ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

I. Arisi, M. D'Onofrio, R. Brandi, A. Di Mambro, A. Felsani, S. Capsoni, G. Drovandi, G. Felici, E. Weitschek, P. Bertolazzi, A. Cattaneo

LOGIC CLASSIFICATION AND FEATURE SELECTION FROM GENE EXPRESSION PROFILE IN THE BRAIN OF THE AD11 ANTI-NGF MICE MODEL OF ALZHEIMER'S DISEASE AT DIFFERENT STAGES OF NEURODEGENERATION

R. 10-02 2010

I. Arisi – European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, i.arisi@ebri.it.

M. D'Onofrio – European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, mara.donofrio@ebri.it.

R. Brandi – European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, r.brandi@ebri.it.

A. Di Mambro – European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy.

A. Felsani – European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, armando.felsani@ebri.it.

S. Capsoni – European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, s.capsoni@ebri.it.

G. Drovandi – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Roma, Italy, and Università degli studi Roma Tre guido.drovandi@iasi.cnr.it.

G. Felici – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Roma, Italy, giovanni.felici@iasi.cnr.it.

E. Weitschek – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Roma, Italy, and Università degli studi Roma Tre emanuel.weitschek@iasi.cnr.it.

P. Bertolazzi – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Roma, Italy, paola.bertolazzi@iasi.cnr.it.

A. Cattaneo – European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, and Scuola Normale Superiore, Piazza dei Cavalieri, 56126 Pisa, Italy, a.cattaneo@ebri.it.

ISSN: 1128–3378

Collana dei Rapporti
Istituto di Analisi dei Sistemi ed Informatica “Antonio Ruberti”
Consiglio Nazionale delle Ricerche

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

LOGIC CLASSIFICATION AND FEATURE SELECTION FROM GENE EXPRESSION PROFILE IN THE BRAIN OF THE AD11 ANTI-NGF MICE MODEL OF ALZHEIMER'S DISEASE AT DIFFERENT STAGES OF NEURODEGENERATION

I. Arisi, M. D'Onofrio, R. Brandi, A. Di Mambro, A. Felsani, S. Capsoni, G. Drovandi, G. Felici, E. Weitschek, P. Bertolazzi and A. Cattaneo

Ivan Arisi - European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, i.arisi@ebri.it
Mara D'Onofrio - European Brain Research Institute "Rita Levi-Montalcini", 00143 Roma, Italy, mara.donofrio@ebri.it
R. Brandi - European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, r.brandi@ebri.it
A. Di Mambro - European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy
Armando Felsani - European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, armando.felsani@inmm.cnr.it
S. Capsoni - European Brain Research Institute "Rita Levi-Montalcini", Roma, Italy, s.capsoni@ebri.it
Guido Drovandi - Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Roma, Italy, and Università degli studi Roma Tre, guido.drovandi@iasi.cnr.it
Giovanni Felici - Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Roma, Italy, giovanni.felici@iasi.cnr.it
Emanuel Weitschek - Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Roma, Italy, and Università degli studi Roma Tre, emanuel.weitschek@iasi.cnr.it
Paola Bertolazzi - Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Roma, Italy, paola.bertolazzi@iasi.cnr.it
Antonino Cattaneo - European Brain Research Institute "Rita Levi-Montalcini", 00143 Roma, Italy and Scuola Normale Superiore, Piazza dei Cavalieri, 56126 Pisa, Italy, a.cattaneo@ebri.it

Abstract

The gene expression profiles of the AD11 transgenic mouse model of neurodegeneration were investigated, at different time ages, by whole genome microarray analysis (2-color Agilent platform), coupled with ad-hoc logic data mining techniques. The AD11 Alzheimer-like model, expressing an anti-NGF (Nerve Growth Factor) antibody, was obtained by crossing mice with the light chain and mice with the heavy chain (VH) of the antibody. The AD11 mouse model develops a progressive neurodegenerative Alzheimer-like phenotype, including intra- and extra-cellular Aβ peptide accumulation,, intracellular neurofibrillary tangles, major synaptic remodeling and neuronal loss. The gene expression profile of brain regions at an early, presymptimatic, stage of the Alzheimer's like neurodegeneration in the AD11 model was recently characterized by microarray analysis and validated by qRT-PCR of 243 significant candidates. Wide changes in gene expression profiles occur already at 1 month of age. Interestingly, the most significantly affected clusters of mRNAs are linked to inflammation and immune response, mainly genes for complement factors, as well as to Wnt signaling. This gene expression pattern highlights that an early event in AD11 neurodegeneration is represented, together with neurotrophic deficits and synaptic remodeling, by an inflammatory response and an unbalance in the immunotrophic state of the brain. The goal of the present work was to study the gene profiles related to the disease progression and to identify a limited set of genes able to discriminate AD11 and control mice, and their functional relations. To reach this goal, following the progression of the neurodegeneration process, different neuronal areas (basal forebrain, cortex and hippocampus) of the AD11 and control mice were characterized at 1, 3, 6 and 15 months of age by microarray analysis. The data analysis was performed on two main subsets of 60 samples each, 1-3 months and 6-15 months, using both standard statistical methods and more advanced logic data mining techniques, able to efficiently cope with the biological and experimental noise in the system. A small number of "fingerprinting" logic formulas were isolated, encompassing mRNAs whose expression levels were able to discriminate, with a high degree of accuracy, between diseased and control mice. Such a subset was studied by functional analysis using online and manual annotation to characterize the progression of the neurodegenerative disease from the early (1-3 months) to the advanced stage (6-15 months).

1. Introduction

Alzheimer's Disease (AD) is in the majority of cases characterized by a late-onset, progressive neurodegeneration, clinically characterized by short term memory loss and cognitive dementia. A small proportion of AD patients has a genetic basis (early onset familial AD, EOFAD) (Delacourte, 2006), while the vast majority is sporadic (LOAD, or Late Onset AD), and the causes are unknown. No consensus exists, to describe a clear pathway causally linking some identified early cause or events to the final AD phenotypic triad (amyloid pathology, tau tangles and cholinergic deficit), in sporadic AD (Hardy and Selkoe, 2002). Deficits in NGF signaling, transport or processing have been suggested as one possible class of upstream events ultimately leading to LOAD (Cattaneo et al., 2008). Animal models could help dissecting mechanistic aspects of the disease. The majority of animal models for AD presently available is based on individual or multiple transgenic mice expressing mutant forms of human EOFAD genes (Games et al., 2006; Oddo et al., 2003), and is unlikely to be instructive for the early phases of sporadic AD. Only a few models for sporadic AD have been described (Bons et al., 2006): the AD11 anti-NGF model (Capsoni et al., 2000; Ruberti et al., 2000) belongs to this class. AD11 transgenic mice express postnatally a recombinant neutralizing anti-NGF antibody and develop a progressive neurodegeneration, characterized by a severe cholinergic and behavioural impairment, extensive neurofibrillary pathology linked to tau protein, aberrant processing of APP and intra- and extra-cellular deposition of Abeta-peptides (from the endogenous mouse APP gene) and a significant impairment of cortical and hippocampal synaptic plasticity (Capsoni et al., 2002a, 2002b; Capsoni et al., 2000; Origlia et al., 2006; Pesavento et al., 2002; lagostena et al J. neuroscience 2010) (see **Table 1**). The progressive comprehensive neurodegenerative phenotype observed in AD11 mice demonstrates a direct causal link, in mice, between an abnormal or unbalanced NGF signaling and the activation of the amyloidogenic cascade, as well as of the neurofibrillary pathology (Capsoni and Cattaneo, 2006; Cattaneo et al., 2008). This pathway could recapitulate, in mice, some events occurring also in the human brain undergoing AD neurodegeneration of the sporadic type.

Phenotypic Markers	**Tissue**	**Age (months)**							
		1	2	4	6	9	10	12	15
ChAT reduction	Basal forebrain	-	+	+	++	++	++	++	++
Hyperphosphorylated tau*	Entorhinal cortex	-	+	+	++	++	++	+++	+++
	Occipital cortex	-	-	+	+	+	++	++	+++
	Parietal cortex	-	-	+	+	++	++	+++	+++
	Hippocampus	-	-	+	+	++	++	++	+++
Dystrophic neurites (tau-positive)	Cerebral cortex	-	-	-	-	+	++	++	+++
Neurofibrillary tangles	Cerebral cortex	-	-	-	-	-	-	+	+++
PHFs	Cerebral cortex	-	-	ND	ND	ND	ND	ND	++
MAP-2 altered distribution	Cerebral cortex	-	+	+	++	++	++	+++	+++
	Hippocampus	-	-	+	++	++	++	+++	++
Accumulation of Aβ in dystrophic neurites	Hippocampus	-	-	-	+	++	++	+++	+++

β- amyloid plaques	Hippocampus	-	-	-	-	-	-	++	+++
Neuronal loss	Cerebral cortex	-	-	-	-	+	+	++	++
DNA fragmentation	Cerebral cortex	-	-	-	-	-	-	-	+
Cortical LTP deficit	Cerebral cortex	-	+	ND	++	+++	ND	ND	ND
Hippocampal nicotine enhancement failure		ND	ND	ND	+	ND	ND	ND	ND
Object recognition test deficit		+/-	+/-	+	+	++	+++	+++	+++
Radial Maze (retention)		ND	-	+	-	ND	ND	ND	ND
Morris Water Maze deficit (acquisition)		ND	-	-	-	+	++	++	++
Morris Water Maze deficit (retention)		ND	-	-	-	-	++	++	++
Anti-NGF antibody		-	+	+	+	+	+	+	+

Table 1 Progression of neurodegeneration in the AD11 mouse model. The symbol + indicates a qualitative measure of each phenotypic marker (+ means light, + + means mild , + + + means severe). Corresponding quantitation is in the referred papers. The table reports the phenotypic markers that show a significant difference with respect to age-matched control littermates.

(*) Revealed by silver impregnation and immunohistochemistry with monoclonal AT8 antibodies raised against hyperphosphorylated tau and tangles.

The early molecular events in this initial phase were already investigated by the authors, with the objective of identifying key molecular pathways likely to play a role in the incipient neurodegeneration process (D'Onofrio et al, 2009). The mRNA gene expression profile of AD11 brain areas was already characterized, by microarray and quantitative real-time PCR (qRT-PCR) analysis, in the early phases of the neurodegeneration process (at 1 and 3 months of age), in comparison to transgenic controls. Data showed that wide changes in gene expression profiles occurred in AD11 mouse already at 1 month of age, when no overt neuropathology is evident. Several mRNAs appeared to be differentially expressed, including species related to inflammation and immune response, to neurotrophic response, synaptic neurotransmission and different signaling pathways, with major differences in the Wnt pathway. These expression data, in the early phases of neurodegeneration, are therefore characterized by a striking overall "immunotrophic", neurotrophic and synaptic unbalance, that is broader than what would be expected on the basis of NGF neutralization per se, and is likely to be related to the specific mode of NGF neutralization, in these mice, with anti-NGF antibodies in the brain.

The AD11 model was exploited, in this study, to investigate changes of gene expression in brain regions (hippocampus, basal forebrain and cortex), during different phases of their progressive neurodegeneration. For this purpose, the gene profile of AD11 mice brain areas (hippocampus, basal forebrain and cortex) at different age, 1, 3, 6 and 15 months of age, was analyzed by microarray analysis associated with data mining techniques.

In order to analyze mRNA expression data, we adopted a "fingerprinting" method (described and applied in Felici and Truemper, 2002, Bertolazzi at al. 2009a, 2009b, 2010) that enables to isolate a small number of logic formulas, based on the expression levels of few genes that are able to discriminate with a high degree of accuracy between diseased and control mice. The analysis led to the identification of logic formulas encompassing a subset of genes, whose functional properties were further analyzed by online and manual annotation, to characterize the progression of the

neurodegenerative disease from the early (1-3 months) to the advanced stage (6-15 months).

2. Methods

Anti-NGF AD11 mouse model

AD11 transgenic mice (Ruberti et al., 2000) express a recombinant version of the monoclonal antibody mAb αD11 that specifically recognizes and neutralizes NGF (Cattaneo et al., 1988; Covaceuszach et al., 2008). A total of 30 female AD11 mice (5 per group), 30 female AD11-VH (5 per group) and 3 nontransgenic littermate control mice were used for this study. As in previous studies (Capsoni et al., 2000), AD11-VH mice were used as transgenic negative control mice, that are consistently negative with respect to all the neurodegeneration markers (Table 1). Each AD11, AD11-VH or control mouse was individually tested by transgene genotyping (for VH and VK transgenes). In addition, the level of transgenic anti-NGF antibodies was determined in the serum of each mouse, as described (Ruberti et al., 2000). The mRNA for the VH and VK antibody chains was also determined in each AD11 and AD11-VH mouse, by qRT-PCR. Mice were kept under a 12 hours dark to light cycle, with food and water *ad libitum*. Experiments were performed according to the European law for laboratory animal welfare and experimentation n. 86/609.

RNA isolation, amplification and labelling

Hippocampus (HP), cortex (CTX) and basal forebrain (BF) of the right hemisphere were dissected from the brains of freshly sacrificed mice. Total RNA was isolated from these brain areas using Trizol (Invitrogen) and DNAse treated by Qiagen columns. Quality and integrity of each sample was checked using the Agilent BioAnalyzer 2100 (Agilent RNA 6000 nano kit): samples with a RNA Integrity Number (RIN) index lower than 8.0 were discarded. All the experimental steps involving the labelling, hybridization and washing of the samples were done following the Agilent protocol[1]. Aliquots from the same RNA sample, prepared (and pooled) from 2 whole brains of wild type mice of the same strain (C57BL x SJLF2), were used in all hybridizations as a Reference sample, to reduce the experimental variability.

Hybridization of oligonucleotide mouse microarrays

The gene expression profiling was performed using a two-color protocol by Agilent[2] with reference experimental design: AD11, AD11-VH samples and the reference sample were always labelled with Cy5 and Cy3, respectively. Cy3 and 5-labelled cRNA were hybridized to Agilent 4x44k whole mouse genome oligonucleotide microarrays (G4122F).

[1] http://chem.agilent.com
[2] www.chem.agilent.com/enUS/Products/Instruments/dnamicroarrays/Pages/default.aspx

Scanning, feature extraction and statistical analysis

Post-hybridization image acquisition was accomplished using the Agilent scanner G2564B, equipped with two lasers (532 nm and 635 nm). Images were analyzed by Agilent Feature Extraction. Data filtering was performed in Microsoft Excel by discarding spots close to the background level. Basic data analysis was performed with Agilent GeneSpring GX and Microsoft Excel. Differentially expressed mRNAs were identified by the SAM technique (Tusher et al., 2001). The analysis of over- and under- represented functional annotations was performed using the Panther database of Biological Processes[3] and the DAVID web tool[4] (Huang et al., 2009), in the latter case selecting the following categories both for functional clustering (high stringency) and charts: goterm_bp_4, goterm_bp_5, goterm_cc_4, goterm_cc_5, goterm_mf_4, goterm_mf_5, panther_bp_all, panther_mf_all, interpro, pir_superfamily, smart, panther_family, panther_subfamily, kegg_pathway, panther_pathway, chromosome, cytoband, cog_ontology, sp_pir_keywords, up_seq_feature. Gene hierarchical clustering analysis was done using MeV 4.4[5] (Saeed et al., 2006).

Logic mining analysis

The purpose of this analysis of microarray data is to discover genes whose expression or co-expression strongly characterizes the AD11 models. Combinations of genes, whose biological role could be investigated and interpreted, are identified whose expression level determine an effective separation between diseased and control mice (in the following also referred as the two *classes* of the samples). Microarray data sets are characterized by a large number of genes in every sample (in the range of tens of thousands); it is therefore very important to adopt methods that are able to extract a subset of genes able to characterize the AD11 model among the exponential number of potential ones. A logic data mining method is used, which is composed of three main steps: (i) the application of discrete cluster analysis (DCA), an efficient gene expression clustering method; (ii) the selection of the most relevant clusters of genes (feature selection); (iii) the identification of the logic formulas that best characterize the AD11 samples versus the AD11-VH ones (formula extraction).

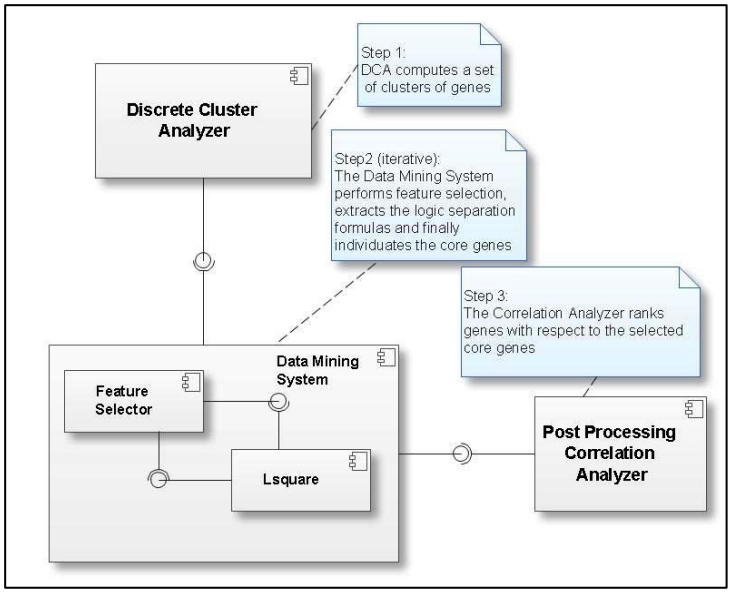
Discrete cluster analysis (DCA). *First*, a mapping into integer values of the real value of the gene expressions is performed. Given a feature f, its mean μ and variance σ are used to create a number of equally sized intervals symmetrical with respect to μ and proportional in size to σ. Good intervals are those that contain a large proportion of samples of the same class; thus, class entropy (i.e., a measure of the concentration of the samples in one of the two classes, diseased or control) of the distribution of the samples in each interval is measured and adjacent intervals are merged if they are empty or their merging does not increase their class entropy. The resulting interval mapping is, for each feature, associated with an integer encoding; two or more genes are merged into the same cluster when their integer encoded expression over the intervals are the same for each sample. Finally, a gene for each cluster is elected as its representative (clusters composed of a single genes may also be present).

[3] www.pantherdb.org/tools/genexAnalysis.jsp
[4] http://david.abcc.ncifcrf.gov
[5] MultiExperiment Viewer, TM4 suite , http://www.tm4.org/mev/

Feature selection. Feature selection is a typical problem in data mining and data analysis. A method based on an integer optimization formulation that selects a given number of features (say β) that maximizes the lower bound of the discriminating power between each pair of samples of different class is proposed. This optimization problem is of large size and grows quadratically with the number of samples; for its solution a fast heuristic algorithm based on greedy randomized search is adopted. A more detailed description can be found in (Bertolazzi et al. 2009a, 2009b, 2010).

Formula extraction. Each feature determined in the previous step is used by the logic data mining system Lsquare to extract the formulas. At this step of the analysis a feature is associated with a gene being above or below a given threshold, and a logic formulas is composed with conjuctions ("and") and disjunctions ("or") of more features. The classification formulas are determined through the solution of an integer optimization problem based on a minimum cost satisfiability problem (MINSAT). After obtaining the logic formulas the feature selection and the formula extraction step are repeated to acquire all the separating formulas, adding a constraint that is able to avoid the selection of the previously chosen features. This two steps are iterated until no more separation is found. A complete description of the Lsquare method can be found in Felici and Truemper, 2002, 2006).

Post processing correlation analysis. Once all the formulas that are able to distinguish between diseased and control mice are identified, the core set of the genes used to express these formulas are selected and denoted with S, while the remaining genes are ordered with respect to their correlation with *S*. This last step creates the final output that is composed by the core genes and the most relevant ones, which are



considered to be those with an absolute correlation greater than 0.45.

The Microarray Logic Mining software system (MLM). A data mining software has been designed specifically for microarray data analysis applications. MLM implements the methods described above: clustering, iterative feature selection, logic formulas identification and ranking by correlation.

3. Results

In order to elucidate the biological mechanisms, particularly related to specific gene activation, at the basis of the progression of the neurodegenerative process in the AD11 model, the whole ensemble of experimental samples was divided into two groups: the first contains the mRNA expression data for the three analyzed neuronal tissues (hippocampus (HP), cortex (CTX) and basal forebrain (BF)) at 1 and 3 months of age, the second contains the mRNA expression data for the three analyzed neuronal tissues at 6 and 15 months of age. Each of these two groups is composed by 60 samples with over 20.000 filtered and normalized genes for each sample. The data analysis was performed on each group as a whole, in order to extract more general information: in this way, the focus is on genes that can discriminate between the early and late stage of the neurodegeneration, without entering into the details of the difference in the expression profile between the tissues or between the specific time points, and, eventually, on genes that characterize the disease during the whole life span.

The application of the clustering method DCA shrinks the gene set down to 3656 for 1-3 months and to 3615 for 6-15 months. A large cluster of almost 7000 genes for both aging intervals is created, containing those genes that do not show any property of interest for the analysis. After DCA, we apply the feature selection step requiring to select a set of genes of very small dimension. In the first runs a single gene is selected only if it is able to distinguish alone between diseased and control samples. After a gene is chosen and the logic formulas are obtained we remove the selected gene from the data, reiterating this procedure until no separation with only one gene is possible. The adoption of this simple approach results in the discovery of 7 (resp. 9) core genes in the 1-3 (resp. 6-15) months data set, each one able to separate the diseased from control mice. It is interesting to point out that these genes do not belong to proper clusters (they are singletons after the application of DCA).

The method led to the identification of few single-gene formulas **(Table 2)**: every formula has independently the property of being able to correctly classify a gene expression sample and assign it either to the AD11 or the control class, within the group from which the formula was extracted, either 1-3 months or 6-15 months. These formulas correspond to two small sets of 7 and 9 genes out of the sample groups 1-3 months and 6-15 months, called $\{core\}_{1\text{-}3m}$ and $\{core\}_{6\text{-}15m}$ respectively (**Fig.1**): every gene is thus highly discriminating between the AD11 and the control class. Three genes are in common between the two groups, though inserted in slightly different formulas. The absolute value of the standard correlation between the expression vectors (one element per sample) of these genes, shown on the right side of Table 2, is about 0.84 on average for both sets, thus based upon a similar expression pattern we assume that there is a close functional connection between them.

Formulas at 1-3 months of age

Predicted Class	Gene	Condition
AD11	gene1	< 0.76
control	gene1	≥ 0.76
control	***gene2***	< 0.70
AD11	***gene2***	≥ 0.70
control	gene3	< 1.49
AD11	gene3	≥ 1.49
control	gene4	< 0.69
AD11	gene4	≥ 0.69
control	***gene5***	< 1.47
AD11	***gene5***	≥ 1.47
control	gene6	< 0.47
AD11	gene6	≥ 0.47
control	***gene7***	< 0.87
AD11	***gene7***	≥ 0.87

Standard correlation

	gene1	***gene2***	gene3	gene4	***gene5***	gene6	***gene7***
gene1	1.00	-0.75	-0.86	-0.86	-0.88	-0.73	-0.84
gene2	-0.75	1.00	0.89	0.88	0.86	0.86	0.85
gene3	-0.86	0.89	1.00	0.94	0.89	0.80	0.93
gene4	-0.86	0.88	0.94	1.00	0.90	0.81	0.94
gene5	-0.88	0.86	0.89	0.90	1.00	0.74	0.85
gene6	-0.73	0.86	0.80	0.81	0.74	1.00	0.73
gene7	-0.84	0.85	0.93	0.94	0.85	0.73	1.00

Formulas at 6-15 months of age

Predicted Class	Gene	Condition
control	gene8	< 0.54
AD11	gene8	≥ 0.54
control	***gene2***	< 0.62
AD11	***gene2***	≥ 0.62
control	gene9	< 1.36
AD11	gene9	≥ 1.36
control	gene10	< 0.47
AD11	gene10	≥ 0.47
control	***gene5***	< 1.86
AD11	***gene5***	≥ 1.86
AD11	gene11	< 0.79
control	gene11	≥ 0.79
control	***gene7***	< 0.80
AD11	***gene7***	≥ 0.80
control	gene12	< 1.09
AD11	gene12	≥ 1.09
control	gene13	< 0.4
AD11	gene13	≥ 0.4

Standard correlation

	gene8	***gene2***	gene9	gene10	***gene5***	gene11	***gene7***	gene12	gene13
gene8	1.00	0.84	0.80	0.82	0.83	-0.84	0.89	0.78	0.86
gene2	0.84	1.00	0.90	0.90	0.83	-0.87	0.83	0.81	0.86
gene9	0.80	0.90	1.00	0.92	0.79	-0.91	0.88	0.74	0.92
gene10	0.82	0.90	0.92	1.00	0.76	-0.93	0.84	0.79	0.92
gene5	0.83	0.83	0.79	0.76	1.00	-0.75	0.80	0.82	0.73
gene11	-0.84	-0.87	-0.91	-0.93	-0.75	1.00	-0.85	-0.74	-0.89
gene7	0.89	0.83	0.88	0.84	0.80	-0.85	1.00	0.74	0.88
gene12	0.78	0.81	0.74	0.79	0.82	-0.74	0.74	1.00	0.73
gene13	0.86	0.86	0.92	0.92	0.73	-0.89	0.88	0.73	1.00

Table 2. Right: formulas that discriminate AD11 and control samples in experimental data. Formulas are extracted by logic mining from the 1-3 months and 6-15 months sample groups. Each formula is represented by one gene and an inequality on a normalized value threshold. Genes are numbered from 1 to 13, gene 2,5 and 7 are highlighted being in both groups, though with different conditions. Left: standard correlation between genes: they are either highly positively correlated or highly negatively correlated.

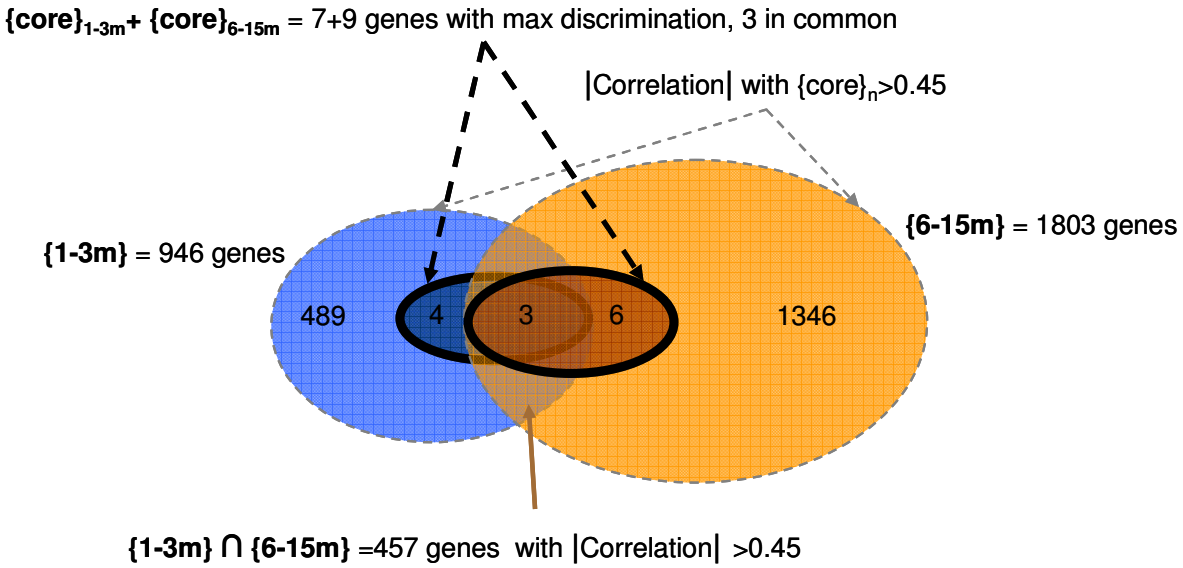


Fig. 1 Venn diagram of discriminating genes. The two inner subsets $\{core\}_{1\text{-}3m}$ and $\{core\}_{6\text{-}15m}$ contain the genes most effective in discriminating AD11 samples and control samples (see Methods for details). The outer sets {1-3m} and {6-15m} contain the genes that show an average absolute correlation value > 0.5 with the genes in the corresponding inner subsets, across all the samples. The intersection between $\{core\}_{1\text{-}3m}$ and $\{core\}_{6\text{-}15m}$ contains 3 genes, while the intersection between {1-3m} and {6-15m} contains 457 genes.

In this work, the experimental RNA samples were extracted from animal tissues, where it is expected to find a more complex combination of different pathological and compensatory mechanisms, compared to simpler cell models. Furthermore, since AD is a multifactorial disease, it also expected that the molecular bases of the neurodegeneration in this mouse AD-like pathology should be looked for more in the general biological context than in the expression pattern of few individual, though relevant, genes. Therefore to investigate the cellular mechanism triggered by or simply connected to the genes in $\{core\}_{1\text{-}3m}$ and $\{core\}_{6\text{-}15m}$ sets, considered as the "tip of the iceberg", we built two longer lists of genes based on the {core} sets, obtaining the new larger sets called {1-3m} and {6-15m} containing 946 and 1803 genes respectively, of which 457 are in the intersection, that show a maximum absolute value >0.45 for the correlation between each of their gene elements and any gene of the $\{core\}_{1\text{-}3m}$ and $\{core\}_{6\text{-}15m}$ sets **(Fig. 1)**. A first validation of these lists was to test their ability to correctly separate AD11 and control samples in the different tissues and time points, that is in each experiment: using hierarchical clustering we could indeed show a clear separation between the samples in all the cases **(Fig.2)**. A second validation step was the comparison of these 2 new ordered sets with the 12 list of differentially expressed genes obtained by the SAM algorithm (data not shown), one per each tissue and age of the mouse: though both {1-3m} and {6-15m} sets must be compared with 6 lists of differential genes each, the overlap is around 40% on average and tends to concentrate on top-ranked genes in the SAM lists (Tusher et al., 2001), which is relevant because the two sets should capture what is really relevant for all these lists at the same time.

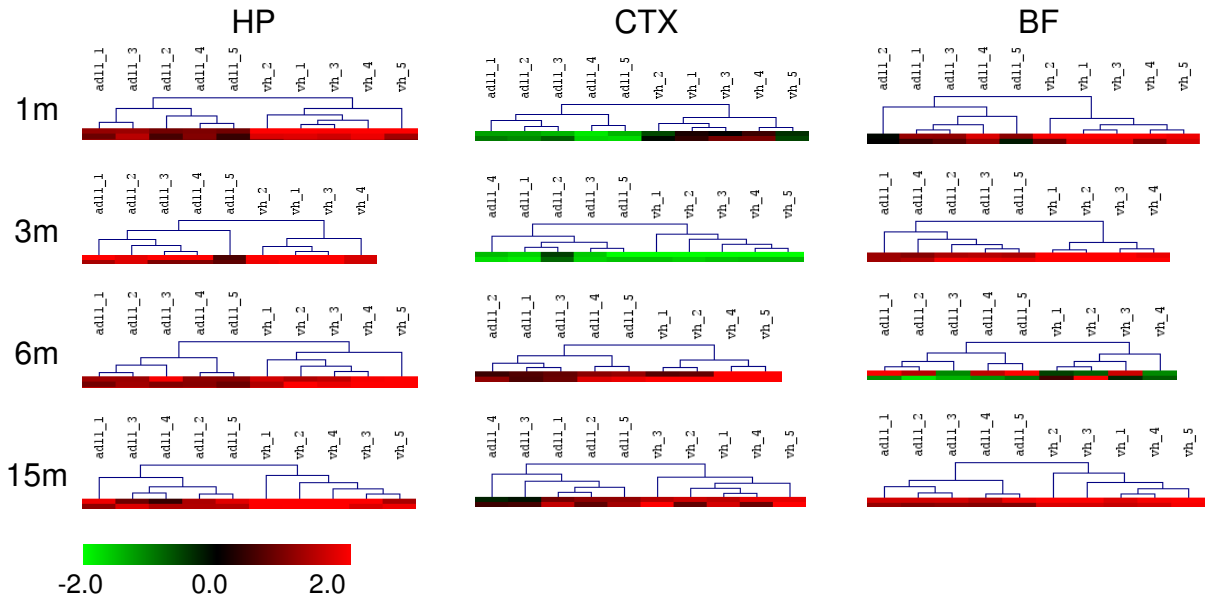


Fig. 2. Hierarchical clustering, with Euclidean distance and average linkage, of samples in the three analyzed tissues at 1, 3, 6, 15 months of age. The clusters are based on the gene lists shown in Fig.2: {1-3m} for 1 and 3 months of age, and {6-15m} for 6 and 15 months of age. Only the top of clusters are shown, so as to highlight that the two gene lists lead to a clear separation between AD11 and VH (control) samples. The data are lowess normalized and Log2 transformed. Only the top of the clusters are shown, where the AD11 samples and the AD11-VH control samples are labelled. The samples are clearly divided in two classes AD11 and AD11-VH (control) in all the experiments.

The {1-3m} and {6-15m} sets were then studied using the DAVID online tool for functional analysis of gene lists (Dennis et al., 2003), confirming the involvement of important gene categories in the neurodegeneration of AD11 mice. It should be underlined that, because of the mathematical pooling of the samples from the three brain tissues, HP, CTX, and BF, and from two time points for each of the two analyzed sets, the statistical significant functional categories tend to reflect a kind of average behaviour of the brain as a whole.

The data show that there are significant identical or similar functional categories common to both sets, thus spanning almost the entire life of the mice, about 24 months, and confirm the previous observation, on the involvement of the inflammatory and immune system genes in the AD11 mouse model not only at an early stage of the neurodegeneration (1-3 months) but also at a later stage (6-15 months) **(Table 3)**.

It has already been shown by the authors (D'Onofrio et al, 2009) that, at the early stage of the neurodegenerative process, genes involved in inflammation and immune response (both innate and adaptive) are widely represented; in that case a combination of Panther, DAVID and manual functional annotation was used. Regulation of such genes is a very prominent event also in quantitative terms, since mRNAs in such functional classes are among those showing the highest fold variation. Differential expression of genes coding for proteins of the complement system is particularly noteworthy because of the recently demonstrated role of the

complement system in mediating CNS synapse elimination, during normal brain development. It is remarkable to see that the cytoband 17-B1, the most significant for {1-3m} but also highly statistical significant for the {6-15m} set, contains mRNAs for different complement factors and histocompatibility complex in both cases. We may hypothesize that:

- one or more transcription factors targeting this chromosome region are likely to be responsible for this effect;
- the involvement of central components of the immune response could have a role also in later stage of the disease;

therefore, it would be important to further characterize the contribution of specific genes to the starting and the progression.

Another process present in both sets, but as the most significant one in the {6-15m} set and showing a striking modulation in the brain, is the cation transport pathway. These data are in agreement with a recent work on the AD11, showing that this model exhibits at 6 months of age an alteration of chloride homeostasis in the hippocampus, leading to a shift of GABA effect from the hyperpolarizing to the depolarizing direction (Lagostena et al., 2010).

mRNA categories concerning DNA remodelling, transcription regulation and splicing also appear in both lists of Table 3, suggesting that fundamental steps in gene expression such as mRNA transcription and processing are profoundly altered, contributing to the synthesis of aberrant protein compounds. As to the closely related categories of gene translation, protein expression, protein modification, transport and degradation, they look significantly affected by the NGF deprivation in this model in both lists, which is noteworthy in a model of AD where, among the major cellular hallmarks, are aggregations of misfolded or aberrantly processed proteins.

Concerning cell development, cell structure and proteins linked to the cytoskeleton, again these categories seem to include almost the entire lifespan on the animals, from 1 to 15 months, while the apoptosis is not included in the 1-3 months list, that is long before the first sign of neuronal loss appear (see Table 1).

{1-3m}

Category	Term	Count	Benjamini corrected Pvalue
CYTOBAND	17 B1	34	2.97E-11
PANTHER_BP_ALL	BP00143:Cation transport	156	1.70E-04
PANTHER_BP_ALL	BP00071:Proteolysis	200	1.74E-04
PANTHER_BP_ALL	BP00063:Protein modification	117	7.12E-04
PANTHER_BP_ALL	BP00151:MHCII-mediated immunity	57	7.36E-03
PANTHER_BP_ALL	BP00273:Chromatin packaging and remodeling	44	9.56E-03
PANTHER_BP_ALL	BP00044:mRNA transcription regulation	326	1.30E-02
PANTHER_BP_ALL	BP00138:Protein targeting	27	1.64E-02
PANTHER_MF_ALL	MF00034:Voltage-gated potassium channel	58	2.12E-02
PANTHER_BP_ALL	BP00224:Cell proliferation and differentiation	49	2.31E-02
PANTHER_MF_ALL	MF00262:Non-motor actin binding protein	102	2.32E-02
PANTHER_BP_ALL	BP00150:MHCI-mediated immunity	83	2.54E-02
PANTHER_MF_ALL	MF00255:Non-motor microtubule binding protein	51	2.81E-02
PANTHER_BP_ALL	BP00193:Developmental processes	78	3.41E-02
PANTHER_MF_ALL	MF00283:Ubiquitin-protein ligase	36	3.50E-02
PANTHER_MF_ALL	MF00261:Actin binding cytoskeletal protein	59	3.92E-02
PANTHER_MF_ALL	MF00232:Interleukin	35	4.02E-02
PANTHER_BP_ALL	BP00036:DNA repair	65	4.06E-02
PANTHER_BP_ALL	BP00242:Embryogenesis	7	4.09E-02
PANTHER_BP_ALL	BP00066:Protein acetylation	28	4.12E-02
PANTHER_BP_ALL	BP00142:Ion transport	84	4.33E-02

Table 3 – part I: Functional analysis of {1-3m} and gene sets, using the DAVID database.

{6-15m}

Category	Term	Count	Benjamini Corrected Pvalue
PANTHER_BP_ALL	BP00143:Cation transport	283	2.43E-07
PANTHER_BP_ALL	BP00071:Proteolysis	357	2.05E-06
GOTERM_BP_4	GO:0006412~translation	68	3.67E-05
GOTERM_BP_4	GO:0015031~protein transport	84	4.67E-05
PANTHER_BP_ALL	BP00044:mRNA transcription regulation	611	2.86E-05
PANTHER_BP_ALL	BP00276:General vesicle transport	87	2.73E-05
GOTERM_BP_4	GO:0045184~establishment of protein localization	87	5.42E-05
SP_PIR_KEYWORDS	alternative splicing	282	6.50E-05
PANTHER_MF_ALL	MF00262:Non-motor actin binding protein	193	7.60E-05
PANTHER_BP_ALL	BP00066:Protein acetylation	55	1.64E-04
SP_PIR_KEYWORDS	Mitochondrion	77	4.04E-04
PANTHER_BP_ALL	BP00138:Protein targeting	47	6.67E-04
CYTOBAND	17 B1	31	3.15E-03
PANTHER_BP_ALL	BP00151:MHCII-mediated immunity	96	1.08E-03
PANTHER_MF_ALL	MF00141:Hydrolase	75	1.62E-03
PANTHER_BP_ALL	BP00285:Cell structure and motility	110	1.83E-03
SP_PIR_KEYWORDS	spliceosome	20	2.42E-03
PANTHER_BP_ALL	BP00112:Calcium mediated signaling	41	2.22E-03
PANTHER_BP_ALL	BP00199:Neurogenesis	88	2.55E-03
PANTHER_BP_ALL	BP00129:Endocytosis	32	4.82E-03
PANTHER_BP_ALL	BP00193:Developmental processes	138	5.26E-03
PANTHER_MF_ALL	MF00006:Interleukin receptor	55	6.35E-03
PANTHER_MF_ALL	MF00231:Microtubule binding motor protein	86	6.69E-03
PANTHER_BP_ALL	BP00206:Chromosome segregation	44	8.95E-03
PANTHER_MF_ALL	MF00267:Membrane traffic protein	44	9.54E-03
PANTHER_MF_ALL	MF00073:Translation elongation factor	34	9.95E-03
GOTERM_BP_5	GO:0006464~protein modification process	148	1.29E-02
PANTHER_MF_ALL	MF00018:Chemokine	18	1.68E-02
PANTHER_MF_ALL	MF00264:Microtubule family cytoskeletal protein	56	1.88E-02
PANTHER_BP_ALL	BP00179:Apoptosis	58	2.02E-02
PANTHER_BP_ALL	BP00273:Chromatin packaging and remodeling	66	2.23E-02
PANTHER_BP_ALL	BP00114:MAPKKK cascade	21	2.45E-02
PANTHER_MF_ALL	MF00275:Transcription cofactor	47	2.69E-02
PANTHER_BP_ALL	BP00149:T-cell mediated immunity	107	2.77E-02
PANTHER_MF_ALL	MF00033:Voltage-gated calcium channel	67	3.21E-02
PANTHER_MF_ALL	MF00283:Ubiquitin-protein ligase	57	3.36E-02
PANTHER_MF_ALL	MF00074:Translation release factor	15	3.50E-02
PANTHER_MF_ALL	MF00218:Calmodulin related protein	23	4.38E-02
PANTHER_MF_ALL	MF00269:SNARE protein	19	4.44E-02

Table 3 – part II: Functional analysis of {6-15m} gene sets, using the DAVID database. mRNAs without a gene symbol annotation were excluded from the analysis. Categories are sorted according to the corresponding Benjamini corrected p-value. Some very general and poorly informative items were excluded from these tables.

4. Conclusions

The logic mining techniques reported in this work, integrated by more standard statistical analysis, have been applied here for the first time to a large ensemble of

high quality gene expression data obtained from the same biological system, a mouse model of progressive neurodegeneration, with several biological replicates for each experimental condition. The aim of the expression profile analysis was twofold: first, to select a set of genes able to reliably distinguish AD11 from control animals and, second, to shed light on the biology underlying the progression of the pathology. The standard statistics, including clustering, is very useful for data inspection and visualisation but may sometimes be somewhat approximate when dealing with the selection of differentially expressed genes, especially with a large number of samples. The logic mining approach proved here to be very useful and reliable in selecting two groups of very few, highly discriminating, genes at 1-3 months and 6-15 months of age, each of them worth to be further analyzed by a careful manual annotation, including analysis of the gene and promoter sequences and literature mining. The class prediction using these genes was very correct, as a consequence of this it was decided to expand this two small sets to have larger gene lists able to undergo a functional analysis using online functional annotation databases. These new gene lists, beyond correctly dividing AD11 and control samples in hierarchical clustering, contain significant subsets of genes belonging to functional categories either already known to be affected in this model or biologically fully compatible with the pathological framework. We highlight that the logic mining was very effective for studying this model of neurodegeneration in an unbiased manner and has the potential of providing a crucial contribution both for basic research and for diagnostic and prognostic purposes in the human AD.

Acknowledgments

This research was supported by the following grants: FIRB RBIN04H5AS, FIRB RBLA03FLJC from the Italian Ministry of Higher Education and Scientific Research, IIT Grant from the Italian Institute of Technology, MEMORIES Specific Targeted Research Project from the EU 6th Framework Program, Telethon Foundation (GGP05234), Alzheimer Association Grant. We thank Agilent Technologies for generously providing to EBRI part of the technological microarray platform and for technical support. Precious advice from Dr. Sebastiano Cavallaro (ISN-CNR), during the early phases of the project, is gratefully acknowledged.

References

Bertolazzi P., Felici G., Lancia G.: Application of feature selection and classification to computational molecular biology, Biological data Mining, Chen J.K., Lonardi S. eds., 257-294, 2010

Bertolazzi P., Felici G., Weitschek E.: Learning to classify species with barcodes, BMC Bioinformatics, 2009

Bertolazzi P., Felici G., Festa P.: Logic Based Methods for SNPs Tagging and Reconstruction, Computers & Operations Research, 2009

Bons, N., Rieger, F., Prudhomme, D., Fisher, A., Krause, K.H., 2006. Microcebus murinus: a useful primate model for human cerebral aging and Alzheimer's disease? Genes Brain Behav. 5, 120–130

Capsoni, S., Cattaneo, A., 2006. On the molecular basis linking Nerve Growth Factor (NGF) to Alzheimer's disease. Cell. Mol. Neurobiol. 26, 619–633

Capsoni, S., Giannotta, S., Cattaneo, A., 2002a. Beta-amyloid plaques in a model for sporadic Alzheimer's disease based on transgenic anti-nerve growth factor antibodies. Mol. Cell. Neurosci. 21, 15–28

Capsoni, S., Giannotta, S., Cattaneo, A., 2002b. Early events of Alzheimer like neurodegeneration in anti-nerve growth factor transgenic mice. Brain Aging 2, 24–43

Capsoni, S., Ugolini, G., Comparini, A., Ruberti, F., Berardi, N., Cattaneo, A., 2000. Alzheimer-like neurodegeneration in aged antinerve growth factor transgenic mice. Proc. Natl. Acad. Sci. U.S.A. 97, 6826–6831

Cattaneo, A., Capsoni, S., Paoletti, F., 2008. Towards non invasive nerve growth factor therapies for Alzheimer's disease. J. Alzheimers Dis. 15, 255–283

Delacourte, A., 2006. From physiopathology to treatment of Alzheimer's disease. Rev. Neurol. (Paris) 162, 909–912

Dennis G. Jr, Sherman B.T., Hosack D.A., Yang J., Gao W., Lane H.C., Lempicki R.A., DAVID: Database for Annotation, Visualization, and Integrated Discovery, Genome Biol. 2003;4(5).

D'Onofrio, M., Arisi, M., Brandi, R., Di Mambro, A., Felsani, A., Capsoni, S., Cattaneo, A. Early inflammation and immune response mRNAs in the brain of AD11 anti-NGF mice. Neurobiol of aging, 2009.

Felici G., Truemper K.: A Minsat approach for learning in logic domains, INFORMS Journal on computing 2002

Felici G., Truemper K..: The Lsquare System for Mining Logic Data, in: Encyclopedia of Data Warehousng and Mining, Wang J. ed., Idea Group Reference, 693-697, 2006

Games, D., Buttini, M.,Kobayashi, D., Schenk, D., Seubert, P., 2006. Mice as models: transgenic approaches and Alzheimer's disease. J. Alzheimers Dis. 9, 133–149

Hardy, J., Selkoe, D.J., 2002. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science 297, 353–356

Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57. PubMed PMID: 19131956.

Lagostena L., Marcelo R.S., D'Onofrio M., Brandi R., Arisi I., Capsoni S., Franzot J., Cattaneo A., and Cherubini E. "In the adult hippocampus, chronic NGF deprivation shifts GABAergic signaling from the hyperpolarizing to the depolarizing direction", J. Neurosci. 2010 30: 885-893; doi:10.1523/JNEUROSCI.3326-09.2010.

Oddo, S., Caccamo, A., Shepherd, J.D., Murphy, M.P., Golde, T.E., Kayed, R., Metherate, R., Mattson, M.P., Akbari,Y., LaFerla, F.M., 2003. Tripletransgenic model of Alzheimer's disease with plaques and tangles: intracellular Abeta and synaptic dysfunction. Neuron 39, 409–421

Origlia, N., Capsoni, S., Domenici, L., Cattaneo, A., 2006. Time window in cholinomimetic ability to rescue long-term potentiation in neurodegenerating anti-nerve growth factor mice. J. Alzheimers Dis. 9, 59–68

Pesavento, E., Capsoni, S., Domenici, L., Cattaneo, A., 2002. Acute cholinergic rescue of synaptic plasticity in the neurodegenerating cortex of anti-nerve-growth-factor mice. Eur. J. Neurosci. 15, 1030–1036

Ruberti, F., Capsoni, S., Comparini, A., Di Daniel, E., Franzot, J., Gonfloni, S., Rossi, G., Berardi, N., Cattaneo, A., 2000. Phenotypic knockout of nerve growth factor in adult transgenic mice reveals severe deficits in basal forebrain cholinergic neurons, cell death in the spleen, and skeletal muscle dystrophy. J. Neurosci. 20, 2589–2601

Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. TM4 microarray software suite. Methods Enzymol. 2006;411:134-93. Review. PubMed PMID: 16939790.

Tusher, V.G., Tibshirani, R. and Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 98, 5116-5121.