompertz's curves. The parameters of these curves were estimated for each of 24 SNPs, and the null-hypotheses about similarity of mortality rates for carriers and non-carriers of each of these genetic variants were tested. The results of these analyses are shown in Tables 2 and 3. They indicate that difference in mortality patterns between carriers and non-carriers of corresponding genetic variants is highly statistically significant. The Kaplan-Meier estimates of four pairs of survival functions for carriers and non-carriers of four genetic variants are shown in Fig. A. Survival curves for other variants have similar patterns. One can see from this figure that carriers of each of 24 detected SNPs have lower values of survival.

The detected genetic variants are linked with genes whose expressions are crucial for maintaining organism's functioning (Table 4). Detected variants individually and jointly are associated with survival. The "genetic dose" index has a strong and significant effect on lifespan in the presence and in the absence of observed covariates. As any other genetic variant detected in the GWAS, our variants do not necessarily have negative effects on lifespan in any individual. Difference in personal genetic background or in the exposure to external conditions may influence the effect sign and size. These differences also explain why changes in the sample call rates result in different sets of selected genetic variants. The high level of sample call rate reduces the number of individuals eligible for GWAS. In the case of relatively small frequencies

of corresponding SNP alleles the estimated association becomes sensitive to the balance between positive and negative effects of the variant in the population of study participants.

Typical arguments of opponents are that the findings are likely to be false positives because they are not confirmed by other GWA studies. Our objection against this argument is that genetic associations with longevity traits depend to a great extent on the environment, which activates some genes and suppresses others. So the populations exposed to different environments are likely to show involvement of different genetic variants in lifespan (Yashin et al, 2012).

The chances that detected associations are random are likely to be small: all effects are genome wide significant, are all negative, and show associations with lifespan for males and females in separate analyses.

The negative contribution of genetic factors in lifespan at late ages are in line with predictions of the two evolutionary theories of aging MA and AP. Some findings including evolutionary conserved genes and genetic pathways involved in regulation of aging are difficult to explain using MA hypothesis of aging. Such conservation would be hard to make if mutations would happen randomly and kept in the genomes during evolutionary time in various species (Mitteldorf, 2012). However, living organisms might develop universal non-specific mechanisms coping with aging to guarantee reproduction in the presence of various unpredictable genetic disturbances. Then this machinery is used to cope for post reproductive survival to compensate consequences of bad mutations with late manifestation. More studies are needed to clarify this problem.

References

Bartke, A.; Coschigano, K.; Kopchick, J.; Chandrashekar, V.; Mattison, J.; Kinney, B.; and Hauck, S. "Genes That Prolong Life: Relationships of Growth Hormone and Growth to Aging and Life Span." *Journal of Gerontology Series A: Biological Sciences and Medical Sciences* 56, no. 8 (2001): B340–B349.

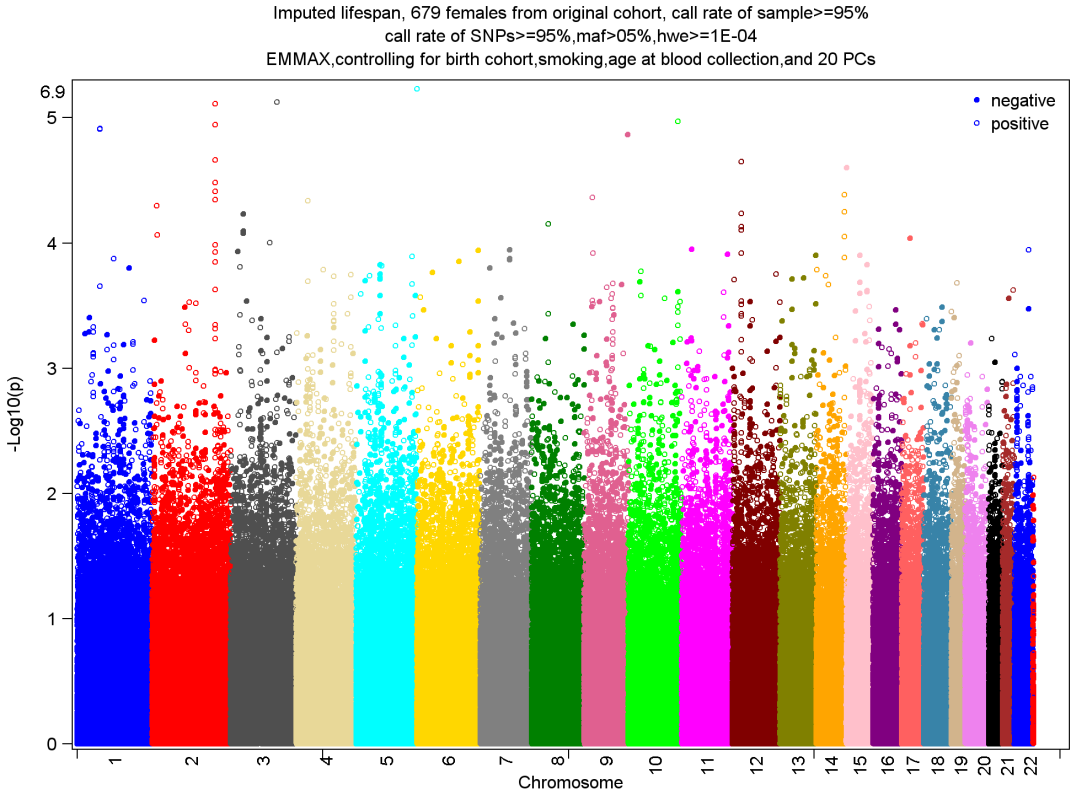
Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R. A C. elegans mutant that lives twice as long as wild type Nature 1993;366:461-4.

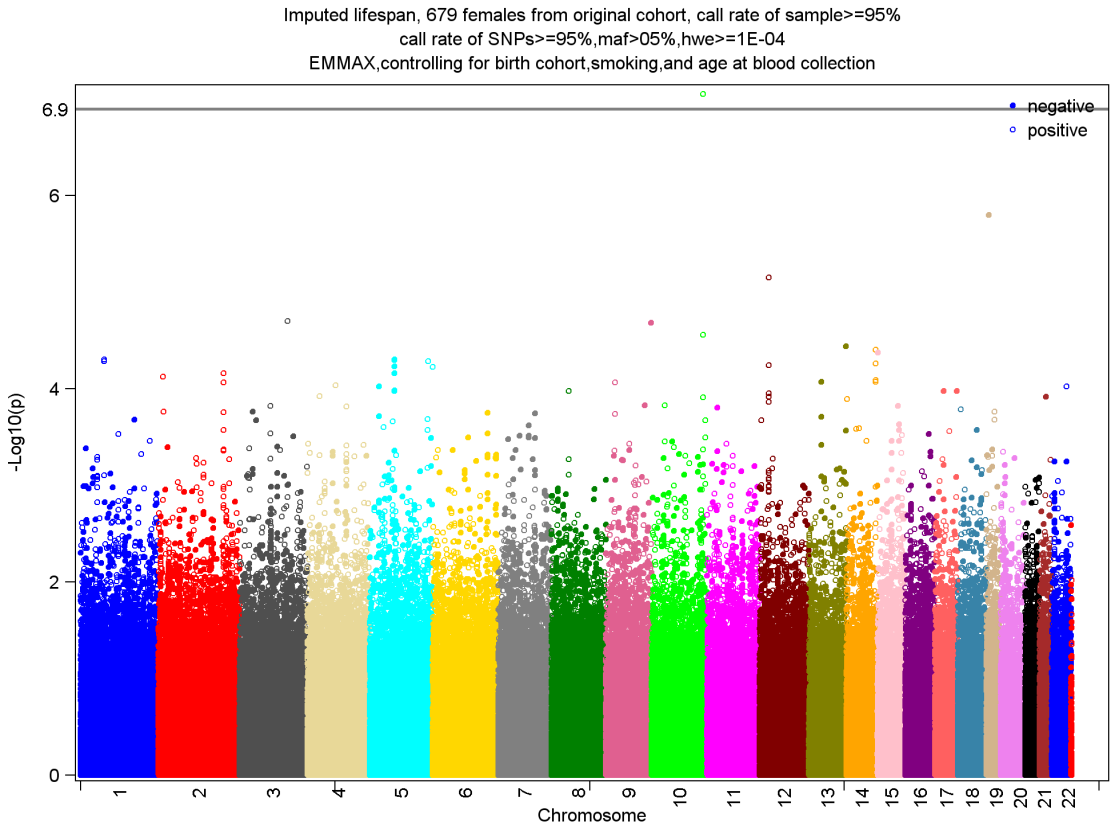
Martin GM. Modalities of gene action predicted by the classical evolutionary biological theory of aging. Ann N Y Acad Sci. 2007 Apr;1100:14-20. Review. PubMed PMID: 17460162.

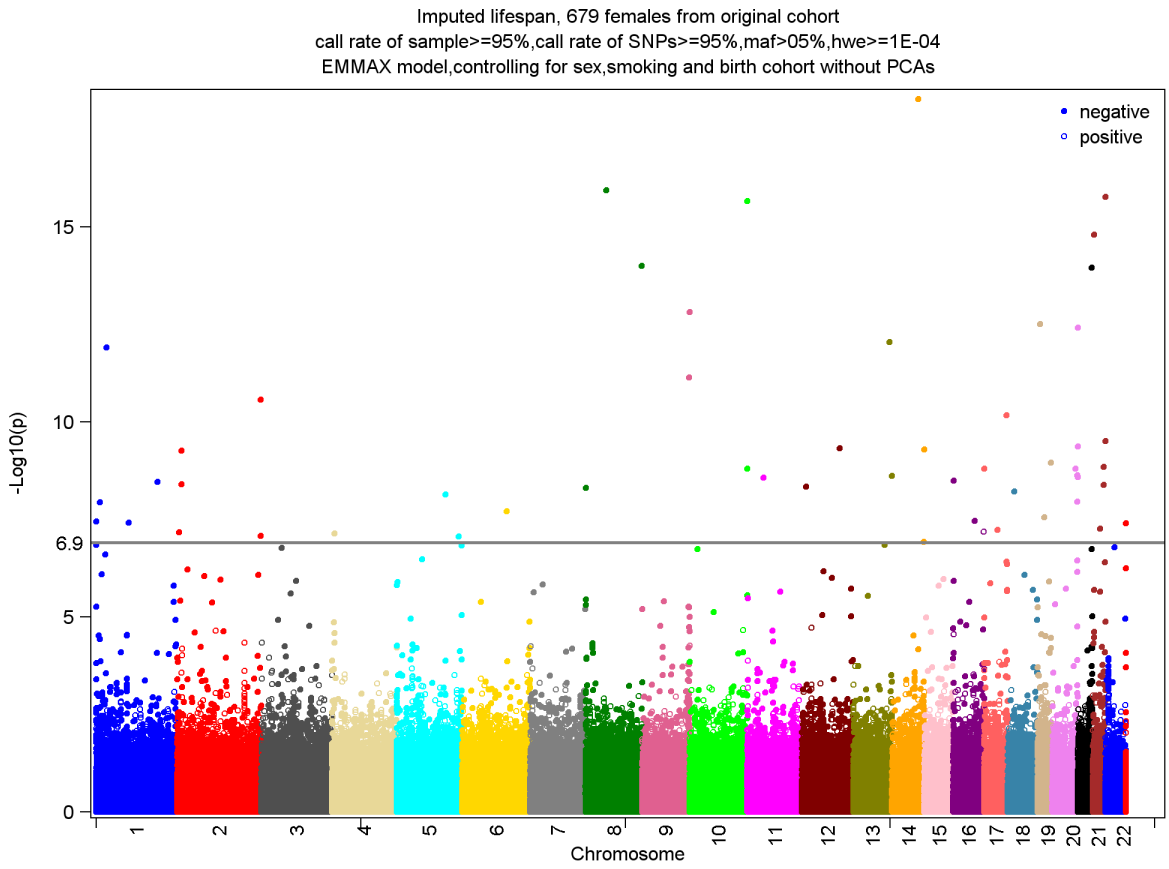
Martin GM. The biology of aging: 1985-2010 and beyond. FASEB J. 2011 Nov;25(11):3756-62. Review. PubMed PMID: 22046003.

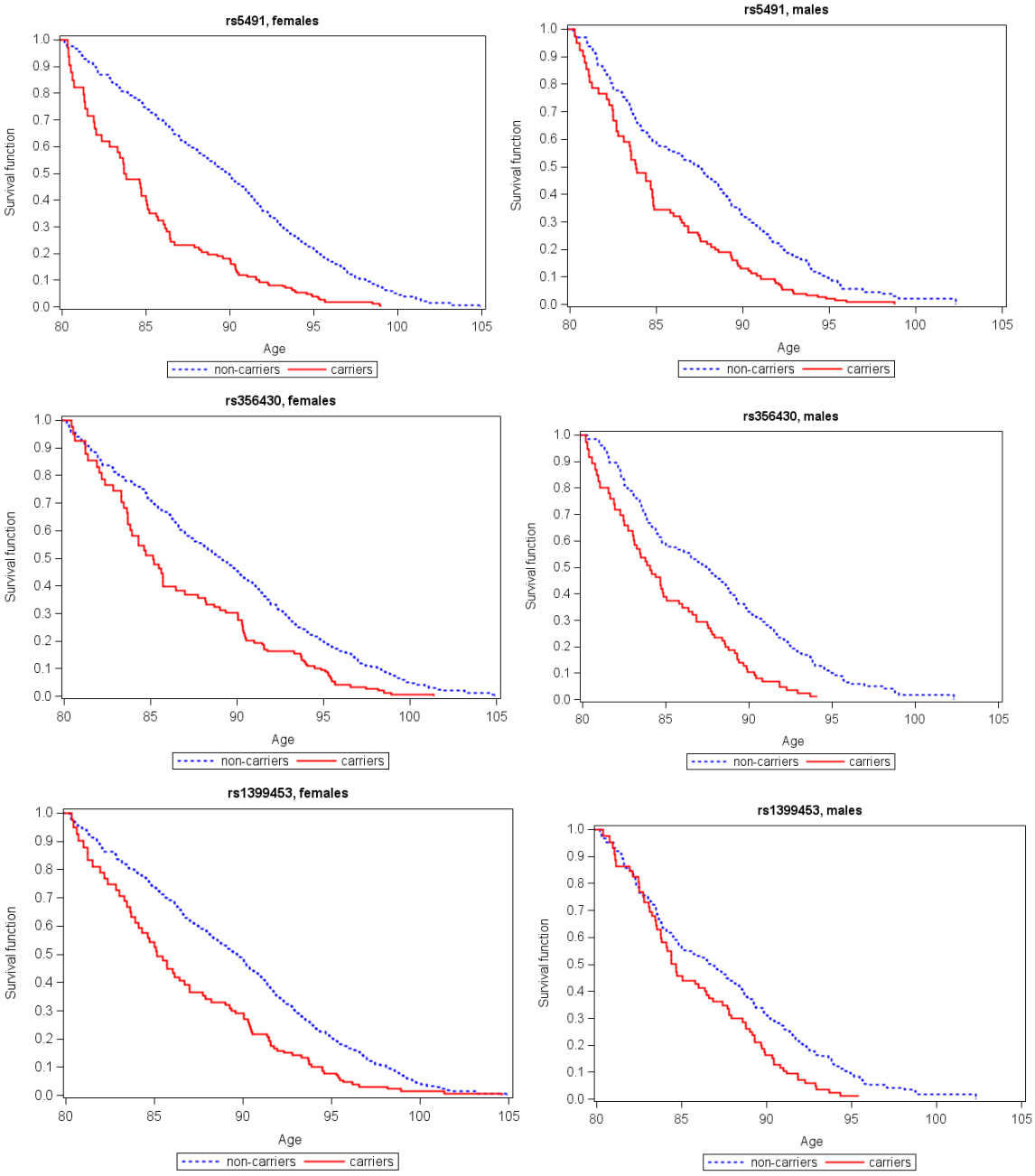
Zwaan BJ: The evolutionary genetics of ageing and longevity. Heredity 1999; 82: 589– 597.

Supplemental information

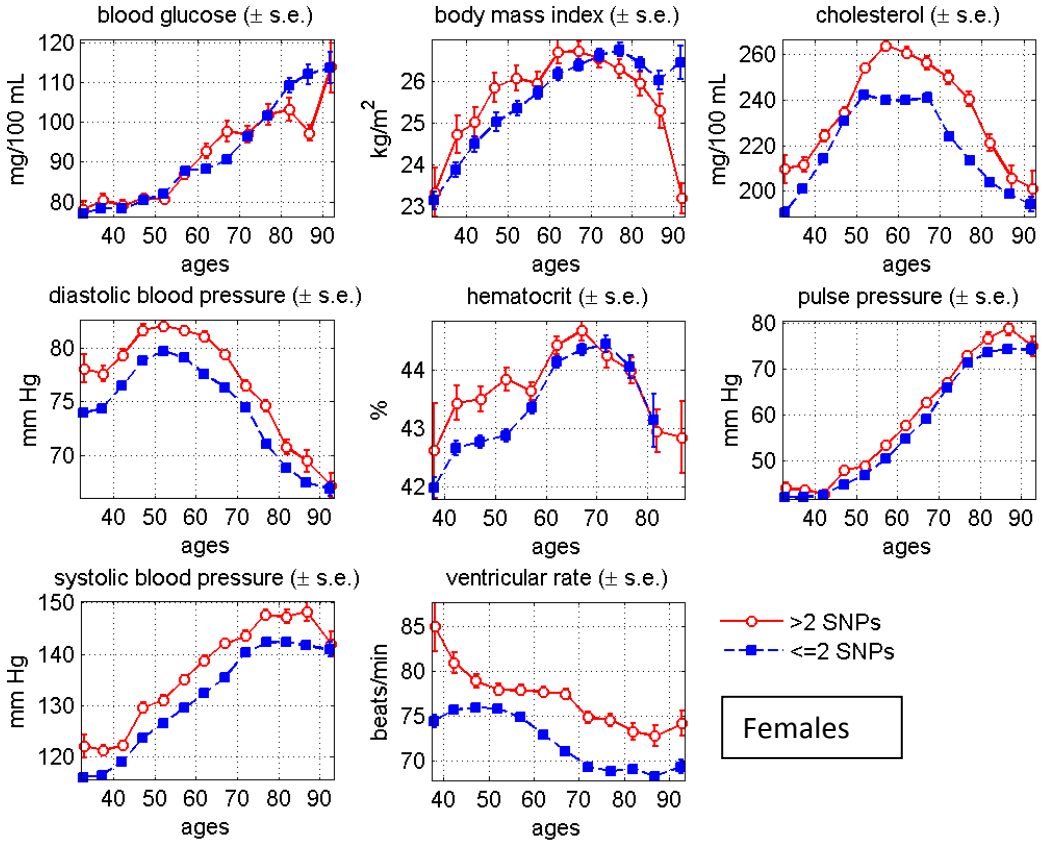

Imputed lifespan, 679 females from original cohort, call rate of sample>=95%
call rate of SNPs>=95%,maf>05%,hwe>=1E-04
EMMAX,controlling for birth cohort,smoking,age at blood collection,and 20 PCs
negative
positive
-Log10(p)
6.9
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

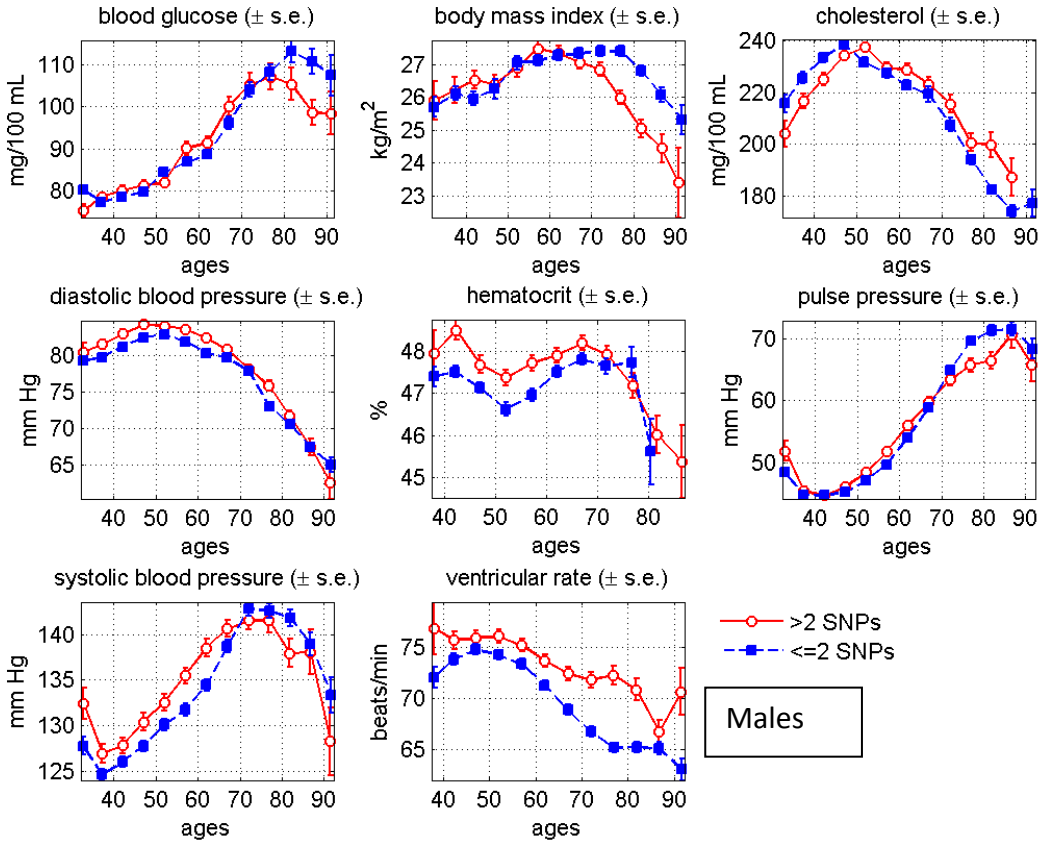

Imputed lifespan, 679 females from original cohort, call rate of sample>=95%
call rate of SNPs>=95%,maf>05%,hwe>=1E-04
EMMAX,controlling for birth cohort,smoking,and age at blood collection
negative
positive
6.9
6
4
2
0
-Log10(p)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


Imputed lifespan, 679 females from original cohort
call rate of sample>=95%,call rate of SNPs>=95%,maf>05%,hwe>=1E-04
EMMAX model,controlling for sex,smoking and birth cohort without PCAs
negative
positive
-Log10(p)
15
10
6.9
5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome



Survival functions after age 80 for male and female carriers and non-carriers of minor alleles for 3 selected SNPs. Survival functions for other alleles look similar.


blood glucose (± s.e.)
mg/100 mL
ages
body mass index (± s.e.)
kg/m2
ages
cholesterol (± s.e.)
mg/100 mL
ages
diastolic blood pressure (± s.e.)
mm Hg
ages
hematocrit (± s.e.)
%
ages
pulse pressure (± s.e.)
mm Hg
ages
systolic blood pressure (± s.e.)
mm Hg
ages
ventricular rate (± s.e.)
beats/min
ages
>2 SNPs
<=2 SNPs
Females


blood glucose (± s.e.)
mg/100 mL
ages
body mass index (± s.e.)
kg/m2
ages
cholesterol (± s.e.)
mg/100 mL
ages
diastolic blood pressure (± s.e.)
mm Hg
ages
hematocrit (± s.e.)
%
ages
pulse pressure (± s.e.)
mm Hg
ages
systolic blood pressure (± s.e.)
mm Hg
ages
ventricular rate (± s.e.)
beats/min
ages
>2 SNPs
<=2 SNPs
Males

Essential findings about 24 SNPs (the SNPs were selected for the association of the minor "frailty" allele with decreased survival at oldest ages (80+)):

1) To understand biological relevance of the 24 "frailty" SNPs to aging and common diseases (which are major determinants of lifespan) we annotated respective SNPs and reviewed current evidence about the functional effects of genes closest to these SNPs using NCBI/Entrez bases, GO, Ensembl, PANTHER, and other relevant online resources. We focused on biological functions of genes and regulatory regions, for which we detected SNPs located within such regions, or linked to nearby genes (see Table).

We found that from the list of 20 genes linked to the 24 SNPs, at least 3 genes (15%) were involved in ***Golgi*** *vesicular transport and membrane*: ARF1, CORO7, ARFGAP1.

The Golgi apparatus is a membranous structure in cell that processes and packs proteins made by the endoplasmic reticulum, before sending them out to their destination; in particular for secretion, using the secretory vesicle.

From the genes linked to our 24 SNPs and involved in the Golgi processing, two (ARFGAP1 and ARF1) turned out to be very closely interacting: The ARFGAP1 promotes hydrolysis of ARF1-bound GTP, which is required for the fusion of protein vesicles with Golgi compartments. The fact that our analysis identified SNPs in functionally closely connected genes, which are located on *different* chromosomes, tells in support of the real association between cellular processes regulating the protein traffic in the Golgi apparatus and survival at oldest old ages.

There is limited number of studies that suggest potential mechanisms of this intriguing connection. Cho et al. (2011) reported that the structure of the Golgi complex is significantly altered in senescent cells, and this can disturb normal protein secretion by such cells. The disturbed secretion may in turn partly explain often excessive and unbalanced release of pro-inflammatory factors by the senescent cells (Campisi et al. 2011), which in turn may potentially contribute to both physiological aging associated changes and pathology. The Golgi network was also linked to mTOR signaling; and aberrant Golgi trafficking was implicated in metastatic cancers (particularly in prostate cancer and melanoma) (e.g., Sánchez-Laorden et al. 2009; Abraham et al. 2009; Millarte and Farhan, 2012).

For additional information, we also tested our list of genes for over- or under-representation of Gene Ontology (GO) terms related to particular biological processes, using GeneTrail online software for Gene Set Enrichment/Overrepresentation analysis (Backes et al. 2007). We compared our list of genes (corresponding to 19 of 24 SNPs that were located in genes, or linked to genes) with the reference set of about 14.5K genes related to ~156K SNPs located within genes, and belonging to Affymetrix 500K SNP array (~437K after quality control). The GeneTrail detected overrepresentation of genes related to Golgi apparatus and membrane (4 observed vs. 0.6 expected), with p-value 0.03. There was also specific overrepresentation of genes involved in Golgi transport vesicle coating: 2 observed genes (ARFGAP1 and ARF1) vs. 0.013 expected, with p-value 0.007. The false positive detection rate method was used for multiple testing adjustment in this software. This analysis strongly supports results of our review

of gene functions emphasizing the role of Golgi apparatus in normal cell functioning and organism's chances to survive the oldest old age.

2) SNP rs4904670 is located in intron of NRDE2 (a.k.a. C14orf102). **N.B.** One of the earlier found 39 pro-survival SNPs, rs2282032 (Yashin et al. 2010), is located in another intron of the same NRDE2 gene. NRDE2 codes *"NRDE-2, necessary for RNA interference, domain containing protein"* and is poorly studied so far. Some SNPs of this gene, including the rs2282032 are, however, in LD with neighbor gene - PSMC1 - proteasome 26S subunit, ATPase, 1.

3) The rs1794108 is located in exon of PSMD13 gene involved in protein degradation by the proteasome. PSMD13 is in strong LD with neighbor SIRT3 gene, of the Silent information regulator 2 (Sir2) family of histone deacetylases (sirtuin HDACs). Sirtuins (SIRT1-7) are mammalian homologues of the Sir2 gene in yeast and play a central role in epigenetic gene silencing, DNA repair and recombination, cell-cycle, microtubule organization, and in the regulation of aging (Mahlknecht et al. 2011). SIRT3 also has a role as a tumor promoter or tumor suppressor, depending on context (Alhazzazi et al. 2011). There are significantly different PSMD13-SIRT3 haplotype pools between centenarians and younger people (Bellizzi et al. 2007). Close relation of the identified SNP to the candidate aging/longevity genes supports its true association with

4) For most evaluated SNPs, the minor allele displays a trade-off effect on female (but not male) survival. That is, it is associated with worsened survival at oldest old ages (around 85), but with better survival before.

Figure below shows that absence of the minor "frailty" alleles in persons' genomes significantly improves survival beyond the old age (80+). Paradoxically, the same genetic background may also reduce chances of the non-carriers to reach that old age in first place.

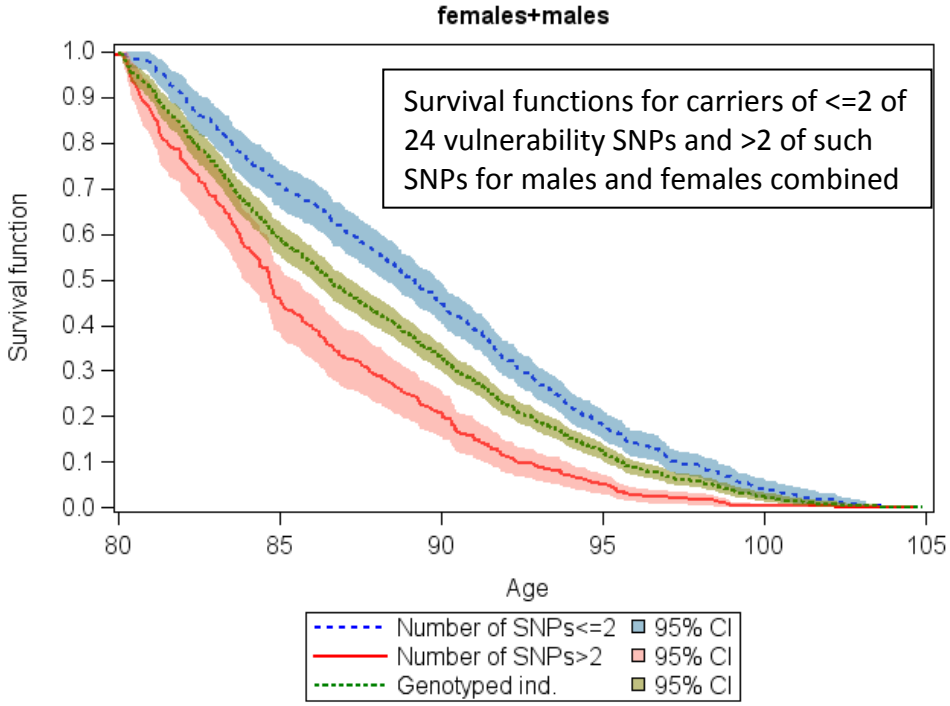


Table4. Essential characteristics of the 24 SNPs associated with lower survival at oldest ages (80+) and relevant genes

SNP name	Chr	MAF	In/Out gene	Closest gene	Gene/protein function
rs3738682	1q42		intronic	**ARF-1** ADP-ribosylation factor 1	a small GTP-binding protein having a central role in intra-Golgi vesicular protein transport. Modulates vesicle budding and uncoating. N.B. The hydrolysis of ARF1-bound GTP is mediated by **ARFGAP**s. Also: member of the RAS superfamily; role in controlling cell proliferation (Boulay et al. 2011); the downstream target of PI3K (Nishida et al. 2011).
rs356430	5q31.2		probably LincRNA	**CTB-35F21.1** (LincRNA)	CTB-35FF21.1 region may contain LincRNA. LincRNA work in complexes with proteins and perform regulatory functions such as inhibiting transcription and translation (Yoon et al. 2012)
rs17067605	5q34		intergen		
rs2353447	8q11.1		intergen	RP11-783P22.2	
rs4565533	9q34.3		intergen	near RXRA	N.B. no LD (r2>0.5) found between rs4565533 and SNPs in RXRA (retinoid X receptor, alpha).
rs7894051	10q26.2		intronic	ECHS1 enoyl CoA hydratase, short chain, 1, mitochondrial	ECHS1 catalyzes the hydration of 2-trans-enoyl-coenzyme A (CoA).
rs1440483	11q25		intronic	B3GAT1 beta-1,3-glucuronyltransferase 1	B3GAT1 is the key enzyme in a glucuronyl transfer reaction during the biosynthesis of the carbohydrate epitope HNK-1 (human natural killer-1, also known as CD57 and LEU7).
rs1794108	11p15.5		exon - nonsyn coding	PSMD13 proteasome (prosome, macropain) 26S subunit, non-ATPase, 13	PSMD13 acts as a regulatory subunit of the 26S proteasome involved in the ATP-dependent degradation of ubiquitinated proteins. There is high LD between PSMD13-SIRT3. Sirtuins (SIRT1-7) play a central role in epigenetic gene silencing, DNA repair, cell-cycle, microtubule organization, and in aging (Mahlknecht et al. 2011).
rs5743998	11p15.5		intronic	TOLLIP toll interacting protein	TOLLIP regulates inflammatory signaling and is involved in interleukin-1 receptor trafficking and in the turnover of IL1R-associated kinase. Inhibits cell activation by microbial products. Inhibits IRAK1 phosphorylation and kinase activity
rs9971555	11p13		intronic	ABTB2 ankyrin repeat and BTB (POZ) domain containing 2	
rs1399453	12q23.1		intronic	ANO4 anoctamin 4	member of a family of Ca2+-activated Cl- channels (Tian et al. 2012)
rs10845099	12p13		intergen (linked to KLRD1 gene)	linked to KLRD1 - KILLER CELL LECTIN-LIKE RECEPTOR subfamily D, member 1	Located between GABARAPL1 (GABA(A) receptor-associated protein like 1) and KLRD1. KLRD1 plays role as a receptor for the recognition of MHC class I HLA-E molecules by NK cells and some cytotoxic T-cells.

rs4904670	14q32.11		intronic	NRDE2 (a.k.a. C14orf102);	NRDE2 codes NRDE-2, necessary for RNA interference, domain containing. It is linked no next gene - PSMC1 - proteasome (prosome, macropain) 26S subunit, ATPase, 1.
rs9925881	16p12.2		intergen	between TRNAL7 and EEF2K	N.B. No proxy snps (by LD) were found by SNAP near (+- 500K distance) of this SNP. Closest is EEF2K - a highly conserved protein kinase in the calmodulin-mediated signaling pathway that links activation of cell surface receptors to cell division.
rs9928967	16p13.3		Exon-nonsyn coding	CORO7 - coronin 7; a.k.a. CORO7-PAM16 readthrough	CORO7 plays a role in Golgi complex morphology and function. PAM16 is suspected to be involved in increased rates of anaerobic metabolism, resistance to apoptosis and altered growth-factor sensitivity.
rs5491	19p13.3		Exon-nonsyn coding	ICAM1 - intercellular adhesion molecule 1; a.k.a. CD54	Acell surface glycoprotein which is typically expressed on endothelial cells and cells of the immune system. Mediates intracellular adhesion.
rs2586484	17q21.33		intergen (linked to several genes)	near COL1A1; in LD with FAM117A, CACNA1G	The SNP is in perfect LD with 7 surrounding SNPs located in or near a synteny block of genes with evolutionary conserved order, especially in FAM117A and CACNA1G.
rs8081943	17p11.2		intronic	RAI1 retinoic acid induced 1	located within the Smith-Magenis syndrome region. May function as a transcriptional regulator. Regulates transcription through chromatin remodeling.
rs9896996	17p13.3		intergen (linked to SMG6)	located between MIR212 and MIR132. In LD with SMG6 (smg-6 homolog, nonsense mediated mRNA decay factor)	In LD with rs62067977 (r2= 0.62; D'= 0.79), which is intronic SNP of SMG6 gene. Its protein is part of the telomerase ribonucleoprotein complex and binds single-stranded DNA at the telomeres. SMG6 also participates in mRNA decay.
rs11536959	20q11.23		intronic	LBP - lipopolysaccharide binding protein	Binds to the lipid A moiety of bacterial lipopolysaccharides (LPS), a glycolipid present in the outer membrane of all Gram-negative bacteria. The LBP/LPS complex seems to interact with the CD14 receptor.
rs6090342	20q13.33		intronic	ARFGAP1 - ADP-ribosylation factor GTPase activating protein 1	Promotes GTP hydrolysis on the small G protein Arf-1 on Golgi membranes. Involved in membrane trafficking and /or vesicle transport.
rs2838566	21q22.3		intergen	between LRRC3 - and TSPEAR	N.B. No proxy snps (in LD with rs2838566) were found within +- 500K distance.
rs6007952	22q13.3		intronic	GRAMD4 - GRAM domain containing 4	Mitochondrial effector of E2F1 (MIM 189971)-induced apoptosis. Plays a role as a mediator of E2F1-induced apoptosis in the absence of TP53/p53.

rs8135777	22q13.3		intronic	SHANK3 - SH3 and multiple ankyrin repeat domains 3	Shank3 is a large scaffold postsynaptic density protein implicated in dendritic spinematuration and synapse formation, regulates the structural organization of neurotransmitter receptors in post-synaptic dendritic spines making it a key element in chemical binding crucial to nerve cell communication.

References

NCBI Entrez cross-database search. http://www.ncbi.nlm.nih.gov/sites/gquery

AceView: a comprehensive cDNA-supported gene and transcripts annotation, Genome Biology 2006, 7(Suppl 1):S12

Abraham RT. GOLPH3 links the Golgi network to mTOR signaling and human cancer. Pigment Cell Melanoma Res. 2009 Aug;22(4):378-9. doi: 10.1111/j.1755-148X.2009.00596.x. PubMed PMID: 19624311.

Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Müller R, Meese E, Lenhof HP. GeneTrail--advanced gene set enrichment analysis. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W186-92. Epub 2007 May 25. PubMed PMID: 17526521; PubMed Central PMCID: PMC1933132.

Campisi J, Andersen JK, Kapahi P, Melov S. Cellular senescence: a link between cancer and age-related degenerative disease? Semin Cancer Biol. 2011 Dec;21(6):354-9. doi: 10.1016/j.semcancer.2011.09.001. Epub 2011 Sep 10. Review. PubMed PMID: 21925603; PubMed Central PMCID: PMC3230665.

Kähler AK, Djurovic S, Rimol LM, Brown AA, Athanasiu L, Jönsson EG, Hansen T, Gústafsson O, Hall H, Giegling I, Muglia P, Cichon S, Rietschel M, Pietiläinen OP, Peltonen L, Bramon E, Collier D, St Clair D, Sigurdsson E, Petursson H, Rujescu D, Melle I, Werge T, Steen VM, Dale AM, Matthews RT, Agartz I, Andreassen OA. Candidate gene analysis of the human natural killer-1 carbohydrate pathway and perineuronal nets in schizophrenia: B3GAT2 is associated with disease risk and cortical surface area. Biol Psychiatry. 2011 Jan 1;69(1):90-6. doi: 10.1016/j.biopsych.2010.07.035. Epub 2010 Oct 15. PubMed PMID: 20950796.

Lake SL, Jmor F, Dopierala J, Taktak AF, Coupland SE, Damato BE. Multiplex ligation-dependent probe amplification of conjunctival melanoma reveals common BRAF V600E gene mutation and gene copy number changes. Invest Ophthalmol Vis Sci. 2011 Jul 29;52(8):5598-604. doi: 10.1167/iovs.10-6934. PubMed PMID: 21693616

Liu X, Feng R, Du L. The role of enoyl-CoA hydratase short chain 1 and peroxiredoxin 3 in PP2-induced apoptosis in human breast cancer MCF-7 cells. FEBS Lett. 2010 Jul 16;584(14):3185-92. doi: 10.1016/j.febslet.2010.06.002. Epub 2010

Jun 10. PubMed PMID: 20541551.

Millarte V, Farhan H. The Golgi in cell migration: regulation by signal transduction and its implications for cancer cell metastasis. ScientificWorldJournal. 2012;2012:498278. doi: 10.1100/2012/498278. Epub 2012 May 1. Review. PubMed PMID: 22623902; PubMed Central PMCID: PMC3353474.

Sánchez-Laorden BL, Herraiz C, Valencia JC, Hearing VJ, Jiménez-Cervantes C, García-Borrón JC. Aberrant trafficking of human melanocortin 1 receptor variants associated with red hair and skin cancer: Steady-state retention of mutant forms in the proximal golgi. J Cell Physiol. 2009 Sep;220(3):640-54. doi: 10.1002/jcp.21804. PubMed PMID: 19452503; PubMed Central PMCID: PMC2705480.

Cribbs DH, Berchtold NC, Perreau V, Coleman PD, Rogers J, Tenner AJ, Cotman CW. Extensive innate immune gene activation accompanies brain aging, increasing vulnerability to cognitive decline and neurodegeneration: a microarray study. J Neuroinflammation. 2012 Jul 23;9:179. doi: 10.1186/1742-2094-9-179. PubMed PMID: 22824372; PubMed Central PMCID: PMC3419089.

Pimentel-Nunes P, Gonçalves N, Boal-Carvalho I, Afonso L, Lopes P, Roncon-Albuquerque R Jr, Henrique R, Moreira-Dias L, Leite-Moreira AF, Dinis-Ribeiro M. Helicobacter pylori induces increased expression of Toll-like receptors and decreased Toll-interacting protein in gastric mucosa that persists throughout gastric carcinogenesis. Helicobacter. 2013 Feb;18(1):22-32. doi: 10.1111/hel.12008. Epub 2012 Sep 3. PubMed PMID: 23061653.

Tian Y, Schreiber R, Kunzelmann K. Anoctamins are a family of Ca2+-activated Cl- channels. J Cell Sci. 2012 Nov 1;125(Pt 21):4991-8. doi: 10.1242/jcs.109553. Epub 2012 Sep 3. PubMed PMID: 22946059.

Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, Huarte M, Zhan M, Becker KG, Gorospe M. LincRNA-p21 suppresses target mRNA translation. Mol Cell. 2012 Aug 24;47(4):648-55. doi: 10.1016/j.molcel.2012.06.027. Epub 2012 Jul 26. PubMed PMID: 22841487; PubMed Central PMCID: PMC3509343.