Report Number: WUCSE-2007-31

2007

DNA repair in incipient Alzheimer's disease

Monika Ray and Weixiong Zhang

Alzheimer’s disease (AD) is a progressive neurodegenerative disorder currently with no cure. Understanding the pathogenesis in the early stages of late-onset AD can help gain important mechanistic insights into this disease as well as aid in effective drug development. The analysis of incipient AD is steeped in difficulties due to its slight pathological and genetic differences from normal ageing. The difficulty also lies in the choice of analysis techniques as statistical power to analyse incipient AD with a small sample size, as is common in pilot studies, can be low if the proper analytical tool is not employed. In... Read complete abstract on page 2.

DNA repair in incipient Alzheimer's disease

Monika Ray and Weixiong Zhang

Complete Abstract:

Alzheimer’s disease (AD) is a progressive neurodegenerative disorder currently with no cure. Understanding the pathogenesis in the early stages of late-onset AD can help gain important mechanistic insights into this disease as well as aid in effective drug development. The analysis of incipient AD is steeped in difficulties due to its slight pathological and genetic differences from normal ageing. The difficulty also lies in the choice of analysis techniques as statistical power to analyse incipient AD with a small sample size, as is common in pilot studies, can be low if the proper analytical tool is not employed. In this study, we propose the use of a new method of significant genes selection, multiple linear regression, which uses the cognitive index (MiniMental Status Examination (MMSE)) and pathological characteristic (neurofibrillary tangles (NFT)), along with gene expression profiles, to select genes. The data consists of 7 incipient AD affected subjects and 9 age-matched normal controls. The analysis resulted in 686 significant genes with a false discovery rate of 0.2. Among the various biological processes previously known to be associated with AD, we discovered a set of 14 DNA repair genes that had statistically elevated or lowered levels of mRNA expression. Many key players involved in the defense against DNA damage were present in this list of 14 genes. In this article we report the status of DNA repair activity in incipient AD. From this study we conclude that the much observed apoptosis in AD may also be due to the activity of DNA repair genes. These findings have not been previously reported with respect to incipient AD and may shed new light onto its pathogenesis. This is the first study that has incorporated multiple clinical phenotypes of AD affected individuals in order to select statistically significant genes. It is also the first in analysing DNA repair genes in the context of AD via microarray gene expression analysis.

2007-31

DNA repair in incipient Alzheimer's disease

Authors: Monika Ray and Weixiong Zhang

Corresponding Author: zhang@cse.wustl.edu

Type of Report: Other

DNA repair in incipient Alzheimer's disease

Monika Ray[1] and Weixiong Zhang[1,2,†]
[1]Department of Computer Science and Engineering
[2]Department of Genetics
Washington University in Saint Louis
Saint Louis, MO 63130-4899, USA
email: mray@cse.wustl.edu, zhang@cse.wustl.edu

†: Corresponding author: zhang@cse.wustl.edu.

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder currently with no cure. Understanding the pathogenesis in the early stages of late-onset AD can help gain important mechanistic insights into this disease as well as aid in effective drug development. The analysis of incipient AD is steeped in difficulties due to its slight pathological and genetic differences from normal ageing. The difficulty also lies in the choice of analysis techniques as statistical power to analyse incipient AD with a small sample size, as is common in pilot studies, can be low if the proper analytical tool is not employed. In this study, we propose the use of a new method of significant genes selection, multiple linear regression, which uses the cognitive index (MiniMental Status Examination (MMSE)) and pathological characteristic (neurofibrillary tangles (NFT)), along with gene expression profiles, to select genes. The data consists of 7 incipient AD affected subjects and 9 age-matched normal controls. The analysis resulted in 686 significant genes with a false discovery rate of 0.2. Among the various biological processes previously known to be associated with AD, we discovered a set of 14 DNA repair genes that had statistically elevated or lowered levels of mRNA expression. Many key players involved in the defense against DNA damage were present in this list of 14 genes. In this article we report the status of DNA repair activity in incipient AD. From this study we conclude that the much observed apoptosis in AD may also be due to the activity of DNA repair genes. These findings have not been previously reported with respect to incipient AD and may shed new light onto its pathogenesis. This is the first study that has incorporated multiple clinical phenotypes of AD affected individuals in order to select statistically significant genes. It is also the first in analysing DNA repair genes in the context of AD via microarray gene expression analysis.

Keywords: gene expression, incipient alzheimer's disease, late-onset alzheimer's disease, DNA repair genes, multiple linear regression, gene selection

Alzheimer's disease (AD) is a complex progressive neurodegenerative disorder of the brain and is the commonest form of dementia, with 50-70% of all clinically presented cases being histopathologically confirmed at post mortem [1]. Advancing age is the major contributing factor for increased susceptibility to AD and the old are the fastest-growing segment of the United States population and have the highest prevalence of dementia [2]. AD has a complex aetiology with strong genetic and environmental determinants. Pathologically AD is characterised by the presence of neurofibrillary tangles (NFT) in the neurons of the cerebral cortex and hippocampus as well as the deposition of beta amyloid (Aβ) plaques in the entorhinal cortex, hippocampus, amygdala and association areas of frontal, temporal, parietal and occipital cortex. Several processes have been associated with AD, such as inflammation, loss of neurons, synaptic pathology, calcium dysregulation , cholesterol synthesis, re-entry into the cell cycle, oxidative stress, to mention a few. However, the molecular processes that initiate these processes are still unclear.

As there is no cure for late-onset alzheimer's disease (LOAD), treatment focuses on relieving and slowing down the progression of the symptoms. Hence, early diagnosis of AD can help in effective treatment strategies. Understanding the pathology present in incipient AD cases will help understand the progression and aetiology of the disease. Microarray data have many layers of information. The information revealed depends on the kind of analysis tools employed. Although the data used in this study has been previously analysed by Blalock et al. [3], the authors' main objective is to re-mine the data with a different set of techniques in order to gain new mechanistic insights into incipient AD.

The diagnosis of AD is made clinically by the finding of progressive memory loss with increasing inability to participate in daily activities. The MiniMental Status Examination (MMSE) or Folstein test is a quantitative measure of cognitive performance [4]. MMSE varies within the population by age and educational level. The normal MMSE score for individuals 80 years of age and older is 25 and higher. A pathological hallmark of AD is NFTs which are protein aggregates found within neurons. Tangles are formed by hyperphosphorylation of a microtubule-associated protein, tau, causing it to aggregate in an insoluble form. This phenomenon is normal in ageing, however, it is much more pronounced in AD brains, resulting in the loss of synapses and eventual neuronal death. The amount of NFTs is determined from postmortem brain specimens. Patients affected by AD have higher NFT scores. Based on the MMSE and NFT scores, there are four diagnoses (personal communication with Eric Blalock)- (a) High MMSE and low NFT $\rightarrow$ normal ageing; (b) low MMSE and high NFT $\rightarrow$ alzheimer's disease; (c) low MMSE and low NFT $\rightarrow$ another dementing pathology and (d) high MMSE and high NFT $\rightarrow$ 'cognitive reserve hypothesis' which suggests that individuals with a high intellect can withstand a large pathological insult and maintain their cognitive prowess [5]. In this paper, we use the MMSE and NFT scores of normal controls and the AD affected subjects along with gene expression profiles to address the differences between normal and AD affected individuals at the gene expression level.

Figure 1 shows the MMSE and NFT scores of sixteen subjects, 9 controls and 7 diagnosed with incipient AD, from a dataset originally analysed by Blalock et al. [3]. As can be seen from the figure, the separation between the controls and incipient cases is not distinct as there is some overlap between the 2 groups. Furthermore, it is evident that one variable without the other does not explain the AD diagnosis. Although all the controls have high MMSE and very low NFT scores, the incipient cases are spread across the entire MMSE range. If NFT is a good indicator of AD, then the samples labelled 1,2 and 3 should be classified as AD affected. If MMSE is a sufficient indicator of AD, then samples 4 and 5 should be labelled as normal subjects. However, that is not how the samples have been labelled, indicating that MMSE and NFT scores taken *together* influence the clinical diagnosis. The choice of clinical measures used to define microarray-based transcriptional profiles has a great impact on downstream analysis. Different clinical metrics will lead to different set of results.

In the method described here, we calculate the strength of the association between each gene's expression profile, and the MMSE and NFT phenotypes of each patient using multivariate linear regression. The assessment of the relationship between the gene expression, and clinical and histopathological phenotypes would be more relevant to better understand the underlying biological structure rather than correlating the expression to a single abstract phenotype, such as class labels. Instead of pooling into a class, the incorporation of individual samples' characteristics into the analysis will lead to a better study design, such as taking both MMSE and NFT for gene selection in this study. Using the p-values obtained from the linear regression model, multiple hypothesis testing is carried out to calculate the corresponding q-value for each gene. This is the most crucial step in the entire analysis as it plays a significant role in the false discovery rate (FDR) calculation as well as interpretation of downstream results. Based on the q-values, transcripts that are significant at a particular FDR are taken for further analysis.

We obtained a set of 686 significant genes at a FDR of 0.20, which included 14 genes, some of which are involved in initiating DNA repair or in the recruitment of other products involved in repair, and others that are involved in cell cycle check-point as a response to DNA damage. As the state of DNA repair activity in human AD has not been studied via microarray analysis, it prompted us to further analyse this set of 14 genes. This is the first report in which multiple macro-level phenotypes of AD were taken into account to select differentially expressed significant genes. Furthermore, to the best of our knowledge, this is also the first study that has focused on the expression levels of DNA repair genes in incipient AD. Previous expression studies on AD have focused on other characteristics of AD [3, 6, 7, 8, 9, 10, 11].

Materials and Methods

Data. The dataset consists of hippocampal specimens of 16 individuals [3]. There are 7 affected patients with incipient AD and 9 age-matched normal controls. Each subject had a MiniMental Status Examination (MMSE) score which varied from 20 (affected) to 30 (normal). The MMSE is a continuous measure that has been used as a reliable index of AD-related cognitive status. The patients were classified as 'incipient' based on their MMSE scores [3]. More details on this data can be found in [3].

Significant transcripts selection. The data was normalised using GCRMA as it has the best balance between precision and accuracy [12]. Probesets were mapped to genes using DAVID [13]. Probesets that didn't map to any gene name as well as those matching to hypothetical proteins with no known functions were removed. When multiple probesets mapped to the same gene, only the probeset with the highest mean was selected. This preprocessing resulted in 11543 unique genes.

In order to measure the strength of association between two independent (explanatory) variables taken simultaneously and one dependent variable, we use multivariate linear regression as described in [14]. In this analysis, MMSE and NFT are the two independent variables and the gene's mRNA expression level is the dependent variable. An in-depth explanation of multiple linear regression can be found in [15, 16, 17].

Let mRNA expression level be denoted by y_{ij} for $i = 1, 2, ..., n$ genes and $j = 1, 2, ..., m$ subjects or individuals. Let the total number of covariates be k, with $k = 2$ in this study, and $x = < x_1, x_2 >$ where x_1 is the MMSE value and x_2 is the NFT value. Then the linear regression model is given by

$$y_{ij} = b_0 + b_1 x_{1j} + b_2 x_{2j} + \epsilon_{ij} \qquad (1)$$

where b_0 is the regression constant, b_1 and b_2 are regression coefficients, and ϵ_{ij} is the random error that is assumed to be i.i.d normal distribution with zero mean and constant variance.

Equation 1 is estimated by least squares, which yields parameter estimates such that the sum of squares of errors is minimised. The resulting equation is

$$\widehat{y_{ij}} = \widehat{b_0} + \widehat{b_1} x_{1j} + \widehat{b_2} x_{2j} \quad (2)$$

where $\wedge$ denotes estimated values. There is no error term as the true model is unknown. Therefore, after the model has been estimated, the regression residuals r are defined as

$$r_{ij} = y_{ij} - \widehat{y_{ij}} \quad (3)$$

where $r_{ij} = \widehat{\epsilon_{ij}}$, y is the observed value and $\widehat{y}$ is the predicted value.

As the residuals are correlated and have variances, they are normalised to have zero mean and constant variance (homoscedastic). The normalised residual, $normr$, has a Student's t distribution with (total number of samples - total number of random variables - 1) degrees of freedom. The total number of random variables in this study is 3, i.e., gene expression, MMSE and NFT.

The sum-of-squares-error (SSE) (also known as the residual sum of errors) is the sum of squares of the residuals. SSE is defined by $SSE = \sum_{j=1}^{m} (normr_j)^2$. The smaller SSE, the better the approximating function fits the data. The regression-sum-of-squares (RSS) is the amount of variability in the response that is accounted for by the regression model. RSS is given by $RSS = \sum_{j=1}^{m} (\widehat{y}_j - \overline{y})^2$ where $\overline{y} = \frac{1}{m} \sum_{j=1}^{m} y_j$ is the average gene expression across all individuals. The total amount of variability in the response is referred to as the total-sum-of-squares (TSS) and is defined as $TSS = \sum_{j=1}^{m} (y_j - \overline{y})^2$. It is the amount of variation in the data that cannot be accounted for by the regression model. In other words, the RSS is the difference between the TSS and SSE.

The R^2 statistic or the coefficient of determination is given by $R^2 = 1 - \frac{SSE}{TSS}$ and it ranges from 0 to +1. It is the proportion of variability in a data set that is accounted for by a statistical model and is a measure of the global fit of the model. As the number of covariates in the model increases, R^2 increases, however it does not decrease due to the addition of noisy covariates. In order to account for noisy covariates being included in the model, the adjusted R^2 is calculated, which is the same as R^2 except that it penalises R^2 by the number of variables used in the model. The adjusted R^2 is given by

$$R^2 = 1 - (1 - R^2) \frac{m - 1}{m - k - 1} \quad (4)$$

The estimator of error variance (EV) is defined as $EV = (\frac{normr}{\sqrt{m-k}})^2$, where m is the total number of subjects and k is the total number of covariates and $(m - k) \geq 0$. $(m - k)$ is the residual degrees of freedom. Therefore, the F statistic for regression is defined for $k > 1$ and given by $F = \frac{(\frac{RSS}{k-1})}{EV}$. The F ratio estimates the statistical significance of the regression equation. It incorporates sample size and number of predictors in the assessment of significance of the relationship. This is the advantage of using the F ratio over R^2 as a model can have a high R^2 and still not be statistically significant if the sample size is not large compared to the number of predictors in the model. The significance probability p for regression is $p = 1-$ (F cumulative distribution function with (k - 1) and (m - k) degrees of freedom at the values in F).

As microarrays result in the measurement of several thousand probes, the individual p-values are not a reliable measure of significance. The individual p-values are corrected for multiple testing by calculating each gene's q-value using the Benjamini and Hochberg method of FDR calculation [18].

A variation of this gene selection strategy was published during the preparation of this manuscript [20]. On further analysis, the authors of [20] only use R^2 as the criteria for gene selection. No F ratio or q-value calculation was performed.

Results and Discussion

Unlike the analysis performed in [3] which used Pearson correlation coefficient, we carried out the comparison between 9 normal controls and 7 incipient AD using multiple regression analysis (see Methods). Statistical power in the analysis of datasets, depends on factors such as sample size and the tool employed for the analysis. Different analytical tools make different assumptions about sample distribution, population distribution, etc. Other differences include null hypothesis, sensitivity and specificity of the tool. We compared our multiple regression approach to select differentially expressed (DE) transcripts with SAM [19]. SAM is an open-source software which uses a modified t-statistics approach to identify DE genes. We ran SAM on the data with class labels - control and affected. The lowest FDR achieved by SAM was 0.50. On the other hand, multiple linear regression on the same dataset resulted in 52 DE genes at a FDR of 0, 303 genes at a FDR of 0.05 and 426 genes at a FDR of 0.10. This indicates that the multiple regression approach is probably a better method to analyse this data if one requires a large set of significant genes with a low FDR. This approach took into account two variables, i.e. cognitive index and pathological, - MMSE and NFT, respectively - associated with each subject, along with the observed gene expression to select DE genes. A variation of this idea was recently published during the writing of this manuscript [20].

From Figure 1 it seems that the difference in the gene expression levels between controls and incipient AD cases would be subtle as there is some overlap. After correcting for multiple testing, 686 genes were considered significant with a FDR of 0.20 as opposed to 89 genes identified to be correlated with both MMSE and NFT in Blalock et al.'s study [3]. The entire list of 686 genes is provided in supplemental information. Statistically significant biological processes were identified using EASE (http://niaid.abcc.ncifcrf.gov/home.jsp). A few of the statistically significant biological processes present in our significant genes (SG) list is shown in Table 1. Quite a few of the SG have been identified in previous microarray studies and reported to be associated with AD, such as calcium channel dysregulation, amyloid processing genes, apoptosis genes etc. As shown by Blalock et al., there are indeed many transcriptional and tumour suppressor responses [3]. As we used a different gene selection method, we were hunting for genes that have previously not been associated with AD via microarray studies. We discovered that DNA replication and repair biological process was also present in the SG list. When the list of 686 SG was compared to the genes described in a study on human DNA repair [21] and to DAVID [13], 14 DNA repair genes were present in our list (see Table 2). Since the study of DNA repair in AD is still in its nascent stage and has not been investigated in depth, we decided to further analyse this set of genes involved in the defense against DNA damage.

Genomes are subject to damage by chemical and physical agents in the environment and by free radicals or alkylating agents endogenously generated in metabolism. Mature neurons in the mammalian brain cannot divide and are highly metabolically active. Due to the high oxygen consumption rate by the brain, reactive oxygen species (ROS) can contribute to neuronal damage. Oxidative stress in neurons in human neurodegenerative diseases such as AD has been documented in previous reports [22, 23]. ROS attack of DNA can lead to DNA-DNA and DNA-protein cross linking, re-entry into cell-cycle by mature neurons, DNA strand breaks, production of oxidized base adducts, modification of DNA bases leading to problems in DNA replication and altered protein synthesis, and sister chromatid exchange and translocation in nuclear DNA [23]. The maintenance of genome integrity is essential and particularly important to neurons as they are among the longest living cells in the body. In response to DNA damage, cells activate multiple signalling pathways, leading to the accumulation of proteins in complex multisubunit nuclear foci, that represent sites of DNA replication arrest or sites of DNA repair [24]. To deal with DNA damage, cells have evolved a repertoire of cell-cycle check-point and DNA repair processes. In order to repair DNA damage, three main DNA repair pathways are present - base excision repair, nucleotide excision repair, and mismatch repair. An excellent survey on DNA repair in neurons is [25].

Non-homologous end-joining (NHEJ) is the predominant pathway used to repair double-strand breaks in DNA and is evolutionarily conserved. Ligase IV (LIG4) and protein kinase, DNA-activated, catalytic polypeptide (PRKDC) are involved in NHEJ [26]. LIG4 and XRCC4 form a ligation complex in the cell and play an important role in NHEJ [27]. In the repair process, LIG4 joins broken nucleotides together by catalysing the formation of an internucleotide ester bond between the phosphate backbone and the deoxyribose nucleotides [17]. Furthermore, it has been shown that PRKDC negatively regulates LIG4 protein stability [28]. LIG4 expression level was severely decreased in the AD subjects and PRKDC level was only slightly elevated. Study in [28] shows that LIG4 can facilitate PRKDC binding to the LIG4-XRCC4 complex and in the absence of LIG4, PRKDC and XRCC4 do not bind efficiently. Increased occurrence of double-strand breaks in DNA due to the oxidative damage requires increased activity of NHEJ components. Deficiency of LIG4 has been shown to be associated with extensive neuronal apoptosis [29, 30]. Deficiency in any of the NHEJ components can lead to chromosomal instability [28].

The mismatch excision repair (MMR) system is responsible for repairing the erroneous insertion, or deletion of bases that can arise during DNA replication and recombination, as well as repairing some forms of DNA damage. Repair is carried out by excising the wrongly incorporated base and replacing it with the correct nucleotide. The MMR system is composed of several protein complexes. Muts homolog 2, colon cancer, nonpolyposis type 1 (MSH2), postmeiotic segregation increased 1 (PMS1) and polymerase (DNA directed), epsilon (POLE) are involved in DNA mismatch excision repair. In particular, MSH2 expression was decreased in the AD subjects. MSH2 is a key mammalian mismatch repair gene that initiates the recognition of a base mispair and subsequently recruits additional MMR proteins involved in the repair. Normal neurons exposed to neurotoxins resulted in an increased production of MSH2 [31]. Cells deficient in MMR genes have increased susceptibility to genomic instability and cancer. It has been documented that cancer and neurodegenerative diseases may share a common pathway for the progression of the neurodegenerative disease [32]. PMS1 expression level was elevated in the AD individuals in our study. MMR proteins have also been known to regulate cellular response to DNA damage by signalling apoptosis [33]. The exact role of all MMR genes in response to DNA damage still remains unclear. It is hypothesised that just like protein degradation and nuclear export of p53 are blocked by DNA damage leading to increased levels of intranuclear p53, DNA damage induces the accumulation of human PMS1 through ataxia-telangiectasia-mutated (ATM)-mediated protein stabilisation [33]. POLE has been implicated in mismatch repair, nucleotide excision repair (NER) and base excision repair (BER) [21, 34]. AD subjects showed increased POLE expression level. Although extensive details about its role in DNA repair has not been well documented, it has been shown to play a vital role in the NER system in the presence of proliferating cell nuclear antigen, replication factor C (RFC), replication protein A, and DNA ligase I [35]. RFC was also present in our list of SG and showed elevated levels of expression in AD subjects. RFC is a five-subunit protein complex that is required for DNA replication. A recent study has suggested that BRCA1-associated complex (BASC) is key to recognising and repairing DNA damage. Among other components of this complex are MSH2, BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase (UCH)) (BAP1), and RFC [36, 13]. BAP1 level was decreased in the AD brains in our study. BAP1, which is a tumour suppressor gene, is required for transcription-coupled DNA repair [37]. Post-meiotic segregation increased 2-like 2 (PMS2L2) expression level was increased in AD subjects and is thought to be associated with DNA repair via sequence similarity but lacks experimental evidence [21].

ROS attack on DNA leads to a variety of modifications of purine and pyrimidine bases. BER prevents mutations by removing the oxidative lesions from the DNA. 7,8-dihydro-8-oxoguanine (8-oxoG) is an important mutagenic lesion. Nth endonuclease III-like 1 (NTHL1) is involved in BER and participates in the removal of 8-oxoguanine from 8-oxoguanine/guanine mispairs in DNA [38]. However, the increase in NTHL1 was very subtle in the AD cases.

H2a histone family, member x (H2AFX) is required for checkpoint mediated arrest of cell cycle progression and for efficient repair of DNA double strand breaks. H2AFX helps in the recruitment of repair and signalling proteins to the sites of DNA damage [39, 40]. Its expression level in the AD individuals was only slightly elevated. Fanconi anemia, complementation group g (FANCG) (*alias: XRCC9*) is associated with hypersensitivity to DNA-damaging agents, chromosomal instability (increased chromosome breakage), and defective DNA repair [41]. It is a part of the RAD6 pathway which is sensitive towards a variety of genotoxic agents. As a DNA repair protein, it may operate in postreplication repair or in a role of the cell cycle checkpoint guard [13]. Although not much has been documented about FANCG, it has been shown to be involved in protection against oxidative DNA damage [41]. FANCG levels was also increased in the AD affected subjects. Ubiquitin specific peptidase 1 (USP1) has been implicated in DNA repair and cell cycle regulation [42]. USP1 regulates the Fanconi anemia (FA) pathway. The FA pathway is required for the normal cellular response to DNA damage. Ubiquitination of the FA protein, Fanconi anemia D2 (FANCD2), is a critical event in DNA damage repair [42]. Deficiency of USP1 results in an accumulation of monoubiquitinated FANCD2. USP1 is necessary to deubiquitinate FANCD2 after the repair of specific DNA damage sites, in order to avoid overall-deleterious effects on genome integrity [42]. USP1 expression level was increased in the AD affected individuals.

Casein kinase 1 (CK1) contain a family of highly related serine/threonine protein kinases. Members of the CK1 family, such as casein kinase 1, epsilon (CSNK1E), have been shown to be sensitive to DNA damage and involved in chromosomal maintenance [43, 44]. CSNK1E level was decreased in the AD subjects in our study. Decreased levels of CSNK1E results in a significant increase in transforming growth factor (TGF)-β-induced transcription [45]. TGF-β signalling pathway suppresses the cell cycle and is defective in cancerous cells. Hence, decrease in CSNK1E levels leads to an increased activity of tumour suppressor genes.

Growth arrest and DNA-damage-inducible, gamma (GADD45G) is a cell cycle control gene and was increased in the AD brains in our study. GADD45G transcript levels is increased following stressful growth arrest conditions and treatment with DNA-damaging agents [13]. Expression of the GADD45G induces p38/JNK activation and apoptosis. Stress-responsive p38 and JNK mitogen-activated protein kinase (MAPK) pathways regulate cell cycle and apoptosis [46].

Conclusion

AD has been characterised by extensive cell death. Cell death can occur by injury (necrosis) or by suicide (apoptosis). Sometimes cells are induced to commit suicide in order to preserve genomic integrity. Two factors induce a cell to commit suicide - the withdrawal of positive signals, that is, signals needed for continued survival, and the receipt of negative signals. Positive signals include growth factors for neurons, and negative signals include increased levels of oxidants within the cell, DNA damage by these oxidants and the accumulation of misfolded proteins. In AD brains the negative signals clearly overshadow the positive signals. If the rate of DNA damage is greater than the rate of DNA repair, errors accumulate resulting in early senescence, apoptosis or cancer. Therefore, DNA repair rate is an important determinant of cell pathology [17] (see Figure 2).

Some of the 14 DNA repair genes present in our SG list are involved in the early stages of DNA damage detection and some of them help in the recruitment of other downstream proteins to help in DNA repair. A few of them are also involved in cell cycle control. LIG4 and PRKDC are the most well studied and documented repair genes in our SG list. The level of LIG4 was significantly decreased in the AD subjects. On the other hand, there were repair genes that had elevated levels of expression in the AD subjects. However, repair genes sometimes “repair” by inducing cell death. Hence, it cannot be assumed that increased levels of expression would necessarily prevent apoptosis or that decreased levels would automatically induce apoptosis. In our study we found that genes that should have had decreased levels of expression displayed increased levels, and the vice versa was true for genes that should have had increased levels of expression. In our list of 14 DNA repair genes, a few of them signal apoptosis pathways when present at elevated levels. On the other hand, deficiency in the expression of certain other repair genes also lead to significant cell death, as in the case with LIG4. Therefore, it is the orchestrated expression of many repair genes that can eventually lead to cell survival. This orchestrated action was disrupted in AD subjects resulting in DNA repair genes themselves inducing apoptosis. Further validation of these DNA repair genes as well as more investigation into the roles of currently known DNA repair genes would shed more light into the DNA repair process in AD affected individuals.

As cell cycle check-point is also a defence mechanism against DNA damage, there were many cell cycle related genes in our list of 686 transcripts. However, the difficulty lies in identifying those genes that respond to DNA damage and help in DNA repair versus those that are being switched on as a result of DNA damage, such as a mutation, and do not actually help in repair. In the set of 14 repair genes described in this article, the few that were involved in cell cycle have been shown in literature to be also involved in DNA repair. As more research is carried out in DNA repair, it is likely that the complete functions of more genes as well as the exact mechanism of DNA repair by some of the genes listed in this paper will be elucidated.

Instead of applying a cut-off value for increased or decreased levels of expression, we chose the significant genes for further analysis based on their statistical significance. We also felt that incipient AD subjects would have very subtle differences in their expression profile when compared to controls, and therefore allowed a FDR of 0.2. Although we used a linear model to correlate the gene expression levels to MMSE and NFT scores, a simple extension would be a non-linear correlation. However, AD is a difficult disease to study due to its close relationship to ageing. How can we be sure that certain gene changes are normal due to ageing while others are abnormal and more likely to be associated with AD pathology? Designing an experiment that can clearly delineate ageing from AD would help further the understanding of this complex disease.

This study is the first in which multiple variables obtained at a macro-level of AD individuals were taken into account simultaneously, as opposed to binary class phenotypes, along with gene expression to select genes perturbed in LOAD. Multiple testing correction was also performed to avoid single inference errors. The method applied in this article was significantly different from the one employed by Blalock et al. in their work [3] and had higher statistical power. Furthermore, this is also the first study that focused on the expression levels of DNA repair genes in AD affected individuals via microarray analysis. This is a small-scale study designed to test a new analytical tool, however, we hope that its results will motivate a larger, more sophisticated study to investigate the status of DNA repair in AD patients.

Acknowledgment

This research was funded in part by NSF grants EIA-0113618 and IIS-0535257, and Monsanto Co. The authors would also like to thank Johannes Freudenberg at the Biomedical Informatics Department in Cincinnati Children's Hospital, Cincinnati OH, for helping with the data normalisation. The authors would like to especially thank Eric M. Blalock at the Department of Molecular and Biomedical Pharmacology, University of Kentucky College of Medicine, Lexington, KY for his patience and excellent advice in analysing this data.

References

[1] Burns, A, Byrne, E J, Maurer, K (2002) *Lancet* **360**, 163-165.

[2] Schneider, E (1999) *Science* **283**, 796797.

[3] Blalock, E M, Geddes, J W, Chen, K C, Porter, N M, Markesbery, W R, Landfield, P W (2004) *Proc Natl Acad Sci USA* **101** , 2174-2178.

[4] Folstein, M F, Folstein, S E, McHugh, P R (1975) *J. Psychiat, Res.* **12**, 189-198 .

[5] Snowdon, D A (2003) *Annals of Internal Medicine* **139(5)**, 450-454.

[6] Ricciarelli, R, D'Abramo, C, Massone, S, Marinari, U M, Pronzato, M A, Tabaton, M (2004) *IUBMB Life* **56(6)**, 349-354.

[7] Small, S A, Kent, K, Pierce, A, Leung, C, Min, S K, Okada, H, Honig, L, Vonsattel, J, Kim, T (2005) *Annals of Neurology* **58(6)**, 909-919.

[8] Maes, O C, Xu, S, Yu, B, Chertkow, H M, Wang, E, Schipper, H M (2006) *Neurobiology of Aging* .

[9] Loring, J F, Wen, X, Lee, J M, Seilhamer, J, Somogyi, R (2001) *DNA , Cell Biology* **20(11)**, 683.

[10] Walker, P R, Smith, B, Liu, Q Y, Famili, F, Valdes, J, Liu, Z, Lach, B (2003) *Artificial Intelligence in Medicine* .

[11] Dunckley, T, Beach, T G, Ramsey, K E, Grover, A, Mastroeni, D, Walker, D G, LaFleur, B J, Coon, K D, Brown, K M, Caselli, R, Kukull, W, Higdon, R, McKeel, D, Morris, J C, Hulette, C, Schmechel, D, Reiman, E M, Rogers, J, Stephan, D A (2006) *Neurobiology of Aging* **27(10)**, 1359-1371.

[12] Irizarry, R A, Wu, Z, Jaffee, H A (2006) *Bioinformatics* **22(7)**, 789-794.

[13] Dennis, Jr, G, Sherman, B T, Hosack, D A, Yang, J, Gao, W, Lane, H C , Lempicki, R A (2003) *Genome Biology* **4(5)**, 3.

[14] Chatterjee, S, Hadi, A S (1986) *Statistical Science* **1(3)**, 379-416.

[15] Draper, N, Smith, H (1981) *Applied Regression Analysis, 2nd edition, Wiley*.

[16] *How to Read the Output From Simple Linear Regression Analyses (http://www.tufts.edu/ gdallal/slrout.htm)*.

[17] Wikipedia *en.wikipedia.org/wiki/DNA_repair*.

[18] Benjamini, Y, Hochberg, Y (1995) *Journal of the Royal Statistical Society. Series B (Methodological)* **57(1)**, 289-300.

[19] Tusher, V G, Tibshirani, R, Chu, G (2001) *Proc Natl Acad Sci USA* **98**, 5116 - 5121.

[20] Matsui, S, Ito, M, Nishiyama, H, Uno, H, Kotani, H, Watanabe, J, Guilford, P, Reeve, A, Fukushima, M, Ogawa, O (2007) *Bioinformatics* **23(6)**, 732-738.

[21] Wood, R D, Mitchell, M, Sgouros, J, Lindahl, T (2001) *Science* **291**, 1284-1289.

[22] Davydov, V, Hansen, L A, Shackelford, DA (2003) *Neurobiology of Aging* **24**, 953-968.

[23] Markesbery, W R, Lovell, M A (2006) *Antioxid Redox Signal* **8(11-12)**, 2039-2045.

[24] Gregory, R C, Taniguchi, T, DAndrea, A D (2003) *Seminars in Cancer Biology* **13(1)**, 77-82.

[25] Brooks, P J (2002) *Mutation Research* **509(1-2)**, 93-108.

[26] Budman, J, Kim, S A, Chu, G (2007) *J. Biol. Chem.* **282(16)**, 11950-11959.

[27] Jim Haber Lab at Brandeis University *(www.bio.brandeis.edu/haberlab/jehsite/nhej.html)*.

[28] Wang, Y, Nnakwe, C, Lane, W, Modesti, M, Frank, K M (2004) *J. Biol. Chem.* **279(36)**, 3728237290.

[29] Gao, Y, Sun, Y, Frank, K M, Dikkes, P, Fujiwara, Y, Seidl, K J, Sekiguchi, J M, Rathbun, G A, Swat, W, Wang, J, Bronson, R T, Malynn, B A, Bryans, M, Zhu, C, Chaudhuri, J, Davidson, L, Ferrini,R, Stamato, T, Orkin, S H, Greenberg, M E, Alt, F W (1998) *Cell* **95**, 891902.

[30] Barnes, D E, Stamp, G, Rosewell, I, Denzel, A, Lindahl, T (1998) *Current Biology* **8**, 13951398.

[31] Belloni, M, Uberti, D, Rizzini, C, Jiricny, J, Memo, M (1999) *Journal of Neurochemistry* **72**, 974979.

[32] Uberti, D, Ferrari, T G, Memo, M (2003) *Toxicology Letters* **139(2-3)**, 99-105.

[33] Luo, Y, Lin, F, Lin, W (2004) *Molecular, Cellular Biology* **24(14)**, 6430-6444.

[34] Fuss, J, Linn, S (2002) *J. Biol. Chem.* **277(10)**, 8658-8666.

[35] Shivji, M K, Podust, V N, Hubscher, U, Wood R D (1995) *Biochemistry* **34(15)**, 5011-5017.

[36] Wang, Y, Cortez, D, Yazdi, P, Neff, N, Elledge, S J, Qin, J (2000) *Genes, Development* **14(8)**, 927-939.

[37] Rauscher III, F J Patel, G, Jensen, D E, Proctor, M, Sekido, Y, Minna, J, Wilkinson, K D, Avrutskaya, A V, Leadon, S A (2000) *Cancer Detection, Prevention* **24(suppl 1)**.

[38] Matsumoto, Y, Zhang, Q, Takao, M, Yasui, A, Yonei, S (2001) *Nucleic Acids Res.* **29(9)**, 19751981.

[39] Bewersdorf, J, Bennett, B T, Knight, K L (2006) *Proc Natl Acad Sci USA* **103(48)**, 18137-18142.

[40] Gallmeier, E, Winter, J M, Cunningham, S C, Kahn, S R, Kern, S E (2005) *Carcinogenesis* **26(10)**, 1811-1820.

[41] Futaki, M, Igarashi, T, Watanabe, S, Kajigaya, S, Tatsuguchi, A, Wang, J, Liu, J M (2002) *Carcinogenesis* **23(1)**, 67-72.

[42] Nijman, S M, Huang, T T, Dirac, A M, Brummelkamp, T R, Kerkhoven, R M, D'Andrea, A D, Bernards, R (2005) *Mol. Cell* **17(3)**, 331-339.

[43] Behrend, L, Milne, D M, Stter, M, Deppert, W, Campbell, L E, Meek, D W, Knippschild, U (2000) *Oncogene* **19(47)**, 5303-5313.

[44] Fish, K J, Cegielska, A, Getman, M E, Landes, G M, Virshup, D M (1995) *J. Biol. Chem.* **270(25)**, 14875-14883.

[45] Waddell, D S, Liberati, N T, Guo, X, Frederick, J P, Wang, X (2004) *J. Biol. Chem.* **279(28)**, 29236-29246.

[46] Online Mendelian Inheritance in Man (OMIM) *http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=604949*.

Table 1: Statistically significant Biological Processes present in 686 genes

GO Biological Process	**Ease Score**
Regulation of cellular physiological process	5.59E-08
Primary metabolism and nucleobase, nucleoside, nucleotide and nucleic acid metabolism	4.76E-07
Regulation of transcription, DNA-dependent and primary metabolism	1.50E-05
Negative regulation of cellular physiological process and cellular process	2.14E-05
Transcription, DNA-dependent	2.76E-05
Biopolymer metabolism	0.017361586
Protein modification and DNA replication	0.020111636
Dna replication and cellular macromolecule metabolism	0.024086171
Actin filament capping and actin cytoskeleton organisation and biogenesis	0.043723268
Actin filament depolymerisation and actin filament capping	0.043723268

Table 2: 14 DNA repair genes in the list of 686 significant genes with FDR ≤ 0.20

RefSeq ID	**Gene name**	**Activity**
NM 004629	fanconi anemia, complementation group g (FANCG)(alias:XRCC9)	RAD6 pathway
NM 002105	h2a histone family, member x (H2AFX)	Editing and processing nucleases
NM 002312	ligase IV, DNA, atp-dependent (LIG4)	Non-homologous end-joining
NM 006904	protein kinase, dna-activated, catalytic polypeptide (PRKDC)(alias:XRCC7)	Non-homologous end-joining
NM 000251	muts homolog 2, colon cancer, nonpolyposis type 1 (e. coli)(MSH2)	Mismatch excision repair
NM 000534	postmeiotic segregation increased 1 (PMS1)	Mismatch excision repair
NM 006231	polymerase (dna directed), epsilon (POLE)	Mismatch repair, nucleotide excision repair
NM 002528	nth endonuclease III-like 1 (NTHL1)	Base excision repair
NM 152221	casein kinase 1, epsilon (CSNK1E)	–
NM 001017415	ubiquitin specific peptidase 1 (USP1)	–
NM 006705	growth arrest and DNA-damage-inducible, gamma (GADD45G)	–
NM 002916	replication factor c (activator 1) 4, 37kda (RFC4)	Mismatch repair
NM 002679	postmeiotic segregation increased 2-like 2 (PMS2L2)	Mismatch repair
NM 004656	brca1 associated protein-1 (ubiquitin carboxy-terminal hydrolase) (BAP1)	–

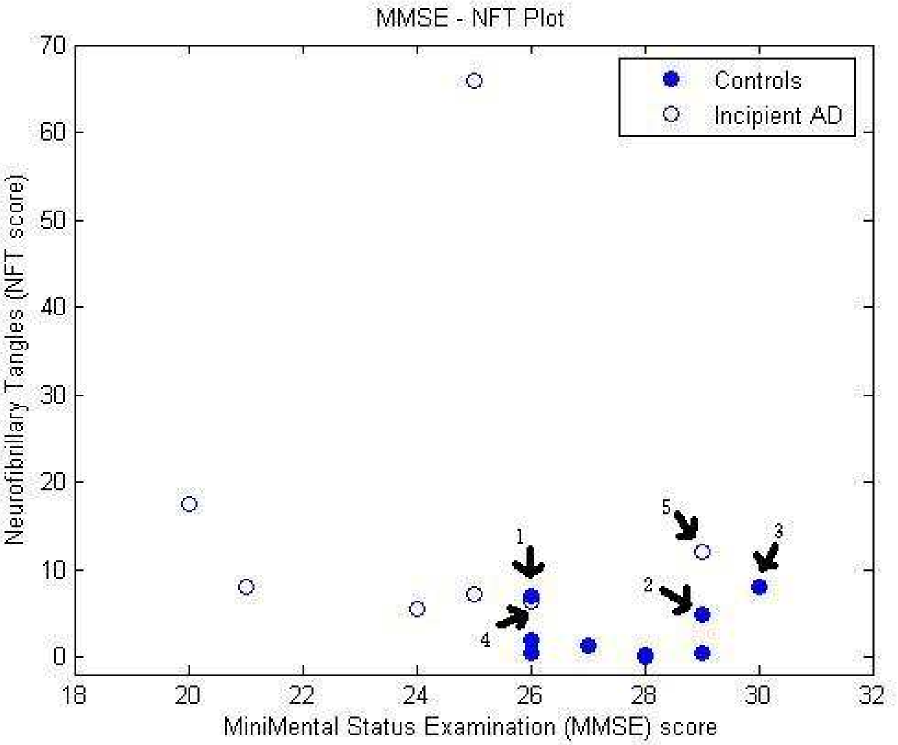

Figure 1: The MMSE-NFT plot for 9 normal controls and 7 AD affected cases. The ontrols are shown in solid circles while affected cases are in white circles. While the controls occupy a small compact space in the lower right corner, the affected samples spread out over the entire MMSE range. As can be seen there is an overlap between a normal and affected subject, i.e. samples 1 and 4, indicating no clear class distinction based on MMSE and NFT scores. Furthermore, just one variable (either MMSE or NFT) without the other does not explain the AD diagnosis. If NFT is the sole indicator of AD, then the samples labelled 1,2 and 3 should be classified as AD affected. If MMSE is a sufficient indicator of AD, then samples 4 and 5 should be labelled as normal subjects. Since both MMSE and NFT played a role in the AD diagnosis, *both* variables were taken into account in the gene selection algorithm. Due to the overlap of control and incipient case, a subtle change in gene expression level was expected resulting in the choice of a FDR of 0.2.

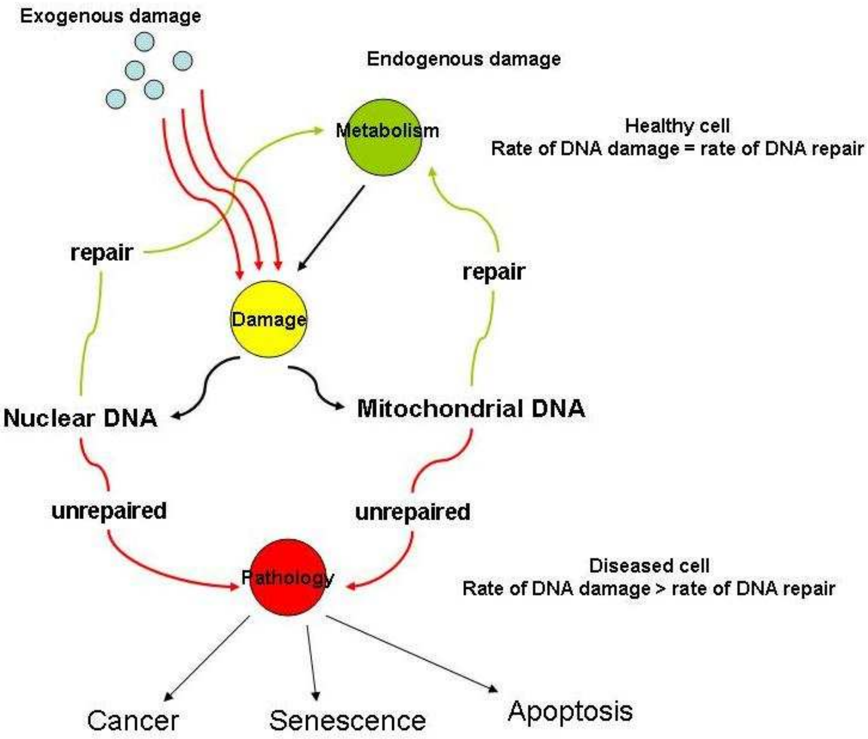

Figure 2: Cartoon depicting a healthy and diseased cell based on the accumulation of errors in the cell. Mature neurons are highly metabolically active and produce large amounts of ATP in order to generate action potentials. This results in high rates of DNA damage. Healthy cells have a proper balance of the DNA repair and DNA damage mechanisms. DNA repair rate is an important determinant of cell pathology. Figure taken from [17]