**Selection of New Model Organisms for Complex Genetic Diseases:**

**Alzheimer Disease**

Joseph D. Romano1 and Indra Neil Sarkar1,2,3

1Department of Microbiology and Molecular Genetics;

2Center for Clinical and Translational Science; and

3Department of Computer Science

University of Vermont

Burlington, VT 05405

To Whom Correspondence Should be Addressed:

Indra Neil Sarkar, PhD, MLIS

Center for Clinical and Translational Science

University of Vermont

89 Beaumont Avenue

Given Courtyard N309

Burlington, VT 05405 USA

Email: [neil.sarkar@uvm.edu](mailto:neil.sarkar@uvm.edu)

Phone: +1-802-656-8283

Facsimile: +1-802-656-4589

**Description of Authors**

**Joseph D. Romano** is an undergraduate honors student majoring in Molecular Genetics at the University of Vermont.

**Indra Neil Sarkar** is the Director of Biomedical Informatics in the Center for Clinical and Translational Science, in addition to being an Assistant Professor in Microbiology and Molecular Genetics as well as Computer Science at the University of Vermont. His research involves the development of biomedical informatics methods across the entire spectrum of life, from molecules to populations.

**Competing Interests**

The authors declare that they have no competing interests.

ABSTRACT

INTRODUCTION

The life sciences have relied on the use of “model” organisms to observe and study conserved functions across species. Two significant reasons to use model organisms in place of human subjects include: (1) the obstacles presented by the ethics of performing research on human subjects; and (2) the slowness and technical obstacles associated with acquiring results from human subjects in a controlled setting. The second of these is partially an artifact of the lifespan of humans, the low yield in offspring, and the difficulty to adequately inbreed human populations to allow for controlled genetic experiments. Model organisms have enabled researchers to circumvent these kinds of challenges while providing an insightful estimate of human physiology and genetics.

It is current practice to select a well-established model organism for research in comparative genomics. However, it might be worthwhile to consider whether or not a chosen model is best suited for modeling the particular human function of interest. Furthermore, it may be necessary to identify a new model species for research of a particular trait. The selection of such a new species must be subject to scrutiny by the biomedical community, and should be chosen using a carefully selected set of criteria.

A model organism for biomedical research must possess multiple traits to be useful to researchers. The availability of data prior to adopting the model is the first criterion, required to mitigate the time of adoption by the research community. Second, the proposed model organism needs to be one that is easy to work with, and therefore easy to observe for target traits and genotypes/phenotypes. The third major consideration is whether the proposed model organism’s genetics and molecular physiology are similar enough to those of a human to allow for generation of useful and relatable data. The promise of the future applications of biomedical research increasingly suggest that clinical diagnoses and procedures will enable the leverage of genomic knowledge in the context of medicine 1. In support of this, there will be an increasing need to identify model systems that can provide additional views of biological processes in light of biomedicine. Such perspectives will only be possible with the expansion of model organisms beyond the current choices (predominantly *M. musculus* and *R. norvegicus*).

Although the molecular mechanics of genetics compose a relatively new field of knowledge, humans have been manipulating the genomes of domesticated animals for tens of thousands of years 2. Through successive generations of selection for desirable traits (e.g., dairy output), the genomes of these species have been effectively standardized. Inbreeding of cattle, swine, and sheep have thus resulted in genomic standardization necessary for successful agriculture. Animal species of agricultural significance may thus be suitable candidates to consider as potential models.

To explore the validity of this statement, we have developed a method to define a set of organisms as potential models for a specific complex genetic disease, and subsequently assign scores to each, indicating the degree to which it accurately models the disease. Our methods make use entirely of freely accessible utilities and databases for gene/protein sequence analysis in addition to scripts and utilities written by the authors and available for public use under the Creative Commons license [will do before submitting for review! – on Github or similar code repository]. The disease we have used to test our algorithm is Alzheimer Disease – one of the most clinically significant and well-characterized complex genetic disorders [CITE].

MATERIALS & METHODS

Databases and Software

All databases and software not created by the authors are freely accessible by the public, and many are likewise open-source. Initial iterative BLAST and PSI-BLAST alignments use the Basic Local Alignment Search Tool (BLAST) algorithm published by the National Center for Biotechnology Information (NCBI). Nucleotide and protein sequences used in all analyses were from NCBI’s nr (non-redundant) database, covering all organisms indexed by GenBank and GenPept. MUltiple Sequence Comparison by Log-Expectation (MUSCLE) was used to generate multiple sequence alignments. Phylogenetic trees are constructed using the maximum parsimony method.

All scripts and programs created by the authors are written in Ruby and make use of the BioRuby class for fetching and manipulating sequences.

Selection of Potential Model Organisms and Candidate Homologous Sequences

OMIM was used to identify polymorphisms that have been shown to correlate with an increased risk for developing AD in humans. Of those polymorphisms, the ones corresponding to a specific known gene product were selected for the following analyses. The PACIP1 gene was excluded from this analysis, due to it containing 2 N-terminal and 3 C-terminal BRCA1 domains [CITATION FROM OMIM] that make identification of significant homologies very difficult. The BCHE (Butyrylcholinesterase) gene was added to this list manually, due to recent preliminary studies that have not yet been reflected in OMIM [FIND CITATION] suggesting a correlation.

The nucleotide sequences of these genes were each run against two sequence alignment algorithms: NCBI’s Position Specific Iterative BLAST (PSI-BLAST) and a custom scripted iterative BLAST. PSI-BLAST translates the query sequence into an amino acid sequence and builds a position specific scoring matrix (PSSM) containing similar sequences. Successive iterations of PSI-BLAST search the subject database for sequences that align with the PSSM, and subsequently add those sequences to the PSSM for successive iterations. The custom iterative BLAST begins with running BLASTx (translating the query into an amino acid sequence and searching for homologous protein sequences), building a list of the result sequences, and then iterating through every sequence in the results with BLASTp (using the protein query to search for homologous protein sequences), and adding any new gene results to the end of the result list. For both methods, the database used was NCBI’s nr (non-redundant) database, which contains sequences from all indexed species. Additionally, the Expect (E-value) threshold for each method was restricted to 0, to ensure that reported alignments represent actual homologies and are not due to chance sequence similarities. The resulting lists generated for each AD gene are sequences that are highly homologous to those genes, searched across a database of all organisms indexed by NCBI.

Potential model organisms were then determined according to the criterion that they are represented in the list of homologies for every AD gene. The resultant list of organisms (excluding *Homo sapiens*) includes 19 distinct species, represented in [MAKE A TABLE?]. As may be expected evolutionarily, all of these species are mammals. Additionally, 5 of the 19 species are primates, supporting the expectation that potential models will be evolutionarily closely related to *H. sapiens*. It is worth noting that *Mus musculus* and *Rattus norvegicus* – the current “gold standard” mammalian models in most laboratory settings – are both included.

Verification of Evolutionary Significance

For each organism in the list of potential models, a single gene was selected from the combined PSI-BLAST / custom iterative BLAST to be compared to each of the *H. sapiens* AD genes, resulting in a selected list of 19 homologous genes to each AD gene. In most instances, the BLAST results for each gene had multiple homologies in each potential model organism. When this was the case, the “best” homology was selected by using each of these genes as subjects in BLASTp (where the corresponding AD gene is the query), and that sequence was the one used.

To further ensure the selected genes are good candidates, sequence lengths were observed for wide deviations from the length of the human AD gene to which they are homologous. Such an analysis is intended to alleviate the risk of having selected small gene fragments or large portions of chromosomes, either of which would not be considered appropriate homologs. Additionally, similarity matrices for each group of genes homologous to an AD gene were created, with the metric of similarity being the expect value (E-value) returned by a BLAST2 alignment of the two protein sequences. If the selected sequences are both of a similar length and return BLAST2 E-values equal to or near zero (less than 1e-100), the sequence is similar enough to be used in additional analyses.

Two existing databases of orthologous sequence groups (InParanoid7 and OrthoMCL DB) were queried for each of the human AD genes to see if they returned any additional species that supplement the 19 already identified by our methods.

Phylogenetic Analysis and Ranking Potential Models

MUSCLE was used to generate multiple sequence alignments of all genes related to each AD gene, resulting in 10 alignments (one for each AD gene) of 20 sequences (one for H. sapiens and one for each of the 19 potential model species). As of completing this draft of the manuscript, this is the current extent of completed work. The complete study will involve the generation of phylogenetic trees for each multiple sequence alignment using the maximum parsimony method. Additionally, the alignments will be combined mathematically to generate a numerical score of “model-ness”, based on a vector-space approach for determining similarity.

RESULTS

As of the time of writing, the study has not yet concluded. However, preliminary results show a trend suggesting a likely outcome. A list of potential model species has been created, based upon data generated by PSI-BLAST results combined with the results of the custom iterative BLAST model described previously. Excluding humans, the list of 19 identified organisms is as follows:

|  |
| --- |
| Pan troglodytes |
| Nomascus leucogenys |
| Macaca mulatta |
| Papio anubis |
| Saimiri boliviensis boliviensis |
| Callithrix jacchus |
| Ailuropoda melanoleuca |
| Felis catus |
| Otolemur garnettii |
| Bos taurus |
| Ovis aries |
| Oryctolagus cuniculus |
| Tupaia chinensis |
| Rattus norvegicus |
| Mus musculus |
| Cavia porcellus |
| Pteropus alecto |
| Cricetulus griseus |
| Orca orca |

Each of these species shares at least one strongly homologous gene with each of the following 10 H. sapiens genes that are implicated in affecting susceptibility to Alzheimer Disease:

|  |
| --- |
| Alpha-2-macroglobulin (A2M) |
| Angiotensin I-converting enzyme (ACE) |
| Amyloid beta A4 precursor protein-binding, family B, member 2 (APBB2) |
| Amyloid beta A4 precursor protein (APP) |
| Butyrylcholinesterase\* (BCHE) |
| Bleomycin hydrolase (BLMH) |
| Myeloperoxidase (MPO) |
| Nitric oxide synthase 3 (NOS3) |
| Plasminogen activator, urinary (PLAU) |
| Sortilin-related receptor (SORL1) |

(\*Not currently listed in OMIM as a disease gene for AD)

For each AD gene, a similarity matrix including itself and all homologs was created (10 matrices total). Since the initial iterative BLAST and PSI-BLAST both specified an E-value cutoff of 0, most values in the similarity matrices likewise returned 0 - any non-zero values were small enough to indicate that they still are likely homologous (less than 1e-125 in all cases). In comparisons of AD gene sequence lengths to each of their homologs, no abnormal lengths (determined subjectively) were observed. These two comparisons serve as an additional manual screening measure to ensure all groups of homologs are of the highest possible similarity, with no selected sequences representing small gene fragments or large chromosomal assemblies.

Further analyses of these results are currently occurring, as per stated protocol detailed in the Methods section of this study.

DISCUSSION

It is likely of great significance that all organisms that were selected by our algorithm are mammalian species. Additionally, 5 of the 19 species are primates, supporting the expectation that potential models will be evolutionarily closely related to *H. sapiens*. It is worth noting that *Mus musculus* and *Rattus norvegicus* – the current “gold standard” mammalian models in most laboratory settings – are both included.

APP - the gene most commonly associated with AD susceptibility in humans - is known to exist even in many species very distantly related to humans, such as sea anemones and others (CITE). Additionally, many of the other genes implicated in AD susceptibility are understood to interact with either APP or the amyloid products of APP (namely; APBB2, ACE, PLAU, SORL1, A2M, BLMH). Other genes in the list have at least been hypothesized to do so (such as MPO), due to deposition in amyloid plaques of patients with AD. With this knowledge available, new hypotheses may be posed to describe a specific molecular mechanism through which Alzheimer Disease originates and progresses clinically. For investigating such hypotheses in a research setting, our list of ranked model organisms for AD may prove essential. Virtually equivalent disorders to AD - resulting in β-amyloid deposition in the brain accompanied with severe cognitive impairment and memory loss – are already known to exist in some of the organisms on our list of models3, as well as in organisms that are not on the list 4. It should be attempted to characterize similar disorders in all of the model organisms on the list. Based on the findings of this study, equivalent diseases in more highly ranked model organisms should most accurately reflect AD in humans, and those organisms should be considered first.

However, researchers must keep in mind that the results of our algorithm are purely based on sequence analysis, and although similarity in genes and gene products often correlates to similarity in metabolism and physiology, this is often not the case. This is the rationale behind referring to our results as potential model organisms instead of simply model organisms. Particularly if non-computational laboratory analyses are to be done one these species, many additional factors will dictate suitability as a model for the disease, such as physiological similarities to humans and the relative ease of working with the species in a laboratory setting. This fact in itself illustrates the need to identify alternative model organisms in the first place.

For example, *M. musculus* has a mammalian genome that is comparable to humans [24], however there are several notable differences that pose potential limitations to research. For example, size and life span differ greatly between mice and humans. Additionally, *M. musculus* metabolism and human metabolism differ significantly [25]. *M. musculus* has also been evolved to reproduce at a high rate over a short lifespan, while humans reproduce less frequently over a longer lifespan. Finally, the difference between the *M. musculus* and human immune system necessitated the need to develop transgenic *M. musculus* strains to allow for limited generation of useful immunologic data for clinical application to human health [26,27].

In addition to gene homologies with human orthologues, the metabolism and anatomy of *B. taurus* (or of other mammalian agricultural species, such as *Ovis aries*, *Sus scrofa*, or *Equus ferus*) may be reasons to consider it as a suitable model organism. While model organisms should represent estimations of human systems, a significant factor for selection has historically been ease-of-use in the laboratory. However, while short lifespans and easily observable phenotypes may be essential in early studies of disease processes, there may be merit in advancing subsequent studies to include those that have more physiological characteristics similar to humans. To this end, humans and many domesticated or agricultural species are closer in physical size, reproduce at controlled rates (indicative of “stabilizing” reproduction), and genetically are evolved to live in stable environments (e.g., in well-managed farm facilities) 5. By contrast, many current model organisms (e.g., mice) are significantly different in size than humans, have a short lifespan, designed to reproduce in very large numbers quickly (“opportunistic” reproductive strategy), and genetically are evolved to live in highly dynamic environments 6.

Logistically, implementing a given agricultural species as an additional model organism should be simpler than other types of non-model animals (e.g., wild animals). Agricultural animals are already domesticated and confined within farms and agricultural facilities, so researchers could gather blood and tissue samples on an as-needed basis, and make observations of the specimens as dictated by the specific research endeavor. Tangentially, the involvement of farmers in enrolling in such a program could yield economic benefits, such as a monetary incentive for providing researchers with samples. The bioethical impact of this process could be designed to be minimal – the life of animal specimens would not need to be compromised, and their living conditions should not need to be altered. Nonetheless, it is still important to underscore that mice and other rodents are essential as initial mammalian model organisms for many reasons, including the ability to control laboratory environments. Mice and other rodents also have certain genetic traits that have been highly selected for through many generations of inbreeding and documentation of phenotypic expression7. Agricultural species have undergone a similar process of selection, with the aim of isolating traits with agriculturally beneficial traits, and are similarly well-observed 8. Agricultural species may prove to supplement model organism based studies before studies involving human subjects. The addition of agricultural animals to the list of model organisms could thus serve to bridge the gap between knowledge inferred by studies on mice and better enable the translation of that knowledge to the clinical realm 9,10. This is aptly demonstrated by the aforementioned fact that transgenic *M. musculus* strains are required for studies in human immunological processes. While species such as *B. taurus* have not yet demonstrated to have a “more human” immune system, an intermediate species could alleviate current efforts required to make these inferences between mice and humans 11.

REFERENCES

1. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genetics in medicine : official journal of the American College of Medical Genetics [Internet] 2007;9:665-674. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=18073579&retmode=ref&cmd=prlinks

2. Zeder MA. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proceedings of the National Academy of Sciences of the United States of America [Internet] 2008;105:11597-11604. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=18697943&retmode=ref&cmd=prlinks

3. Gunn-Moore DA, McVee J, Bradshaw JM, Pearson GR, Head E, Gunn-Moore FJ. Ageing changes in cat brains demonstrated by beta-amyloid and AT8-immunoreactive phosphorylated tau deposits. Journal of feline medicine and surgery [Internet] 2006;8(4):234-42. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=16603401&retmode=ref&cmd=prlinks

4. Chambers JK, Uchida K, Harada T, et al. Neurofibrillary tangles and the deposition of a beta amyloid peptide with a novel N-terminal epitope in the brains of wild Tsushima leopard cats. PloS one [Internet] 2012;7(10):e46452. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23056312&retmode=ref&cmd=prlinks

5. Fadiel A, Anidi I, Eichenbaum KD. Farm animal genomics and informatics: an update. Nucleic acids research [Internet] 2005;33:6308-6318. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=16275782&retmode=ref&cmd=prlinks

6. Pianka E. On r-and K-selection. The American Naturalist [Internet] 1970;Available from: http://www.jstor.org/stable/10.2307/2459020

7. [Internet]. The Jackson Laboratory; Available from: http://jaxmice.jax.org/list/cat481365.html

8. [Internet]. University of Missouri Extension; Available from: http://extension.missouri.edu/p/G2911

9. Attfield KE, Dendrou CA, Fugger L. Bridging the gap from genetic association to functional understanding: the next generation of mouse models of multiple sclerosis. Immunological reviews [Internet] 2012;248:10-22. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22725951&retmode=ref&cmd=prlinks

10. Schmouth J, Bonaguro RJ, Corso-Diaz X, Simpson EM. Modelling human regulatory variation in mouse: finding the function in genome-wide association studies and whole-genome sequencing. PLoS genetics [Internet] 2012;8:e1002544. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22396661&retmode=ref&cmd=prlinks

11. Watanabe Y, Takahashi T, Okajima A, et al. The analysis of the functions of human B and T cells in humanized NOD/shi-scid/gammac(null) (NOG) mice (hu-HSC NOG mice). International immunology [Internet] 2009;21:843-858. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19515798&retmode=ref&cmd=prlinks