# Alzheimer related genes show accelerated evolution

Anne Nitsche[a], Kristin Reiche[**,b], Uwe Ueberham[**,c], Christian Arnold[**,a,d,e], Jörg Hackermüller[*,f,g], Friedemann Horn[*,d], Peter F. Stadler[*,a,d,h,i,j,k,l], Thomas Arendt[*,c]

[a]Bioinformatics Group, Department of Computer Science, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[b]Bioinformatics Unit, Department of Diagnostics, Fraunhofer Institute for Cell Therapy and Immunology – IZI, Perlickstraße 1, D-04103 Leipzig, Germany
[c]Paul-Flechsig-Institut für Hirnforschung, Jahnallee 59, D-04107 Leipzig, Germany
[d]Department of Diagnostics, Fraunhofer Institute for Cell Therapy and Immunology – IZI, Perlickstraße 1, D-04103 Leipzig, Germany
[e]Computational EvoDevo, Department of Computer Science, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[f]Young Investigators Group Bioinformatics and Transcriptomics, Department Molecular Systems Biology, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany
[g]Department of Computer Science, University Leipzig, Leipzig, Germany
[h]Interdisciplinary Center for Bioinformatics, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[i]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
[j]Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
[k]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark
[l]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

## Abstract

Alzheimer's disease (AD) is a neurodegenerative disorder of unknown cause with complex genetic and environmental traits. Here, we show that gene structures of loci, that show AD-associated changes in their expression, evolve faster than the genome at large. This phylogenetic trait of AD suggests a critical pathogenetic role of recent adaptive evolution of human brain and might have far reaching consequences with respect to the appropriateness of model systems and the development of disease-modifying strategies.

*Key words:* Alzheimer's disease, ncRNA, tiling array, splice site conservation, brain evolution, adaptation

Alzheimer's disease (AD) is an age-related, chronic neurodegenerative disorder, neuropathologically characterized by fibrillar aggregates of Aβ-peptides and the microtubule-associated protein tau. While AD is extremely prevalent in human elderly, both Aβ and tau pathology are less common in non-primate mammals, and even non-human primates develop only an incomplete form of the disease [1]. This specificity of AD to human clearly implies a phylogenetic aspect of the disease and indicates that adaptive changes of cerebral structure and function that have occurred during human

evolution may have rendered the human brain sensitive to AD [2]. Still, the evolutionary dimension of the AD pathomechanism remains difficult to prove and has not been established unequivocally so far.

To prove the contribution of brain evolution towards the AD pathomechanism, here, we established the AD-associated genome-wide RNA profile comprising both protein-coding (cRNA) and non-protein-coding (ncRNA) transcripts. For the first time, we also applied a systematic analysis on the conservation of splice sites based on multiple alignments [3] across 18 vertebrates of homologs of AD-associated protein-coding and non-coding genes [Supplement Sec. 1.8].

To this end, we designed a custom array comprising 931, 898 probes derived from Agilents Whole Human Genome Oligo array, long non-coding RNA (lncRNA) probes extracted from public databases, computationally predicted loci of structured RNAs, and lncRNA probes experimentally identified by transcriptome-wide expression variation studies based on the Affymetrix

---

*Corresponding authors
**Contributed equally

*Email addresses:* anne@bioinf.uni-leipzig.de (Anne Nitsche), kristin.reiche@izi.fraunhofer.de (Kristin Reiche), uwe.ueberham@medizin.uni-leipzig.de (Uwe Ueberham), christian.arnold@embl.de (Christian Arnold), joerg.hackermueller@ufz.de (Jörg Hackermüller), friedemann.horn@izi.fraunhofer.de (Friedemann Horn), studla@bioinf.uni-leipzig.de (Peter F. Stadler), thomas.arendt@medizin.uni-leipzig.de (Thomas Arendt)

Human Tiling 1.0 array comparing AD patients with control samples [Supplementary Methods]. Applying this custom array to 19 AD patients and 22 control samples, we identified a differential expression of 154 multi-exonic cRNAs with a total of 4, 162 splice sites and 141 multi-exonic lncRNAs with a total of 1, 297 splice sites.

Genome-wide studies that systematically analyze the evolutionary age of protein-coding and non-protein-coding AD-associated genes have not been performed previously. While major evolutionary changes might have occurred at the transcriptomic level, they appear to be particularly pronounced for lncRNAs [3, 4]. As shown by analyses of sequenced genomes of a large variety of species, the relative amount of non-coding sequence increases consistently with complexity [5]. Thus, lncRNAs, most likely constitute a critical layer of gene regulation in complex organisms that has expanded during evolution [6]. However, the evolutionary histories of lncRNAs have been notoriously hard to study due to their usually low level of sequence conservation. This not only hampers comprehensive homology-based annotation efforts but also makes it nearly impossible to obtain the high fidelity sequence alignments that are required for in depth studies into their evolution. Alternatively, the conservation of gene structure and particularly the conservation of splice sites may also be used to establish homology of lncRNAs [3]. Splice sites therefore leave "phylogenetic footprints", and conserved patterns of splice sites may be used to predict novel transcripts from multiple genome alignments [7, 8]. Although lncRNAs are clearly ancient components of vertebrate genomes, they exhibit a rapid turnover of their intron/exon structures [3] that may be indicative of functional adaptation.

While the disease-relevance of lncRNAs is increasingly recognized, previous systematic gene expression profiling studies in AD nevertheless focussed predominantly on protein-coding genes. Consequently, so far, only a few individual AD-associated ncRNAs have been identified and functionally characterized [9].

In order to compare the conservation of genes at a structural level, we classify the data by the "degree of conservation" $c$, which is the fraction of conserved splice junctions per gene. We ask – for a fixed value of $c$ – whether loci that are differentially expressed in AD patients show signs of accelerated evolution compared to the set of genes contained in the Gencode v14 annotation of the human genome. Different thresholds of $c$ highlight different aspects of conservation and evolutionary change: At $c > 0\%$ we assay only presence or absence of a gene, and thus its evolutionary origin. The other extreme, $c = 100\%$ focusses on the precise con-

servation of the gene structure. Protein-coding genes and non-protein-coding genes were independently investigated for their conservation [Supplement, Fig. S1].

Nearly all AD-associated protein-coding genes are evolutionarily old (Fig. 1C). There were no differences in conservation rate at $c > 0\%$ between AD-associated and all protein-coding genes, i.e., AD-associated protein-coding genes did not originate later in evolution than other protein-coding genes. In line with previous reports [6], lncRNAs are much less well conserved and many have emerged in the course of mammalian evolution. The fraction of conserved lncRNAs thus decreases rapidly with evolutionary time (Fig. 1A,B). As for protein-coding sequences we did not observe a significantly younger origin of AD-associated genes.

While there is no recognizable difference in the evolutionary age of origin between AD-associated genes compared to the transcriptome as a whole (Fig. 1C), we observed significant, albeit more subtle differences in the evolution of AD-associated and general lncRNAs, concerning the changes in gene structure. With an increasing degree of conservation $c$, the conservation rate of AD-associated non-coding genes decreases and becomes significantly ($p < 0.05$) distinguishable from the background level (Fig. 1A,B) not only for distantly related mammals but even primates, when complete conservation of gene structure is considered ($c = 100\%$). In other words, the fraction of transcripts that have the entirety of their splice sites conserved is smaller amongst AD-associated ncRNAs than amongst non-coding genes at large. AD-associated ncRNAs hence show an accelerated evolution of their gene structure. This is indicative of a more rapid functional adaptation of AD-associated non-coding genes.

Although protein-coding genes are much better conserved than lncRNAs we observed the same increase of splice site turnover in AD at $c = 100\%$. In fact, the relative effect is even stronger compared to non-protein-coding loci ($\approx$ 30-40% *versus* $\approx$ 5-15% difference, shown as red lines in Fig. 1B and 1D, respectively). Since the same fraction of transcripts is already detectable at low conservation degrees, while the conservation rate decreases with higher $c$, we conclude that splice sites are systematically less conserved in human AD-associated regions compared to the typical behavior of the transcriptome.

While protein-coding loci exhibit an enhanced rate of small changes in their gene structure, we observe large changes in lncRNAs, again with a significantly enhanced rate in the AD-associated ncRNAs. This suggests that in particular AD-associated non-coding genes

2

play an important, as yet largely unexplored, role in the AD pathomechanisms.

We have shown here that gene structures of AD-associated loci evolve faster than the genome at large, while there is no evidence that AD-associated genes originated particularly late in evolution. In order to capture the evolution of lncRNAs, we focused on gene structure, i.e., the conservation of splice sites because this approach makes it possible to separate the evolution of the transcripts from other selective constraints such as regulatory DNA elements that may affect sequence conservation [3]. Changes in gene structure can be expected to have in general larger functional effects than point mutations. The enhanced rate of gene structure evolution in AD-related genes hints a relation of AD to recent adaptive evolution, presumably in relation to the rapid evolution of the human brain. Importantly, replacing the background set by only gene expressed in brain did not affect the conclusions [Supplement, Sec. 2.3].

Major phenotypic brain changes that have occurred in the course of recent human evolution, in particular between human and chimpanzee, appear to be mostly the result of an increase in gene expression and are, thus, reflected at the transcriptomic level. [10–12]. Genes whose expression has increased in human brain are mainly related to growth and differentiation [13] and frequently are involved in transcriptional regulation and RNA processing [10, 11]. The most significant differences in gene expression between the human and non-human primate brain have been observed in the association cortex [10, 14], i.e., brain areas that have expanded during hominid evolution [15] and are affected in AD most early and most constantly [16]. Evolutionary expansion of the neocortex, and in particular phylogenetic shaping of association areas, is associated with a developmental deceleration and an extended period of high neuronal plasticity into adulthood [13]. The presence of these neurons which remain structurally immature throughout their lifespans might provide the prerequisite both for the human adaption to the "cognitive niche" and for a high vulnerability towards factors that lead to the development of AD [17–19].

Our data support the concept that neuronal vulnerability in AD is a result of the evolutionary legacies that have occurred during the course of evolution of the human brain, making AD an example of antagonistic pleiotropy. This evidence for a phylogenetic trait of AD highlights the necessity to reconsider our approaches to define the molecular pathology of AD and the appropriateness of current animal model systems [20] to develop disease-modifying strategies.

# References

[1] Perez SE, et al. Alzheimer's disease pathology in the neocortex and hippocampus of the western lowland gorilla (Gorilla gorilla gorilla). Journal of Comparative Neurology 521, 2013:4318–4338.

[2] Rapoport S. Hypothesis: Alzheimer's disease is a phylogenetic disease. Medical Hypotheses 29, 1989:147–150.

[3] Nitsche A, Rose D, Fasold M, Reiche K, Stadler PF. Comparison of splice sites reveals that long non-coding RNAs are evolutionarily well conserved. RNA 21, 2015:801–812. Doi: 10.1261/rna.046342.114.

[4] Necsulea A, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505, 2014:635–640. doi:10.1038/nature12943.

[5] Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 29, 2007:288–299.

[6] Mattick JS. The functional genomics of noncoding RNA. Science 309, 2005:1527–1528.

[7] Hiller M, et al. Conserved introns reveal novel transcripts in *Drosophila melanogaster*. Genome Res 19, 2009:1289–1300.

[8] Rose D, Hiller M, Schutt K, Hackermüller J, Backofen R, Stadler PF. Computational discovery of human coding and non-coding transcripts with conserved splice sites. Bioinformatics 27, 2011:1894–1900.

[9] Zhang Z. Long non-coding RNAs in alzheimer's disease. Curr Top Med Chem 16, 2016:511–519.

[10] Cáceres M, et al. Elevated gene expression levels distinguish human from non-human primate brains. Proceedings of the National Academy of Sciences 100, 2003:13030–13035.

[11] Uddin M, et al. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. Proceedings of the National Academy of Sciences of the United States of America 101, 2004:2957–2962.

[12] Fukuda K, et al. Regional DNA methylation differences between humans and chimpanzees are associated with genetic changes, transcriptional divergence and disease genes. Journal of Human Genetics 58, 2013:446–454.

[13] Somel M, et al. Transcriptional neoteny in the human brain. Proceedings of the National Academy of Sciences 106, 2009:5743–5748.

[14] Cáceres M, Suwyn C, Maddox M, Thomas JW, Preuss TM. Increased cortical expression of two synaptogenic thrombospondins in human brain evolution. Cerebral Cortex 17, 2007:2312–2321.

[15] Stephan H. Evolutionary trends in limbic structures. Neuroscience & Biobehavioral Reviews 7, 1983:367–374.

[16] Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathologica 82, 1991:239–259.

[17] Arendt T. Alzheimer's disease as a loss of differentiation control in a subset of neurons that retain immature features in the adult brain. Neurobiology of aging 21, 2000:783–796.

[18] Bianchi S, et al. Synaptogenesis and development of pyramidal neuron dendritic morphology in the chimpanzee neocortex resembles humans. Proceedings of the National Academy of Sciences 110, 2013:10395–10401.

[19] Bufill E, Blesa R, Agustí J. Alzheimer's disease: an evolutionary approach. Journal of Anthropological Sciences 91, 2013:1–23.

[20] Zahs KR, Ashe KH. 'too much good news'–are Alzheimer mouse models trying to tell us how to prevent, not cure, Alzheimer's disease? Trends in Neurosciences 33, 2010:381–389.

*Author contributions.*

TA, UU, KR, CA, and JH designed and performed the tiling array and custom array experiments. KR, CA, and JH performed bioinformatic analysis of the expression data. AN developed and performed the computational method to assess the conservation of splice sites.
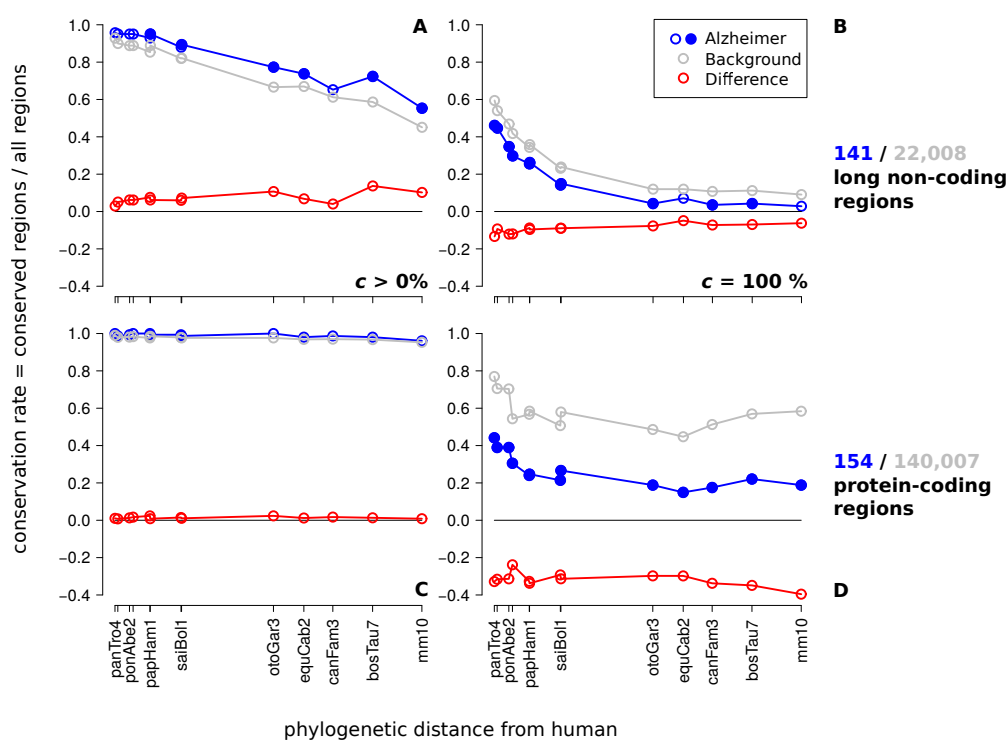
Figure 1: **Conservation rates of human AD-associated non-protein-coding (A,B) and protein-coding (C,D) regions** comparing presence ($c > 0\%$) and precisely conserved gene structure ($c = 100\%$). On the horizontal axis mammalian species are indicated (denoted by the UCSC abbreviations) at their phylogenetic distance from human. Distinct data points are connected by lines to guide the eye. Variations in assembly and alignment quality cause some non-monotonicity in the curves, the overall decrease of conservation with phylogenetic distance is nevertheless clearly visible. Statistical significance of differences is computed independently for each species. Filled circles indicate $p < 0.05$. The fraction of detectable conserved AD-associated non-coding genes is marginally higher than the conservation of the background set non-coding transcripts if only presence/absence of a transcript is considered (**A**). In contrast, if conservation of the entire gene structure is considered, AD-associated genes are significantly less conserved than the control. This is true for both lncRNAs (**B**) and protein-coding genes (**D**). The Supplement provides further conservation rate results for conservation degree $c > 60\%$. Additional controls against possible confounding effects e.g. of alignment quality in Supplement Fig. S4 and Fig. S5 show that the trends shown here are robust.