

# SCIENTIFIC REPORTS



OPEN

## Leveraging existing data sets to generate new insights into Alzheimer's disease biology in specific patient subsets

Received: 20 October 2014

Accepted: 24 August 2015

Published: 23 September 2015

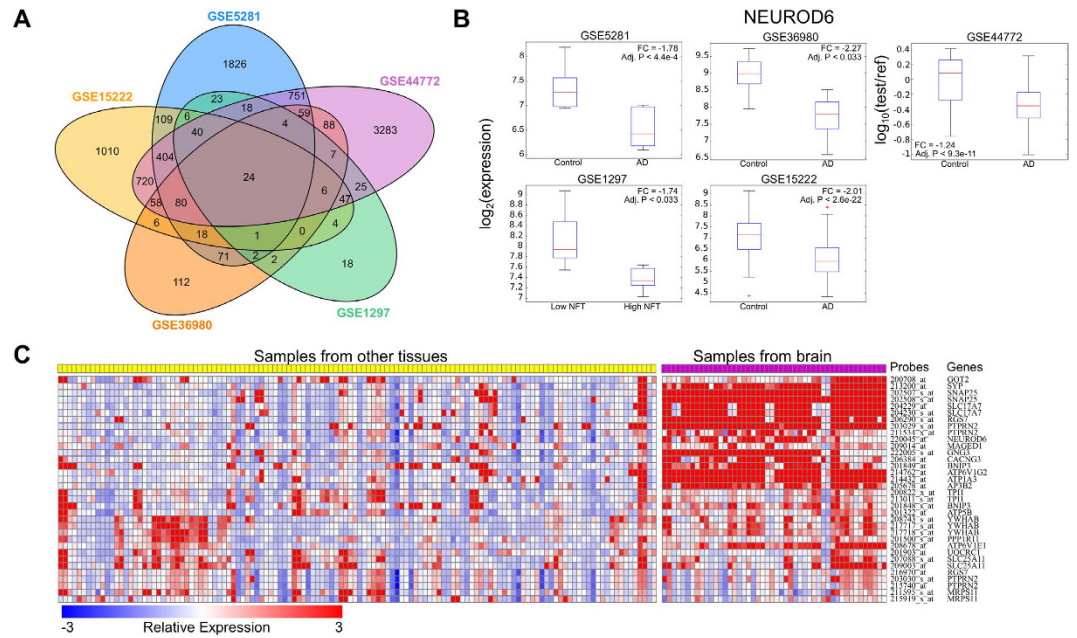
Kevin D. Fowler<sup>1,\*</sup>, Jason M. Funt<sup>1,\*</sup>, Maxim N. Artyomov<sup>1,2</sup>, Benjamin Zeskind<sup>1</sup>, Sarah E. Kolitz<sup>1,\*</sup> & Fadi Towfic<sup>1,\*</sup>

To generate new insights into the biology of Alzheimer's Disease (AD), we developed methods to combine and reuse a wide variety of existing data sets in new ways. We first identified genes consistently associated with AD in each of four separate expression studies, and confirmed this result using a fifth study. We next developed algorithms to search hundreds of thousands of Gene Expression Omnibus (GEO) data sets, identifying a link between an AD-associated gene (NEUROD6) and gender. We therefore stratified patients by gender along with APOE<sub>4</sub> status, and analyzed multiple SNP data sets to identify variants associated with AD. SNPs in either the region of NEUROD6 or SNAP25 were significantly associated with AD, in APOE<sub>4</sub>+ females and APOE<sub>4</sub>+ males, respectively. We developed algorithms to search Connectivity Map (CMAP) data for medicines that modulate AD-associated genes, identifying hypotheses that warrant further investigation for treating specific AD patient subsets. In contrast to other methods, this approach focused on integrating multiple gene expression datasets across platforms in order to achieve a robust intersection of disease-affected genes, and then leveraging these results in combination with genetic studies in order to prioritize potential genes for targeted therapy.

In recent years, many investigators have thoughtfully applied genetics and genomics approaches to investigate the biology of Alzheimer's Disease (AD)<sup>1–3</sup>. These efforts have yielded a rich collection of gene expression and single-nucleotide polymorphism (SNP) data sets, along with extensive analyses of particular data sets. The availability of such studies provides the opportunity to generate fresh insights into the biology of AD, independently of prevailing hypotheses, by integrating existing data sets in novel and innovative ways.

A number of studies have examined ways to do this. For example, Krauthammer *et al.* used seed genes of known importance in AD to identify additional candidate genes using genetic linkage and text-mined protein-protein interaction data via a graph-theoretic method<sup>4</sup>. Chen *et al.* reported a method to rank AD-related genes by importance based on database protein-protein interaction data<sup>5</sup>. Liu and colleagues interpreted genomic and proteomic data using a Bayesian statistical framework with the aim of prioritizing candidate genes, and found that this approach was able to identify known candidate genes for Alzheimer's<sup>6</sup>. Soler-López *et al.* utilized an initial list of known AD genes to identify additional genes of interest based on combining protein-protein interaction data with criteria of AD-associated genomic locations or changes in gene expression<sup>7</sup>. Caberlotto *et al.* obtained a list of seed genes from a gene expression dataset, SNP data, as well as genes previously identified as potential AD drug targets, and

<sup>1</sup>Immuneering Corporation, Cambridge, Massachusetts, United States of America. <sup>2</sup>Department of Immunology and Pathology, Washington University in St. Louis, St. Louis, Missouri, United States of America. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to S.K. (email: skolitz@immuneering.com)



**Figure 1. Genes downregulated in AD consistently across 5 data sets.** (a) Venn diagram illustrating the intersection of significantly downregulated genes across multiple data sets. (b) Box plots of NEUROD6 expression in each of the 5 data sets. (c) Heat map of the 24 consistently downregulated genes showing specificity for brain tissue.

used database protein-protein interaction data to investigate potential underlying biology represented among these genes<sup>8</sup>.

These methods all utilized protein-protein interaction data to help predict which genes might be important in disease. By contrast, the approach reported here uses intersections across multiple datasets to filter for more robust candidate genes, obtaining an intersection of genes from expression datasets that enables targeted mining of SNP data to identify those candidate genes more likely to be causal.

Methodologies for combining heterogeneous data sets to identify robust biological signals are not well established, particularly in the area of integrating gene expression and SNP data from separate cohorts. Differentially expressed genes alone are of limited utility, since they include both downstream signals resulting from disease pathology and upstream signals that may be more causative. Incorporating SNP signals into an analysis can help identify causative signals that may represent more direct targets for new medicines.

In conducting such integrative analyses, it is also critical to consider patient stratification. Today, studies of many different diseases are increasingly finding subsets of patients with distinct patterns of biology<sup>9,10</sup>. It is plausible that not all AD patients have identical mechanisms driving their common symptoms and manifestations of the disease. To the extent that AD patients may differ in certain aspects of the biology underlying their disease, it stands to reason that certain medicines may be more, or less, effective in particular subsets of patients.

Therefore, we sought to integrate publicly available gene expression and SNP data sets as a means to generate new insights into the biology of AD, stratify these patients, and generate hypotheses for new subset-specific medicines.

## Results

**Differential expression analysis between AD patients and healthy controls.** We first performed an integrated analysis of existing Alzheimer's gene expression data sets. We identified genes with significantly differing expression levels between healthy controls and AD patients (as defined by overall diagnosis or NFT score) in each of four data sets. Taking the intersection, 25 genes were downregulated significantly with disease in all four data sets. We subsequently obtained a fifth data set (Zhang *et al.*), and observed that in this study 24 out of these 25 genes were also downregulated significantly with disease. We thus identified 24 genes that were significantly downregulated with disease in each of 5 data sets (Fig. 1a and Table 1). Box plots of NEUROD6 (Fig. 1b) and SNAP25 (Supplementary Fig. 1) illustrate the consistent downregulation of Table 1 genes across each of the 5 data sets. Additional details on fold changes and p-values for the 24 genes appear in Supplementary Table 1 (lists of genes differentially expressed in 4 out of the 5 datasets are also provided, in Supplementary Table 8). While establishing such a stringent criterion may eliminate some relevant genes, we reasoned that the resulting genes would be unambiguously associated with AD. The intent of taking the intersection between multiple data sets, each

Downregulated in AD vs control in all 5 data sets			
AP3B2	C14orf132	MRPS11	SLC25A11
ATP1A3	C14orf2	NEUROD6	SNAP25
ATP5B	CACNG3	PPP1R11	SYP
ATP6V1E1	GNG3	PTPRN2	TPI1
ATP6V1G2	GOT2	RGS7	UQCRC1
BNIP3	MAGED1	SLC17A7	YWHAB

**Table 1.** List of 24 genes downregulated in AD consistently across 5 data sets.

imperfect and bearing its own peculiarities, is to compensate for shortcomings in any single data set. Indeed, the observed overlap in gene expression effects is striking, especially given that the analysis integrates data across multiple brain compartments and disease timepoints, as well as microarray platforms.

Pathway analysis was performed using DAVID<sup>11</sup> on the list of 24 genes. These genes are enriched significantly in 19 GO CC pathways, including those related to mitochondria, membrane, and vesicles, specifically synaptic vesicle membrane, and 11 GO MF pathways, relating mainly to ATPase activity (Supplementary Table 6). Eight out of the 24 genes are annotated in GO CC as “mitochondrion” (GO:0005739): ATP5B, ATP6V1E1, BNIP3, C14orf2, GOT2, MRPS11, SLC25A11, and UQCRC1. Several additional genes are related to mitochondria, as will be discussed below. Four genes are annotated in GO MF as “ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism” (GO:0015662): ATP5B, ATP6V1G2, ATP6V1E1, and ATP1A3. Three genes are annotated as “synaptic vesicle membrane” (GO:0030672): ATP6V1G2, SLC17A7, and SYP. Several genes on the list also have ties to glutamate (including CACNG3, which regulates AMPA-sensitive glutamate receptors; SLC17A7, which mediates glutamate uptake into synaptic vesicles; SLC25A11, a mitochondrial oxoglutarate carrier; and GOT2, mitochondrial glutamic-oxaloacetic transaminase 2), which is interesting because both impairments in glutamatergic transmission and excitotoxicity are thought to play a role in AD<sup>12</sup>.

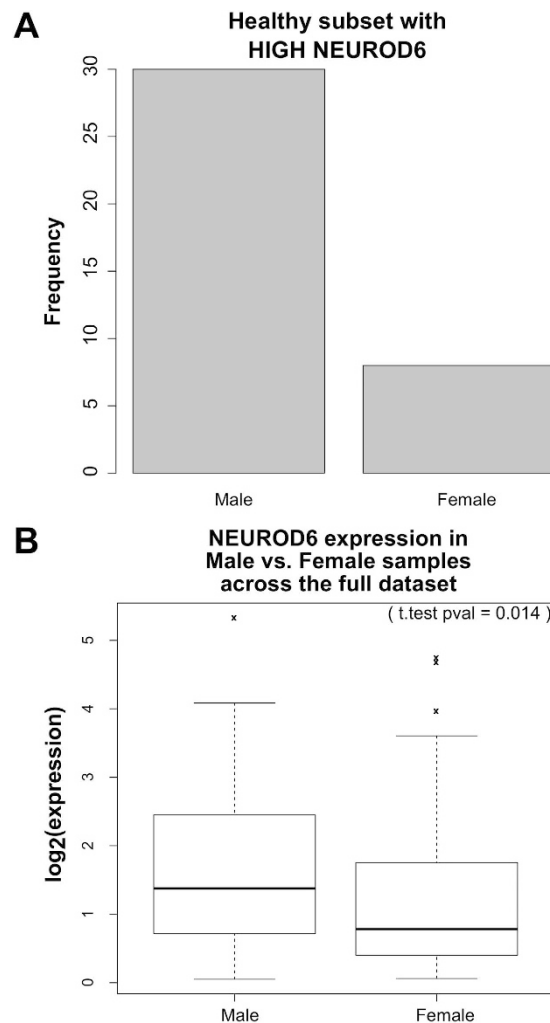
**Degree of brain-specific expression of identified genes using BioGPS database.** To determine the degree to which these genes were expressed preferentially in the brain, we utilized publicly available tissue-specific array data, as described in Methods. Overall, for the 24 genes identified, we found a high degree of specificity for expression in brain tissue (Fig. 1c).

**Searching the Gene Expression Omnibus detects gender differences in identified genes.** To generate further insight into the list of expression-identified genes, we developed a method for identifying patterns in the Gene Expression Omnibus (GEO), a large database of publicly available gene expression data sets including over 500,000 human samples. We developed a Wilcoxon-test-based algorithm for comprehensively searching GEO to identify those samples in which each of the 24 genes was significantly modulated relative to other genes.

The most striking finding was related to gender, a factor that is suggested to play a role in AD<sup>13</sup>. As shown in Supplementary Table 2, the CACNG3, GNG3, and NEUROD6 genes showed a gender-based pattern. For example, we observed in a dataset from healthy brain, GSE11882 (“Gene expression changes in the course of normal brain aging are sexually dimorphic”) that 38 samples were significantly enriched (FDR adjusted p-value < 0.05) for high expression of NEUROD6 (Supplementary Table 3). Strikingly, 30 of the 38 samples with enriched NEUROD6 expression were from males (Fig. 2a). This difference was highly significant, with hypergeometric p-value  $1.71 \times 10^{-4}$ . In contrast, the entire GSE11882 data set was well-balanced by gender (173 samples with 82 female and 91 male) (Supplementary Fig. 2). We analyzed the pattern of NEUROD6 expression across the full data set to search for expression differences by gender in NEUROD6, and found that NEUROD6 expression was significantly higher in males than females with a nominal p-value of 0.014 (Fig. 2b). By dividing the data set into individual compartments, we found that NEUROD6 was significantly differentially expressed in two of the four compartments (with nominal p-values 0.0052 and 0.007; Supplementary Fig. 3). The finding that NEUROD6 differs in expression level between healthy men and women is particularly intriguing given that (a) NEUROD6 expression is downregulated with disease, and (b) gender may play a role in AD<sup>13</sup>.

Examining the GSE11882 data set with regard to the SNAP25 gene, two out of three expression probesets representing this gene were significantly lower in women than in men, in an individual compartment in the GSE11882 data set (nominal p-values 0.018 and 0.032; Supplementary Table 4). Note that GSE11882 was analyzed for differential gene expression between genders, but provided no Alzheimer-specific information.

To see whether these gender-related differences could have been recovered immediately from the original gene expression datasets, we also performed differential expression directly on each of these five AD datasets to examine differences in expression between genders (see Supplementary Table 9). We were not able to recover differences (beyond nominal p-values) between expression of NEUROD6 and



**Figure 2.** Possible link between NEUROD6 and gender. (a) Gender distribution of NEUROD6-high samples in a data set of healthy controls. (b) Comparison of NEUROD6 expression by gender across the entire data set of healthy controls.

SNAP25 in males and females in all five datasets (stratified by Alzheimer's status, see Supplementary Table 11). In the largest dataset (GSE15222), we observed nominally significant differences in SNAP25 expression between females and males among AD patients but not among controls. In the second largest dataset (GSE44772), SNAP25 differed at least nominally by gender among AD patients and among controls. In one other dataset, GSE1297, SNAP25 differed at least nominally between M and F among controls. Fisher's exact tests comparing the counts of samples originating from males and females, stratifying by disease status, determined that 4 out of the 5 datasets do not show a significant gender bias (see Supplementary Table 12). These results examining this more subtle effect are not surprising given the study sizes and different platforms across studies, and underscore the higher sensitivity of the pipeline described in this work as compared to traditional approaches for analysis.

We chose to focus on NEUROD6 and SNAP25 on the basis of these gender-related patterns and the genetic analyses described below.

**Utilizing SNP studies to identify putatively causative genes.** To distinguish between downstream signals resulting from disease pathology, and upstream signals that may be more causative and therefore better targets for therapy, we utilized single-nucleotide polymorphism (SNP) data in conjunction with the gene expression data. The established method for combining these two types of data, expression quantitative trait locus (eQTL) analysis<sup>14</sup>, requires both gene expression and SNP data from the same cohort of patients. Since most of the available gene expression and SNP data came from separate cohorts, we developed another approach to identify converging lines of evidence for disease-causing genes from these disparate but complementary types of genomic data.

We examined the regions in and around (details as described below and in Methods) the 24 genes identified as downregulated in AD (Table 1) for disease-associated SNPs in three data sets: the ADNI1

cohort, the National Institute on Aging Late-Onset Alzheimer's Disease Family Study ("LOAD study")<sup>15</sup>, and a recent study by Zhang *et al.* ("Cell study")<sup>3</sup>. Limiting the SNPs of interest to targeted regions reduces the risk of false positives. This narrow focus supports the use of a more lenient p-value threshold greater than the typical threshold applied for GWAS studies (often  $p < 5 \times 10^{-8}$ ). In order to relax the threshold from an overly conservative genomewide significance cutoff, yet maintain a desired level of stringency, we chose to use a p-value threshold of  $p < 5 \times 10^{-4}$ . The size of the region surrounding each gene was taken to be 200 kb up- and downstream, as many enhancer regions for transcriptional start sites, in general, are known to extend to such distances<sup>16</sup>. Indeed, in some cases, such elements are located as far as 1 Mb away from the affected gene; for example, in the case of the genes SHH<sup>17</sup> and POU3F4<sup>18</sup>. For this reason, we performed an additional analysis using less stringent distance and p-value parameters of 1 Mb and  $p < 10^{-2}$ , respectively. For both the strict and less stringent analyses, we defined a result to be of particular interest if a significant SNP from the GWAS studies comparing AD versus healthy control was located in the region of one of the AD-downregulated genes, in at least two out of three SNP studies.

In the case of the least stringent distance and p-value parameters (1 Mb and  $p < 10^{-2}$ ), for 15 out of 24 genes, significant SNPs were located within the region of interest in all three studies, as shown in Supplementary Table 7a, and the regions surrounding all but three of the expression-identified genes contained significant SNPs in at least two out of three studies. For further investigation, we utilized the results of the stringent filtering ( $p < 5 \times 10^{-4}$ , 200 kb; shown in Supplementary Table 7b) to help prioritize genes, as discussed in further detail below.

GWAS studies to date have identified APOE as the gene most consistently associated with AD. Carriers of the APOE  $\epsilon 4$  (APOE4) allele have a significantly increased risk of AD<sup>19</sup> for reasons that are incompletely understood, and several clinical studies have found putative differences in response to therapy based on APOE4 status<sup>20,21</sup>. In addition, recent work suggests that APOE4 imparts an increased risk of conversion to AD in women<sup>22</sup>. Given this information as well as the gender differences discussed above, in addition to examining the regions of interest in all patients, we also performed subset-specific analyses, based on both gender and apolipoprotein E (APOE) status. We reasoned that different subsets of AD patients might have differences in the biological factors driving their disease.

SNPs in the region of NEUROD6 were associated with AD specifically in APOE4+ women in both the ADNI1 and LOAD cohorts (Supplementary Table 7b). These NEUROD6 SNPs are illustrated in Fig. 3a,c. We utilized a newly developed propensity plotting method to visualize the specific influence of these SNPs (Fig. 3b,d). Status at these SNPs was highly associated with disease propensity. In contrast, SNPs in the region of SNAP25 were associated with AD specifically in APOE4+ men in both the LOAD and Cell data sets (Supplementary Table 7b and Supplementary Fig. 4). SNPs were also identified near MRPS11 for APOE4+ women, and near GOT2 for women as well as for the whole cohort, for the chosen distance and p-value parameters. We narrowed our focus to NEUROD6 and SNAP25 since their respective identified SNPs were closer to these genes than to MRPS11 and GOT2, though all of these genes may have intriguing biological connections, as discussed below.

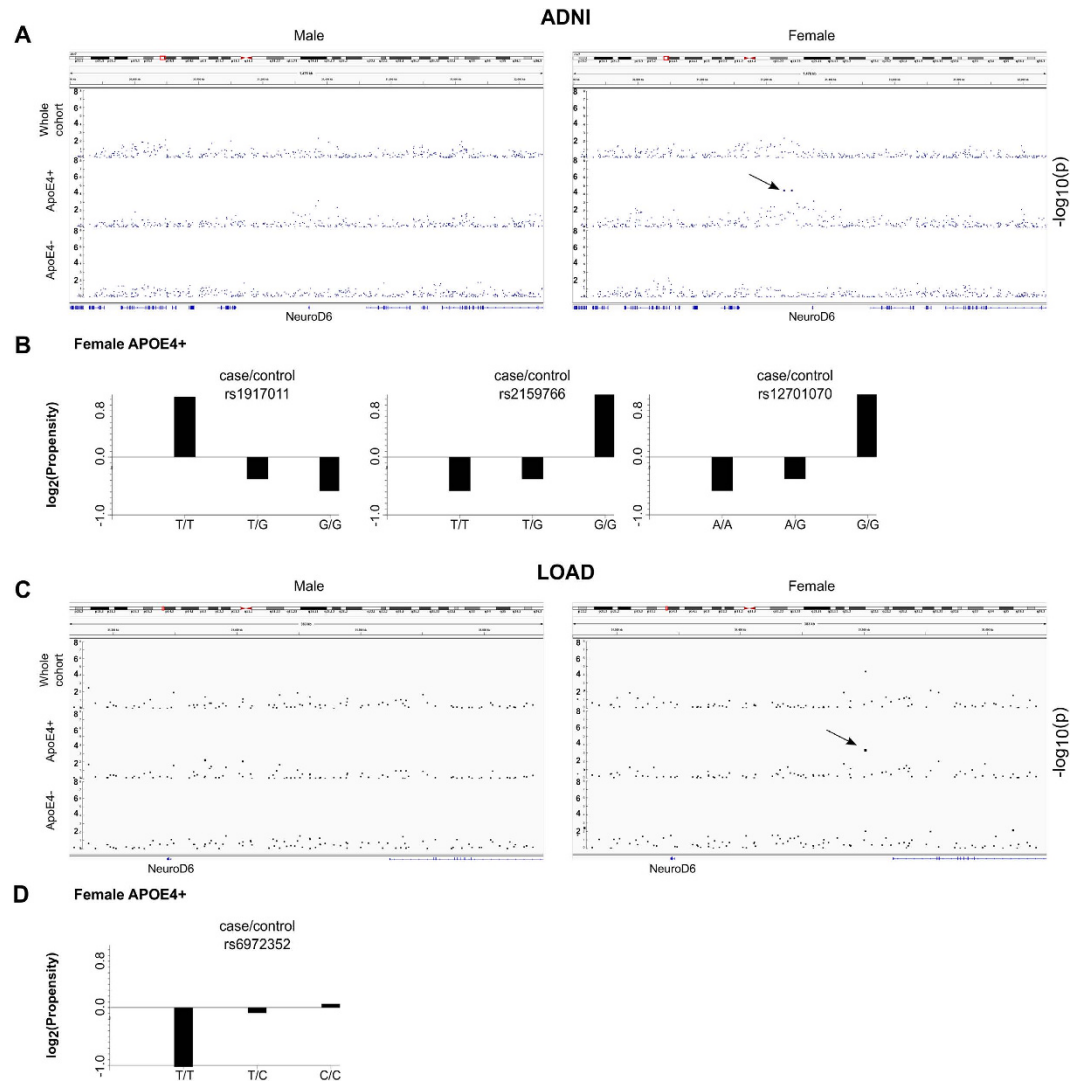
**Identifying medicines to restore expression of causative genes, using CMAP data.** We next sought to determine which medicinal drugs and compounds could restore expression of NEUROD6 or SNAP25 shown to be downregulated in AD. We developed an enrichment algorithm to identify specific drugs in the Connectivity Map (CMAP)<sup>23</sup> databases that induced significantly higher expression of NEUROD6 or SNAP25 in culture. We identified 34 unique compounds that upregulate NEUROD6, and nine that upregulate SNAP25 (Table 2). Note that this hypothesis-generating analysis serves to identify drugs of interest for further examination, and these compounds are not expected to be highly specific for these genes. The full list of drugs with their known gene/pathway targets is available in Supplementary Table 10.

## Discussion

A number of the genes identified in the expression analysis as downregulated with AD have striking common biology. The list of 24 genes enriched significantly for GO pathways including those related to mitochondria, membrane, and vesicles, specifically synaptic vesicle membrane, and ATPase activity (Supplementary Table 6). Several genes on the list also have ties to glutamate, which is of particular interest since both impairments in glutamatergic transmission and excitotoxicity are thought to play a role in AD<sup>12</sup>. To evaluate which genes on this list were likely to be causal in disease, we examined SNP data and determined that in particular patient cohorts, significant SNPs were found in multiple data sets in the regions around NEUROD6 and SNAP25.

NEUROD6 is a transcription factor involved in neuronal differentiation, and has been shown to increase mitochondrial mass and play a role in response to oxidative stress<sup>24</sup>. This is intriguing because the aging process has a negative impact on mitochondrial function<sup>25</sup> and leads to an increase in mitochondrial DNA mutations<sup>26</sup>, and rates of Alzheimer's increase dramatically with age<sup>27</sup>.

APOE also has ties to the mitochondria<sup>28</sup>. The APOE  $\epsilon 4$  isoform has been shown to cause mitochondrial damage specifically in neurons<sup>28</sup>. APOE  $\epsilon 4$  also has lower antioxidant capability than other isoforms, and amyloid beta induces oxidative stress to a greater extent when APOE  $\epsilon 4$  is present<sup>29</sup>. Oxidative stress may also induce hyperphosphorylation of tau<sup>30</sup>, another key factor in AD. Impairing the transport



**Figure 3. NEUROD6-related SNPs are associated with AD specifically in APOE4+ female patients.** (a) Plot showing a “cone” of disease associated SNPs around NEUROD6 in APOE4+ female patients, but not in APOE4+ male, APOE4- female, or APOE4- male patients. Top AD-associated SNPs related to NEUROD6 in APOE4+ female patients in ADNI1 were: rs1917011 ( $p < 3.82e-5$  in female APOE4+ patients,  $p < 0.692$  in male APOE4+,  $p < 0.844$  in female APOE4-), rs2159766 ( $p < 3.82e-5$  in female APOE4+,  $p < 0.771$  in male APOE4+,  $p < 0.624$  in female APOE4-), and rs12701070 ( $p < 3.82e-5$  in female APOE4+,  $p < 0.561$  in male APOE4+,  $p < 0.624$  in female APOE4-). (b) Propensity plots showing the disease risk (positive values) or protection (negative values) in APOE4+ female patients as a function of a patient’s status for each of three top NEUROD6 SNPs. (c) Plot showing disease associated SNP near NEUROD6 in APOE4+ female patients, but not in APOE4+ male, APOE4- female, or APOE4- male patients from the LOAD data set. This SNP was rs6972352 ( $p < 0.00049$  in female APOE4+,  $p < 0.2247$  in male APOE4+, and  $p < 0.010$  in female APOE4-). (d) Propensity plots showing the disease risk (positive values) or protection (negative values) in APOE4+ female patients as a function of a patient’s status for each of three top NEUROD6 SNPs.

of mitochondria into axons was shown to enhance tau phosphorylation and neurodegeneration<sup>31</sup>. Several studies in a variety of systems have shown that oxidative stress induces upregulation of BACE1<sup>32</sup>, an enzyme critical for the production of amyloid beta. In other systems, oxidative stress has been shown to increase production of amyloid precursor protein<sup>33</sup>. Because NEUROD6 confers tolerance to oxidative stress<sup>24</sup>, it has the potential to mitigate some of this damage. As NEUROD6 expression is lower in women, and APOE4+ individuals have lowered tolerance for ROS damage<sup>28,29</sup>, it stands to reason that a SNP associated with further impairment of NEUROD6 may put APOE4+ females at particular risk of damage due to oxidative stress.

CMAP compound name	p.val	adj.p.val
<b>(A) NEUROD6-elevating compounds</b>		
sodium phenylbutyrate	8.03E-05	4.23E-02
arachidonic acid	8.22E-05	4.23E-02
2-deoxy-D-glucose	8.59E-05	4.23E-02
fasudil	8.76E-05	4.23E-02
nordihydroguaiaretic acid	1.04E-04	4.23E-02
monastrol	1.09E-04	4.23E-02
tacrolimus	1.12E-04	4.23E-02
quercetin	1.12E-04	4.23E-02
sulindac	1.14E-04	4.23E-02
troglitazone	1.17E-04	4.23E-02
staurosporine	1.17E-04	4.23E-02
troglitazone	1.22E-04	4.23E-02
thalidomide	1.26E-04	4.23E-02
CP-944629	1.35E-04	4.23E-02
mercaptopurine	1.40E-04	4.23E-02
haloperidol	1.49E-04	4.23E-02
exisulind	1.57E-04	4.23E-02
sirolimus	1.71E-04	4.23E-02
tanespimycin	1.71E-04	4.23E-02
suramin sodium	1.74E-04	4.23E-02
genistein	1.76E-04	4.23E-02
erastin	1.78E-04	4.23E-02
clofibrate	1.80E-04	4.23E-02
LY-294002	1.92E-04	4.23E-02
tanespimycin	1.93E-04	4.23E-02
LY-294002	1.97E-04	4.23E-02
prednisolone	1.99E-04	4.23E-02
fulvestrant	2.01E-04	4.23E-02
meteneprost	2.05E-04	4.23E-02
monorden	2.17E-04	4.23E-02
tretinoin	2.22E-04	4.23E-02
nifedipine	2.30E-04	4.23E-02
sulindac sulfide	2.32E-04	4.23E-02
wortmannin	2.36E-04	4.23E-02
MK-886	2.46E-04	4.29E-02
PF-01378883-00	2.59E-04	4.38E-02
monorden	2.82E-04	4.65E-02
iloprost	3.06E-04	4.91E-02
<i>Note: 34 unique compounds. Repeats due to differing cell lines or repeated experiments.</i>		
<b>(B) SNAP25-elevating compounds</b>		
valproic acid	2.20E-05	1.91E-02
guanabenz	9.14E-05	3.81E-02
karakoline	8.89E-05	3.81E-02
tetracycline	1.03E-04	4.01E-02
diloxanide	1.28E-04	4.45E-02
metoprolol	1.38E-04	4.52E-02
yohimbic acid	1.59E-04	4.75E-02
azapropazone	1.63E-04	4.75E-02
proguanil	1.93E-04	4.92E-02

**Table 2. Drug candidates determined based on patient subset analysis.** (a) Compounds that induce enriched expression of NEUROD6. (b) Compounds that induce enriched expression of SNAP25.

SNAP25 has an important role in synaptic function as part of the Soluble NSF Attachment Protein Receptor (SNARE) complex, which is involved in synaptic vesicle exocytosis, and has been tied to neurodegeneration<sup>34,35</sup>. SNARE proteins are known to be sensitive to oxidative stress, with SNAP25 being the most sensitive, which has been proposed to relate mitochondrial dysfunction to reduced synaptic activity in neurodegeneration<sup>36</sup>. Indeed, in a mouse model of AD using male mice, lower SNAP25 levels were observed in the hippocampus, along with lowered levels of glutamate, ATP, and mitochondrial membrane potential; the authors attributed the observed deficiency in glutamatergic neurotransmission (which requires a high level of energy consumption) to dysfunction of mitochondrial bioenergetics<sup>37</sup>.

The other two genes identified for the chosen distance and p-value parameters for certain patient subsets (female APOE4+, and female as well as whole cohort, respectively) were MRPS11 and GOT2. Intriguingly, both are mitochondrial genes. MRPS11 encodes a component of the small subunit of the mitochondrial ribosome. Mutations in other mitochondrial small subunit ribosomal proteins are known to cause severe disruption in the respiratory chain<sup>38</sup>. GOT2 encodes a mitochondrial kynurenine transaminase, which acts on kyurenine to produce kynurenic acid (KYNA), a neuroprotective NMDA and nicotinic receptor antagonist that may also have antioxidant properties<sup>39</sup> and can affect glutamatergic neurotransmission in multiple ways<sup>12</sup>. Interestingly, lower KYNA levels have been observed in AD patients<sup>40</sup>, and raising KYNA levels via inhibition of kyurenine-3-monooxygenase (KMO; another enzyme that acts on kyurenine) ameliorates neurodegeneration *in vivo*<sup>41</sup>.

Several of the compounds identified via our CMAP search as upregulating NEUROD6 or SNAP25 expression have shown promise in published mouse models of AD, as described below. The highest ranked compound for its ability to upregulate NEUROD6 is sodium phenylbutyrate, which has recently been proposed as a therapeutic for neurodegenerative diseases due to its ability to increase neurotrophic factors in brain cells, along with the fact that it is safe, orally delivered, and crosses the blood brain barrier<sup>42</sup>. Another study using a mouse model of AD demonstrated that independent of A $\beta$  levels, sodium phenylbutyrate decreased spatial learning and memory impairment by upregulating markers for synaptic and dendritic growth<sup>43</sup>. The third most highly ranked compound is 2-Deoxy-D-Glucose, which has been shown to reduce pathology in a female mouse model of AD by inducing ketogenesis and enhancing mitochondrial capacity<sup>44</sup>. We propose that these drugs warrant additional investigation as AD therapeutics, particularly in APOE4+ women.

Why is NEUROD6 expression most significant in females? Estrogen signaling appears to stimulate the production of enzymes such as glutathione peroxidase that protect the mitochondria against oxidative stress<sup>13</sup>, so the loss of estrogen upon age could leave women more susceptible to mitochondrial damage associated with impairment of NEUROD6 production and the resultant loss of protective effects. In fact, among the list of compounds from CMAP2 that most significantly elevate expression of NEUROD6 (Table 2) is genistein, which has been proposed as a means to replace the protective effect of estrogen on mitochondria in aging women<sup>13</sup>.

The top drug shown to significantly elevate SNAP25 expression, valproic acid, is known to have neuroprotective properties<sup>45</sup>. In recent studies in mouse models of AD, valproic acid was demonstrated to protect against loss of neurons<sup>46</sup> and limit A $\beta$  production and behavioral deficits<sup>47</sup>. Karakoline is a nicotinic receptor agonist that was recently shown to improve cognitive function in a mouse model of AD<sup>48</sup>. Tetracycline was shown to protect from A $\beta$  toxicity in *C. elegans*<sup>49</sup>, and its derivatives are actively being explored as potential therapeutics in mouse models of AD<sup>50</sup>. We hypothesize that these compounds warrant further investigation for treatment of AD, particularly in APOE4+ men.

We developed an innovative pipeline for extracting new insights from existing data sets, utilizing a breadth of separate, publicly available resources (as illustrated in Fig. 4). This pipeline is innovative in particular because of the focus on integrating multiple gene expression datasets across various platforms in order to achieve a robust intersection of disease-affected genes, then leveraging these results in combination with genetic studies in order to identify those that are more likely causal. This approach has identified NEUROD6 and SNAP25 as important factors in AD, especially in APOE4+ women and men, respectively. Both of the genes identified have solid connections to the neuronal oxidative damage response. We have also identified subset-specific drug candidates for modulation of these genes. More broadly, our analyses suggest that further exploration of gender-specific biology could lead to effective new medicines for neurodegeneration, along with special considerations for corresponding clinical trials, and that patient stratification by gender, genetics or other factors could facilitate data-driven identification of new therapeutic options for a variety of diseases.

## Methods

**Identification, Processing and Analysis of Expression Data Sets.** All expression data sets used in this analysis were obtained from the Gene Expression Omnibus (GEO) site (<http://www.ncbi.nlm.nih.gov/geo/>) and consisted of GEO project accession numbers: GSE5281<sup>51</sup>, GSE1297<sup>52</sup>, GSE36980<sup>53</sup>, GSE15222<sup>54</sup>, and GSE44772<sup>3</sup>. A summary of these data sets, including all sample IDs used in the analysis, array type, and the compartment of the brain that the samples were collected from can be found in Supplementary Table 5. Each data set was processed and analyzed independently. We identified genes differentially expressed between AD and control in each using LIMMA (Version 3.18.13)<sup>55</sup>, taking a



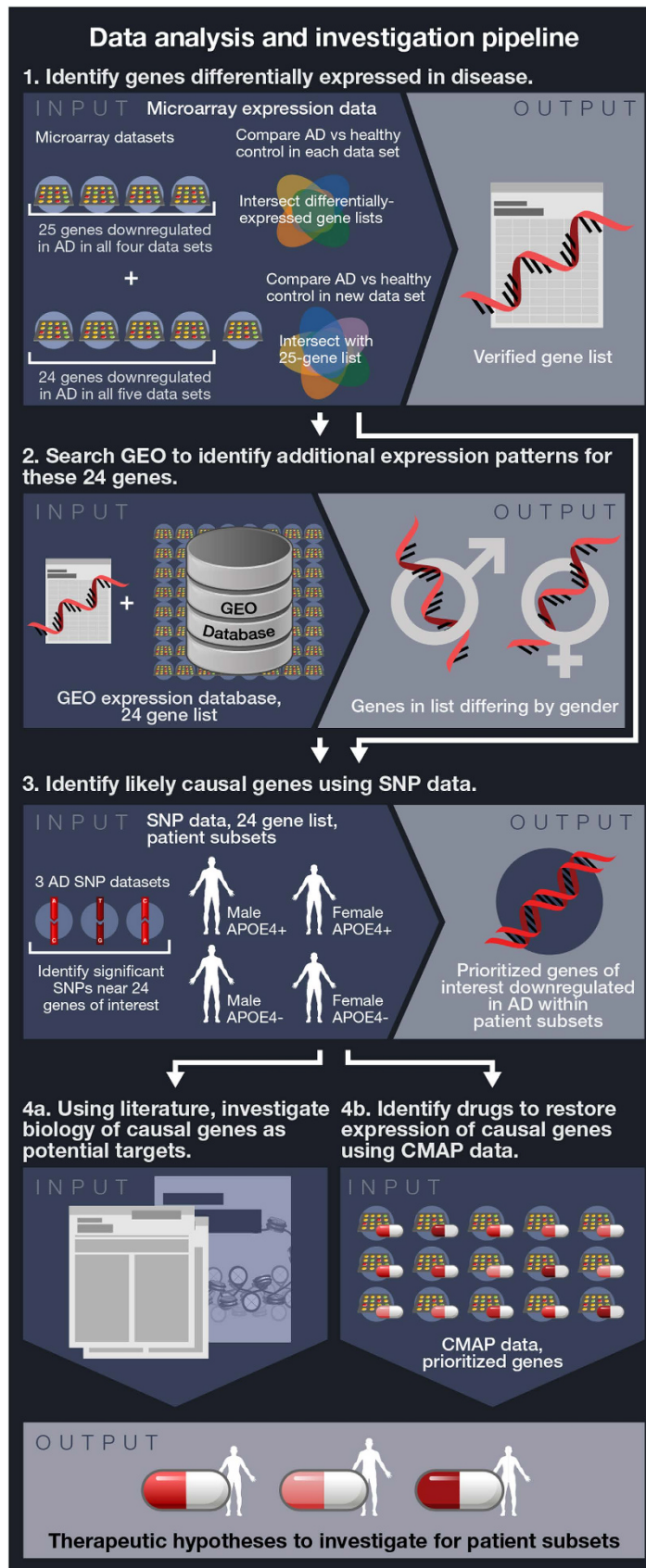


Figure 4. For legend see next page.

**Figure 4. Schematic of analysis pipeline.** (Art credit: Kate Mahan.) We developed an innovative pipeline for extracting new insights from existing publicly available resources, including five gene expression datasets, three SNP datasets, and multiple databases, in order to generate hypotheses for potential treatment strategies in subsets of AD patients. (1) We identified genes differentially expressed in AD in multiple gene expression data sets, leading to a list of 24 genes downregulated in AD. (2) We searched all of GEO to identify expression patterns for these genes in additional (not necessarily AD) datasets, revealing gender differences for some genes. (3) For patient subsets of interest (defined by gender and APOE4 status), we performed GWAS in each of three separate SNP data sets (AD vs healthy control) in regions up and downstream of 24 genes, allowing prioritization of genes having nearby SNPs. (4) The genes thus identified provide insights that can help facilitate development of new therapeutics. For example, searching the CMAP database identified existing drugs warranting further investigation for specific subsets of AD patients. In addition, further elucidating the biology of these genes can allow identification of entirely new therapeutic targets.

significance threshold of FDR-corrected  $p$ -value  $< 0.05$ . We further identified the genes significant in multiple data sets by taking the intersection of the individual differentially expressed gene lists across studies.

The general analysis method employed for expression data was as follows. CEL files were obtained from GEO (except where noted) and the data were RMA normalized<sup>56</sup>. The probesets determined to be significantly different between the two conditions of interest (in nearly all cases, AD vs control, except for GSE1297 where this condition was based on low versus high NFT score, as noted below) were calculated using LIMMA. For chips with available MM/PM (mismatch/perfect match) probe data (Affy U133A or U133 Plus 2.0 chips), results were filtered according to Presence/Absence (P/A) calls computed in R using the “affy” package<sup>57</sup>, such that the average P/A call for all of the samples in the condition of interest with the higher average value (e.g., controls, for probesets downregulated in disease) was at least marginal.

LIMMA comparisons were used as a metric for assessing the signal to noise tradeoff in each brain compartment for data sets with multiple brain compartments (visual cortex excluded). The compartment with the greatest number of significant probesets (FDR  $p$ -value  $< 0.05$ ) was chosen for analysis. For Affymetrix chips, the presence of Affymetrix control probesets among the top significant differential expression results was considered a QC flag; compartments having Affymetrix control probesets as either  $> 0.25\%$  of significant differential expression results, or in any of the top ten results, were discarded.

GSE5281 consists of samples collected from 6 compartments of the brains of AD patients along with controls and measured using the Affymetrix Human Genome U133 Plus 2.0 Array. We used the samples collected from the Entorhinal Cortex (EC), per the QC procedure described above.

GSE1297 contains mini-mental state examination (MMSE) scores and neurofibrillary tangle (NFT) scores for all individuals in the study, with gene expression measurements from the hippocampus made using the Affymetrix Human Genome U133A Array. We found roughly 5 times as many significant probesets when comparing the individuals with the 9 highest NFT scores versus the individuals with the 9 lowest NFT scores than attained when comparing Severe AD patients (based on MMSE score) vs. controls. For this reason, we used samples with the top 9 NFT scores versus the lowest 9 NFT scores in this analysis. The lowest 9 NFT scores included both controls and patients labeled as having incipient AD. The highest 9 NFT scores included patients labeled as having severe AD, moderate AD, and incipient AD.

GSE36980 is comprised of samples from 3 compartments of the brains of AD patients and controls, measured using the Affymetrix Human Gene 1.0 ST Array. We used the data from the hippocampus as the compartment with the most probesets significantly differing in expression between AD patients and controls (as described above).

GSE15222 consists of samples collected from multiple cortical compartments of the brains of AD patients and controls, measured using the Sentrix HumanRef-8 Expression BeadChip. Raw data was not available from this data set, so we used the processed data available on the corresponding author’s laboratory website<sup>58</sup>. The authors rank-invariant normalized their data and then filtered it in two ways. First, transcripts were only considered if they were detected in at least 90% of AD patients or 90% of controls. Second, transcript expression intensities were only considered in analyses for a given sample if their Illumina detection scores were  $> 0.99$ <sup>58</sup>. Since these data had been normalized across all measured compartments of the brain, we considered all the samples together.

The net result of these analyses was a list of probesets significantly affected by Alzheimer’s disease for each data set. To obtain a robust list of affected genes, we determined the intersection of the gene names from all four lists. We performed separate intersections for genes upregulated in each study and genes downregulated in each study. After obtaining a list of 25 genes downregulated in disease relative to control across all four studies, we obtained a fifth data set (GSE44772).

GSE44772, which we obtained after analyzing the other four data sets, consists of samples collected from three compartments in the brains of both AD patients and controls, measured using a custom Rosetta/Merck Human 44k 1.1 microarray. Using the same QC procedure described above, we chose to utilize the data from the prefrontal cortex. Raw data was not available from this data set, so we used

the processed data available in the Series Matrix File on GEO. The data was in the form of normalized log<sub>10</sub> ratios between the test sample and a pooled reference sample. LIMMA was used to determine the probesets that were significantly different between the AD patients and the controls. 24 out of the 25 downregulated genes from the four-data set intersection were also significantly downregulated with disease in this data set.

**Brain specificity heat maps.** For the 24 genes of interest, all probesets were mapped to tissue-specific arrays<sup>59</sup>, available via BioGPS<sup>60</sup>. The probesets were clustered hierarchically with a metric of Pearson correlation, and displayed via subtracting by the median and dividing by the absolute deviation (both in GENE-E, <http://www.broadinstitute.org/cancer/software/GENE-E/>). Tissues were annotated by whether or not they were brain-associated, and sorted and grouped accordingly.

**Stratified GWAS.** We conducted targeted gene association testing from the SNP data sets listed in the text using PLINK<sup>61</sup>, with patient subsets defined by gender and APOE4 status. Results were visualized using the Integrative Genomics Viewer (IGV)<sup>62</sup>.

In order to obtain the strongest possible AD-relevant signals, we restricted the analysis of the LOAD data set to only those patients with an AD diagnosis confirmed by autopsy. Similarly, in analysis of the Zhang *et al.* data set we excluded controls and patients having a diagnosis of Huntington's disease.

Of the 757 patients in ADNI1 who were genotyped via Illumina 610 Quad array SNP chip, 389 were categorized as either AD patients or healthy controls, and the rest (those with mild cognitive impairment, MCI) were excluded from further consideration. GWAS was also run on patient sub-cohorts, stratified by APOE4 status, gender, and the combination of both. Effect sizes for each group are as follows: unstratified = 175 AD, 214 healthy controls (HC); APOE4- = 58 AD, 156 HC; APOE4+ = 117 AD, 58 HC; Female APOE4- = 31 AD, 73 HC; Female APOE4+ = 51 AD, 26 HC; Male APOE4- = 27 AD, 88 HC; Male APOE4+ = 66 AD, 32 HC; Male all = 93 AD, 115 HC; Female all = 82 AD, 99 HC.

The LOAD data set included 1985 patients and 2058 controls genotyped via Illumina Human 610 Quad v1B SNP chip. Patients included in the analysis were limited to those who had an AD diagnosis confirmed by autopsy. Unstratified, subjects included in the analysis were 440 patients and 2058 controls. Numbers of subjects within subsets were: APOE4- = 99 AD, 1256 HC; APOE4+ = 341 AD, 802 HC; Female APOE4- = 74 AD, 773 HC; Female APOE4+ = 230 AD, 483 HC; Male APOE4- = 25 AD, 483 HC; Male APOE4+ = 111 AD, 319 HC; Male all = 136 AD, 802 HC; Female all = 304 AD, 1256 HC.

The Zhang *et al.* (or "CELL") data set included 374 patients and 366 controls genotyped via Illumina HumanHap650Y SNP chip. We excluded subjects with a diagnosis of Huntington's disease from the analysis. Unstratified, 371 patients and 159 controls were included in the analysis. Numbers of subjects in subsets were as follows: APOE4- = 209 AD, 130 HC; APOE4+ = 162 AD, 29 HC; Female APOE4- = 129 AD, 34 HC; Female APOE4+ = 90 AD, 5 HC; Male APOE4- = 80 AD, 96 HC; Male APOE4+ = 72 AD, 24 HC; Male all = 152 AD, 120 HC; Female all = 219 AD, 39 HC. In this data set we removed from consideration patients and controls with a diagnosis of Huntington's disease. Therefore, smaller numbers of controls were available, especially in the highly stratified subsets. Specifically, the female APOE+ cohort contained only five controls, which may explain why SNPs calculated to be significant in female APOE4+ patients in the other two data sets did not achieve significance in this subset.

**Propensity plotting.** We adapt a measure of bias for identifying amino acids on the surface that are likely to be part of an interface domain from the proteomics field<sup>63-65</sup> to measure the preference of a particular SNP genotype to case vs. control subsets of the data set. Briefly, we define SNP propensity to be the fraction of patients diagnosed with AD given a particular SNP variant to the control fraction with this SNP variant. Specifically, it is defined as:

$$\log_2\left(\text{Propensity}_{rsX=i}\right) = \frac{\frac{\text{CASE}_i}{\text{CASE}}}{\frac{\text{CASE}_i + \text{CONTROL}_i}{\text{CASE} + \text{CONTROL}}} \quad (1)$$

Where  $\text{CASE}_i$  is the fraction of patients diagnosed with AD and having SNP variant  $i$ .  $\text{CONTROL}_i$  is the fraction of control individuals with SNP variant  $i$ .  $\text{CASE}$  is the total number of patients diagnosed with AD in the data set, and  $\text{CONTROL}$  is the total number of control individuals in the data set.

**GEO-search analysis.** We searched all human data sets across all platforms in GEO as of July 2013. We used a one-sample Wilcoxon test to measure the significance of differential expression for probesets annotated to a gene of interest against all other probesets in the sample, correcting for FDR using Benjamini-Hochberg.

Supplementary Table 2 was constructed by counting the number of samples,  $m$ , from the GSE11882 data set that passed our adjusted  $p$ -value threshold of 0.05 for the GEO search algorithm described above. Of the samples that passed our adjusted threshold of 0.05, we counted the number of samples,  $n$ , that came from males. We then calculated the significance of obtaining  $\geq n$  male samples if we drew  $m$  samples randomly from a total of 173 samples (which is the total number of samples in GSE11882)

with 91 males (total number of male samples in GSE11882) based on the hypergeometric distribution. We omitted the significance calculation if the number of significant samples,  $m$ , was less than 15 (<10% of the data set).

**CMAP search.** CMAP (Connectivity Map) is a large collection of microarray-based transcriptional signatures for 7000 expression profiles from cultured cells treated with 1,309 compounds. We obtained the full CMAP (builds 01 and 02)<sup>66,23</sup> data sets and utilized a one-sample Wilcoxon test to identify expression profiles from compounds that significantly increased expression of a gene of interest in culture after treatment. We then adjusted for multiple hypothesis testing (FDR  $p$ -value < 0.05).

**ADNI-data acquisition.** The SNP data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([adni.loni.ucla.edu](http://adni.loni.ucla.edu)). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a \$60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see [www.adni-info.org](http://www.adni-info.org).

## References

- Bertram, L. & Tanzi, R. E. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. *Nat. Rev. Neurosci.* **9**, 768–778 (2008).
- Cruchaga, C. *et al.* GWAS of Cerebrospinal Fluid Tau Levels Identifies Risk Variants for Alzheimer's Disease. *Neuron* **78**, 256–268 (2013).
- Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).
- Krauthammer, M., Kaufmann, C. A., Gilliam, T. C. & Rzhetsky, A. Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proceedings of the National Academy of Sciences* **101**, 15148–15153 (2004).
- Chen, J. Y., Shen, C. & Sivachenko, A. Y. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac Symp Biocomput* 367–378 (2006).
- Liu, B. *et al.* Exploring candidate genes for human brain diseases from a brain-specific gene network. *Biochemical and Biophysical Research Communications* **349**, 1308–1314 (2006).
- Soler-Lopez, M., Zanzoni, A., Lluís, R., Stelzl, U. & Aloy, P. Interactome mapping suggests new mechanistic details underlying Alzheimer's disease. *Genome Research* **21**, 364–376 (2011).
- Caberlotto, L., Lauria, M., Nguyen, T.-P. & Scotti, M. The Central Role of AMP-Kinase and Energy Homeostasis Impairment in Alzheimer's Disease: A Multifactor Network Analysis. *PLoS ONE* **8**, e78919 (2013).
- Khairbek, M. A., Klemenhagen, K. C., Sahay, A. & Hen, R. Neurogenesis and generalization: a new approach to stratify and treat anxiety disorders. *Nat Neurosci* **15**, 1613–1620 (2012).
- Kulesza, P., Ramchandran, K. & Patel, J. D. Emerging concepts in the pathology and molecular biology of advanced non-small cell lung cancer. *Am. J. Clin. Pathol.* **136**, 228–238 (2011).
- Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucl. Acids Res.* **37**, 1–13 (2009).
- Zádori, D. *et al.* Glutamatergic Dysfunctioning in Alzheimer's Disease and Related Therapeutic Targets. *J. Alzheimers Dis.* (2014). doi: 10.3233/JAD-132621
- Viña, J. & Lloret, A. Why women have more Alzheimer's disease than men: gender and mitochondrial toxicity of amyloid-beta peptide. *J. Alzheimers Dis.* **20** Suppl 2, S527–533 (2010).
- Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
- Lee, J. H. *et al.* Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: implication of additional loci. *Arch. Neurol.* **65**, 1518–1526 (2008).
- Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
- Lettice, L. A. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12**, 1725–1735 (2003).
- Naranjo, S. *et al.* Multiple enhancers located in a 1-Mb region upstream of POU3F4 promote expression during inner ear development and may be required for hearing. *Human Genetics* **128**, 411–419 (2010).
- Thambisetty, M., Beason-Held, L., An, Y., Kraut, M. A. & Resnick, S. M. APOE epsilon4 genotype and longitudinal changes in cerebral blood flow in normal aging. *Arch. Neurol.* **67**, 93–98 (2010).
- Salloway, S. *et al.* A phase 2 multiple ascending dose trial of bapineuzumab in mild to moderate Alzheimer disease. *Neurology* **73**, 2061–2070 (2009).

21. Rogers, M. B. GammagardTM Misses Endpoints in Phase 3 Trial. *Alzheimer Research Forum* (2013). at <<http://www.alzforum.org/new/detail.asp?id=3485>> (Date of access: May 2013)
22. Altmann, A., Tian, L., Henderson, V. W. & Greicius, M. D. Alzheimer's Disease Neuroimaging Initiative Investigators. Sex modifies the APOE -related risk of developing Alzheimer disease: Sex and APOE -Related AD Risk. *Annals of Neurology* **75**, 563–573 (2014).
23. Lamb, J. The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer* **7**, 54–60 (2007).
24. Uittenbogaard, M., Baxter, K. K. & Chiaramello, A. The neurogenic basic helix-loop-helix transcription factor NeuroD6 confers tolerance to oxidative stress by triggering an antioxidant response and sustaining the mitochondrial biomass. *ASN Neuro* **2**, e00034 (2010).
25. De la Monte, S. M., Luong, T., Neely, T. R., Robinson, D. & Wands, J. R. Mitochondrial DNA damage as a mechanism of cell loss in Alzheimer's disease. *Lab. Invest.* **80**, 1323–1335 (2000).
26. Wang, Y. *et al.* Muscle-specific mutations accumulate with aging in critical human mtDNA control sites for replication. *PNAS* **98**, 4022–4027 (2001).
27. Kawas, C., Gray, S., Brookmeyer, R., Fozard, J. & Zonderman, A. Age-specific incidence rates of Alzheimer's disease The Baltimore Longitudinal Study of Aging. *Neurology* **54**, 2072–2077 (2000).
28. Chen, H.-K. *et al.* Apolipoprotein E4 domain interaction mediates detrimental effects on mitochondria and is a potential therapeutic target for Alzheimer disease. *J. Biol. Chem.* **286**, 5215–5221 (2011).
29. Lauderback, C. M. *et al.* Apolipoprotein E modulates Alzheimer's A $\beta$ (1–42)-induced oxidative damage to synaptosomes in an allele-specific manner. *Brain Research* **924**, 90–97 (2002).
30. Melov, S. *et al.* Mitochondrial Oxidative Stress Causes Hyperphosphorylation of Tau. *PLoS ONE* **2**, e536 (2007).
31. Iijima-Ando, K. *et al.* Loss of Axonal Mitochondria Promotes Tau-Mediated Neurodegeneration and Alzheimer's Disease-Related Tau Phosphorylation Via PAR-1. *PLoS Genet* **8**, e1002918 (2012).
32. Tong, Y. *et al.* Oxidative stress potentiates BACE1 gene expression and Abeta generation. *J Neural Transm* **112**, 455–469 (2005).
33. Frederikse, P. H., Garland, D., Zigler, J. S. & Piatigorsky, J. Oxidative Stress Increases Production of  $\beta$ -Amyloid Precursor Protein and  $\beta$ -Amyloid (A) in Mammalian Lenses, and A Has Toxic Effects on Lens Epithelial Cells. *J. Biol. Chem.* **271**, 10169–10174 (1996).
34. Burgoyne, R. D. & Morgan, A. Chaperoning the SNAREs: a role in preventing neurodegeneration? *Nat. Cell Biol.* **13**, 8–9 (2011).
35. Sharma, M. *et al.* CSP $\alpha$  knockout causes neurodegeneration by impairing SNAP-25 function. *EMBO J* **31**, 829–841 (2012).
36. Keating, D. J. Mitochondrial dysfunction, oxidative stress, regulation of exocytosis and their relevance to neurodegenerative diseases. *J. Neurochem.* **104**, 298–305 (2008).
37. Cassano, T. *et al.* Glutamatergic alterations and mitochondrial impairment in a murine model of Alzheimer disease. *Neurobiology of Aging* **33**, 1121.e1–1121.e12 (2012).
38. Haque, E. *et al.* The effect of mutated mitochondrial ribosomal proteins S16 and S22 on the assembly of the small and large ribosomal subunits in human mitochondria. *Mitochondrion* **8**, 254–261 (2008).
39. Lugo-Huitrón, R. *et al.* On the antioxidant properties of kynurenic acid: Free radical scavenging activity and inhibition of oxidative stress. *Neurotoxicology and Teratology* **33**, 538–547 (2011).
40. Gulaj, E., Pawlak, K., Bien, B. & Pawlak, D. Kynurenic acid and its metabolites in Alzheimer's disease patients. *Advances in Medical Sciences* **55**, 204–211 (2010).
41. Zwilling, D. *et al.* Kynurenic acid 3-Monooxygenase Inhibition in Blood Ameliorates Neurodegeneration. *Cell* **145**, 863–874 (2011).
42. Corbett, G. T., Roy, A. & Pahan, K. Sodium Phenylbutyrate Enhances Astrocytic Neurotrophin Synthesis via Protein Kinase C (PKC)-mediated Activation of cAMP-response Element-binding Protein (CREB) IMPLICATIONS FOR ALZHEIMER DISEASE THERAPY. *J. Biol. Chem.* **288**, 8299–8312 (2013).
43. Ricobaraza, A. *et al.* Phenylbutyrate ameliorates cognitive deficit and reduces tau pathology in an Alzheimer's disease mouse model. *Neuropsychopharmacology* **34**, 1721–1732 (2009).
44. Yao, J., Chen, S., Mao, Z., Cadenas, E. & Brinton, R. D. 2-Deoxy-D-Glucose Treatment Induces Ketogenesis, Sustains Mitochondrial Function, and Reduces Pathology in Female Mouse Model of Alzheimer's Disease. *PLoS ONE* **6**, e21788 (2011).
45. Loy, R. & Tariot, P. N. Neuroprotective properties of valproate: potential benefit for AD and tauopathies. *J. Mol. Neurosci.* **19**, 303–307 (2002).
46. Long, Z. *et al.* Valproic acid attenuates neuronal loss in the brain of APP/PS1 double transgenic Alzheimer's disease mice model. *Curr Alzheimer Res* **10**, 261–269 (2013).
47. Qing, H. *et al.* Valproic acid inhibits Abeta production, neuritic plaque formation, and behavioral deficits in Alzheimer's disease mouse models. *J. Exp. Med.* **205**, 2781–2789 (2008).
48. Nie, H. *et al.* New nicotinic analogue ZY-1 enhances cognitive functions in a transgenic mice model of Alzheimer's disease. *Neurosci. Lett.* **537**, 29–34 (2013).
49. Diomedea, L. *et al.* Tetracycline and its analogues protect *Caenorhabditis elegans* from  $\beta$  amyloid-induced toxicity by targeting oligomers. *Neurobiol. Dis.* **40**, 424–431 (2010).
50. Ferretti, M. T., Allard, S., Partridge, V., Ducatenzeiler, A. & Cuello, A. C. Minocycline corrects early, pre-plaque neuroinflammation and inhibits BACE-1 in a transgenic model of Alzheimer's disease-like amyloid pathology. *J Neuroinflammation* **9**, 62 (2012).
51. Liang, W. S. *et al.* Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set. *Physiological Genomics* **33**, 240–256 (2008).
52. Blalock, E. M. *et al.* Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. USA* **101**, 2173–2178 (2004).
53. Nakabeppu, Y. & Hokama, M. Expression data from post mortem Alzheimer's disease brains (GSE36980). (2013). at <<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36980>> (Date of access: May 2013).
54. Webster, J. A. *et al.* Genetic Control of Human Brain Transcript Expression in Alzheimer Disease. *The American Journal of Human Genetics* **84**, 445–458 (2009).
55. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).
56. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
57. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).
58. Myers, A. Laboratory of Functional Neurogenomics Website at <<http://labs.med.miami.edu/myers/LFuN/data.html>> (Date of access: March 2013).
59. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
60. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology* **10**, R130 (2009).

61. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
62. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
63. Pérez-Cano, L. & Fernández-Recio, J. Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins* **78**, 25–35 (2010).
64. Kim, O. T. P., Yura, K. & Go, N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.* **34**, 6450–6460 (2006).
65. Liu, Z.-P., Wu, L.-Y., Wang, Y., Zhang, X.-S. & Chen, L. Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics* **26**, 1616–1622 (2010).
66. Lamb, J. *et al.* The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **313**, 1929–1935 (2006).

## Acknowledgements

We thank Valur Emilsson, Chris Gaiteri, Solly Sieberts, and Bin Zhang for their generous assistance with the data associated with their publication. We thank Kate Mahan for excellent scientific illustration (Fig. 4). **ADNI acknowledgment:** Data collection and sharing for the ADNI (SNP) data utilized in this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health ([www.fnih.org](http://www.fnih.org)). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514. All ADNI members with full affiliations are listed at: [https://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf). **LOAD dataset acknowledgment:** Funding support for the “Genetic Consortium for Late Onset Alzheimer’s Disease” was provided through the Division of Neuroscience, NIA. The Genetic Consortium for Late Onset Alzheimer’s Disease includes a genome-wide association study funded as part of the Division of Neuroscience, NIA. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by Genetic Consortium for Late Onset Alzheimer’s Disease. (Dataset accession: phs000168.v1.p1). For the Alzheimer’s Disease Neuroimaging Initiative -- A subset of SNP data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database ([adni.loni.ucla.edu](http://adni.loni.ucla.edu)). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found in Supplementary Information and at [http://adni.loni.ucla.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Authorship\\_List.pdf](http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf).

## Author Contributions

K.D.F., J.M.F., M.N.A., B.Z., S.E.K. and F.T. analyzed data, wrote the manuscript and prepared the figures.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** All authors are employees, advisors, and/or stockholders of Immuneering. Immuneering has applied for patent protection on the reported methods/findings.

**How to cite this article:** Fowler, K. D. *et al.* Leveraging existing data sets to generate new insights into Alzheimer’s disease biology in specific patient subsets. *Sci. Rep.* **5**, 14324; doi: 10.1038/srep14324 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>