BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 7 2009, pages 875–881
doi:10.1093/bioinformatics/btp073

Gene expression

Meta-analysis of age-related gene expression profiles identifies common signatures of aging

João Pedro de Magalhães[1,*,†], João Curado[2] and George M. Church[1]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA and [2]Escola Superior de Biotecnologia, 4200 Porto, Portugal

Received on September 15, 2008; revised on January 11, 2009; accepted on January 31, 2009

Advance Access publication February 2, 2009

Associate Editor: David Rocke

ABSTRACT

Motivation: Numerous microarray studies of aging have been conducted, yet given the noisy nature of gene expression changes with age, elucidating the transcriptional features of aging and how these relate to physiological, biochemical and pathological changes remains a critical problem.

Results: We performed a meta-analysis of age-related gene expression profiles using 27 datasets from mice, rats and humans. Our results reveal several common signatures of aging, including 56 genes consistently overexpressed with age, the most significant of which was *APOD*, and 17 genes underexpressed with age. We characterized the biological processes associated with these signatures and found that age-related gene expression changes most notably involve an overexpression of inflammation and immune response genes and of genes associated with the lysosome. An underexpression of collagen genes and of genes associated with energy metabolism, particularly mitochondrial genes, as well as alterations in the expression of genes related to apoptosis, cell cycle and cellular senescence biomarkers, were also observed. By employing a new method that emphasizes sensitivity, our work further reveals previously unknown transcriptional changes with age in many genes, processes and functions. We suggest these molecular signatures reflect a combination of degenerative processes but also transcriptional responses to the process of aging. Overall, our results help to understand how transcriptional changes relate to the process of aging and could serve as targets for future studies.

Availability: http://genomics.senescence.info/uarrays/signatures.html

Contact: jp@senescence.info

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Changes in gene expression are associated with numerous biological processes, cellular responses and disease states. The availability of microarrays has made it possible to study gene expression in a high-throughput fashion and gather insights about biology and disease. In recent years, a massive amount of microarray studies have been conducted. To compile and organize the numerous datasets generated, resources like the Gene Expression Omnibus (GEO) (Barrett *et al*., 2007) have been established. The availability of microarray data from multiple experiments opens up new research opportunities. By eliminating idiosyncrasies of individual platforms and enhancing the signal-to-noise ratio, comparing profiles across platforms and species may reveal conserved molecular signatures that would otherwise be obscure in single datasets (Moreau *et al*., 2003; Ramasamy *et al*., 2008). Indeed, meta-analyses of gene expression profiles integrating multiple microarray studies have been particularly useful to identify conserved genetic signatures of cancer (Rhodes *et al*., 2004).

Aging is major biological process and a risk factor for many diseases. To gain new insights into the process of aging and identify potentially important genes and biomarkers, many microarray studies have been conducted in several species, including humans, either by comparing young and old tissues (Edwards *et al*., 2007; Ida *et al*., 2003) or by comparing samples across the lifespan (Lu *et al*., 2004; Rodwell *et al*., 2004). To collect and store the large body of gene expression data in aging, a database of gene expression aging studies was recently assembled entitled Gene Aging Nexus (GAN) (Pan *et al*., 2007). Aging gene expression studies, however, have been typically noisy with often few genes found to be differentially expressed with age and of these even fewer found to overlap different tissues (Weindruch *et al*., 2002) and species (McElwee *et al*., 2007). Therefore, elucidating the transcriptional features of aging and how these relate to physiological, biochemical and pathological changes with age remains a critical problem.

Considering the number of aging gene expression studies conducted to date in different tissues and organisms, it may be possible to employ combinatorial approaches to identify common molecular signatures of the normal aging process. Because the underlying molecular mechanisms of aging remain a subject of debate, however, the mere existence of transcriptional programs driving aging is a contentious issue, and whether independent transcriptional programs can drive aging in different tissues is unknown. Previous results suggest that most genes differentially expressed with age in a given tissue are not genes specifically expressed in that tissue (Rodwell *et al*., 2004), suggesting that only a small fraction of transcriptional responses are tissue-specific and hence that molecular signatures of aging might overlap different tissues. Nonetheless, molecular signatures of aging can be subject to different interpretations rather than as an active program of aging.

*To whom correspondence should be addressed.

†Present address: School of Biological Sciences, University of Liverpool, Biosciences Building, Crown St., Liverpool L69 7ZB, UK.

For example, they may represent compensatory mechanisms (de Magalhaes and Toussaint, 2004).

In this work, our goal was to identify common molecular signatures of aging. Such signatures, which we define as distinguishing features of molecular changes with age, may be associated with and play a biological role in the physiological decline that characterizes aging. We obtained data from 27 publicly available studies in mice, rats and humans from GAN and GEO, and performed a meta-analysis of age-related gene expression profiles. Many methods exist for the statistical and functional annotation of microarray data (Hong and Breitling, 2008; Ramasamy *et al*., 2008; Slonim, 2002; Verducci *et al*., 2006). Herein, we developed a simple methodology to compare age-related gene expression data across platforms and species that in order to cope with the noisy nature of age-related gene expression profiles emphasizes sensitivity. Our results reveal several signatures of aging most notably involving an activation of inflammation/immune response genes and an underexpression of mitochondrial genes. We interpret our signatures in the context of known physiological and biochemical age-related alterations. Our results further reveal many previously unknown transcriptional changes with age in genes, processes and functions that could serve as targets for future studies and help to paint a better picture of how transcriptional changes relate to the process of aging at different levels.

2 METHODS

2.1 Data selection and processing

Microarray data was primarily downloaded from GAN version 2.0 (http://gan.usc.edu/) (Pan *et al*., 2007), a repository of gene expression data in aging with a few datasets also downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/) (Barrett *et al*., 2007). This gene expression data is already normalized with background subtracted (Pan *et al*., 2007). All experiments reporting age-related expression profiles in mammals were downloaded, including studies comparing young and old samples and studies reporting gene expression profiles at more than two age groups. The vast majority of datasets consisted of single-channel intensity data from Affymetrix microarrays. One study was performed using spotted cDNA microarrays (Ida *et al*., 2003), but was also included since the signal intensity from the young and old pairs of specimens compared in the original report were available.

Only age-related data from healthy, adult, non-treated specimens was analyzed and data from specific diseases, treatments and mutants were excluded. For example, in caloric restriction studies we only took data from young and old controls, not from the calorie-restricted animals. Experiments with less than three samples for either young or old specimens (but including pooled samples examined using the same microarray) were excluded. Since aging gene expression profiles can be detected early in adult life (Lu *et al*., 2004; McCarroll *et al*., 2004), all datasets with more than two adult time points were included, even if the oldest animals were middle-aged. Studies in which data was obtained from a set of genes selected in a highly biased fashion (e.g. custom-arrays featuring only genes associated with a given pathway) were excluded. Although we cannot perform a comprehensive evaluation of the quality of each experiment, a meta-analysis is in its essence a technique to eliminate poor quality data.

Overall, 12 experiments from mice, 11 from rats and 4 from humans were analyzed (Supplementary Table S1), comprising almost 5 million gene expression measurements from over 400 individual samples. As described (Pan *et al*., 2007), genes in different platforms are linked by their UniGene IDs. In GEO, gene annotation is derived from the Entrez Gene and UniGene databases using sequence identifiers (Barrett *et al*., 2007). If more than 30% of measurements for a given probe contained nulls or missing values, the probe was excluded. Otherwise, null values were replaced by the probe's average (row average method) and probes targeting the same gene were averaged.

2.2 Detecting genes with age-related expression profiles

To avoid the problems of comparing microarray data obtained using different platforms and experimental systems (e.g. in the number and type of samples), we discarded effect sizes and instead employed a meta-profiling method that compares statistical measures obtained from each dataset, a variant of the value counting procedure initially applied to study cancer (Rhodes *et al*., 2002, 2004). A flow diagram of our method is available as Supplementary Figure S1.

For each dataset, we first tested the hypothesis that the expression of a given gene is associated with age. Data were log2-transformed and we performed a linear regression for each gene using the equation:

$$Y_{ij} = \beta_{0j} + \beta_{1j}\mathrm{Age}_i + \varepsilon_{ij} \quad (1)$$

where Y_{ij} is the signal intensity of gene j in sample i, Age_i is the age of the specimen from which sample i was obtained, and ε_{ij} is the error term. The coefficients β_0 and β_1 were estimated by least squares.

Statistical significance of the differential expression was estimated with a two-tailed F-test to determine whether the slope of the curve is different than zero, which herein would indicate an association between the expression signal and age. Genes with a P-value below 0.05 were considered putatively age dependent. Using this 0.05 cutoff, we obtained between 0.91% and 9.84% of putative over- and underexpressed genes with age for each experiment with, on average, 4.52% of genes overexpressed in each experiment and 4.62% underexpressed. Admittedly, this is a relatively relaxed cutoff threshold and, considering most experiments studied thousands of genes, emphasizes sensitivity rather than specificity. In fact, 13 793 genes passed our F-test at least once, which represents roughly half of all tested genes. As described below, however, we were careful to correct for multiple hypothesis testing when determining statistical significance of the combined profiles to minimize false positives. Calculations were performed using the R language (R Development Core Team, 2008).

Where possible, we compared our results with those originally published with the datasets to verify that, in spite of our more relaxed threshold, there was a considerable overlap between the results. In the case of the human kidney dataset, we noticed a discrepancy between our results and those reported by the authors (Rodwell *et al*., 2004), particularly for the results obtained in kidney medulla, and thus we decided to use only the data from kidney cortex. In the original report with this dataset, the cortex had considerably more genes differentially expressed with age than the medulla (Rodwell *et al*., 2004), so despite this procedure our results were still largely representative of those initially reported for this dataset.

2.3 Identifying common signatures of aging

To identify genes consistently under- or overexpressed during aging, we tried to find the genes with the largest number of putatively age-related signals in our multiple datasets. To calculate the probability of observing an equal or higher than number of under- or overexpressed gene occurrences, we employed the cumulative binomial distribution:

$$P(X \geqslant k) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j} \quad (2)$$

where the probability P of a gene being overexpressed with age is 0.0452 and the probability of it being underexpressed with age is 0.0462 (as detailed above), the number of occurrences (k) is the number of experiments in which the gene is putatively under- or overexpressed and the number of trials (n) is the number of experiments in which the gene's expression was measured.

If available, we used HomoloGene (Wheeler *et al*., 2008) to obtain human homologs of rodent genes and our results are, with rare exceptions, displayed using the Entrez Gene ID of the human homolog and the corresponding HUGO Gene Nomenclature Committee (HGNC) symbol.

Fisher's inverse chi-square approach was used to serve as a comparison with the value counting method. Succinctly, for each gene we calculated the sums of the logarithm of the *P*-values across k studies for over- and underexpression separately and compared this test statistic against a χ^2-distribution with $2k$ degrees of freedom, as described (Hong and Breitling, 2008; Ramasamy *et al*., 2008).

To identify enriched functional groups present in our top genes, we employed the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis *et al*., 2003). Because this analysis focused on enriched functional categories rather than individual genes, we used a more liberal cutoff threshold for selecting the top genes, as detailed below. DAVID was run with default options.

We used Gene Ontology (GO) annotation, which describes how gene products behave in a cellular context (Ashburner *et al*., 2000), to further identify pathways, processes and functions significantly altered by aging. To identify GO categories that tend to be associated with genes under- or overexpressed with age, we used Equation (2), yet the number of occurrences (k) and the number of trials (n) for each category, rather than referring to a single gene, refer to all genes associated with that category. In other words, for each of the 8293 GO categories present, we calculated the number of times each gene associated with that category was over- or underexpressed and then determined statistical significance using the cumulative binomial distribution. We used GO annotation from each of the species and then combined the results for all GO categories using the results from the three species used. Our algorithm was implemented in the Perl language.

In order to employ a set of profiles as diverse as possible, we tried to obtain datasets from different tissues. In the few cases two or more experiments from the same tissue and species were available, we only counted them as one if the experiments yielded convergent results and discarded them if they yielded divergent results (i.e. a given gene being underexpressed with age in one experiment and overexpressed in another). In one instance regarding human muscle datasets, we merged the results from two separate experiments since these were conducted by the same group using the same methods and platform (Welle *et al*., 2003, 2004).

2.4 False discovery rate simulations

To estimate the number of false positives in both individual genes and GO categories, we performed 1000 simulations using random permutations and the exact same procedure described above. In other words, for each dataset the gene identifiers corresponding to each expression profile were randomly permutated using the Fisher–Yates shuffle algorithm, but the total number of genes, the gene names and the expression profiles remained the same. This allowed us to estimate, by chance, how many genes and GO categories above a given threshold we were expected to find. Based on our simulations, we calculated the false discovery rate (FDR) (Q), defined as the number of expected false positives over the number of significant results (Storey and Tibshirani, 2003), for each gene and GO category. We set our significance threshold at $Q < 0.1$, which though admittedly arbitrary (Storey and Tibshirani, 2003) has been used in similar studies (Rhodes *et al*., 2004). Our full results are available as Supplementary Material and on our website (http://genomics.senescence.info/uarrays/signatures.html) if others wish to perform analyses with different criteria. For $Q < 0.1$ we set the cutoff *P*-value at $P < 0.0007$ for overexpressed genes, $P < 0.0002$ for underexpressed genes, $P < 0.006$ for GO categories overrepresented in overexpressed genes and $P < 0.003$ for GO categories overrepresented in underexpressed genes. For the FDR analysis using DAVID, we relaxed the threshold to $Q < 0.5$ which resulted in a cutoff *P*-value of $P < 0.02$ for overexpressed genes and $P < 0.009$ for underexpressed genes. The simulations were performed using the Perl language.

2.5 Comparing age-related gene expression signals between datasets

Microarray data were obtained from different studies, under different conditions. To minimize biases and idiosyncrasies of individual experiments when comparing gene expression profiles, we normalized the log2 signals to common young and old ages for each species: respectively, 25 and 75 years in humans, 3 and 25 months for mice and 5 and 30 months for rats. In spite of these differences in lifespan, it is reasonable to think that the process of aging in humans and rodents shares at least some mechanisms, only timed at different paces, and results in common molecular signatures. From the regression coefficients, we calculated the log2 ratio of the expression signal old/young using the abovementioned normalized ages. The meta-signature using these values was rendered using TreeView (Eisen *et al*., 1998).

3 RESULTS

3.1 Identifying genes differentially expressed with age

Normalized microarray data, by and large single-channel intensity data from Affymetrix microarrays, was downloaded primarily from GAN but also GEO. After data selection (see Section 2), we obtained 27 different experiments from mice, rats and humans comprising almost 5 million gene expression measurements from over 400 individual samples. One can imagine advantages to human-only datasets, since there are likely differences in aging between organisms and this is an area of great controversy (McCarroll *et al*., 2004; McElwee *et al*., 2007; Zahn *et al*., 2007), but considering the few microarray studies conducted in aging humans, and possible advantages in meta-comparisons, we chose to expand our analysis to primates plus rodents.

There are several meta-analysis methods for detecting differentially expressed genes in microarray experiments (Hong and Breitling, 2008; Moreau *et al*., 2003; Ramasamy *et al*., 2008). To avoid the difficulties of comparing microarray data obtained in considerable different conditions, we compared the statistical parameters of differential expression obtained from the individual datasets instead of comparing the gene expression signals between the platforms, a value counting method previously employed in the meta-analysis of cancer (Rhodes *et al*., 2002, 2004). Because gene expression changes with aging tend to be more subtle than in specific diseases (Pan *et al*., 2007), we employed a relatively relaxed threshold with $P < 0.05$ to test whether the expression of each gene in each experiment is associated with age (see Section 2). Afterwards, to identify genes consistently differentially expressed with age, we used the binomial distribution to test whether the number of times a given gene is putatively under- or overexpressed with age is higher than expected by chance, after adjusting our *P*-values based on FDR analyses to select genes with a FDR (Q) below 0.1 (see Section 2). A flow diagram of our meta-analysis method is available as Supplementary Figure S1.

Overall, we identified 56 genes across the studies consistently overexpressed with age, using a cutoff of $P < 0.0007$. Simulations in which genes were randomly assigned to gene expression signals in each experiment showed that five significant genes would be expected by chance for a FDR below 10%. Therefore, the genes we identified represent an intersection of genes that share an expression profile significantly associated with aging (Fig. 1A). The gene with the lowest *P*-value was *APOD* or apolipoprotein D, previously associated with neurodegenerative diseases (Kalman *et al*., 2000). In addition, numerous genes overexpressed with age play roles in

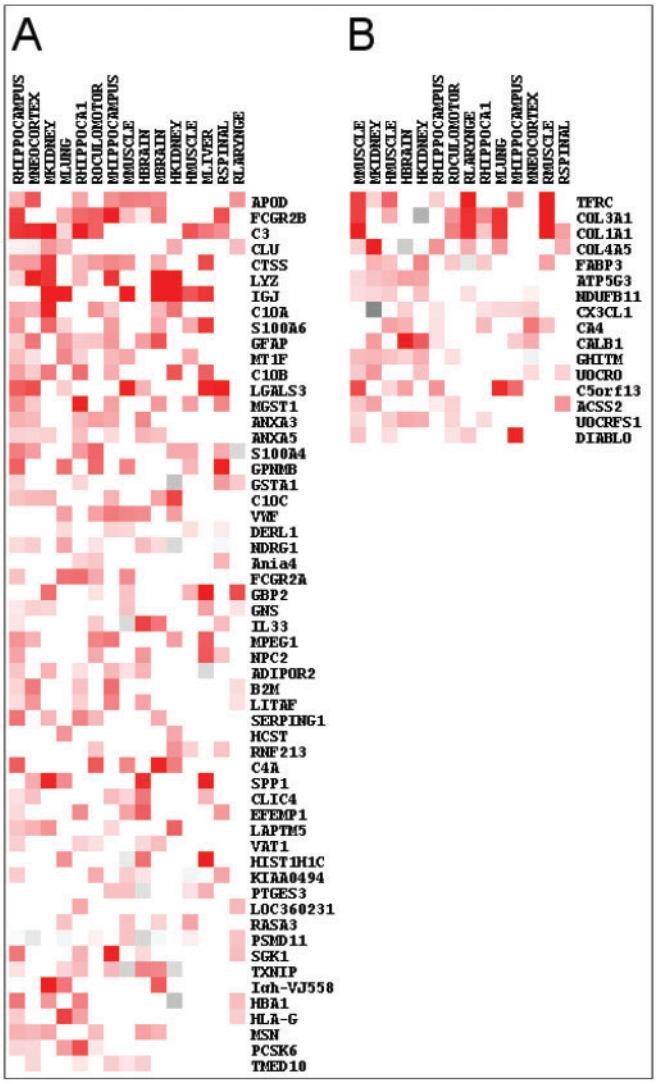

Fig. 1. Meta-signature of top genes consistently differentially expressed with age. (**A**) Genes consistently overexpressed with age. Fifteen datasets were selected for the figure. Red indicates genes overexpressed with age with the intensity proportional to the expression signal old/young adjusted for a common age (see Section 2). Black to gray indicates genes underexpressed with age. (**B**) Fourteen datasets were selected for the figure. Red indicates genes underexpressed with age with the intensity proportional to the expression signal young/old adjusted for a common age (see Section 2). Black to gray indicates genes overexpressed with age. For both A and B, white indicates either not studied or non-significant and genes are ordered from most to least significant.

inflammation, such as *CTSS*, *FCGR2B*, *IGJ*, *C3*, *C1QA* and *C1QB*. Other genes consistently overexpressed with age included lysozyme (*LYZ*), clusterin (*CLU*), microsomal glutathione S-transferase 1 (*MGST1*), glutathione S-transferase A1 (*GSTA1*), S100 calcium binding protein A4 (*S100A4*) and A6 (*S100A6*), and annexin A3 (*ANXA3*) and A5 (*ANXA5*).

We found only 17 genes consistently underexpressed with age at a cutoff of $P < 0.0002$ (from our simulations an average of 1.5 would be expected by chance). These results are shown in Figure 1B and include four genes encoding mitochondrial proteins (*ATP5G3*, *NDUFB11*, *UQCRQ* and *UQCRFS1*) and three collagen genes (*COL3A1*, *COL1A1* and *COL4A5*). The top gene was the transferrin receptor *TFRC*.

Interestingly, nine genes overexpressed and four underexpressed from our meta-signature have been validated experimentally, mostly by direct measurement of mRNA levels by qRT-PCR (Supplementary Table S2), which demonstrates that our method can detect biologically meaningful results.

To further assess the power of our method, we compared the results to those obtained with Fisher's inverse chi-square approach. Among the most significant genes, the overlap was considerable with 8 out of the 10 most significant genes using our method also being statistically significant using Fisher's inverse chi-square approach. Although we obtained a larger number of significant genes using this approach (112 versus 73), this was mostly due to single experiments with a small number of samples having a biased weight on the test statistic (Supplementary Tables S5 and S6).

Table 1. Top functional annotation clusters of significant differentially expressed genes

Cluster	Enrich. score	No. of annot.	No. of genes
Overexpressed genes ($n = 236$ with $Q < 0.5$)			
Immune response, complement activation	6.88	41	86
Lysosome	6.48	7	16
Plasma, extracellular region	5.41	5	37
Signal, glycoprotein	4.55	6	80
Negative regulation of apoptosis	2.75	16	53
Underexpressed genes ($n = 141$ with $Q < 0.5$)			
Mitochondrion	5.49	52	70
Oxidative phosphorylation	3.57	79	82
Cytoplasm	3.19	5	108
Hydroxylysine, hydroxylation, collagen	2.83	43	47

Clusters from DAVID with an enrichment score above 2.5 are displayed. Cluster titles were selected based on the broadest of the top annotations in the cluster.

3.2 Functional annotation clustering of top genes

To identify the biological processes associated with gene expression changes with age, we first evaluated our top genes differentially expressed with age using the functional annotation tools in DAVID, a web-accessible set of tools that allow researchers to infer the biological meaning behind large lists of genes (Dennis *et al*., 2003). Because our focus was on enriched functional categories rather than on individual genes, we employed a more liberal criterion to select the genes for functional clustering, instead using a FDR of 50% (i.e. $Q < 0.5$). Among the 236 genes overexpressed with age (118 would be expected by chance), the top cluster was related to immune responses. Also noteworthy were clusters related to the lysosome and apoptosis (Table 1).

For the 141 genes underexpressed with age at the more liberal threshold (69 were expected by chance), the top clusters were mostly related to mitochondria and oxidative phosphorylation as well as collagen (Table 1).

3.3 GO categories overrepresented in age-related transcriptional profiles

To further identify and characterize the biological processes and functions associated with gene expression changes with age, we identified GO categories overrepresented among genes over- or underexpressed with age. Instead of identifying categories enriched in the genes with the strongest signal (as is common in microarray studies and for which we used DAVID), in our meta-analysis we counted the number of occurrences of all genes in all experiments associated with a given GO category and, using the binomial distribution, determined the probability of obtaining an equal or higher number of putative under- or overexpressed gene occurrences (see Section 2). At our FDR-adjusted cutoff ($P < 0.006$ for $Q < 0.1$), we found 175 GO categories enriched for genes overexpressed with age when 17 would be expected by chance. The top categories were largely in accordance with the functional annotation clusters obtained from the top genes, such as categories related to immune response, like complement activation (GO:0006958 and GO:0006956) and antigen processing (GO:0019886 and GO:0002504), the lysosome (GO:0005764), and apoptosis (GO:0006915) and anti-apoptosis (GO:0006916).

The results from our GO analysis not only reinforced those obtained with DAVID but also allowed us to associate many other potentially interesting functions and processes with aging. GO categories overrepresented for overexpressed genes included phagocytosis (GO:0050766 and GO:0006911), lysozyme (GO:0003796), detoxification of copper ion (GO:0010273), cadmium ion binding (GO:0046870), transcription repressor activity (GO:0016564) and negative regulation of transcription (GO:0045892), tau protein binding (GO:0048156), insulin-like growth factor binding (GO:0005520), retinoid binding (GO:0005501) and glutathione (GO:0004364 and GO:0006749). Although interpreting these results is not straightforward as many could represent adaptations to aging, others could play some mechanistic role in aging and their potential utility to further studies is discussed below.

We found 84 GO categories enriched for genes underexpressed with age at a cutoff of $P < 0.003$ (eight would be expected by chance). Again, the top categories were largely consistent with the functional annotation clusters identified through DAVID with several categories related to mitochondria (GO:0005759, GO:0005743, GO:0005739, etc.), electron transport chain (GO:0006120, GO:0005747, GO:0005746, etc.) and NADH dehydrogenase activity (GO:0008137 and GO:0003954). There were other categories related to energy metabolism such as tricarboxylic acid cycle (GO:0006099), glycolysis (GO:0006096), aerobic respiration (GO:0009060) and even the broader category metabolic process (GO:0008152). Lastly, some categories were related to collagen (GO:0005586, GO:0032964 and GO:0005581).

4 DISCUSSION

One major difference between cancer and aging microarray studies is that while in cancer many genes tend to be differentially expressed and the challenge is to identify the most important ones, few genes tend to be differentially expressed with age and the challenge is to identify significant ones. To address this, we modified the meta-profiling algorithm previously successfully employed to study cancer (Rhodes *et al*., 2002, 2004). Our method is a two-step process that first evaluates—using a relaxed threshold to emphasize sensitivity—whether genes are associated with age in individual datasets and then constructs a meta-profile from the aggregate datasets using the binomial distribution and setting cutoff thresholds based on FDR simulations.

By integrating gene expression profiles from several studies we were able to identify genes that tend to be consistently over- or underexpressed with age, a meta-signature of aging in mammals. These genes overlapped with those obtained using Fisher's inverse chi-square approach and though some are novel, others have been validated by direct methods, showing that our method is adequate for dealing with the idiosyncrasies of aging gene expression profiles, such as the high heterogeneity of the datasets.

Although gene expression changes may follow or drive the process of aging, these differentially expressed genes may serve as a basis for further studies, for example, for deriving reliable biomarkers of aging. In addition, we were able to associate biological processes and functions with this meta-signature. One major pathway that we found upregulated with aging was the immune/inflammatory response, which is in line with what is known about the physiology of aging. It is well-established that inflammatory levels increase with age and inflammatory processes have been associated with various age-related diseases (Bruunsgaard *et al*., 2001). Besides, given the systemic nature of the immune system, it is reasonable that changes with age in inflammatory and/or immune response would have an effect on different tissues that could be detected as common molecular signatures of aging.

Interestingly, we also found evidence of pathways consistently altered during aging in multiple tissues that involve mechanisms intrinsic to cells. As aforementioned, however, interpreting aging microarray experiments is no trivial task (Clarke *et al*., 2008). Available expression data cannot dissect out the age-related responses of different cell types in a complex sample, some cell types of which may be dying while others may be growing, while others are simply quiescent. Besides, genes differentially expressed with age may indicate a transcriptional response to aging rather than an underlying mechanism or transcriptional program causing degeneration (de Magalhaes and Toussaint, 2004). For example, *APOD* appears to play a role in protection from oxidative stress and, in fact, overexpression of human *APOD* in flies extends lifespan (Muffat *et al*., 2008). Therefore, we hypothesize that the upregulation of *APOD* with age may not be a deleterious mechanism associated with the physiological decline characteristic of aging but rather may be a response to the process of aging. Many other genes overexpressed with aging, like *MGST1* which is known to protect cells from oxidative stress (Siritantikorn *et al*., 2007), might also fall into this category.

Our results suggest an overexpression of genes related to lysosomes, such as cathepsins (*CTSS*, *CTSH* and *CTSZ*), and the lysosomal membrane. Lysosomes degrade many macromolecules, including proteins, and biochemical changes in these organelles have been described with aging (Cuervo and Dice, 2000). One hypothesis is that the overexpression of genes associated with lysosomal function, as well as that of genes related to phagocytosis, is a cellular response to the accumulation of abnormal proteins with age. In this context, adaptive aging gene expression changes could help pinpoint changes at other levels. We also found an overexpression with age of anti-apoptotic genes and cell-cycle

regulators like granulin (*GRN*) and annexins. It is possible that some of these genes are not upregulated due to their role in apoptosis but rather as a part of other functions. For example, clusterin (*CLU*) is an extracellular chaperone that could curtail the effects of misfolding and aggregation of proteins (Kumita *et al*., 2007).

Interestingly, we found evidence of overexpression of genes previously found overexpressed in senescent cells, such as fibronectin (*FN1*) (Kumazaki *et al*., 1991) and p21 (*CDKN1A*), though both were slightly above our cutoff threshold (respectively, $Q = 0.11$ and $Q = 0.21$). The upregulation of p21 with age was previously validated by western blot in muscle (Edwards *et al*., 2007) and could be related to an increased proportion of senescent or growth arrested cells with age, an area of extensive other studies and interest (de Magalhaes and Faragher, 2008). Increased levels of *CLU* have too been associated with cellular senescence and shown to protect cells from cytotoxic insults (Dumont *et al*., 2002), and *APOD* expression also increases in senescent fibroblasts (Provost *et al*., 1991). These results suggest that senescent biomarkers detected *in vitro* may be important biomarkers during mammalian aging *in vivo*. Considering that inflammatory processes can induce senescence and that senescent cells can secrete inflammatory cytokines (Kuilman *et al*., 2008), these cellular biomarkers may well be related and/or contribute to systemic factors, emphasizing the need for integrative models of aging (de Magalhaes and Faragher, 2008).

Genes underexpressed with age may be simpler to interpret since, not only we found fewer than those overexpressed, but most fall into energy metabolism categories and are less likely to represent transcriptional responses to aging. Significantly, we found an underexpression of mitochondrial genes, including of genes associated with the electron transport chain, which is in agreement with known biochemical and physiological observations suggesting a mitochondrial functional decline with age (Ames *et al*., 1995). In particular, a respiratory failure with age has been reported in high energy-consuming tissues like the brain (Navarro *et al*., 2002) and muscles (Trounce *et al*., 1989), which make up a considerable number of our datasets. Extracellular matrix and collagen were also found to be underexpressed with age, mostly due to the underexpression of different forms of collagen. Though slightly above our cutoff threshold ($Q = 0.18$), elastin (*ELN*) was found to be underexpressed with age. Age-related changes in collagen and elastin, such as reduced collagen deposition, is typical of aged tissues such as the skin (Uitto, 1986). Our results suggest that these may represent common biochemical age-related changes.

Our functional annotation analyses of top genes using DAVID and of GO categories using a value counting method reinforced each other, thus demonstrating the power and accuracy of the new meta-profiling method employed in the latter. Moreover, the analyses of GO categories revealed many other biological processes of potential interest for associating molecular changes during aging with physiological changes, including processes (to our knowledge) not previously associated with aging in gene expression studies. Succinctly, we found evidence of upregulation of blood coagulation (GO:0007596), which is in agreement with the reported hypercoagulability of aged individuals (Mari *et al*., 1995). Since blood coagulation potential increases from an early age (Andrew *et al*., 1992), another possible interpretation of gene expression changes with age related to developmental mechanisms that continue throughout adulthood (de Magalhaes and Church, 2005). On another note, we found evidence of upregulation of transcription repressor activity (GO:0016564) and negative regulation of transcription (GO:0045892), suggesting that transcriptional activity decreases with age. This is in line with previous results indicating a decrease in RNA and protein synthesis with age. Because total RNA and protein content do not appear to decrease with age, one hypothesis is that RNA and protein turnover decrease with age and might be a factor in the age-related accumulation of abnormal proteins (Van Remmen *et al*., 1995). Also, we found an upregulation with age of detoxification pathways, such as xenobiotic catabolic process (GO:0042178) and detoxification of copper ion (GO:0010273), which again may suggest a transcriptional response to the process of aging.

The breadth of GO categories we identified as significantly associated with aging opens avenues for future studies by, for example, a more careful analysis of whether changes in these processes parallel other age-related changes and pathologies and even by gene manipulation experiments in model systems to test whether such processes might drive aging. Though it is impossible for us to discuss all the significant functional categories that we found differentially expressed with aging, many may merit further attention to understand transcriptional changes during aging and our full results are available as Supplementary Tables S7 and S8 and on our website (http://genomics.senescence.info/uarrays/signatures.html).

Recently, the Atlas of Gene Expression in Mouse Aging Project (AGEMAP) reported gene expression profiles with age for 8932 genes in 16 mouse tissues (Zahn *et al*., 2007). We chose not to include this large dataset in our meta-analysis because then our work would be considerably biased towards the AGEMAP results. Nonetheless, it is interesting to note that there is a considerable overlap among the top functional categories identified in AGEMAP and in our study with genes related to mitochondrial electron transport chain found underexpressed, and cell cycle and immune response/inflammation genes found overexpressed in AGEMAP (Zahn *et al*., 2007). These similarities are noteworthy as they emphasize the quality and utility of meta-analyses for aging research and how one can obtain meaningful global signatures of aging using a cost-effective computational method.

5 CONCLUSIONS

Although other studies have compared age-related microarray datasets from different species and in long-lived mutants and conditions (e.g. in caloric restriction) (McCarroll *et al*., 2004; McElwee *et al*., 2007; Swindell, 2008), our work is the first to perform a comprehensive meta-profiling of aging in a systematic way to identify conserved signatures of aging. By integrating multiple gene expression profiles and employing a novel method that emphasizes sensitivity, we were able to identify genes and processes altered by aging at the transcriptional level with unprecedented power, and our work reveals previously unknown transcriptional changes during aging, in particular a surprisingly large number of GO categories. These genes and functions could represent targets for future studies in helping to define biomarkers of aging, testing their mechanistic role experimentally and in helping to develop the emerging discipline of computational systems biology of aging by increasing our understanding of transcriptional regulation during aging (Kriete, 2006). Indeed, we suggest that these molecular signatures of aging not only reflect a mix of degenerative processes

but also transcriptional responses to aging as healthy cells adapt to degeneration. As the aging transcriptome continues to be increasingly better characterized, meta-profiling and integrative approaches will be increasingly more useful to understand the aging process.

ACKNOWLEDGEMENTS

The authors would like to thank John Aach for valuable comments on previous drafts of the article, Xianghong Jasmine Zhou and Chun-Chi Liu for their assistance with GAN, Graham Rockwell for assistance with the R language and useful comments on previous drafts and Jorge Ivan Velez for his help with the R language.

Funding: National Institutes of Health-National Human Genome Research Institute Centers of Excellence in Genomic Science (to G.M.C.); Fundação Luso-Americana (to J.C.).

Conflict of Interest: none declared.

REFERENCES

Ames,B.N. *et al*. (1995) Mitochondrial decay in aging. *Biochim. Biophys. Acta*, **1271**, 165–170.

Andrew,M. *et al*. (1992) Maturation of the hemostatic system during childhood. *Blood*, **80**, 1998–2005.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Barrett,T. *et al*. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

Bruunsgaard,H. *et al*. (2001) Aging and proinflammatory cytokines. *Curr. Opin. Hematol.*, **8**, 131–136.

Clarke,R. *et al*. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat. Rev. Cancer*, **8**, 37–49.

Cuervo,A.M. and Dice,J.F. (2000) When lysosomes get old. *Exp. Gerontol.*, **35**, 119–131.

de Magalhaes,J.P. and Church,G.M. (2005) Genomes optimize reproduction: aging as a consequence of the developmental program. *Physiology (Bethesda)*, **20**, 252–259.

de Magalhaes,J.P. and Faragher,R.G. (2008) Cell divisions and mammalian aging: integrative biology insights from genes that regulate longevity. *Bioessays*, **30**, 567–578.

de Magalhaes,J.P. and Toussaint,O. (2004) How bioinformatics can help reverse engineer human aging. *Ageing Res. Rev.*, **3**, 125–141.

Dennis,G. Jr. *et al*. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Dumont,P. *et al*. (2002) Overexpression of apolipoprotein J in human fibroblasts protects against cytotoxicity and premature senescence induced by ethanol and tert-butylhydroperoxide. *Cell Stress Chaperones*, **7**, 23–35.

Edwards,M.G. *et al*. (2007) Gene expression profiling of aging reveals activation of a p53-mediated transcriptional program. *BMC Genomics*, **8**, 80.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Hong,F. and Breitling,R. (2008) A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, **24**, 374–382.

Ida,H. *et al*. (2003) Age-related changes in the transcriptional profile of mouse RPE/choroid. *Physiol. Genomics*, **15**, 258–262.

Kalman,J. *et al*. (2000) Apolipoprotein D in the aging brain and in Alzheimer's dementia. *Neurol. Res.*, **22**, 330–336.

Kriete,A. (2006) Biomarkers of aging: combinatorial or systems model? *Sci. Aging Knowledge Environ.*, **2006**, pe1.

Kuilman,T. *et al*. (2008) Oncogene-induced senescence relayed by an interleukin-dependent inflammatory network. *Cell*, **133**, 1019–1031.

Kumazaki,T. *et al*. (1991) Fibronectin expression increases during in vitro cellular senescence: correlation with increased cell area. *Exp. Cell Res.*, **195**, 13–19.

Kumita,J.R. *et al*. (2007) The extracellular chaperone clusterin potently inhibits human lysozyme amyloid formation by interacting with prefibrillar species. *J. Mol. Biol.*, **369**, 157–167.

Lu,T. *et al*. (2004) Gene regulation and DNA damage in the ageing human brain. *Nature*, **429**, 883–891.

Mari,D. *et al*. (1995) Hypercoagulability in centenarians: the paradox of successful aging. *Blood*, **85**, 3144–3149.

McCarroll,S.A. *et al*. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.*, **36**, 197–204.

McElwee,J.J. *et al*. (2007) Evolutionary conservation of regulated longevity assurance mechanisms. *Genome Biol.*, **8**, R132.

Moreau,Y. *et al*. (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.*, **19**, 570–577.

Muffat,J. *et al*. (2008) Human ApoD, an apolipoprotein up-regulated in neurodegenerative diseases, extends lifespan and increases stress resistance in Drosophila. *Proc. Natl Acad. Sci. USA*, **105**, 7088–7093.

Navarro,A. *et al*. (2002) Behavioral dysfunction, brain oxidative stress, and impaired mitochondrial electron transfer in aging mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, **282**, R985–R992.

Pan,F. *et al*. (2007) Gene Aging Nexus: a web database and data mining platform for microarray data on aging. *Nucleic Acids Res.*, **35**, D756–D759.

Provost,P.R. *et al*. (1991) Apolipoprotein D transcription occurs specifically in nonproliferating quiescent and senescent fibroblast cultures. *FEBS Lett.*, **290**, 139–141.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramasamy,A. *et al*. (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.*, **5**, e184.

Rhodes,D.R. *et al*. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, **62**, 4427–4433.

Rhodes,D.R. *et al*. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA*, **101**, 9309–9314.

Rodwell,G.E. *et al*. (2004) A transcriptional profile of aging in the human kidney. *PLoS Biol.*, **2**, e427.

Siritantikorn,A. *et al*. (2007) Protection of cells from oxidative stress by microsomal glutathione transferase 1. *Biochem. Biophys. Res. Commun.*, **355**, 592–596.

Slonim,D.K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.*, **32**(Suppl.), 502–508.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Swindell,W.R. (2008) Comparative analysis of microarray data identifies common responses to caloric restriction among mouse tissues. *Mech. Ageing Dev.*, **129**, 138–153.

Trounce,I. *et al*. (1989) Decline in skeletal muscle mitochondrial respiratory chain function: possible factor in ageing. *Lancet*, **1**, 637–639.

Uitto,J. (1986) Connective tissue biochemistry of the aging dermis. Age-related alterations in collagen and elastin. *Dermatol. Clin.*, **4**, 433–446.

Van Remmen,H. *et al*. (1995) Gene expression and protein degradation. In Masoro,E.J. (ed.) *Handbook of Physiology. A Critical, Comprehensive Presentation of Physiological Knowledge and Concepts, Section 11: Aging*. Oxford University Press, London, pp. 171–234.

Verducci,J.S. *et al*. (2006) Microarray analysis of gene expression: considerations in data mining and statistical treatment. *Physiol. Genomics*, **25**, 355–363.

Weindruch,R. *et al*. (2002) Gene expression profiling of aging using DNA microarrays. *Mech. Ageing Dev.*, **123**, 177–193.

Welle,S. *et al*. (2003) Gene expression profile of aging in human muscle. *Physiol. Genomics*, **14**, 149–159.

Welle,S. *et al*. (2004) Skeletal muscle gene expression profiles in 20-29 year old and 65-71 year old women. *Exp. Gerontol.*, **39**, 369–377.

Wheeler,D.L. *et al*. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.

Zahn,J.M. *et al*. (2007) AGEMAP: a gene expression database for aging in mice. *PLoS Genet.*, **3**, e201.