An Entropy-based Statistical Workflow Provides Noise-Minimizing Biological Annotation for Muscular Aging

Theodoros Koutsandreas, Ioannis Valavanis, Eleftherios Pilalis, Aristotelis Chatziioannou*
Institute of Biology, Medicinal Chemistry and Biotechnology
National Hellenic Research Foundation (NHRF)
Athens, Greece
* Corresponding author (email: achatzi@eie.gr)

***Abstract*— This study aims to expand the efficiency of the interpretation concerning the aging process, by exploring a broad gene set, derived from the analysis of an integrative transcriptomic microarray dataset. The dataset comprises human skeletal muscle samples, obtained from healthy males and females, that were used to derive a gene signature of a high informative content, with respect to its functional association with the aging phenotype. Towards this end, a multilayered computational workflow integrating advanced statistical methodologies for the derivation of reliable confidence measures, distribution-based entropy calculations to examine the informational content of the dataset, enrichment analysis, graph-theoretic methods and intuitive visualization was applied. Specifically, statistical testing revealed differentially expressed genes, while an uncertainty calculation algorithm, exploiting Gene Ontology (GO) terms annotations, extended the list of significant genes from 254 to 2791, namely p-value threshold was increased from 0.0005 to 0.103, while keeping simultaneously noise measurements legitimately low. This rich gene set associated functionally the macroscopic phenotype of muscular aging with highly informative, stably correlated with each other, molecular annotations in the GO database. Finally, a set of 57 reliable genes was identified that comprise a gender-independent aging signature, after incorporating crucial information about genes pivotal regulatory role as inferred by the GO tree. The biological interpretation was highly assisted by the illustration of the functional mappings between genes, cellular location and biological processes through circle packing graphs.**

Keywords—aging; muscle; genes; entropy; enrichment analysis; gene ontology; functional annotation; visualization

I. INTRODUCTION

Aging is a complex biological process, whose hallmarks are functional capacity reduction, disturbances in cellular homeostasis, immune system and exacerbated metabolic decline among others. It is strongly correlated with multiple diseases, such as Alzheimer's disease, atherosclerosis, diabetes mellitus, sarcopenia and osteoporosis [1]. Aging, influenced by a variety of intrinsic and extrinsic factors, has been associated with impairments in genomic or proteomic level [2], however it is not fully elucidated how these changes affect the molecular pathways [3]. Skeletal muscle aging comprises a significant issue for aging process, since muscle accounts for approximately half of the cell mass of the human body, and sarcopenia is a key feature of age-related frailty [4], [5]. Age-related changes in skeletal muscle appear to be influenced by sex, however controversy exists regarding how sex influences each aspect of the aging process of skeletal muscle [6].

DNA microarrays have been exploited in several studies in order to reveal distinct molecular patterns declarative of aging in various tissues, including human skeletal muscle. Specifically, DNA microarrays has been used to study primarily sex-related differences in gene expression in human skeletal muscle [6]–[8], while few studies have focused on gender independent profiling [9]. This can be due to that men and women differ in hormone levels and certain muscle characteristics [10], [11], so that it cannot be assumed that they have identical age-related changes in muscle [12].

In this present study, we combined transcriptional profiling data of human skeletal muscle from young and elderly, male and female subjects. The dataset used here comes from different, publicly available microarray data experiments. Data pre-processing was performed using the statistical programming language R and Bioconductor, which provide various computational tools for the comprehensive statistical analysis of the relevant high-throughput genomic data [13].

As a first step, a significant list of differentially expressed genes between young and old subjects in the unified dataset including both males and females, was derived through the application of the popular statistical t-test. We then applied an integrated novel workflow, programmed in Python, that aims to partition the p-value sorted, cumulative statistical distribution of the genes comprising the dataset. Based on this distribution, we propose selecting a broad subset of genes that estimates and minimizes noise by exploiting gene mappings into concrete functional terms with adequate gene membership, describing the underlying biological process in aging, as opposed to promiscuous functional associations with low membership.

In order to shape the broad, representative gene list related to aging phenotype as described above, we exploited gene-function mapping performed by the Gene Ontology (GO) terms, which are derived through statistical enrichment analysis.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

Enrichment analysis was performed here using the StRAnGER [14] functional analysis web application. To derive this representative, yet highly informative (with noise minimization) gene list, we propose instead of the typical rigid confidence limits, a threshold that depends on the cumulative p-value distribution of the genes pertaining the whole dataset. This process is in line with the conformal prediction framework, aiming to derive reliable confidence limits for the independent gene predictions. Shortly, we propose mapping incremental sets of genes into GO terms, and partitioning the corresponding functional terms into informative, when these possess a substantial gene membership (>= 3 genes) or potentially noisy ones (< 3 genes). In this iterative process, we estimate the information content of the functional annotation terms sets and target to minimize the Shannon entropy of the selected gene set, thus finalizing the gene selection process. In information theory, the Shannon entropy is a measure of the necessary information in order to predict the value of a random variable. Consequently, it could serve as a measure of uncertainty that a variable represents a causal (or random (erroneous or passenger) event. In line with that, we use the Shannon entropy to calculate the uncertainty that the examined gene set is selectively enriched or not, for functional terms, that are considered informative for the aging phenotype. The Shannon entropy has been already introduced as a differential metric to detect isolated disease-related genes with mild differential expression [15], excluding however from the analytical perspective, the biological information content of gene sets. In another study the Shannon entropy is used to calculate the information content derived from an ontological structure, without the extraction of important biological information from a genes dataset [16]. In this sense, it is the first time to the best of our knowledge where the concept of entropy is applied in conjunction with that of semantic topology, and specifically the calculation of the graph compactness, as an optimization problem, solved numerically through an iterative computational framework. In order to enhance the robust biomarker discovery process concerning skeletal muscle aging, we used the GORevenge [17] application that can reveal hidden functional regulatory effects among genes by establishing it onto a system's level interpretation. Finally, the results are vividly illustrated through an intuitive visualization scheme, designed and developed in JavaScript language.

II. DATASET

Microarray gene expression data from vastus lateralis biopsies obtained from healthy young (20-29 years old) and old (65-75 years old), male and female subjects, was used here. The dataset comes from different, MIAME-compliant [18] experiments that are publicly available at the Gene Expression Omnibus (GEO) database with accession numbers GDS287, GDS288, GDS472 and GDS473 (see [4], [12] for more detailed information regarding these hybridizations). The dataset encompassed 30 samples in total: 14 from young male (NYm=7) and female (NYf=7), and 16 from old male (NOm=8) and female (NOf=8) subjects. All experiments were used the Affymetrix® (Santa Clara, CA) Human Genome U133A and U133B oligonucleotide arrays (HG-U133 Set). The HG-U133 Set has about 44,000 probe sets that measure the expression of about 33,000 genes.

III. METHODS

A. Data Pre-Processing

Missing value imputation was applied directly to the available normalized data using the nearest neighbor averaging [19]. Expression values were then filtered out based on a threshold fraction of the present detection calls (derived by the Affymetrix' s MAS5 algorithm), hence increasing the ratio of true positives to false positives [20]. We filtered out probes characterized as not Present by the MAS5 detection call in at least 60% of the samples in at least one group. Each expression value was divided by the mean of all expression values in each array series, in order to be comparable between the experiments. In the unified dataset, including both young or old males and females, a total of 11256 probe IDs were kept after the procedure described above.

B. Genes Selection based on statistics, entropy calculations and GO terms

The statistical t-test was used to assess gene expression differentiation between young and old samples. Traditionally, the significant genes selection is realized based on an arbitrary p-value threshold in the range [0.001...0.05] ([9],[21]-[23]). Although this range of p-value thresholds restricts the selection of random genes, it may obstruct the disclosure of important biological information extracted by the enrichment analysis that follows gene selection in such studies. Moreover even a strict set of differentially expressed genes may fail addressing the issue of taming the impact of biological noise, namely irrelevant functional annotations for a complex biological phenotype like aging in this study. With the view to overcome the contingent restriction of information due to the effect of arbitrary rigid thresholds, an entropy-based algorithm is introduced here, which tries to validate the new biological data, which are proposed by the enrichment analysis using GO terms or potentially other gene nomenclatures.

Initially, the algorithm executes a statistical enrichment analysis through StRAnGER web application, for a highly confident gene set, corresponding to differentially expressed genes at a very strict statistical threshold (e.g. $p < 0.0005$). This set is able to produce an informative collection of GO terms, with high confidence (minimal false positive risk). The specific GO terms collection is then expanded to embrace their parent terms, by exploiting the GO linkage tree. As our goal is to enrich this collection of functional annotations with new significant GO terms, the algorithm performs an iterative process, which applies the same procedure to incrementally formed larger gene sets, which correspond to milder p-value thresholds, up to a maximal gene set corresponding to the point that the information content, is maximized, thus the measurement of the Shannon entropy reaches a local minimum.

Specifically, in each round, the gene list is growing by a fixed amount of probe IDs, corresponding to a possibly variable number of genes. New GO terms encountered are initially considered as potential noise, if their enrichment score in StRAnGER analysis is lower than the value of the genes added. However, if the enrichment score of those terms increases in subsequent steps, as the iterative process advances further to form larger sets, and the respective GO term remains

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

significant according to the applied hypergeometric test in StRAnGER results, then automatically this term is consider to contribute additional biological information, rather than noise. Taking advantage of enrichment analysis, we can calculate the noisy content for each gene. A noisiness measure for a gene is the ratio of noisy GO terms, out of the GO terms annotated to this gene. In the rare case where a gene is not annotated to any GO term, then the noise weight is equal to 1, as it is considered noisy from the functional point of view, in the sense that its role has not been studied thus far and has not been linked to any biological process. The gene set emerging in each round is finally evaluated in terms of uncertainty (bits) using Shannon Entropy [24]. This measurement considers the probabilities of the two extreme hypotheses, noisy vs. representative set of genes, and calculates the uncertainty that the gene set could be representative of the biological information beneath the investigated phenotype. The iterative procedure executes further up to propose a maximal gene set, corresponding to a local minimum of uncertainty. This is the one considered the best choice in terms of functional information content related to the phenotype studied.

Overall, the procedure is able to monitor the incoming quanta of biological information incorporated at every incremental step, validate the effect of noise, and finally expand the significant genes list, up to the point where the information content of the dataset is sullied by the detrimental impact of unreliable functional terms, clearly suggesting a loss of the homeostasis of the dataset.

C. Hub genes selection

The GORevenge algorithm, freely available through the web, was applied following the gene selection process described above. GORevenge uses graph-theoretic methods to exploit the GO tree in order to sort related genes/GO terms according to their regulatory impact, as this is derived from the degree of connectivity that each gene/ term possesses with other terms. Thus, it can aid the elucidation of hidden functional regulatory effects among genes and a system's level interpretation. GORevenge was applied separately once for the "Molecular Function" (MF), once for the "Biological Process" (BP) aspect and once for the "Cellular Component" (CC) aspect using the same parameters (the Resnik semantic similarity metric, the Bubble genes algorithm, and a relaxation equal to 0.15). Out of the GORevenge's output list with the most important genes, only those that were simultaneously contained in the original gene list submitted to the method (the one proposed by the gene selection process in previous sub-section) were selected as gene hubs and were transferred to the visualization module next.

D. Visualization

In order to illustrate gene-function relationships, exploiting GO mapping, between cellular components and molecular processes, a visualization module has been applied that generates scalable vector graphics (svg) image files, based on JavaScript D3 visualization library. This implementation creates a svg graph for each significant GO term, as derived from enrichment analysis. Specifically, a circle packing graph is constructed, being a cross section, where the major circle corresponds to the significant GO term and the interior circles belong to other significant, GO terms, from different GO categories, with common content in genes. In addition, the genes displayed in the internal circles, integrate information about their individual expression with green color for the downregulated and red colour for the overexpressed. In this way, the interpretation is visually assisted, so as to aid the elucidation of complex biological procedures, where a number of metabolic or signaling pathways and cellular components are involved.

IV. Results and Discussion

A. Strict gene list selection and enrichment analysis

Starting from a strict list of 354 probe IDs ($p < 0.0005$), which corresponds to 289 unique genes, a functional analysis with StRanGER was executed. The enrichment analysis links the individual genes with specific biological processes, molecular functions and cellular components. The output comprised a list of 50 significant over-represented GO terms (hypergeometric test $p \leq 0.05$, 90% cut-off percentage, 104 bootstrap iterations). The respective results for "Biological Process" terms are presented in Table I. The majority of significant GO terms corresponds to metabolic processes, related to energy production, involvement of mitochondrial compartments and regulation of gene expression. All these terms possess an established role and have been notably related to aging in previous studies, as described in the following.

Terms like small molecule metabolic process, mitochondrial electron transport, regulation of acetyl-CoA biosynthetic process from pyruvate, and tricarboxylic acid cycle are significant biological processes implicated to aging-related pathways as already reported in [9],[27],[28]. Oxidative capacity, ATP synthesis coupled proton transport and ATP catabolic process are affected by the decline in ATP concentration of aged muscle tissues [25],[26]. Aged tissues suffer from oxidative stress which provides the development of stress granules [29]. Also, the investigated functional decline of aged skeletal muscles explains the significance of terms like muscle filament sliding [28], while blood vessel development is impaired in aged tissues [30]. Concerning to cellular compartments, mitochondrion related genes reduce their expression during the aging process [22], while cytosolic pathways are significant for the degradation of abnormal proteins that are accumulated in aged cellular compartments [31].

B. Extended significant genes list and additional biological annotation

In order to enrich the above biological information and derive a representative set of genes for the muscular tissue aging phenotype, the incremental gene selection process was applied exploiting StRAnGER iteratively, to the ever increasing gene sets. With a gradual increase of 40 probe IDs per set, a total of 191 incrementally formed gene lists were analyzed. Entropy calculation results corresponding to the uncertainty (bits) are presented in Fig. 1, where the respective p-value threshold trend is also plotted. The uncertainty varies from 0.44 to 0.611 bits, oscillating irregularly within these

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

limits, between 354 to 2994 probes and finally presenting a local minimum observed at 3714 probes. Afterwards, there is a remarkable gradual entropy increase parallel to the increase of the number of probes, for the gene lists incrementally formed.

TABLE I

GO Term	Term Description	*p*-value
GO:0044281	small molecule metabolic process	1,43687E-06
GO:0022904	respiratory electron transport chain	3,63629E-06
GO:0044237	cellular metabolic process	4,51712E-06
GO:0006099	tricarboxylic acid cycle	4,9405E-06
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	5,39363E-06
GO:0015986	ATP synthesis coupled proton transport	8,37261E-06
GO:0010510	regulation of acetyl-CoA biosynthetic process from pyruvate	8,7867E-06
GO:0006412	translation	9,76222E-06
GO:0006090	pyruvate metabolic process	1,01334E-05
GO:0042776	mitochondrial ATP synthesis coupled proton transport	1,09668E-05
GO:0055114	oxidation reduction	1,17252E-05
GO:0010467	gene expression	1,30704E-05
GO:0034063	stress granule assembly	1,35473E-05
GO:0008053	mitochondrial fusion	1,39152E-05
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	1,52052E-05
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	1,55798E-05
GO:0006915	apoptosis	1,65811E-05
GO:0006200	ATP catabolic process	1,78186E-05
GO:0045454	cell redox homeostasis	1,84699E-05
GO:0030049	muscle filament sliding	1,92991E-05

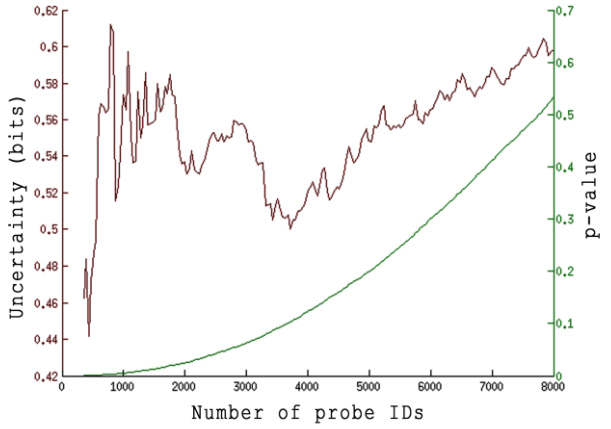

Fig. 1 Entropy (uncertainty) and *p*-value fluctuation as a function of number of probes corresponding to gene lists incrementally formed

Theoretically, the uncertainty graph should contain two different fractions; the initial oscillations, due to the incorporation of new biological information, which initially due to the adoption of stringent thresholds, is considered informative, but as its enrichment increases is accepted as such, and the subsequent, practically monotonic, gradual increase of uncertainty, caused by the introduction of noisy genes, which do not map to informative terms. The significant gene list selection optimal point separating these two fractions, where the accumulation of informative GO terms is minimized and practically from this point on, we observe accumulation of genes not conferring any new, topologically coherent, phenotypic terms. In the specific analysis, this optimal point is more clearly observed as a local minimum, due to the great accumulation of genes into informative terms and a consequent uncertainty reduction, from 2994 to 3714 probe IDs. Naturally, the estimation of the local minimum, is largely relying, to numerous algorithmic settings, as the step of the algorithm, the stringency of the statistical enrichment threshold, the dimensionality of the dataset, among others. , Fine-tuning the performance of the algorithm is an optimization problem for each dataset, but in general the derivation of a global critical point is attainable.

The noisy genes ratio as a function of the number of probes is presented in the bar graph of Fig 2. The blue bars section corresponds to the ratio of genes not mapped to any of the GO terms (briefly this is defined as ratio1) qualified through StRAnGER algorithm, at the given step, while the red bars section (defined as ratio2), is the ratio of genes mapped to qualified GO terms,. The total noisy genes ratio is maximized at 784 probe IDs corresponding to a $p < 0.003$ threshold. At this point the high ratio2 value (=0.059) indicates the integration of multiple new GO terms, which are initially characterized as noise, but in reality they represent important biological information, rather than noise, as it can be surmised from the very low, corrected, enrichment p-value scores. The iterative process validates this assumption as proven by the gradual ratio2 decline in larger gene lists, which means that the noisy GO terms percentage is drastically reduced. On the other hand, the ratio1 is steadily increasing after the point of 3714 probe IDs, suggesting that after this point, at a large amount of promiscuous, unannotated genes is infiltrating. Thus, these genes seem statistically irrelevant with the aging phenotype or are at least represent genes that have not meticulously been studied, therefore elude functional annotation in the Gene Ontology. Therefore, despite of the continuing ratio2 reduction, after this point, we used as cutoff for our significant gene list the point of 3714 probes.

This probe IDs set, which corresponds to 2791 unique genes and $p<0.103$ is selected as the aging significant and informative genes list. The additional biological annotation data, significant for the aging phenotype, which is not correlated with the initial list of GO terms, consists of 53 new GO terms. Out of these, 23 incoming terms correspond to "Biological Process" and are presented in Table II.

Additional GO terms include processes related to protein activation and degradation, viral-host interactions and DNA repair. It is known that an aging muscle has a reduced capacity to synthesize new proteins [32]. Concerning protein catabolic process, a related study mentions that 20S proteasome proteolytic activity is declined in the aged rats muscle tissues [33]. Also, it is proven that all pathways of DNA repair lose their efficiency with age advancement, due to accumulation of loss of function mutations [34]. Wnt signaling pathway is correlated with the signal transduction between the extracellular space and the cell nucleus. Results from the experiments on Caenorhabditis elegans suggest that Wnt signaling regulates aging-intrinsic genetic pathways [35]. Our

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

results corroborate the necessity of Wnt signaling pathway analysis in human aged tissues.

TABLE II

GO Term	Term Description	*p*-value
GO:0006281	DNA repair	1,39404E-07
GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	4,01265E-07
GO:0016032	viral process	5,04072E-07
GO:0019048	modulation by virus of host morphology or physiology	7,84455E-07
GO:0016055	Wnt signaling pathway	1,44435E-06
GO:0000278	mitotic cell cycle	1,54752E-06
GO:0006366	transcription from RNA polymerase II promoter	1,65519E-06
GO:0044255	cellular lipid metabolic process	1,67942E-06
GO:0016567	protein ubiquitination	1,87219E-06
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	1,89895E-06
GO:0006468	protein phosphorylation	1,9476E-06
GO:0006511	ubiquitin-dependent protein catabolic process	2,00238E-06
GO:0000082	G1/S transition of mitotic cell cycle	2,67007E-06
GO:0006470	protein dephosphorylation	2,74105E-06
GO:0031124	mRNA 3'-end processing	2,82219E-06
GO:0006119	oxidative phosphorylation	2,97111E-06
GO:0000380	alternative mRNA splicing, via spliceosome	3,2539E-06
GO:0006376	mRNA splice site selection	3,50225E-06
GO:0006369	termination of RNA polymerase II transcription	4,37063E-06
GO:0000288	nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay	4,38441E-06
GO:0045727	positive regulation of translation	5,40528E-06
GO:0006406	mRNA export from nucleus	6,24915E-06

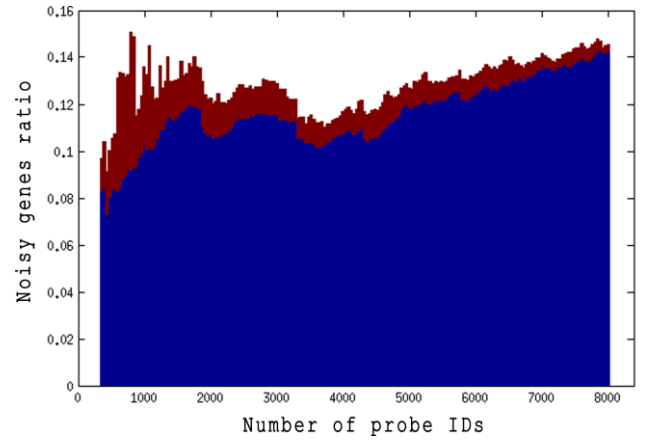

Fig. 2 Noisy genes ratio – blue bar for unannotated genes, red bar for annotated with noisy GO terms genes.

C. Hub genes selected

In the direction of identifying the most critical genes from the extended list of 2791, the GORevenge platform was utilized. An amount of 113 hub genes resulted for both "Biological Processes" and "Molecular Functions". When applying the criterion of existence in the "Cellular Components", the aforementioned hubs set are reduced to 57 significant genes. Only 8 hubs are included in the initial strict gene list of 354 probe IDs with $p < 0.0005$; the vast majority of hubs belong to larger p-values. This set is presented in Table III (the above 8 hubs are illustrated with red colour).

D. Hub genes visualization, Up-down regulation in aging phenotype

With the view to elucidate hub genes activations into specific cellular components and molecular procedures in aging, the visualization method was applied. We chose to illustrate here the visualization of common hub genes activated in the mitochondrial interrmembrane space (randomly chosen), and the correlated, based on common genes activation, biological processes (RNA processing, negative regulation of apoptotic process, oxidation reduction process and small molecules metabolic process). Specifically, the activation of two common differentially expressed genes, CAT and PNPT1, is presented in Fig. 3 and Fig. 4 for young and aged tissues, respectively. PNPT1 takes part only in RNA processing, while its expression declines with age (red in Fig. 3, green in Fig. 4). On the other hand, CAT is activated in negative regulation of apoptotic process, oxidation reduction process and small molecules metabolic process and is up-regulated in aged tissues (green in Fig. 3, red in Fig. 4).

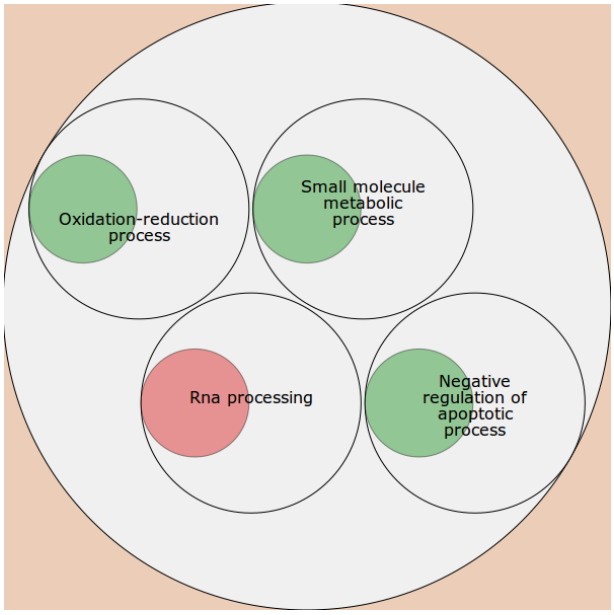

Fig. 3 Young tissue: Up-regulation of PNPT1 in RNA processing (red) and down-regulation (green) of CAT (negative regulation of apoptotic process, oxidation reduction process, small molecules metabolic process).

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

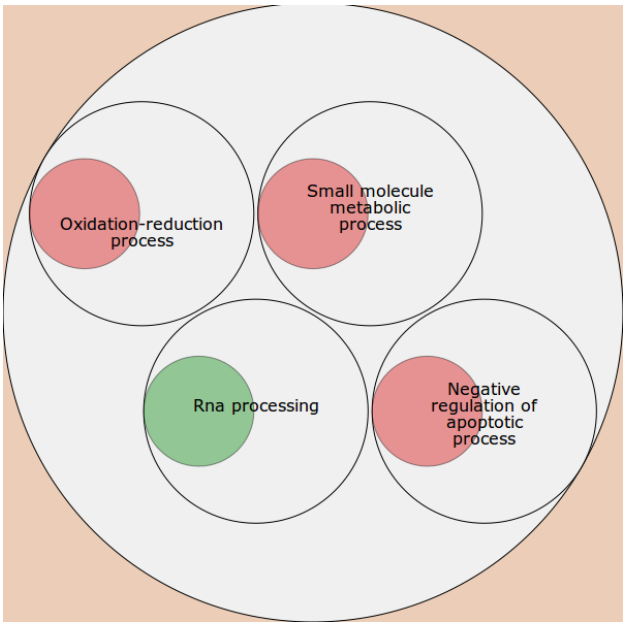

Fig. 4 Old aged tissue: Down-regulation of PNPT1 in RNA processing (green) and up-regulation (red) of CAT (negative regulation of apoptotic process, oxidation reduction process, small molecules metabolic process).

V. Conclusion

In the present work, a novel, highly generic, data-driven, integrated analysis workflow combining statistical methods, entropy calculation, functional analysis and visualization is proposed, in order to identify a gender-independent significant set of genes underlying the aging process. Starting from a transcriptional profiling dataset from different vastus literalis biopsies from young and elderly, male and female objects, we identified a broad set of 2791 genes that correspond to an informative and minimum noisiness list of GO terms in line with aging phenotype. The ultimate list, with potentially high biomarker value for muscular aging, consists of 57 differentially expressed genes that are derived from the aforementioned broad set. These were found to play a central regulatory role in the underlying molecular processes and cellular components, as revealed by the GO tree linkage analysis. Finally, circle packing graphs were used to illustrate the association of important genes for specific molecular processes and cellular compartments. To the best of our knowledge, it is the first time that this type of iterative, automated, relieved from statistical cutoffs, information-based approach is applied for the standardized, processing and interpretation of a voluminous, integrative transcriptomic dataset. The approach manages to exploit the whole dataset by inspecting it overall and partitioning it, to the potentially causal part and the non-informative one for the study of the given phenotype. This is accomplished through a multi-step, procedure. This manages to combine reliable statistical properties and strategies, and information theory (semantic networks and estimation of entropy) in order to extend the efficiency of the interpretation by revealing critical patterns of informational organization that characterize the distribution of values of the dataset.

TABLE III

Gene ID	GOcount [a]	Prune[3] [b]	Prune[6] [b]	Prune[9] [b]
VEGFA	110	82	82	82
WNT5A	108	101	101	101
PTEN	93	80	79	79
SIRT1	89	76	75	75
EGFR	81	72	72	72
SMAD3	74	69	69	69
MAPK1	68	61	61	61
CAV3	58	49	49	49
GSK3B	57	54	54	54
CD36	55	50	50	50
PML	54	50	50	50
TCF7L2	53	49	49	49
SMAD4	52	47	47	47
PTK2	49	45	45	45
TP63	47	44	44	44
MDM2	47	43	43	43
NRP1	44	40	40	40
AQP1	44	36	35	35
PTCH1	43	39	39	39
STAT1	42	35	35	35
ANK2	42	32	31	31
TPR	39	31	31	30
PAFAH1B1	39	37	37	37
INPP5K	39	32	31	31
SMARCA4	37	32	32	32
SKI	37	36	36	36
ERBB4	36	31	31	31
DLG1	36	35	35	35
DAB2	36	34	34	34
RAPGEF2	35	29	29	29
TGFBR3	33	31	31	31
DMD	33	31	31	31
HDAC4	31	27	27	27
GRB2	29	29	29	29
CAPN3	28	25	25	25
C1QBP	28	26	26	26
HSPD1	27	24	23	23
NOS1	26	26	26	26
CTNNA1	26	23	23	23
CAT	25	21	21	21
OGT	24	23	22	22
DNAJA3	24	24	24	24
CREBBP	24	19	18	18
PCNA	23	20	20	20
NCOR1	23	21	21	21
PAM	22	21	21	21
FLNA	21	21	21	21
XRCC5	20	17	17	17
VCP	20	13	13	13
SNW1	20	17	17	17
ADAM10	20	17	17	17
XRCC6	19	15	15	15
TOPORS	19	16	16	16
SQSTM1	19	17	17	17
KCNJ11	19	17	17	17
HSP90AA1	18	15	13	13
DAG1	18	17	17	17

[a] Refers to the number of original GO terms. [b] Refer to number of GO terms remaining after GOrevenge pruning, reflecting the centrality of each gene

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

ACKNOWLEDGMENTS

Authors would like to thank E. Sifakis for his prior participation to the pre-process of aging microarray data.

This work was co-funded by i) the Bilateral Greece-China Research Program of the Hellenic General Secretariat of Research and Technology and the Chinese Ministry of Research and Technology, PROMISE ("Personalization of melanoma therapeutic management through the fusion of systems biology and intelligent data mining methodologies"), ii) THALES Project entitled "Development of Systems Biology and Bioinformatics tools to study the dynamics of cell aging" (MAESTRO), co-financed by the European Union (European Social Fund) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework, iii) the two Entrepreneurial Programs "Competitiveness and Entrepreneurship", Action COOPERATION, entitled "PIK3CA Oncogenic Mutations in Breast and Colon Cancers: Development of Targeted Anticancer Drugs and Diagnostics" (POM) and "Development of novel Angiogenesis-Modulating Pharmaceuticals by screening of natural compounds and synthetic analogues" (DAMP).

REFERENCES

[1] R. D. Kovaiou, D. Herndler-Brandstetter, B. Grubeck-Loebenstein, "Age-related changes in immunity: implications for vaccination in the elderly", Expert Rev Mol Med, vol. 9(3), pp. 1-17, February 2007.

[2] T. B. L. Kirkwood, "Understanding the odd science of aging.," Cell, vol. 120, no. 4, pp. 437–47, Mar. 2005.

[3] J. Vijg and J. Campisi, "Puzzles, promises and a cure for ageing.," Nature, vol. 454, no. 7208, pp. 1065–1071, 2008.

[4] S. Welle, A. I. Brooks, J. M. Delehanty, N. Needler, and C. a Thornton, "Gene expression profile of aging in human muscle.," Physiological genomics, vol. 14, no. 2, pp. 149–59, Jul. 2003.

[5] K. S. Nair, "Aging muscle.," The American journal of clinical nutrition, vol. 81, no. 5, pp. 953–63, May 2005.

[6] D. Liu, M. a Sartor, G. a Nader, E. E. Pistilli, L. Tanton, C. Lilly, L. Gutmann, H. B. Iglayreger, P. S. Visich, E. P. Hoffman, and P. M. Gordon, "Microarray Analysis Reveals Novel Features of the Muscle Aging Process in Men and Women.," The journals of gerontology. Series A, Biological sciences and medical sciences, no. 14, pp. 1–10, Mar. 2013.

[7] S. Welle, R. Tawil, and C. a Thornton, "Sex-related differences in gene expression in human skeletal muscle.," PloS One, vol. 3, no. 1, p. e1385, Jan. 2008.

[8] D. Liu, M. a Sartor, G. a Nader, L. Gutmann, M. K. Treutelaar, E. E. Pistilli, H. B. Iglayreger, C. F. Burant, E. P. Hoffman, and P. M. Gordon, "Skeletal muscle gene expression in response to resistance exercise: sex specific regulation.," BMC genomics, vol. 11, no. 1, p. 659, Jan. 2010.

[9] J. M. Zahn, R. Sonu, H. Vogel, E. Crane, K. Mazan-Mamczarz, R. Rabkin, R. W. Davis, K. G. Becker, A. B. Owen, and S. K. Kim, "Transcriptional Profiling of Aging in Human Muscle Reveals a Common Aging Signature," PLoS Genetics, vol. 2, no. 7, p. e115, 2006.

[10] J. A. Simoneau and C. Bouchard, "Human variation in skeletal muscle fiber-type proportion and enzyme activities," American Journal of Physiology Endocrinology And Metabolism, vol. 257, no. 4, pp. E567–E572, 1989.

[11] R. S. Staron, F. C. Hagerman, R. S. Hikida, T. F. Murray, D. P. Hostler, M. T. Crill, K. E. Ragg, and K. Toma, "Fiber Type Composition of the Vastus Lateralis Muscle of Young Men and Women," Journal of Histochemistry & Cytochemistry, vol. 48, no. 5, pp. 623–629, May 2000.

[12] S. Welle, A. I. Brooks, J. M. Delehanty, N. Needler, K. Bhatt, B. Shah, and C. a Thornton, "Skeletal muscle gene expression profiles in 20-29 year old and 65-71 year old women.," Experimental gerontology, vol. 39, no. 3, pp. 369–77, Mar. 2004.

[13] R. Doerge, "Bioinformatics and Computational Biology Solutions Using R and Bioconductor," Biometrics, vol. 62, no. 4, pp. 1270–1271, 2006.

[14] A. A. Chatziioannou and P. Moulos, "Exploiting Statistical Methodologies and Controlled Vocabularies for Prioritized Functional Analysis of Genomic Experiments: the StRAnGER Web Application," Frontiers in neuroscience, vol. 5, no. January, p. 14, 2011.

[15] K. Wang, C. A. Phillips, G. L. Rogers, F. Barrenas, M. Benson, M. A. Langston, " Differential Shannon entropy and differential coefficient of variation: alternatives and augmentations to differential expression in the search for disease-related genes", Int J Comput Biol Drug Des, vol. 7, no. 0, pp. 183-194, 2014.

[16] A. Warren, J. C. Setubal, "Using Entropy Estimates for DAG-Based Ontologies", Bio-Ontologies, October 2012

[17] K. Moutselos, I. Maglogiannis, and A. Chatziioannou, "GOrevenge: A novel generic reverse engineering method for the identification of critical molecular players, through the use of ontologies.," IEEE Transactions on Biomedical Engineering, no. c, 2011.

[18] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. et. al., "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," Nature genetics, vol. 29, no. 4, pp. 365–371, 2001.

[19] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," Bioinformatics (Oxford, England), vol. 17, no. 6, pp. 520–525, 2001.

[20] J. N. McClintick and H. J. Edenberg, "Effects of filtering by Present call on analysis of microarray experiments.," BMC Bioinformatics, vol. 7, p. 49, Jan. 2006

[21] D. Glass, A. Viñuela, M. N. Davies, A. Ramasamy, L. Parts, et. al. "Gene expression changes with age in skin, adipose tissue, blood and brain", Genome Biology, vol. 14(7): R75, July 2013.

[22] J. P. De Magalhães , J. Curado and G. M. Church, "Meta-analysis of age-related gene expression profiles identifies common signatures of aging" Bioinformatics, vol 25(7), pp. 875-881, January 2009.

[23] G. E. Rodwell, R. Sonu, J. M. Zahn, J. Lund, J. Wilhelmy, L. Wang, W. Xiao, M. Mindrinos, E. Crane, E. Segal, B. D. Myers, J. D. Brooks, R. W. Davis, J. Higgins, A. B. Owen, S. K. Kim, "A transcriptional profile of aging in the human kidney", PLoS Biology, vol. 2(12), pp. 2191-2201, 2004.

[24] C. E. Shannon, "A Mathematical Theory of Communication", Bell System Technical Journal, vol. 27(3), pp. 379-423, 1948.

[25] D. J. Taylor, G. J. Kemp, C. H. Thompson, G. K. Radda, "Ageing: effects on oxidative function of skeletal muscle in vivo", Mol. Cell. Biochem., vol. 174(1-2), pp. 321-324, 1997.

[26] O. Pastoris, F. Boschi, M. Verri, P. Baiardi, G. Felzani, J. Vecchiet, M. Dossena, M. Catapano, "The effect of aging on enzyme activities and metabolite concentrations in skeletal muscle from sedentary male and female subjects." Exp. Gerontol., vol. 35, p.p 95-104, 2000.

[27] A. Navarro, J. M. López-Cepero, and M. J. Sánchez del Pino, "Skeletal muscle and aging.," Frontiers in Bioscience : a journal and virtual library, vol. 6, pp. D26–44, Jan. 2001.

[28] E. Carmeli, R. Coleman, and A. Z. Reznick, "The biochemistry of aging muscle.," Experimental Gerontology, vol. 37, no. 4, pp. 477–489, 2002.

[29] X. J. Lian, I. E. Gallouzi, "Oxidative Stress Increases the Number of Stress Granules in Senescent Cells and Triggers a Parid Decrease in p21waf1/cip1 Translation", The Journal of Biological Chemistry, vol. 284(13), pp. 8877-8887, 2009.

[30] J. M. Edelberg, M. J. Reed, "Aging and Angiogenesis", Front Biocsi., vol. 1(8), pp. 1199-1209, 2003.

[31] P. Johnson, J. L. Hammer, "Cardiac and skeletal muscle enzyme levels in hypertensive and aging rats.", Comp. Biochem. Physiol. Vol. 104(1), pp. 63-67, 1993.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

[32] S. Welle, C. Thornton, M. Statt, “Myofibrillar protein synthesis in young and old human subjects after three months of resistance training”, Am. J. Physiol, vol.268, pp. 422-427.

[33] F. Bardag-Gorce, L. Farout, C. Veyrat-Durebex, Y. Briand, M. Briand, “Changes in 20S proteasome activity during aging of the LOU rat”, Mol. Biol. Rep., vol. 26, pp. 89-93, 1999.

[34] V. Gorbunova, A. Seluanov, Z. Mao, C. Hine, “Changes in DNA repair during aging”, Nucleic Acids Res, vol. 35(22), pp. 7466-7474, December 2007.

[35] M. Lezzerini, Y. Budovskaya, “A dual role of the Wnt signaling pathway during aging in Caenorhabditis elegans”, Aging Cell, vol. 1391), pp, 8-18, February 2014

978-1-4799-7294-4/14/$31.00 ©2014 IEEE