Internal Report CS Bioinformatics Track 14-05 August 2014

# Leiden University
# Computer Science
# Bioinformatics Track

Genotypes-phenotype predictions in patients diagnosed with early onset Alzheimer

Dimitra Zafeiropoulou

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

TU Delft Master's Thesis

August 27th 2014
Pages 1– 15

# Genotypes-phenotype predictions in patients diagnosed with early onset Alzheimer.

Dimitra Zafeiropoulou*

Delft Bioinformatics Lab†, Delft University of Technology, Delft, The Netherlands

Defended on August 27th 2014

**ABSTRACT**

Alzheimer's disease (AD) is the most common form of dementia among older people. Although less prevalent before the age of 65 years old, it is still the most frequent cause of early-onset dementia. Mutations in 3 genes, APP, PSEN1 and PSEN2 (Devi *et al.* (2000),Ertekin-Taner (2007), Bertram and Tanzi (2008), Bird (2008)) , are known to cause early-onset AD, but a large number of familiar cases do not have mutations in these genes and several disease causing genes still need to be identified.

Alzheimer disease causes a variety of clinical symptoms which can be associated with different forms of dementia. Primary progressive aphasia (PPA) and posterior cortical atrophy (PCA) constitute two of the most common representations of atypical AD (von Gunten *et al.*, 2006). Cerebral amyloid angiopathy (CAA), another brain related disorder, is also considered to be closely related to AD (Ghiso and Frangione, 2001). However, the genetic variations that result in the phenotypic variations within patients, are still unknown. In most of the cases, there is also phenotypic overlap among AD cases which makes the distinction between patients even more difficult. It is also not completely clear whether there should be more AD subtypes or not.

The exome sequencing data from 400 Dutch patients in early onset diagnosed with probable AD, PPA, PCA and CAA are used in this study. The aim is to investigate if there is any genetic variation that can explain differences in patients' phenotypes. Another point of interest is the detection of any possible structure both in genotype and phenotype space that may indicate the existence of several subgroups that discriminate between patients diseased with AD.

According to the analysis conducted, it is observed that there is an underlying structure both in genotype and phenotype space because there are clusters detected. However, the clusters are not well defined, probably because the sample population is too homogeneous. Furthermore, there is no genetic variation identified as significantly associated with the phenotypes. Nevertheless, there are 3 genes, HLA-DRB1, HLA-DRB5 and DEFB119 being selected with consistency by most of the methods tested, as those that better explain the phenotypic variation among the patients. HLA-DRB1 and HLA-DRB5 genes have been identified in recent studies as being related to late onset AD (Lambert *et al.*, 2013). Interestingly, they interact with CD74 gene which in turn interacts with APP that acts as a suppressor of amyloid-$\beta$. There is no much evidence about DEFB119 gene which is found to be involved in meningioma, a brain carcinoma. However, it seems to be related to early onset AD as the highest model performance is succeeded when all the 3 genes are used for the predictions.

**Contact:** zafeiropoulou.dimitra@gmail.com

*to whom correspondence should be addressed

†Pattern Recognition and Bioinformatics: http://prb.tudelft.nl

## 1 INTRODUCTION

Alzheimer's disease (AD), the most common form of dementia, is a degenerative disease of the brain affecting the memory and other mental abilities (Berchtold and Cotman, 1998). The disorder usually appears in people older than 65 (late-onset Alzheimer's) and less common in people earlier in adulthood (early-onset Alzheimer's). There are 3 genes APP, PSEN1 and PSEN2 which have been previously identified to be related to early-onset Alzheimer disease (Devi *et al.* (2000),Ertekin-Taner (2007), Bertram and Tanzi (2008), Bird (2008)). Mutations occurring in any of these genes result in accumulation of amyloid-$\beta$(A$\beta$) in brain and in the formation of amyloid plaques which is the main characteristic of the disease. However, there is no significant association detected between known genetic risk factors, including APP, PSEN1 and PSEN2 genes, and the clinical symptoms in AD as it would be expected.

Patients suffering from AD indicate a variability regarding the cognitive profiles. In most of the cases, memory impairment constitutes the main feature, but in some cases patients present atypical symptoms where other cognitive domains such as language are more severely impaired than memory (Stopford *et al.*, 2008). Two of the most common forms of atypical AD are posterior cortical atrophy (PCA) and primary progressive aphasia (PPA) (von Gunten *et al.*, 2006). Cerebral amyloid angiopathy(CAA) is also closely related to Alzheimer disease and it is associated with an increased number of cerebral microbleeds(Charidimou and Werring, 2011).

Currently, there is no drug treatment that provides cure for Alzheimer disease but early diagnosis may prolong patient's life (Leifer, 2003). Clinical diagnosis is usually achieved through a number of clinical examinations (eg. A$\beta$ concentration, ApoE genotype, etc) and neuropsychological tests. There is no evidence about genotypes associated to the clinical phenotypes apart from ApoE phenotypes which are directly measured from specific genetic variations in ApoE gene.

Recent studies have tried to identify additional genes associated with an increased risk of developing AD by conducting genome-wide association studies (GWAS). Such studies look for genomic variation in samples of unrelated patients with AD compared to unrelated control subjects. These GWAS have resulted in several new candidate genes that have been identified as potential risk factors (Bertram *et al.*, 2008). Follow-up studies, further investigate associations between genetic variation within candidate genes of interest and disease state (G *et al.*, 2010).

In contrast to GWAS which look for variations occurring in the whole genome, we concentrate only on variations detected on the coding part of the DNA (exome) as we eventually are only interested in changes as they can be targeted by drugs. This approach is much more powerful because exome sequencing targets all variations in the coding region of the gene whereas GWAS undersample these variations.

In the context of this study, we exome-sequenced 400 Dutch patients diagnosed with early onset AD, PPA, PCA, and CAA. In this way, we focus on variations in a homogeneous group of patients (in disease status as well as ancestry, i.e. only Dutch people) in early onset AD which is more prone to a genetic origin than late onset AD or AD in general. Our main aim is to investigate if genotypes influence phenotypic variation within the patients. Thus, we are interested in detecting variations in genes other than APP, PSEN1 and PSEN2 that may indicate a significant correlation with the different clinical profiles. Even if the patients have been diagnosed with one of the 4 types mentioned above, it is however not known whether more subtypes exist. Because the subtypes are not truly known, it is interesting to see whether more subtypes exist, and if so what the genotype-phenotype relationship would be with respect to all the subtypes.

## 2 MATERIALS AND METHODS

### 2.1 Data

*2.1.1 Clinical Data* The clinical examination data of 400 Dutch patients suffering from Alzheimer disease(AD) were collected. The clinical data consist of 64 different measurements like the amyloid-$\beta$, number of microbleeds in brain, ApoE genotypes, scores of different kind of diagnostic tests, etc (Table S3). All the patients are in early onset, with an onset age less than 68 years old. Each patient is diagnosed as having one or more of the following types of the disease: probable AD, primary progressive aphasia (PPA), cerebral amyloid angiopathy (CAA) and posterior cortical atrophy(PCA). The majority of the patients, 369 out of 400, were diagnosed as probable AD and only 6, 36 and 67 patients were also diagnosed with PPA, PCA and CAA respectively.

*2.1.2 Exome Sequencing Data* All the patients were exome sequenced using genomic blood (DNA) at a coverage ranging from 40-fold to 90-fold using Illumina Hiseq 2000 platform. Prior to variant calling, the dataset is aligned using BWA and hg19 (Kent *et al.*, 2002) reference genome. Picard and Samtools are used to convert, sort, and index the aligned data files. Duplicate reads are marked by Picard. The Genome analysis toolkit (GATK) (McCullagh and Nelder, 1989) is then used to recalibrate the alignments and to call SNPs (by UnifiedGenotyper).

### 2.2 Annotation

SNP calls are annotated through Annovar (Wang *et al.*, 2010). Specifically, SNPs are annotated based on the genomic position, the genomic function (exonic, intronic, UTR, etc) and the effect on the exon (synonymous, nonsynonymous, stopgain, stoploss) (Table S1). Furthermore, nonsynonymous SNPs are annotated based on the effect of altering protein function. For this reason, SIFT (Ng and Henikoff, 2003), a variant effect predictor, was used. SIFT assigns a score that indicates the likelihood of a SNP to be tolerated (i.e normal function) or not (deleterious). SNPs with a SIFT score less than 0.05 are presumed to be damaging.

### 2.3 SNP Selection

Figure 4 gives an overview of the filtering steps that are followed in order to select a subset of likely disease-causing SNPs that can be used in the subsequent association analysis. First, SNPs located on chromosomes X and Y are removed in order to a avoid a sex related bias. Since we are only interested in SNPs that alter protein function we then remove SNPs that are not within the exonic region of the gene or on the splicing sites of the gene, -/+ 10bp before and after the gene. The remaining SNPs are filtered for SNPs that are annotated as nonsynonymous, stopgain and stoploss as they result in an amino acid substitution. SNPs that were predicted as being tolerated without a damaging effect in the protein function are also filtered out (i.e having a SIFT score $>$0.05). As a next step, we use dbSNP in order to filter out common SNPs within the general population having a minor allele frequency(MAF) $>$0.05.

SNPs with more than 70% missing values within the patients are also removed. This step is required because missing values have to be imputed and there is no effective method for imputing variables with high percentage of missing values (Barzi and Woodward, 2004).

Due to the fact that there is no control group and the patients are both Dutch and diseased, we also remove SNPs that are common in $>$90% of the patients. Through this filtering step, we manage to reduce Dutch related bias and to retain SNPs showing variation within the patients.

### 2.4 Gene Scoring

In order to identify genes that may explain the phenotypic variation among the patients, the SNP data is transformed into a gene score. Through the gene name annotation provided by Annovar, SNPs are mapped to the gene that they belong to. SNPs unable to map to any known Ensemble gene or SNPs mapped to more than one gene were not taken into consideration. After mapping SNPs to genes, a score per gene is computed in one of the following ways:

**a.** By counting the SNPs that are mapped to a gene. (eq. 1)

$$\text{score}_{ip} = \frac{\sum_{k=1}^{Ni} \text{SNP}_{kip}}{\sum_{j=1}^{Mi} \text{exon}_{\text{end}}^{ji} - \text{exon}_{\text{start}}^{ji}}, \begin{cases} \text{SNP}_{kip} = 1 & \text{if mutated} \\ \text{SNP}_{kip} = 0 & \text{otherwise} \end{cases} \quad (1)$$

**b.** By summing up scores(score for each SNP is computed based on SIFT score assigned to it) of each SNP in a gene. (eq. 2).

$$\text{score}_{ip} = \frac{\sum_{k=1}^{Ni} \text{SNP}_{kip}}{\sum_{j=1}^{Mi} \text{exon}_{\text{end}}^{ji} - \text{exon}_{\text{start}}^{ji}}, \begin{cases} \text{SNP}_{kip} = 1 - SIFT & \text{if mutated} \\ \text{SNP}_{kip} = 0 & \text{otherwise} \end{cases} \quad (2)$$

,where $\text{score}_{ip}$ is score of gene *i* for patient *p*; $\text{SNP}_{kip}$ represents the k-th SNP of gene *i* mutated or not, in patient *p*; $\text{exon}_{\text{start}}^{ji}$ is the start position of *j*-th exon for gene *i* and $\text{exon}_{\text{end}}^{ji}$ is the end position of *j*-th exon for gene *i*.

In the case of the scoring function of eq. 1, SNPs contribute in the same way to the final gene score. So, when two patients have the same number of SNPs for a particular gene independent to the genomic position or effect their gene score will be the same. On the other hand, when the gene scoring is computed by eq. 2, the damaging effect is also taken into consideration. In this way, two patients with the same number of SNPs for a particular gene may have different gene scores because their SNPs can have varying damaging effect.

## 2.5 Clinical Measurement Selection

The clinical dataset consists of 400 patients and 64 clinical measurements. Clinical measurements with >60% missing values are removed. Moreover, there are several measurements included in the clinical data that are not related to patients' phenotypes (date for first visit, last date checked alive, MRI date, MRI type, etc) and thus those are ignored. Some clinical measurements related to different kind of cognitive tests, although used for clinical diagnosis, tend to vary by age, education and the time the test are taken. As the complete set confounding factors for these test are not known we decided to exclude these measurements also. Finally, 14 out of the initial 64 clinical measurements were left (Table S3).

## 2.6 Imputation of Missing Values

Both clinical and SNP data contain missing values. To deal with this problem, we apply *k* nearest neighbor (Batista and Monard, 2001) imputation, with *k* equal to 1 . The distance measure used to find the nearest neighbor is euclidean distance.

## 2.7 Dissimilarity-based representation

A key part of our method is to represent the patients in a dissimilarity space (Duin *et al.*, 2010) in order to capture their relative differences. In a dissimilarity space objects initially represented by their measured values are re-represented by their distances to the other objects. Formally, given a set of objects $X = \{o_1, o_2, ..., o_n\}$, $i = 1, ..., n$, a mapping function $D(o_i, X) : X \rightarrow \mathrm{R}^n$ is defined in which each object in X is a $n$-dimensional vector, for which each dimension $j$ describes the dissimilarity of object $i$ with object $j$ from the set of objects X, $d(o_i, o_j)$, where $d(.,.)$ is a defined distance measure between two object representations. Hence, every object is re-represented with a $n$-dimensional dissimilarity vector $[d(o_i, o_1)...d(o_i, o_n)]^T$.

Often the Euclidean distance is used as a distance measure to construct the dissimilarity-based representation. That works fine in case of objects with features of a specific data type. But when it comes to features with a mixture of different variable types, the selection of a distance measure is more complicated. Consequently, we decided to use as the dissimilarity function the Gower generalized coefficient of dissimilarity (Chatfield and Collins, 1981), a measure for multivariate data types.

The advantage of using Gower's coefficient is that the appropriate dissimilarity score can be calculated for each variable independently and then combined to give the final value for the dissimilarity score over all variables. The Gower generalized coefficient of dissimilarity between two objects $o_i$ and $o_j$ is defined as follows (Cox and Cox, 2000),(Chatfield and Collins, 1981):

$$\mathrm{s}_{ij} = \frac{\sum_k \mathrm{s}_{ijk}\mathrm{w}_{ijk}}{\sum_k \mathrm{w}_{ijk}} \tag{3}$$

, where: $s_{ijk}$ is the dissimilarity score calculated for variable $k$ and $w_{ijk}$ a weight associated with that variable. For the continuous variables we chose the Euclidean distance measure, and for the ordinal we chose the City Block distance measure. Before calculating the distance measures all variables were normalized to be within the range[0,1]. We chose no special weight for the different variables (i.e. $w_{ijk} = 1$).

**Table 1.** Selected Clinical Features and Variable Types

| Feature Description | Type |
|---|---|
| age of onset | continuous |
| ApoE genotype | ordinal |
| medial temporal lobe (atrophy) right | ordinal |
| medial temporal lobe (atrophy) left | ordinal |
| phosphorylated tau | continuous |
| parietal (atrophy) right | ordinal |
| parietal (atrophy) left | ordinal |
| global atrophy | ordinal |
| amyloid-beta | continuous |
| white matter abnormalities | ordinal |
| holes in tissue | ordinal |
| infracts | continuous |
| microbleeds in brain | continuous |
| total tau | continuous |

As for the genotype space, all the features (gene scores) are continuous. For this reason, the Euclidean distance measure is defined as the dissimilarity function.

## 2.8 Regression

To model the association genotype-phenotype we adopt linear regression (McCullagh and Nelder, 1989):

$$\mathrm{y} = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + ... + \beta_n x_N - 1 + e \tag{4}$$

,where $x_i$'s represent the genotype, in our case the individual gene scores; y represents the phenotype; and the $\beta$'s are the intercepts explaining the importance of the different genes in predicting the right phenotype. It is important to notice that the phenotype $y$ is a one-dimensional representation of the phenotypic relationships between the patients (2.7). To accomplish this, the *n*-dimensional dissimilarity representation of the phenotypic relations between patients is mapped to 1 dimension using a multi-dimensional mapping. In our case we chose to use principal component analysis to do so (Jolliffe, 2002). Hence the principal components of the *n*-dimensional phenotypic dissimilarity representation are determined and the data is mapped onto the first component (the component with the largest eigenvalue) (Fig. 1).

In regression, the $\beta$ coefficients are estimated by solving the following least squares minimization problem:

$$\hat{\beta} = argmin \, \|\mathbf{y} - \beta\mathrm{X}\|_2^2 \tag{5}$$

,where **y** is the 1-dimensional phenotypic representation of a set of training patients, and X represents the genotypes of these training patients.

An alternative *regularized* version of least squares is Lasso (Tibshirani, 1996), which uses an additional penalty factor on the regression weights to enforce robustness against noise in the explanatory variables:

$$\hat{\beta} = argmin \, \|\mathrm{y} - \beta\mathrm{X}\|_2^2 + \lambda \, \|\beta\|_1 \tag{6}$$

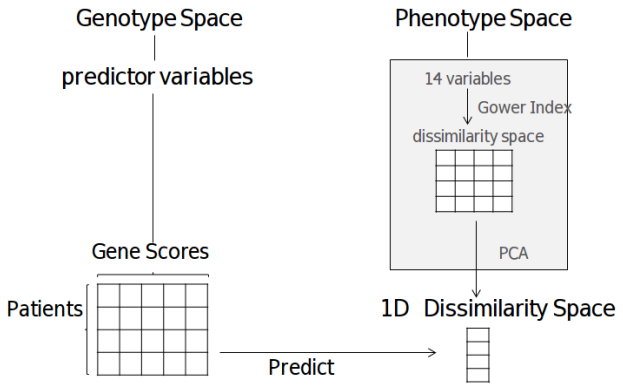



Fig. 1: Genotype - Phenotype association

Lasso regularization forces $\beta$'s to be zero when the explanatory variables ($x_i$'s) are not important for the dependent variable (y), implicitly realizing a variable selection. The amount of variables selected is controlled by the regularization parameter .

In case of highly correlated predictors, Lasso randomly selects one of them and discards the other by setting the coefficient to zero. Elastic net (Zou and Hastie (2005)) allows for grouped selection by inducing a grouping effect during variable selection which results in highly correlated variables to have similar ($\beta$) coefficients. This is realized by a mixture of lasso ($\lambda_1$) and ridge ($\lambda_2$) penalties:

$$\hat{\beta} = argmin \left\| \text{y - } \beta \text{X} \right\|_2^2 + \lambda_1 \left\| \beta \right\|_1 + \lambda_2 \left\| \beta \right\|_2^2 \qquad (7)$$

, where the $\lambda_1$ part of elastic net does variable selection, and the $\lambda_2$ encourages for grouped selection.

## 2.9 Non Linear Modeling

To model non linear genotype-phenotype associations we adopt decision tree regression. A decision tree builds a regression model in the form of a tree structure: it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node (gene) has two or more branches (e.g. gene score values), each representing values for the feature tested. Leaf node (phenotypes) represents a decision on the numerical target.The best split is the one minimizing the MSE of predictions compared to the training data. The final result is a tree with decision nodes and leaf nodes.

## 2.10 Dimensionality Reduction and Feature Selection

Selection of appropriate features is a common problem when fitting a model to data for which the number of features greatly exceeds the number of observations. This is exactly the case for our dataset which consists only of 400 patients, each one characterized by thousands features (genotypes). There are two main methods to deal with this problem: dimensionality reduction and feature selection (Janecek *et al.*, 2008). While in feature selection a subset of the features is extracted, in dimensionality reduction a new feature space is created by a combination of the original features. Both dimensionality reduction and feature selection techniques were applied to the genotypes and compared in terms of prediction performance.

To reduce the dimensionality we opted for two different approaches: *a. Dimensionality reduction by PCA*: a new feature space is constructed by linear transformations of the initial features. The principal components of the genotype space are determined and the $q$ first principal components, ranked on eigenvalues, are selected. *b. Dimensionality Reduction by Supervised PCA*: the same transformation as in (a) is applied only to a preselected number of features. Firstly, linear regression is performed to the whole genotype space and the features are ranked on their $\beta$ values. Then those features that have a $\beta > \theta$, where $\theta$ is a to be determined parameter, are chosen and PCA is performed only on them, afterwhich regression is done again and the $q$ first principal components ranked on their eigenvalues are selected. (Bair *et al.*, 2006)

For feature selection we tested three different methods: *a. Univariate feature ranking*: the features are ranked based on their correlation with the phenotypes. Then the *p* highest ranked features are chosen, i.e. those with the best correlation with the phenotype. *b. Lasso feature selection*: features are selected implicitly by setting the regularization term $\lambda_1$ in Eq.(6). c. *Elastic Net*:features are selected implicitly by setting the regularization term $\lambda_1$ and $\lambda_2$ in Eq. 7.

## 2.11 Parameter Estimation using Double-fold Cross Validation

Both the dimensionality reduction as well as feature selection methods have parameters that determine the complexity of the problem and thus need to be chosen depending on the problem at hand. The optimal parameter of a model is usually determined by cross-validation. In the simplest scenario, the number of patients is split into K equally sized parts, which all but one are used for training the model and the remaining one is used for validation. However in this set up, the samples used for validation are also used for model parameter selection and thus they are not completely independent.

To overcome this problem, we use double-fold cross validation which ensures that parameter selection is independent of the final validation set (Wessels *et al.*, 2005). Briefly (Fig. 2), the data is split in K1 equally sized parts, one is used for validation and the remaining ones are used for model construction. To construct the model, the data K1 (K1-1 parts) is again split in K2 parts. (K2-1) parts are used to train the model for a particular parameter setting $\theta$, which can be evaluated with the left out part (K2-2). Based on the performance estimates in the inner cross validation (K2) we can compare the performances of different parameter settings $\theta$. We then can choose the optimal setting in each inner fold (K2) as the one that minimizes the MSE and estimate the optimal parameter setting as the average of the minima. The performance of that optimal setting is evaluated using the outer cross validation (K1). Note that the parameter setting can be different for each of the outer folds (K1), yet evaluates a whole procedure of optimizing the parameter selection as done in the inner cross validation (K2). To arrive at a final model we can do the parameter selection on the whole data set (i.e. perform only one cross validation to optimize the parameter setting).

To give an example of the procedure consider univariate feature selection (Fig. 2). For each fold of the inner fold (K2), we do univariate feature ranking and then build models with an increasing number of best features in a given range [1,p]. For each parameter setting (*p*-best features) we can evaluate the model of choice (e.g. a linear regression, like in eq. 5). By the end of the inner fold, we obtain K2 different estimations for the performance of the model when selecting *p* features. For each one of the inner folds we detect the optimum number of features, $p_{opt}$, which minimizes the MSE and then we choose as the optimal setting $q$ to be the average of them. By the end of the inner fold, the model is trained with the optimal parameter setting, $q$, identified and the performance is evaluated using the outer fold validation data.

## 3 RESULTS

### 3.1 Annotation and SNP Subset Selection

The total number of SNPs detected in 400 early onset AD patients is 512.292 from which 498.754 are located on chromosomes 1 to 22. After the functional annotation of the variants (Fig. 3), an appropriate subset of SNPs, 35.729 SNPs, is finally selected according to the filtering criteria described in section 2.3.

Missing values of the remaining SNPs are imputed as indicated in section 2.6. The imputation resulted in several SNPs that are not mutated for all the patients and consequently we remove them also. Finally, resulting in 35.729 SNPs (Fig. 3) to be used for the subsequent genotype - phenotype analysis.

The distribution of the number of patients that harbor a unique SNP is depicted on Fig. 5 where it is obvious that several SNPs appear only in one or a few patients. These rare SNPs might not have predictive power on their own but when they mapped to genes with more that one mutations might contribute to the association score between genotype and phenotype thus eventually become interesting in combination with other SNPs.

### 3.2 SNPs to Genes

The 35.729 SNPs are mapped onto 11.867 genes. The distribution of the number of patients that harbor a unique gene being mutated by at least one SNP (Fig. 6) is highly similar to the distribution of the SNPs (Fig. 5). Due to the inclusion of rare SNPs there are 3.190 genes being mutated only once. It is also interesting to note that there are still a few genes that are mutated in the majority of the patients even if SNPs mutated in more than 90% of the patients have been removed.

Among the genes mutated, there are also genes that are related to AD according to previous studies. From a list of 114 genes known as probably related to AD, 67 of them were found to be mutated in the patients (Table: S2). However, in our patient group the majority of those genes are mutated just in a small subset of the patients and not in most of them as it would be expected(Fig. 6). This might be because we concentrate on early onset AD whereas previous studies focussed on a more inhomogeneous set of AD patients.

### 3.3 Outlier detection in genotype space

The genotype space is represented by gene scores (section 2.4). By constructing the dissimilarity matrix of the patients (section 2.7), we find that several patients are very dissimilar to all other patients, suggesting that these patients are outliers (Fig. 7a). After further inspection we found that these patients have a different ancestry (i.e. they themselves or their parents are non-Dutch), so that we decided to excluded them from our dataset as they may lead us to wrong conclusions. There were also 6 patients that gave identical gene scores but had completely different clinical measurements. Although we could not figure out the origin of this phenomena, we chose to remove one of the duplicates for each identical pair of patients detected. As a result of these outlier removals we were left with 358 patients.The dissimilarity scores between the patients are in a similar range (Fig 7b) which is what is expected because the patients are both Dutch and diseased with AD.

After the removal of the genotype-based outlier patients, there were 568 genes detected not to be mutated anymore. Hence, the number of genes dropped to 11.299.

### 3.4 Subtyping patients based on Gene Space

We are interested in investigating if there is any clustering of patients when considering their genotype only. The patients are clustered based on their gene scores using hierarchical clustering with complete linkage and as a distance measure the Euclidean distance.

Fig. 8a shows relative clear clustering of patients in our dataset when using the SNP counting gene score (eq. 1). Even if the different AD subtypes are distributed over all the clusters (Fig. 8b), it seems that there are subgroups of patients at least from a genotypical point of view. It also shows that there are two small clusters that include patients who have a large dissimilarity towards other patients. As these patient have a relatively large distance to all other patients and this is not associated with one of the clinical parameters, this might indicate that these patients also might have a different ancestry from the rest of the patients, although not so large as the previously removed outliers.

We also performed a clustering of the patients when considering the gene scoring that includes the damaging effect of the SNPs (eq. 2). In this case, the clustering is even more pronounced, see Fig. 9a, indicating two prominent large clusters. However, again patients of different AD subtype are distributed (Fig. 9b) all over the clusters. Also in this case there is a small cluster (with the same patients as before) that is dissimilar to all other patients (left most cluster in Fig 10), enforcing the realisation that these patients are genetically different from the rest and thus indeed might have a different ancestry. For this reason, we decided to also remove these 8 patients as probable outliers resulting in the number of patients to be 350.

Finally, we also tried to identify the genes that significantly contributed to the observed clusters (in both representations). To do so we first selected a number of clusters according to the hierarchical dendrogram. Then we associated individual gene scores with the detected clusters. However, there was no gene detected as being significantly correlated with that clustering. Hence, the detected clusters are not driven by a single gene but in a combinatorial way.

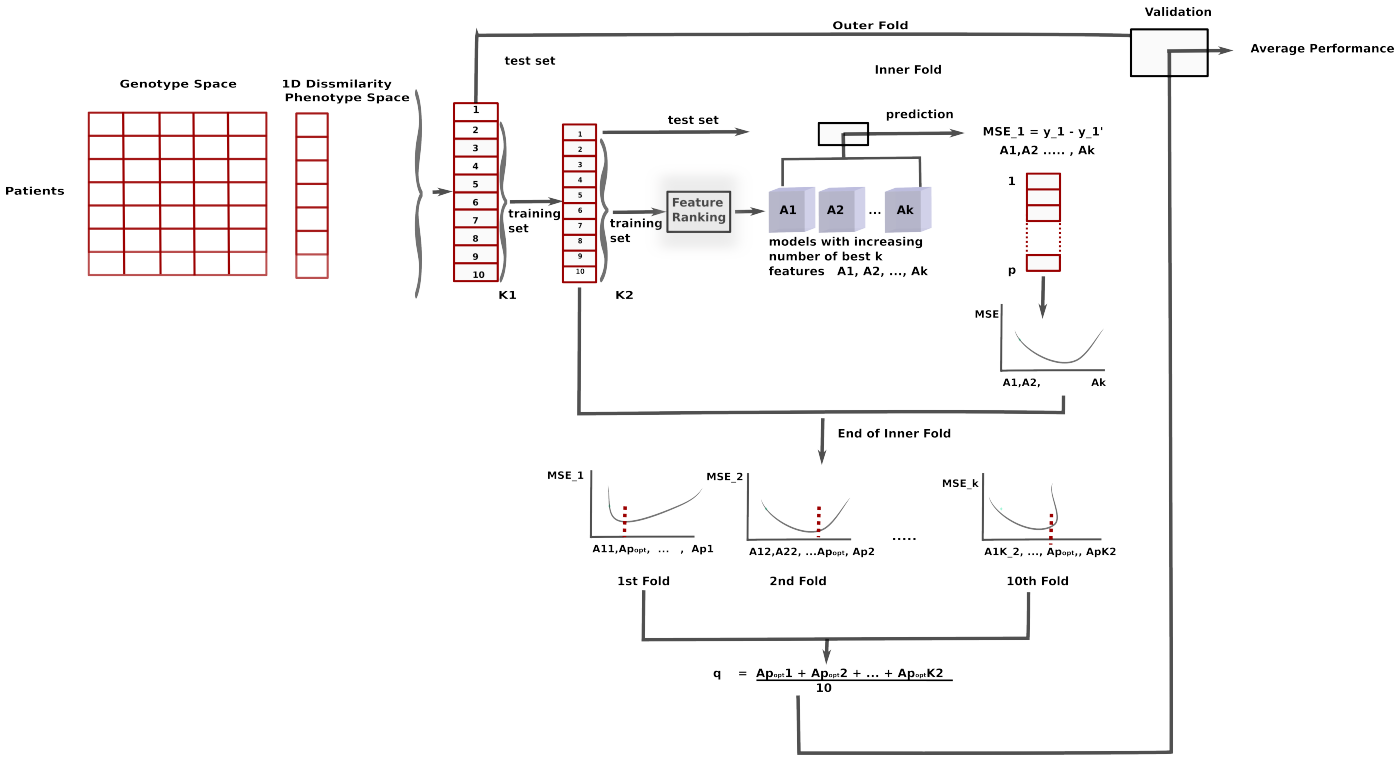



Fig. 2: Feature Selection - Double cross-validation Scheme

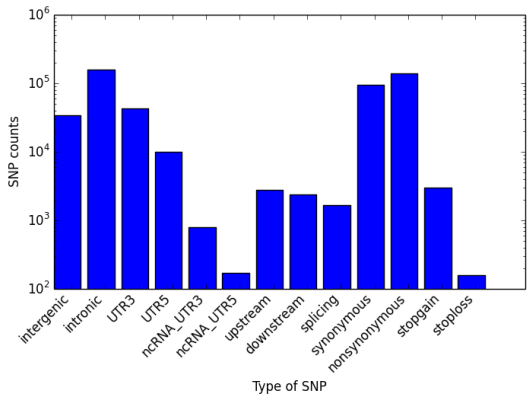



Fig. 3: Distribution of the number of SNPs according to their functional annotation using Annovar.

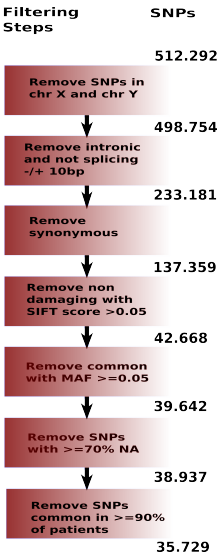



Fig. 4: Filtering steps in order to select a subset of probably disease causing SNPs. The number of the remaining SNPs are shown after each step.

## 3.5 Structure on the Clinical Data

We were also interested in investigating if the patients can be clustered in different groups based on the clinical examination tests. First, the dissimilarity matrix is computed using Gower Index (3), and then hierarchical clustering with complete linkage is performed. The resulting clustering is depicted in Fig. 10a.

Although only 96 out of the 350 patients are diagnosed with PPA, PCA and CAA, we observe a clear cluster yet not overlapping with the AD diagnosis (Fig. 10a). We also do notice that the different groups, even not well defined, are still detected in a two (and one) dimensional representation of the dissimilarity space (Fig. 10b). Further, we need to note that, there is always the risk that some of the clusters formed are related to the imputation of the missing values but we cannot avoid this. We can only eliminate the chance

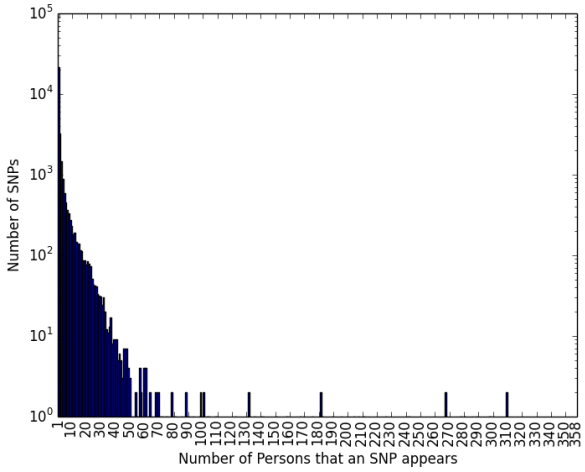



Fig. 5: Distribution of the number of patients with at least one unique SNP mutated

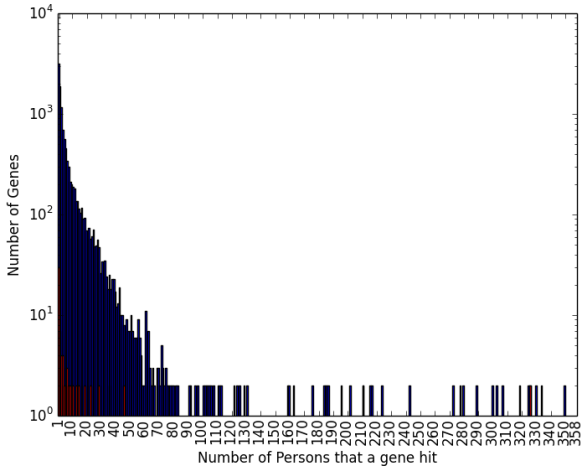



Fig. 6: Distribution of the number of patients that harbor a unique gene being mutated by at least one SNP (red indicates the probable related genes)

that to be the case by removing features having high percentage of missing values.

### 3.6 Predicting phenotype from genotype

Next we examined whether we can predict the observed phenotypes from the genotypes of the patients. For that we make use of either a linear regression, either without regularization or with LASSO or Elastic Net regularization (section 2.8), or a decision tree regression (section 2.9). To determine the proper parameter settings for these models we make use of a double fold cross-validation as indicated in section 2.11. Predictive accuracy is expressed in terms of the pearson correlation of the predicted phenotypic value with the actual value over the test set, or the mean square error between the real and

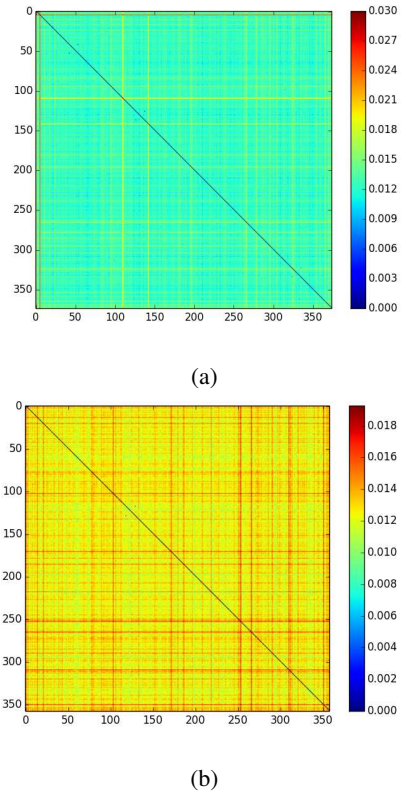



Fig. 7: **Dissimilarity representation of genotypes**: **a**: Patients with outliers. **b**: Patients after removing outliers.

predicted values. Different gene scoring methods (section 2.4) to represent the genotype were tested under the same model settings. Experiments were performed both in the initial genotype space and in the dissimilarity genotype space (section 2.7).

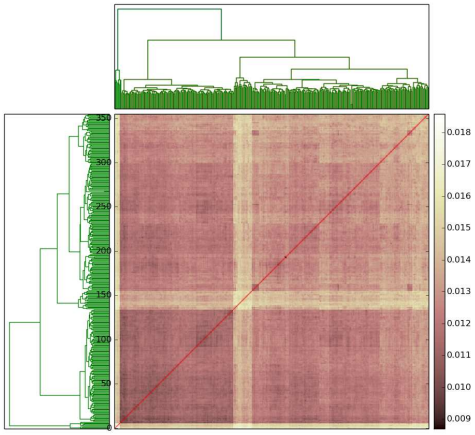
In terms of predictive performance, there was no model that resulted in a really high predictive accuracy. Nevertheless we see that the gene scoring method based on summing SIFT scores gives a better performance than just counting the SNPs. This indicates that the effect of SNP does contribute to the predictive power of a SNP. Furthermore the LASSO and Elastic Net that make use of an implicit feature selection method seem to outperform the other methods as well, hinting towards a better strategy to select the predictive genes.

The average number of features (or dimensions) used to construct the predictive model is relatively low (in the tens with respect to the initial 11.299 genes) with Lasso and Elastic net resulting in the fewest number of features.

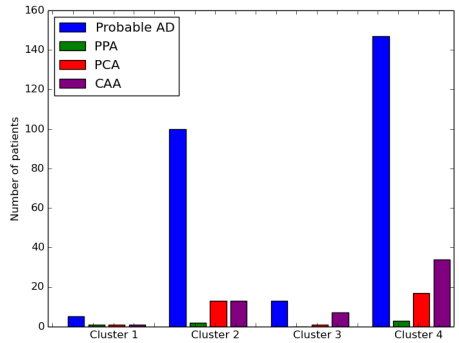
On the one hand, by performing simple linear and decision tree regression with univariate feature selection, we did not observe any consistency between the top-ranked features (genotypes) that were discovered within each fold. On the other hand, among the top features selected by Lasso and Elastic net, three of them, HLA-DRB5, HLA-DRB1 and DEFB119, are always selected, even regardless of chosen gene scoring methodology. None of these features are selected in the case of univariate feature selection. So, although these genes are not useful individually, they are very useful when combined with others.

**Table 2.** Overview of the different experimental setups tested. Performance is estimated in terms of MSE and correlation between real and predicted values.

| Genotype Representation | Prediction method | Feature Selection | | | PCA | | | Supervised PCA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | R | Avg of feat. | MSE | R | Avg of dim. | MSE | R | Avg of dim. |
| Scoring genes by counting SNPs (eq.1) | Linear Regression | 0.2918 | 0.0758 | 77 | 0.2289 | 0.1405 | 39 | 0.2725 | 0.1132 | 52 |
| | Decision Tree Regression | 0.2567 | 0.092 | 68 | 0.3740 | 0.1097 | 44 | 0.2612 | 0.1276 | 55 |
| | Lasso | 0.2179 | 0.1422 | 5 | | | | | | |
| | Elastic Net | 0.214 | 0.18 | 13 | | | | | | |
| Scoring genes by summing SIFT scores of SNPs (eq. 2) | Linear Regression | 2.5238e+24 | 0.1575 | 138 | 0.2249 | 0.1202 | 37 | 0.2132 | 0.1486 | 42 |
| | Decision Tree Regression | 0.2971 | 0.1344 | 186 | 0.3694 | 0.1518 | 45 | 0.2534 | 0.1543 | 58 |
| | Lasso | 0.2165 | 0.20190 | 7 | | | | | | |
| | Elastic Net | 0.214 | 0.17 | 16 | | | | | | |

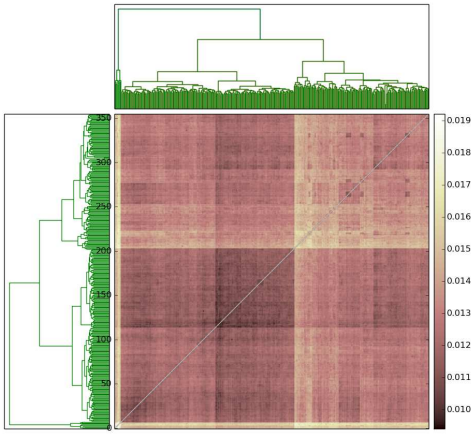


(a)

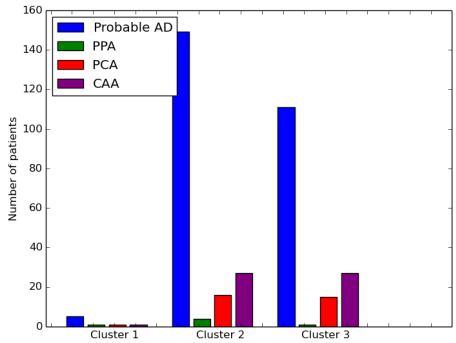



(b)

Fig. 8: **Dissimilarity representation of genotypes using gene scoring of eq. 1**: **a**: Clustering of the genotype space based on counting SNPs (the color shows the euclidean distance of the genes scores according to Eq. 1) **b**: Frequency of the patients' subtypes per cluster (after cutting the tree to 4 clusters)



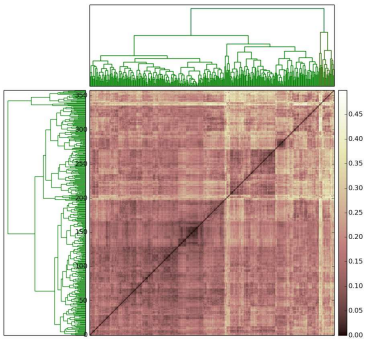
(a)



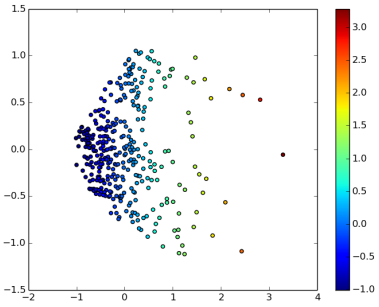

(b)

Fig. 9: **Dissimilarity representation of genotypes using genes score of eq. 2**: **a**: clustering of patients based on counting SIFT scores (the color shows the euclidean distance of the genes scores according to Eq. 2) **b**: Frequency of the patients' subtypes per cluster (after cutting the tree to 3 clusters)

These findings are supported when we use dimensionality reduction to lower the complexity of the regression models. When inspecting the loading factors of the chosen principal components (that indicate the relative importance of the genes in defining these components), we find two highly ranked genes within almost all the



(a)



(b)

Fig. 10: **Clustering of the phenotype space**: **a**: Clustering of the dissimilarity space(color indicates the Gower dissimilarity between patients). **b**: The dissimilarity space projected onto two first principal components colored based on 1D projection of the dissimilarity space.

folds, which are HLA-DRB5 and HLA-DRB1 being in agreement with the results obtained by Lasso and Elastic net.

It is especially interesting that the HLA genes are known to be related to late onset AD in other studies (Lambert *et al.*, 2013). Using STRING (Szklarczyk *et al.*, 2011) we also found that the HLA genes interact with CD74 gene that interact with APP that acts as a suppressor of amyloid-$\beta$ (Fig. 11). On the other hand, DEFB119 is not related to HLA but it is find to be related to meningioma.

We further investigated the performance of the HLA-DRB1, HLA-DRB5 and DEFB119 genes in different combinations. Remarkably we see a small increase in performance when we use all 3 genes when compared to the previous results (Table 4). The highest performance is observed when all 3 genes form the genotype space, indicating that all three genes together are needed to be most informative, again supporting the combinatorial nature of AD (Table 4). The next highest performance is observed when we use a combination of DEFB119 and one of the HLA genes, leading us to conclude that DEFB119 might indeed be related to early onset AD.

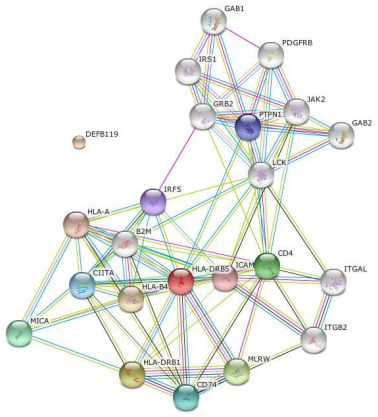



Fig. 11: HLA-DRB1, HLA-DRB2 and DEFB119 interactions

## 4 DISCUSSION AND FUTURE WORK

We investigated if genotypes can be associated with phenotypes of patients diagnosed with early onset AD in order to identify genetic variants that could explain the phenotypic variation among them. We also explored the existence of any possible structure both in phenotypic and genotypic level that may indicate possible new AD subtypes. To identify subgroups of early onset AD patients, we clustered the patients based on genotypes and phenotypes independently. In both cases, we observed a clustering. We could, however, not find any genetic or phenotypic trait that associates well with these clusters. This indicates that there is not one variable explaining the clustering, but that a combination of variables is necessary to define these subtypes.

To adress the problem of associating genotypes to phenotypes, we used several regression methods. A common part of all the approaches is the transformation of the multidimensional phenotype space to a one-dimensional dissimilarity space. As it is not known if there are more subgroups of patients apart from PPA, PCA and CAA, we combined all the phenotypes into a dissimilarity space ensuring that we make use of all the information available and hidden structures may be revealed. Through the regression methods used, we tried to detect the genotypes which best predict the phenotypes as a whole. Dimensionality reduction and feature selection techniques were applied to the genotype space in order to select the most appropriate features (or dimensions) that can explain the phenotypes. From a performance point of view, none of the methods resulted in high predictive accuracy. Non-linear regression methods, such as decision trees regression, have also been tested in case that the low performance is caused by the existence of non linear dependencies, but that did not improve the performance. However, it is particularly interesting that the majority of the methods tested, identified the same subset of genes, HLA-DRB1, HLA-DRB5 and DEFB119, as top ranked genes with consistency within all folds. Both HLA-DRB1 and HLA-DRB5 genes are of great importance as they interact with CD74 gene which in turn interacts with APP that acts as a suppressor of amyloid-$\beta$. Our findings about HLA-DRB1 and HLA-DRB5 are also supported from other studies (Lambert *et al.*, 2013). On the other hand, there is no much evidence for DEFB119 gene (being

**Table 3.** Performance on Dissimilarity Genotype Space

| Initial Representation | Regression | Dissimilarity | MSE | R |
|---|---|---|---|---|
| Gene scores(1) | Linear Regression | euclidean | 0.6579 | 0.1732 |
| | | city block | 3.0762 | 0.0928 |
| | | correlation | 0.2984 | 0.1414 |
| | | cosine | 0.3001 | 0.1450 |
| Gene scores(2) | Linear Regression | euclidean | 0.2902 | 0.1438 |
| | | city block | 0.9617 | 0.1022 |
| | | correlation | 0.2925 | 0.1352 |
| | | cosine | 0.2926 | 0.1383 |

**Table 4.** Model Performance when only HLA-DRB5,HLA-DRB1,DEFB119 genes (or just a combination of them) form the genotype space

| Genes | MSE | R |
|---|---|---|
| HLA-DRB5,HLA-DRB1,DEFB119 | 0.20 | 0.29 |
| HLA-DRB5,HLA-DRB1 | 0.21 | 0.19 |
| HLA-DRB5,DEFB119 | 0.2 | 0.26 |
| HLA-DRB1,DEFB119 | 0.20 | 0.25 |
| HLA-DRB5 | 0.21 | 0.17 |
| HLA-DRB5 | 0.21 | 0.12 |
| DEFB119 | 0.2 | 0.23 |

involved in meningioma, a brain carcinoma) and AD. Nevertheless, by conducting experiments using only these 3 genes and different combinations of them, we do notice that the performance was slightly increased when all 3 genes were used. That indicates that DEFB119 gene may be also of some importance and further investigation is needed. To improve prediction performance and reduce noise in the data, we also opted for an alternative method in which transform the phenotypic space to a new space (5.5). The predictive accuracy was improved but there was no consistency in the genes selected.

Although the results are already interesting by themselves, we would like to emphasize a few notes with respect to the chosen experimental setup and analyses. First, it is really important to include a control group in our analysis. As mentioned, several SNPs that appear in more than 90% of the patients have been removed so as to eliminate mutations biased towards Dutch population (2.3). This step seems reasonable in the current setup but there is a risk of also removing SNPs that are predictive for early onset AD. Consequently, the only way to overcome this problem is to selectively remove SNPs based on a Dutch control group.

Another problem is that the phenotypic information is quite incomplete, in fact there is no measurement for which we have measured values for all patients. In order to deal with this and to be able to impute the missing values, we had to remove a lot of clinical measurements with high percentage of missing information. So, again there is a risk of discarding phenotypic features that may be both descriptive about the disease and strongly correlated with the genotypes. Apart from this, the large number of imputation that we had to do could be a cause of detecting clusters which do not really exist. An option might be to ”remove” the imputation. That is to remove much more data in case there is missing data. Then see whether the data still clusters ”in the same way”. If that is the case than it is not related to the imputation.

Another possible improvement may be related to the gene scoring. An alternative would be to create a score per pathway instead of a score per gene. In this way, the features of the genotype space will be reduced and the prediction task will become easier.

As a last remark, we have also to mention that predicting phenotypes from genotypes is not an easy task. One reason for this is that people may have the same mutation but not necessarily the same phenotypes. In other words, genetics is not the only influence on phenotypic variation: the environment, life history, and many other factors also impact on phenotypic variation.

## REFERENCES


Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101**(473).

Barzi, F. and Woodward, M. (2004). Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, **160**(1), 34–45.

Batista, G. and Monard, M. C. (2001). A study of k-nearest neighbour as a model-based method to treat missing data. In *Proceedings of the Argentine Symposium on Artificial Intelligence*, volume 30, pages 1–9.

Berchtold, N. C. and Cotman, C. W. (1998). Evolution in the conceptualization of dementia and alzheimers disease: Greco-roman period to the 1960s. *Neurobiology of aging*, **19**(3), 173–189.

Bertram, L. and Tanzi, R. E. (2008). Thirty years of alzheimer's disease genetics: the implications of systematic meta-analyses. *Nature Reviews Neuroscience*, **9**(10), 768–778.

Bertram, L., Lange, C., Mullin, K., Parkinson, M., Hsiao, M., Hogan, M. F., Schjeide, B. M., Hooli, B., DiVito, J., Ionita, I., *et al.* (2008). Genome-wide association analysis reveals putative alzheimer's disease susceptibility loci in addition to apoe. *The American Journal of Human Genetics*, **83**(5), 623–632.

Bird, T. D. (2008). Genetic aspects of alzheimer disease. *Genetics in Medicine*, **10**(4), 231–239.

Charidimou, A. and Werring, D. J. (2011). Cerebral microbleeds: detection, mechanisms and clinical challenges. *Future Neurology*, **6**(5), 587–611.

Chatfield, C. and Collins, A. (1981). *Introduction to multivariate analysis*, volume 1. CRC Press.

Cox, T. F. and Cox, M. A. (2000). A general weighted two-way dissimilarity coefficient. *Journal of Classification*, **17**(1), 101–121.

Devi, G., Fotiou, A., Jyrinji, D., Tycko, B., DeArmand, S., Rogaeva, E., Song, Y.-Q., Medieros, H., Liang, Y., Orlacchio, A., *et al.* (2000). Novel presenilin 1 mutations associated with early onset of dementia in a family with both early-onset and late-onset alzheimer disease. *Archives of neurology*, **57**(10), 1454–1457.

Duin, R. P., Loog, M., Pkalska, E., and Tax, D. M. (2010). Feature-based dissimilarity space classification. In *Recognizing Patterns in Signals, Speech, Images and Videos*, pages 46–55. Springer.

Ertekin-Taner, N. (2007). Genetics of alzheimer's disease: a centennial review. *Neurologic clinics*, **25**(3), 611–667.

G, J., AC, N., GW, B., and et al (2010). Meta-analysis confirms cr1, clu, and picalm as alzheimer disease risk loci and reveals interactions with apoe genotypes. *Archives of Neurology*, **67**(12), 1473–1484.

Ghiso, J. and Frangione, B. (2001). Cerebral amyloidosis, amyloid angiopathy, and their relationship to stroke and dementia. *Journal of Alzheimer's Disease*, **3**(1), 65–73.

Janecek, A., Gansterer, W. N., Demel, M., and Ecker, G. (2008). On the relationship between feature selection and classification accuracy. In *FSDM*, pages 90–105. Citeseer.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at ucsc. *Genome research*, **12**(6), 996–1006.

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., *et al.* (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, **45**, 1452–1458.

Leifer, B. P. (2003). Early diagnosis of alzheimer's disease: clinical and economic benefits. *Journal of the American Geriatrics Society*, **51**(5s2), S281–S288.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models.

Ng, P. C. and Henikoff, S. (2003). Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, **31**(13), 3812–3814.

Stopford, C. L., Snowden, J. S., Thompson, J. C., and Neary, D. (2008). Variability in cognitive presentation of alzheimer's disease. *Cortex*, **44**(2), 185–195.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., *et al.* (2011). The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, **39**(suppl 1), D561–D568.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

von Gunten, A., Bouras, C., Kövari, E., Giannakopoulos, P., and Hof, P. R. (2006). Neural substrates of cognitive and behavioral deficits in atypical alzheimer's disease. *Brain Research Reviews*, **51**(2), 176–211.

Wang, K., Li, M., and Hakonarson, H. (2010). Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, **38**(16), e164.

Wessels, L. F., Reinders, M. J., Hart, A. A., Veenman, C. J., Dai, H., He, Y. D., and van't Veer, L. J. (2005). A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**(19), 3755–3762.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.

## 5 SUPPLEMENTARY FIGURES AND TABLES

**Table S1.** Variant Annotation

| Annotation | Explanation |
|---|---|
| exonic | variant overlaps a coding region |
| splicing | variant within a splicing junction |
| intronic | variant overlaps an intron |
| intergenic | variant is in intergenic region |
| ncRNA | variant overlaps a transcript without coding annotation in the gene |
| UTR5 | variant overlaps a 5' untranslated region |
| UTR3 | variant overlaps a 3' untranslated region |
| upstream | variant overlaps 1-kb region upstream of transcription start site |
| downstream | variant overlaps 1-kb region downstream of transcription end site |
| nonsynonymous | a single nucleotide change that cause an amino acid change |
| synonymous | a single nucleotide change that does not cause an amino acid change |
| stopgain | a nonsynonymous variant that lead to the immediate creation of stop codon at the variant site. |
| stoploss | a nonsynonymous variant that lead to the immediate elimination of stop codon at the variant site |

### 5.1 Probable related genes detected mutated to the patients

A list of 114 genes (given by Henne Holstage) that may be involved in AD was also used in our analysis. Only 67 of those genes were detected to be mutated on the patients.

**Table S2.** List of probable related genes found mutated on patients

| Gene List | | | |
|---|---|---|---|
| OR51A4 | ATP2C2 | ABCA7 | FERMT2 |
| NOTCH3 | PSEN1 | NME8 | APOE |
| TREM2 | CST3 | TGFB1 | CD33 |
| TTR | PTK2B | FOXP2 | EIF2AK4 |
| BIN1 | SORL1 | PSEN2 | EPHA1 |
| CMIP | NCSTN | PRND | EIF2AK3 |
| CNTNAP2 | CSF1R | MAPT | HLA-DRB1 |
| CD2AP | HLA-DRB5 | CR1 | CHMP2B |
| FUS | PLD3 | ADAM10 | LRRTM3 |
| CASP9 | MCL1 | TNFRSF1B | BCL2L11 |
| CASP10 | CASP3 | CASP6 | BAK1 |
| RIPK1 | BNIP3L | TNFRSF10B | TNFRSF10D |
| BNIP3 | CASP7 | BAD | BIRC2 |
| APAF1 | BCL2L14 | CRADD | DIABLO |
| LRP1 | MOAP1 | AVEN | TRADD |
| TP53 | BBC3 | CALR | BCL2L1 |
| BIK | APP | ATF4 | ATF6 |
| CASS4 | | | |

### 5.2 Clustering of patients on genotype space when mutation on chrX and chrY are included

In Fig. S1 is depicted the clustering of the patients when mutations detected in both chrX and chrY are included. In this case there are two dominant clusters formed which are related to the sex of the patient (male-female). To avoid a sex related bias, we remove them.

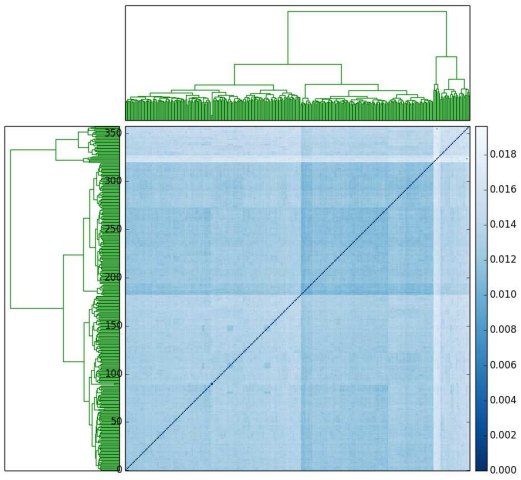



Fig. S1: Clustering of Genotype Space when chrX and chrY are included

### 5.3 Clustering of phenotype space using all the 35 remaining measurements after the removing those having high percentage of missing values

Clustering of the phenotype space (Fig. S2) when no clinical measurements are excluded beforehand apart from those having high percentage of missing values or those that are not relevant (i.e. date of first visit, MRI type, etc.). In the final analysis, several measurements have been removed (mainly brain related measurements) due to argumentation. This selection resulted in a better clustering (10a) compared to the clustering with all the measurements (Fig. S2).

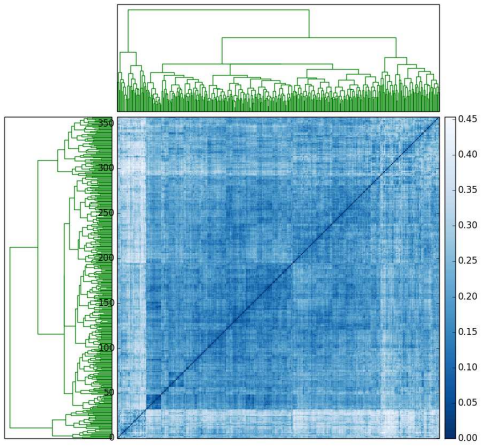



Fig. S2: Clustering of the phenotype space with all the measurements

## 5.4 List of all clinical examination measurements

**Table S3.** All the clinical features measured. Clinical data in green indicate the finla selected measurements that selected to be used in our method heightClinical Feature Description

| | |
|---|---|
| I_dat_1 | date of initial screening |
| I_fam_dem | dementia in family |
| I_fa_psy | psychiatry in family |
| I_die | deceased |
| I_die_dat | date deceased |
| I_die_re | cause of death |
| I_live | last date checked alive |
| I_live | last date checked alive |
| obd_text | text autopsy |
| V_dat | date of test |
| duur van klachten | duration of symptoms |
| VaD | vascular dementia |
| FTLD | frontotemporal lobar degeneration |
| age_of_ons | age of onset |
| V_CAMCOG | cognitive test |
| V_MMSE | cognitive test, especially for memory |
| V_FAB | cognitive test, especially for executive functioning |
| V_GDS | Geriatric Depression Scale |
| V_GDS | Geriatric Depression Scale |
| D_diag | diagnosis |
| D_dat | date of diagnosis |
| D_txt | diagnosis free text |
| ApoE | ApoE genotype |
| L_DATEAFN | liquor date decrease |
| L_MNR | liquor material number |
| L_AB42 | amyloid-beta |
| L_TAU | total tau |
| L_PTAU | phosphorylated tau |
| M_dat | date of MRI test |
| M_scan | type of MRI scan |
| M_MTA_R | Medial temporal lobe (atrophy) right |
| M_MTA_L | Medial temporal lobe (atrophy) left |
| M_parietal_R | parietal (atrophy) right |
| M_parietal_L | parietal (atrophy) left |
| M_atrophy | global atrophy |
| M_Fazekas | a measure of white matter abnormalities |
| M_lacune | holes in tissue |
| M_inf_jn | infarcts |
| M_mbl_to | microbleeds in brain |
| M_txt_ra | radiologist free text |
| N_CRvoor | Neuropsychology, number sequence forward |
| N_CRach | digit sequence backwards |
| N_LDST_90 | etter digit in 90 seconds |
| N_VATA1, N_VATA2 | visual association test A1 and visual association test A2 |
| N_15WT1, N_15WT2, N_15WT3, N_15WT4,N_15WT5 | different word task tests |
| N_15WTui | word Task delayed reproduction |
| N_FL60 | NPO_Fluency in 60 sec |
| N_FLD | NPO_Fluency animal names |
| N_FLA | NPO_Fluency A |
| N_FLT | NPO_Fluency T |
| N_TMTAt | trail making test A |
| N_TMTBt | trail making test B |
| N_NumLoc | NPO_number location |
| N_NumObjDec | NPO_object detection |
| N_Reyuit | NPO_figure of Rey delayed reproduction |
| N_N_ideationalprx | ideational apraxia |
| N_N_ideomotorprx | ideomotore apraxia |
| PPA | primary progressive aphasia |
| PCA | posterior cortical atrophy |
| CAA | cerebral amyloid angiopathy |
| MCI | mild cognitive impairment |
| probable AD | probable Alzheimer disease |
| possible AD | possible Alzheimer disease |
| tau ratio | ratio of tau and amyloid-beta |

## 5.5 Alternative Aproach: Dimensionality Reduction and Supervised output transormation

To control the noise in the data we propose the transformation into a new space that we are able to predict.

*5.5.1 'Preconditioning' variable selection* In each one of the inner folds(2.11) the features that are weakly correlated with the phenotypes above a threshold $\theta$, are selected for further transformations. Specifically, the threshold $\theta$ is set to be 0.1. This is not a high correlation. However for our dataset the highest correlation that is detected is approximately 0.3. By setting the threshold $\theta$ higher than 0.1, there is a risk of detecting no correlated features within the fold. For this reason, we set the correlation threshold to be the low aiming at increasing it in the transformed space.

*5.5.2 Transforming Output based on the selected features* After selecting a subset of features in each inner fold 5.5.1, we use these features to create a new phenotype space. If we plot the correlation of each gene with the phenotypes, we notice that each gene score for a specific gene represents a variety of different phenotypes(Fig. S3). The fact that for one gene score there is wide range of phenotypes, results in a difficult prediction task. For this reason, we reconstruct the phenotype space into N dimensions by following a supervised approach. At this point, the one-dimensional phenotypic space is transformed to N-dimensions, where N equals to the number of genes selected. So,

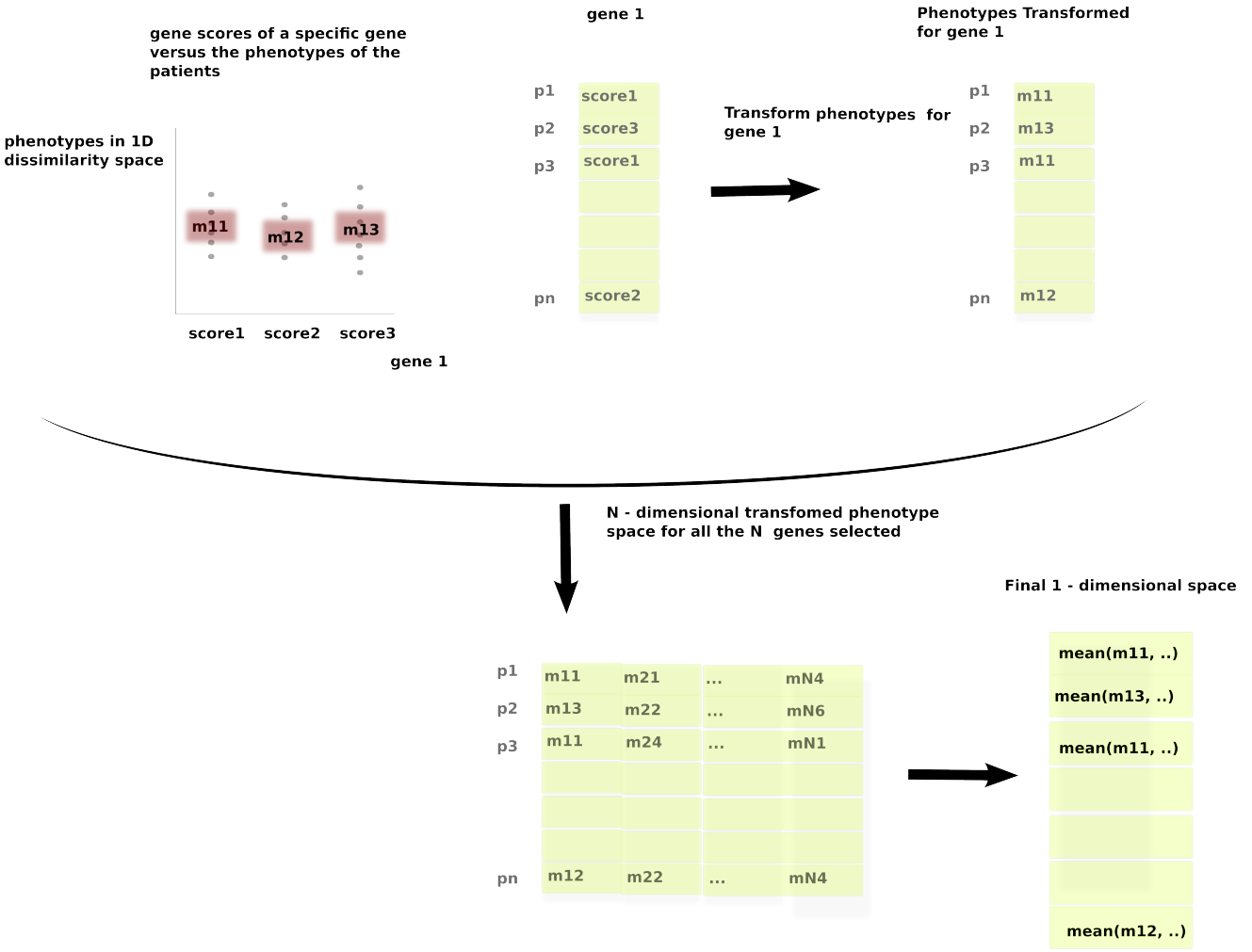



Fig. S3: Supervised Output Transformation. Assuming that we have one gene represented by 3 scores, Then for each score, we compute the mean of the phenotypic values correspond to this score. Afterwards, we build the 1st dimension of the new space. For each patient, the first dimension is setted to be the mean of the phenotypic values for the specific score for the particular gene. The procedure repeated for all the selected genes. At the end, we compute the phenotype of each patient to be the mean of all the features for this patient.

each dimension is reconstructed based on a specific gene. Specifically, for each patient, each dimension in N is setted to be the mean of the phenotypic values for the specific score for the particular gene. By repeating this procedure for each gene, a new phenotype space of N dimensions is formed. At the end, we transform this space to 1 dimensional space by taking the mean for each patient. The test set is transformed based on the transformation found in training set. In case that a gene score appears in the test set but not in the training set, the transformation is based on the first closest gene score (Fig. S3). The best transformation found in the inner fold is the one applied to the outer fold

Due to the fact that the number of features is still high, we follow the same procedure as in (2.10,2.11), in order to reduce the dimensionality and to succeed in predicting the new genotypic space. The new space cannot be easily converted back to the initial phenotype space. But the fact that we know the initial representation of the initial space can lead us to identify the closest phenotypic measurements of the uknown sample patient.

*5.5.3 Results* When genotype space is represented by 1, the average number of principal components predicted is 71, the MSE is 0.143 and the correlation between the real and predicted values is 0.25. For the gene scoring of (2), the number of components selected equals to 45 and the correlation between real and predicted values to 0.42. Nevertheless, there is no consistency in the features selected within the folds.