

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani,^{1*} Nadia Solovieff,¹ Annibale Puca,² Stephen W. Hartley,¹ Efthymia Melista,³ Stacy Andersen,⁴ Daniel A. Dworkis,³ Jemma B. Wilk,⁵ Richard H. Myers,⁵ Martin H. Steinberg,⁶ Monty Montano,³ Clinton T. Baldwin,^{6,7} Thomas T. Perls^{4*}

¹Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA. ²IRCCS Multimedica, Milano, Italy; Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate, 20122, Italy. ³Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA. ⁴Section of Geriatrics, Department of Medicine, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. ⁵Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA. ⁶Departments of Medicine and Pediatrics, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. ⁷Center for Human Genetics, Boston University School of Medicine, Boston, MA 02118, USA.

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.H.P.)

Healthy aging is thought to reflect the combined influence of environmental factors (lifestyle choices) and genetic factors. To explore the genetic contribution, we undertook a genome-wide association study of exceptional longevity (EL) in 1055 centenarians and 1267 controls. Using these data, we built a genetic model that includes 150 single nucleotide polymorphisms (SNPs) and found that it could predict EL with 77% accuracy in an independent set of centenarians and controls. Further in-silico analysis revealed that 90% of centenarians can be grouped into 19 clusters characterized by different combinations of SNP genotypes—or genetic signatures—of varying predictive value. The different signatures, which attest to the genetic complexity of EL, correlated with differences in the prevalence and age of onset of age-associated diseases (e.g., dementia, hypertension, and cardiovascular disease) and may help dissect this complex phenotype into subphenotypes of healthy aging.

The average human lifespan in developed countries now ranges from 80 to 85 years. Environmental factors (lifestyle choices relating to diet, exercise, smoking habits, etc.) as well as genetic factors are believed to contribute to healthy aging. The results of human twin studies suggest that only 20-30% of the variation in survival to an age of about 85 years is determined by genetics (1). Supporting the importance of environmental factors in survival to old age is the 88-year average life expectancy of Seventh-day Adventists (2), who by virtue of their religion have health-related behaviors conducive to healthy aging. Nonetheless, other data—including the observation that exceptional longevity (EL) runs strongly in families—argue that genetic factors play an important contributory role in healthy aging and especially to living 10-30 years beyond the mid-eighties (3).

Based upon the hypothesis that exceptionally old individuals are carriers of multiple genetic variants that influence human lifespan (4), we conducted a genome-wide association study (GWAS) of centenarians. Centenarians are a model of healthy aging, as the onset of disability in these individuals is generally delayed until they are well into their mid-nineties (5, 6). We studied 801 unrelated subjects enrolled in the New England Centenarian Study (NECS) and 926 genetically matched controls. NECS subjects were Caucasians who were born between 1890 and 1910 and had an age range of 95 to 119 years (median age 103 years). Figure S1 in the Supporting Online Material (7) describes the age distribution. Approximately one-third of the NECS sample included centenarians with a first-degree relative also achieving EL, thus enhancing the sample's power (8). Controls included 243 NECS referent subjects who were spouses of centenarian offspring or children of parents who died at the mean age of 73 years, and genome-wide SNP data of 683 subjects selected from the Illumina control database. We selected Illumina controls to match the genetic backgrounds of NECS subjects using an algorithm described in (7), fig. S2. For replication, we used 254 North American Caucasian subjects enrolled by Elixir Pharmaceuticals between 2001-2003. These individuals were born between 1890 and 1910 (age range of 90-114 years) and were unrelated to NECS subjects. Referent subjects ($n = 341$) were identified from the remaining Illumina controls. We analyzed ~295,000 SNPs in the NECS discovery and Elixir replication sets that passed stringent quality control rules using several analytic strategies (Fig. 1).

We conducted standard Bayesian and frequentist single SNP analyses (9) where we identified genome-wide significant SNPs in the discovery set and tested the associations in the replication set [see (7), sect. 5 to 13, and

figs. S3 to S5]. We identified 70 genome-wide significant SNPs in the discovery set and replicated 33 (table S1 and table S6). All associations increased in significance when we analyzed the aggregated discovery and replication data. To address bias due to choice of controls, we repeated the analysis using 867 referent subjects included in a GWAS of Parkinson's disease (PD) (10). All associations were replicated with consistent effects (table S2). We also compared the allele frequencies of the three sources of controls (NECS, Illumina, and PD) and did not find differences that suggested laboratory-specific biases (table S2). This list includes only 5 SNPs associated with common diseases and their risk alleles are significantly less frequent in centenarians than controls (table S3 and fig. S6) (11–16).

The replicated GWAS suggests that many genetic variants contribute to EL, but it does not provide a measure of the combined effects of these variants. We therefore built a genetic risk model to evaluate, *in silico*, the effect of combinations of SNP alleles to predict EL and to explore the hypothesis that subsets of associated SNPs characterize different pathways to EL (4, 6).

We computed the genetic risk associated with a set of SNPs using a Bayesian classification model (17) and designed a search procedure to discover the SNPs to be included [summarized in fig. S7 and (7) sect. 14 and 15]. The procedure builds a series of nested genetic risk models starting with the most significant SNP in the discovery set and incrementally adding one SNP at a time. Each model is used for prediction, and the accuracy of each model to predict EL and average longevity (AL) is evaluated by comparing predicted and observed outcomes (fig. S8). To reduce overfitting, we also used a resampling approach in which we repeatedly split the discovery set into non-overlapping training and test sets that were used, respectively, to estimate the nested genetic risk models and to evaluate their predictive value (figs. S8 and S9). The analysis suggests that up to 150 SNPs should be sufficient to provide accurate risk prediction. table S7 provides complete details of the 150 SNPs, and the probabilities that are used to compute the prediction using the formula in fig. S7. These 150 SNPs are uncorrelated and occur at an average distance of 8Mb; 77 are in known genes with a wide range of functions.

To investigate whether all 150 SNPs are necessary for prediction, we generated a genetic risk profile for each subject by plotting the risk of EL ($p(\text{EL}|\Sigma_k)$, y axis) against the number of SNPs in each of the 150 SNP sets Σ_k (x-axis) and examined their patterns [(7), sect. 18]. Figure 2A shows the profiles from 3 centenarians and a control. In the 106 year old, the first three SNP sets $\Sigma_1 = [\text{rs1036819}]$, $\Sigma_2 = [\Sigma_1, \text{rs9576827}]$, and $\Sigma_3 = [\Sigma_2, \text{rs2075650}]$ predict 48%, 46% and 35% chances of EL. This subject carries genotypes AC, AA, and AG for the 3 SNPs respectively, and because these

genotypes are more common in controls than centenarians, they determine a chance of EL that is lower than the chance of AL. The fourth SNP set, $\Sigma_4 = [\Sigma_3, \text{rs1455311}]$, predicts an almost 65% chance of EL. The subject carries the GG genotype for the SNP rs1455311 that is rare in the general population but much more frequent in centenarians (table S1), and the inclusion of this rare genotype almost doubles the chances of EL. The probability predicted by the next SNP sets with up to 64 SNPs and with more than 133 SNPs ranges between 0.5 and 0.96, while almost none of the SNP sets with 65 to 133 SNPs predicts more than 50% chance of EL. This genetic profile shows that the subject carries some combinations of SNP alleles that are predictive of EL, while other alleles are predictive of AL. However, the overall genetic risk profile determined by all 150 SNP sets makes a strong case for EL. The genetic risk profile of the centenarian who died at age 119 years is dramatically different. With the exception of the first SNP, all SNP sets determine more than 89% chance of EL, and the entire trend of genetic risk predicted by the 150 models provides strong evidence for EL. The profile of the third subject, age 107 years, shows that the first 25 SNPs sets are insufficient to predict the correct outcome, and only the overall trend of genetic risk provides evidence for EL. The fourth plot displays the profile of a control, and shows that this subject carries some longevity associated variants (LAVs); however, the overall trend of genetic risk points to AL rather than EL.

These examples support the hypothesis that EL is determined by varying combinations of LAVs. Consistent with this, an ensemble of the 150 genetic risk models provides 87% specificity and 84% sensitivity in the discovery set (fig. S10), and 77% sensitivity and specificity in the replication set (Fig. 2B). Genetic risk scores based on the number of LAVs or logistic regression with individual genotypes did not reach the same level of accuracy [(7), sect. 17]. Figure 2B also shows that ordering the SNPs by Bayesian significance determines the highest accuracy, and changing the order of the nested SNP sets reduces specificity. SNPs randomly chosen from the 2000 most significant SNPs have almost no predictive value (fig. S9). The specificity of the ensemble of 150 genetic risk models was replicated (83%) in a larger set of 3877 population controls from the Illumina database (fig. S14).

Some genetic risk profiles were recurrent and we speculated that groups of centenarians may have distinct genetic signatures that relate to different sub-types of EL characterized by varying prevalence or age of onset of age-related diseases. To verify this hypothesis, we used a Bayesian model-based clustering procedure (18) to group the genetic risk profiles. We then investigated whether groups of centenarians with particular genetic risk profiles shared specific age-related sub-phenotypes.

Cluster analysis identified 19 groups of 8 or more centenarians with similar genetic risk profiles in the discovery set [(7), sect. 19]. Figure 3 shows the 9 largest clusters while the remaining clusters are shown in fig. S11. No clusters showed enrichment for any European ethnicity [(7), sect. 22]. The prototypical genetic risk profiles associated with each cluster are informative displays of the LAVs, and represent different genetic signatures of EL. These signatures provide a visual representation of the joint effects of LAVs. While the ensemble average provides a global estimate of the probability of EL, the pattern itself provides information about the different sets of LAVs that drive a subject toward this probability. The same cluster analysis identified 11 groups with 6 or more centenarians in the replication set (fig. S12). The signatures of ten of these groups match signatures in the discovery set by trend and predictive accuracy (Fig. 3) and replicate the findings. Furthermore, the signatures are highly specific for EL as shown by the same analysis of matched controls (fig. S13). For example, only 0.6% of controls in the discovery set had a genetic signature most predictive of EL (cluster C1, Fig. 3), and only 8% had genetic signatures that were 69%-98% predictive of EL (clusters C5-C13, fig. S13). These results were replicated in the analysis of all 3,877 Illumina controls (fig. S15). Our finding that about 15% of Illumina controls have signatures with >50% chance of EL is consistent with the suggestion that many more people than previously suspected have the potential, at least genetically, to survive to an exceptional age (19).

We next investigated whether age at death changes between clusters with different genetic signatures. We found that the age distributions in the 19 clusters segregate into two clear groups: clusters C1 through C4 versus the others (Fig. 4A for the 9 largest clusters, fig. S16 for all 19 clusters). More than 75% of the ages of subjects in clusters C1 - C4 were ≥ 106 years old and these four clusters included 46% of the supercentenarians (age ≥ 110 years). The enrichment of LAVs in C1 and C4 (Fig. 3) is consistent with the hypothesis that genes are an important determinant of EL (20). The age distribution of cases in the 9 clusters of the replication set reproduces a pattern that was seen in the discovery set (fig. S16).

To explore the relationship between the different genetic signatures of EL and various age-related diseases, we examined the prevalence of these diseases and their ages of onset in clusters with 30 or more centenarians (Fig. 4, B and C) [(7), sect. 21 and table S4]. Subjects in cluster C1 had a significant delay in the onset of cardiovascular disease, dementia, and hypertension (Fig. 4C). Furthermore, enrichment of LAVs was correlated with a lower prevalence of cardiovascular disease and diabetes (Fig. 4B). Centenarians in clusters C6, C9 and C13 have similar ages at death (median age ~ 103 years) but varying distributions of

ages of onset of dementia and hypertension. Ages of onset of other diseases also differ between other clusters (fig. S17). Interestingly, cluster C19 is composed of 30 centenarians lacking almost all of the LAVs. Although the EL of these subjects may be the result of good health behaviors, or simply just chance, an alternative explanation is that these subjects carry rare genetic variants that were not represented in the SNP array. Examination of 17 centenarians in this cluster for whom we had family data revealed that 59% ($n = 10$) exhibit strong familial longevity (see table S5 and fig. S18). These results suggest that there may be many more genetic modifiers of EL to be discovered and that whole genome sequencing of these subjects may be particularly fruitful.

These genetic signatures confirm that EL is influenced by the combined effects of a large number of SNPs. We also found that a large proportion of the supercentenarians had the greatest enrichment of LAVs, indicating a strong relationship between the number of LAVs and survival to the most extreme ages. While large numbers of LAVs appear to be necessary for extreme survival, we did not observe a substantial difference in the numbers of a large sample of known disease-associated variants carried by centenarians and controls (fig. S6). These preliminary data suggest that EL may be the result of an enrichment of LAVs that counter the effect of disease-risk alleles and contribute to the compression of morbidity and/or disability towards the end of very long lives (6). Other signatures correlated with the prevalence and age of onset of age-related diseases and further investigation is needed to understand how and why they predispose for EL and for specific, different patterns of healthy aging.

The genetic signatures were built by using an ensemble of genetic risk models. The 77% accuracy of these predictions in an independently recruited sample of centenarians shows that genetic data can indeed predict EL without knowledge of any other risk factor. This prediction is not perfect, however, and although it may improve with better knowledge of the variations in the human genome, its limitations confirm that environmental factors (e.g., lifestyle) also contribute in important ways to the ability of humans to survive to very old ages.

References and Notes

1. A. M. Herskind *et al.*, *Hum Genet* **97**, 319 (Mar, 1996).
2. G. E. Fraser, D. J. Shavlik, *Arch Intern Med* **161**, 1645 (Jul 9, 2001).
3. T. T. Perls *et al.*, *Proc Natl Acad Sci U S A* **99**, 8442 (Jun 11, 2002).
4. S. Hekimi, *Nat Genet* **38**, 985 (Sep, 2006).
5. K. Christensen, M. McGue, I. Petersen, B. Jeune, J. W. Vaupel, *Proc Natl Acad Sci U S A* **105**, 13274 (Sep 9, 2008).

6. D. F. Terry, P. Sebastiani, S. L. Andersen, T. T. Perls, *Arch Intern Med* **168**, 277 (Feb 11, 2008).
7. Material and methods are available as supporting material on Science Online.
8. Q. Tan, J. H. Zhao, D. Zhang, T. A. Kruse, K. Christensen, *Am J Epidemiol* **168**, 890 (Oct 15, 2008).
9. P. Sebastiani *et al.*, *Am J Hematol* **85**, 29 (Jan, 2010).
10. N. Pankratz *et al.*, *Hum Genet* **124**, 593 (Jan, 2009).
11. D. Harold *et al.*, *Nat Genet* **41**, 1088 (Oct, 2009).
12. J. C. Lambert *et al.*, *Nat Genet* **41**, 1094 (Oct, 2009).
13. X. Li *et al.*, *J Allergy Clin Immunol* **125**, 328 (Feb, 2010).
14. A. C. Need *et al.*, *Hum Mol Genet* **18**, 4650 (Dec 1, 2009).
15. F. Marroni *et al.*, *Circ Cardiovasc Genet* **2**, 322 (Aug, 2009).
16. J. Dupuis *et al.*, *Nat Genet* **42**, 105 (Feb).
17. D. J. Hand, in *The top ten algorithms in data mining* X. Wu, V. Kumar, Eds. (Chapman and Hall, London, 2009) pp. 163-178.
18. M. F. Ramoni, P. Sebastiani, I. S. Kohane, *Proc Natl Acad Sci U S A* **99**, 9121 (Jul 9, 2002).
19. K. Christensen, G. Doblhammer, R. Rau, J. W. Vaupel, *Lancet* **374**, 1196 (Oct 3, 2009).
20. T. Perls, L. M. Kunkel, A. A. Puca, *J Am Geriatr Soc* **50**, 359 (Feb, 2002).
21. We thank the subjects and family members participating in the New England Centenarian Study and the Elixir Pharmaceuticals Centenarian Study. This work was supported by National Institutes of Health grants: R01 HL087681 (to M.H.S), K24 AG025727 (to T.T.P), R01 AR055115 (to M.M.), R01 AG027216 (to C.B.), R01 NS36711-09 (to R.H.M). T.T.P. thanks the following for previous support that facilitated subject enrollment in the New England Centenarian Study: Ellison Medical Foundation, the American Federation of Aging Research, Alliance for Aging Research, Glenn Foundation for Medical Research, Alzheimer's Drug Discovery Foundation (formerly, the Institute for the Study of Aging), and the Alzheimer's Association. A.P. owns stock in Elixir Pharmaceuticals, a company that conducts aging research.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1190532/DC1
Materials and Methods

Figs. S1 to S18

Tables S1 to S7

References

21 December 2009; accepted 9 June 2010

Published online 1 July 2010; 10.1126/science.1190532

Include this information when citing this paper.

Fig. 1. Schematic showing the methodology used to discover genetic signatures of EL. The analysis included genetic matching to remove population stratification between cases and controls [(7), sect. 4], discovery and replication of single SNP associations [(7), sect. 5 to 13], multivariate genetic risk modeling and generation of predictive genetic profiles [(7), sect. 14 to 18], and cluster analysis of genetic risk profiles to discover genetic signatures of EL [(7), sect. 19 to 23].

Fig. 2. (A) Genetic risk profiles in 4 study subjects (3 centenarians with ages at death 106, 107 and 119 years, and a control). 150 nested SNP sets were used to predict the probability of EL in 4 subjects (y-axis) and were plotted against the number of SNPs in each set (x-axis). Monotonic increasing trends show strong enrichment of LAVs, because the probability of EL increases when the profile includes a new SNP genotype that is more frequent in centenarians than controls [(7), sect. 18]. **(B)** Sensitivity and specificity of ensemble of genetic risk models. The ensemble of genetic risk models uses all 150 nested SNP sets to compute the probability for EL and AL, and the average probability $p(EL | \Sigma_1, \dots, \Sigma_{150}) = \sum_{i=1}^{150} p(EL | \Sigma_i) / 150$ is used for the final prediction. Left: accuracy in the replication set (76.8% centenarians and 77.4% controls were correctly predicted). Middle: accuracy when the top 150 SNPs are ordered at random rather than by Bayesian significance (76.4% centenarians and 62.2% controls were correctly predicted). Right: results when 150 SNPs are randomly selected from the 2,050 most significant SNPs (64.6% centenarians and 42.2% controls were correctly predicted).

Fig. 3. Example of 9 clusters of genetic risk profiles in centenarians of the discovery set and 3 similar clusters in the replication set. In each plot, the x-axis reports the number of SNPs in each genetic risk model (1,...,150), and the y-axis reports the chance of EL predicted by each model. The boxplots (one for each SNP set on the x axis) display the genetic risk profiles of the centenarians grouped in the same cluster. Numbers N in parentheses are the cluster sizes, and the percentage of models that predict the correct outcome: $p(EL | \Sigma_i) > 0.5$. Color coding represents the accuracy of the genetic risks to predict EL (Blue: more than 99% of the genetic risks models predict EL in all subjects; Red: between 68% and 99%; Orange: between 3% and 50%; Green: less than 3%). The full set of 19 clusters is in fig. S11 and includes more than 90% of centenarians in the discovery set. Note that the same analysis identifies 25 groups among the Illumina control subjects included in the discovery set and 52% of these had genetic signatures similar to the green group (C19) (fig. S13). The results were replicated in a set of 3877 population controls (figs. S14 and S15).

Fig. 4. Correlation of genetic signatures with aging sub-phenotypes. **(A)** Age distribution. The boxplots display the

age of subjects in the 9 clusters in Fig. 3. Each box represents 50% of the distribution and the mid- bar is the median age at death. The clusters are ordered by predictive accuracy from the most predictive (C1) to the least predictive (C19). **(B)** Prevalence of age-related diseases in 6 clusters. The analysis shows that genetic signatures are associated with different combinations of age-related diseases. **(C)** Distribution of age of onset of age-related diseases in 5 clusters. The x-axes report age of events, and the y-axes report the event-free survival distribution. Only subjects with events were included in the analysis. The caption below each plot indicates the disease and the p-value to test significance differences using the log-rank test. Median ages of onsets are in the insets.

**801 Centenarians from NECS (discovery)
254 Centenarians from ELIX (replication)
~4,000 population controls (Illumina and NECS)**

Genetic matching

**Discovery Set
801 centenarians
926 population controls**

**Replication Set
254 centenarians
341 population controls**

**1) Single SNP analysis
70 associations**

**Result 1: Replicated
33 associations**

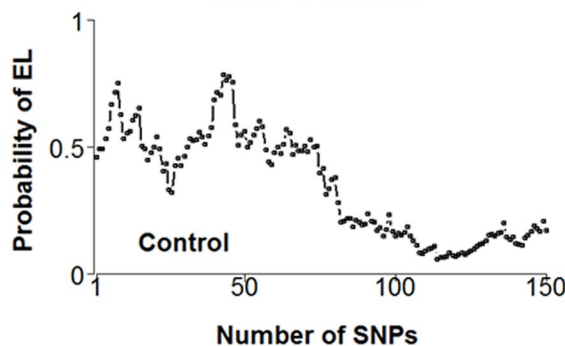
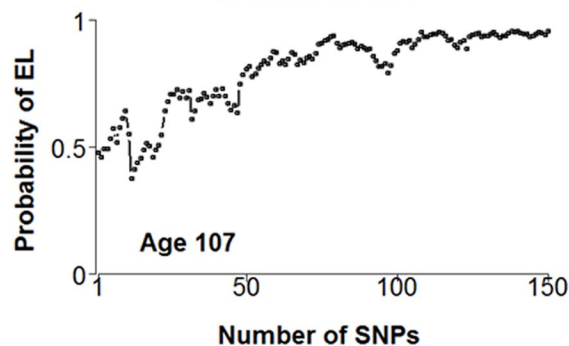
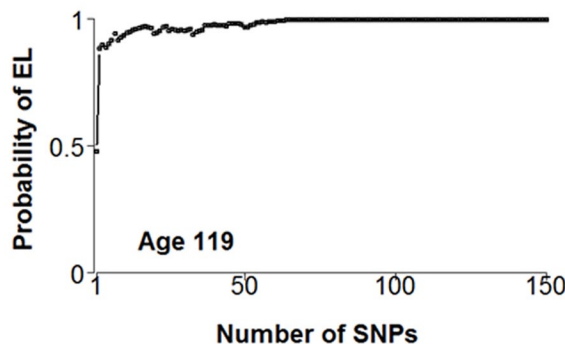
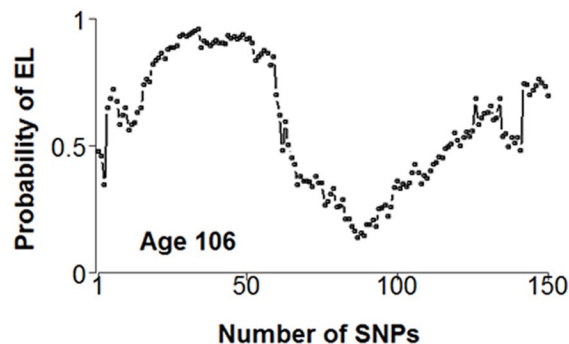
**2) Genetic risk modeling using 150
SNPs**

**Result 2: 77% prediction accuracy
based on independent validation**

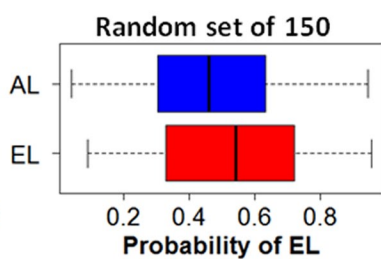
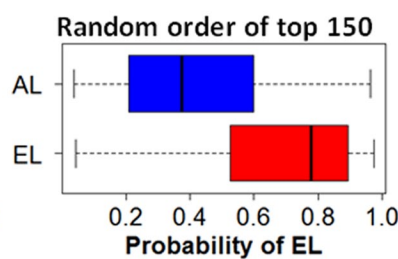
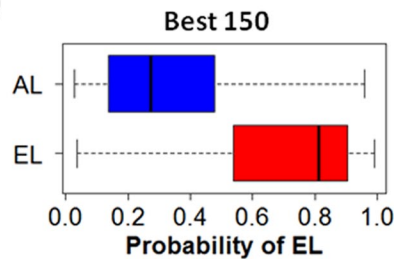
**3) Genetic signature of exceptional
longevity**

**Result 3: Replication of signatures
Correlation with health-span**

A)



B)



Discovery Set

Replication Set

