Mechanisms of Ageing and Development 146 (2015) 23–27

Contents lists available at ScienceDirect

Mechanisms of Ageing and Development

journal homepage: www.elsevier.com/locate/mechagedev

Conservation of pro-longevity genes among mammals

Carter M. Lindborg[a], Kathleen J. Propert[b], Robert J. Pignolo[a,c,*]

[a] Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[b] Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[c] Orthopaedic Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 30 July 2014
Received in revised form 17 March 2015
Accepted 20 March 2015
Available online 21 March 2015

Keywords:
Aging
Longevity
Genetics
Genes
Homology

ABSTRACT

Genes which confer a relative longevity advantage may be regulated at the level of transcription or translation. Alternatively, pro-longevity genes may mediate their effects at the level of protein structure-functional relationships that are beneficially optimized in long-lived species. Longevity associated genes (LAGs) may be operationally defined as genes that confer beneficial effects and are relatively more conserved among long-lived species. Global and local protein sequence alignments of over 10,000 genes across at least 30 mammalian species were examined to identify LAGs. Known LAGs, including growth hormone receptor (*GHR*), and breast cancer 1, early onset (*BRCA1*), have strong associations with maximum lifespan by our analysis. Several common categories of protein function were observed among genes ranked with the strongest associations with MLS identified by all regression models. These genes included those that function in the immune system, cell cycle regulation, and DNA damage response. We provide a ranking of genes with the strongest associations with species maximum lifespan (MLS) by several phylogenetic generalized least squares regression models, including adjustment for confounding variables such as body weight and gestation length.

1. Introduction

Although phenotypic changes related to aging have been abundantly described in both human and other mammalian species, genetic effectors of the aging process are difficult to determine. Current approaches to determining the role of a gene in the aging process include animal models in which either gain or loss of function mutations result in changes in lifespan and comparative human studies of long-lived cohorts (Park, 2011; Hekimi, 2006; Vijg et al., 2001; Cao et al., 2003). With respect to human studies, age-related genes have been identified by a variety of approaches including comparison of differential gene expression in older individuals (particularly centenarians) with younger individuals, longitudinal studies of octogenarians and nonagenarians to identify genes that are differentially expressed in those that reach 100 years of age, and determination of genes in which naturally occurring polymorphisms or mutations result in a change in the aging process (Jylhava and Hurme, 2010).

A novel approach for identification of potentially age-related genes is to examine the conservation of genes in multiple mammalian species and correlate this with the maximum lifespan (MLS) across species. Genetic conservation of DNA or protein sequences can be quantified by alignment methods including the Needleman–Wunsch and Smith–Waterman algorithms, which account for base pair or amino acid matches as well as gaps created to improve overall score and substitution of amino acids with similar functional groups (Smith and Waterman, 1981; Needleman and Wunsch, 1970). Based on standard homology algorithms, the bit score (hereafter referred to as alignment score) for two sequences increases as similarity between the sequences increases. Highly evolutionarily conserved genes are those genes with sequences that are almost identical across species. On the other hand, genes that have a beneficial effect on lifespan may be functionally optimized in longer-lived species based on conservation of DNA or protein sequence information. In this case, it would be expected that these genes and the proteins they encode would be more conserved in a comparison between two longer-lived species and less conserved when comparing between longer-lived and shorter-lived species.

Alignment scores for sequence similarity may thus be used to determine if a gene is longevity-related. By comparing a reference sequence–the genetic sequence in a reference organism–with the homologous sequence in each of the comparison species, a score

* Corresponding author at: Departments of Medicine & Orthopaedic Surgery, University of Pennsylvania School of Medicine, 424B Stemmler Hall, 36th Street & Hamilton Walk, Philadelphia, PA 19104-6081, USA. Tel.: +1 215 898 8653; fax: +1 215 573 2133.
E-mail address: pignolo@mail.med.upenn.edu (R.J. Pignolo).

http://dx.doi.org/10.1016/j.mad.2015.03.004
0047-6374/

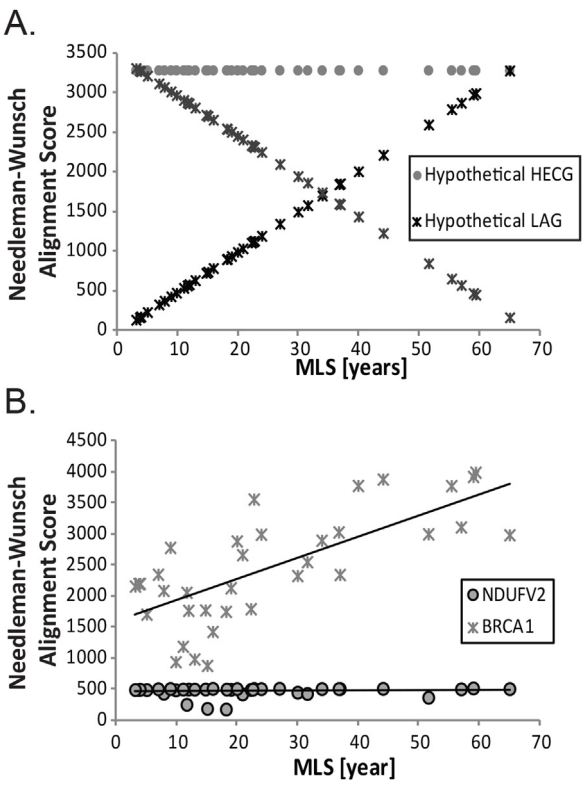

Fig. 1. Relationships of alignment scores for homologous genes and maximum lifespan.
(a) Alignment scores for three hypothetical genes were plotted against the maximum lifespan (MLS) of the species. Hypothetical LAGs may have a steep, positive or negative slope and a correlation approaching ± 1. The hypothetical highly evolutionarily conserved gene (HECG) is a horizontal line indicating that the gene is completely conserved across all species. (b) The alignment scores for the known LAG, *BRCA1*, and the HECG, NADH dehydrogenase (*NDUFV2*), are plotted against species MLS. As can be seen from the figure *BRCA1* has a steeper slope compared to *NDUFV2*.

for the gene in each species may be obtained and compared with the species' maximum lifespan (MLS). As shown in Fig. 1a, three relations can arise when correlating alignment scores with MLS: (1) the gene or protein sequence is near completely or completely conserved across species and independent of MLS, i.e., a Highly Evolutionarily Conserved Gene (HECG), (2) the gene or protein is more similar among longer lived species compared to shorter lived species, and (3) conversely, the gene or protein may be more similar among shorter lived species and less similar among longer lived species. The latter two relationships, at the extremes, represent the cases for longevity associated genes (LAGs).

Here we describe an approach to identify genes related to a trait of interest based on an analysis of their sequence similarity in species where the trait is expressed compared to species where the trait is less favored. By examining the relationship between the extent of conservation of genes across species and the MLS of each species, we report the identification of genes potentially related to longevity and aging (LAGs).

2. Methods

2.1. Data collection

Protein sequences for genes conserved across at least 30 mammalian species were obtained from the OrthoMaM database as FASTA files (Ranwez et al., 2007). Using MATLAB, orthologous sequences were compared using both the Smith–Waterman and Needleman–Wunsch algorithms with BLOSUM62 matrix, gap initiation penalty of 11 and gap extend penalty of 1. Both algorithms generate an alignment score as a measure of sequence alignment and similarity. For each algorithm higher scores indicate a greater degree of conservation and lower scores indicate more divergent sequences. The Needleman–Wunsch algorithm performs a global alignment of sequences and is more sensitive for more closely related sequences (Needleman and Wunsch, 1970) while the Smith–Waterman algorithm performs local alignments and is more sensitive for more distantly related proteins (Smith and Waterman, 1981). Initial analysis indicated ~95% similarity in rankings of genes by the two algorithms. Orthologous genes are closely related by definition so alignment scores produced by the Needleman–Wunsch algorithm were used for analyses due to the greater sensitivity for more closely related sequences. For all comparisons, the human protein sequence was used as the reference sequence and any gene for which there was no human sequence was excluded.

2.2. Phylogenetic generalized least squares (PGLS)

The caper package in R (David Orme et al., 2012; Team RC, 2013) was used to apply the PGLS model to each gene with the similarity score as the dependent variable and the species maximum lifespan as the independent variable (univariable analysis). The PGLS model uses a phylogenetic tree to correct the generalized least squares model for evolutionary relatedness. The mammalian phylogenetic supertree was used for the PGLS model (Bininda-Emonds et al., 2007). The *p*-value for the slope of the regression (p_{MLS}) of each gene was determined. A second PGLS model was applied including gestation length (GL) and body weight (BW) as potential confounding variables (multivariable analysis) and p_{MLS} was determined.

2.3. Residual analysis with phylogenetic correction

For each gene, residuals were calculated from the regression of the alignment scores with a confounding variable (Res_{ASvCV}). Residuals were also calculated from the regression of MLS with a confounding variable (Res_{MLSvCV}). The PGLS model was then applied with Res_{ASvCV} and Res_{MLSvCV} as the dependent and independent variables, respectively, to determine p_{MLS} (Speakman, 2005). This method was applied using BW ($Speakman_{BW}$) and GL ($Speakman_{GL}$) as confounding variables separately. These p-values were not adjusted for multiple comparisons but used primarily to compile a ranking of genes of interest.

2.4. Identification of genes of interest

For each model, genes with a p_{MLS} of <0.01 were identified. The genes with the smallest *p*-values are those orthologous genes that are most divergent when compared between longer and shorter lived species, and thus more conserved among long-lived species. We defined consensus genes as those genes with a $p_{MLS} < 0.05$ identified by all models. The known functions and mutant phenotypes of these consensus genes was determined from the GeneCard gene encyclopedia which compiles various information on genes from multiple sources (Safran et al., 2010).

3. Results

Fig. 1b shows the relationship between alignment score and MLS for *BRCA1*, a known LAG, and NADH dehydrogenase (*NDUFV2*), a known HECG. As can be seen from the figure, the LAG has a steeper slope than the HECG [34.39 for *BRCA1*; 0.58 for *NDUFV2*]. Based on the four regression analyses, *BRCA1* has an average p_{MLS} of 0.004 while *NDUFV2* has an average p_{MLS} of 0.53 where p_{MLS} is the *p*-value