Exploring Complex Disease Gene Relationships Using Simultaneous Analysis

Joseph D. Romano[1], William G. Tharp[2], and Indra Neil Sarkar[1,3,4]

[1]Department of Microbiology and Molecular Genetics;
[2]Division of Endocrinology, Department of Medicine;
[3]Center for Clinical and Translational Science; and
[4]Department 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 S350
Burlington, VT 05405 USA
Email: neil.sarkar@uvm.edu
Phone: +1-802-656-8283
Facsimile: +1-802-656-4589

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.230v1 | CC-BY 4.0 Open Access | received: 1 Feb 2014, published: 1 Feb 2014

Abstract

The characterization of complex diseases remains a great challenge for biomedical researchers due to the myriad interactions of genetic and environmental factors. Adaptation of phylogenomic techniques to increasingly available genomic data provides an evolutionary perspective that may elucidate important unknown features of complex diseases. Here an automated method is presented that leverages publicly available genomic data and phylogenomic techniques. The approach is tested with nine genes implicated in the development of Alzheimer Disease, a complex neurodegenerative syndrome.

The developed technique, implemented through a suite of Ruby scripts entitled "ASAP2," first compiles a list of sequence-similarity based orthologues using PSI-BLAST and a recursive NCBI BLAST+ search strategy, then constructs maximum parsimony phylogenetic trees for each set of nucleotide and protein sequences, and calculates phylogenetic metrics (partitioned Bremer support values, combined branch scores, and Robinson-Foulds distance) to provide an empirical assessment of evolutionary conservation within a given genetic network.

This study demonstrates the potential for using automated simultaneous phylogenetic analysis to uncover previously unknown relationships among disease-associated genes that may not have been apparent using traditional, single-gene methods. Furthermore, the results provide the first integrated evolutionary history of an Alzheimer Disease gene network and identify potentially important co-evolutionary clustering around components of oxidative stress pathways.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.230v1 | CC-BY 4.0 Open Access | received: 1 Feb 2014, published: 1 Feb 2014

Introduction

Classical genetic diseases typically arise due to isolated genetic changes within a single gene or allele (Badano & Katsanis, 2002). Many of these "simple" or "monogenic" diseases follow Mendelian patterns of inheritance. The responsible genetic lesion is often the result of an insertion or deletion event, or the transversion / transposition of a nucleotide. The probability for transmission of simple genetic disorders may thus be easily predicted and generally follow sex-linked or autosomal patterns of heredity. Classic examples of monogenic disorders include cystic fibrosis, sickle cell anemia, and achondroplasia (Velinov et al., 1994; Kerem et al., 1989; Rees et al., 2010). By contrast, complex diseases or disorders may not follow clear hereditary patterns or be diagnosed based on isolated genetic lesions. However, many complex diseases such as cardiovascular disease, type 2 diabetes mellitus, and Alzheimer disease occur with higher frequency among families and close genetic relatives– suggesting that genetic factors may play a central role in their pathogenesis, beyond environmental or behavioral factors (Sillén et al., 2006). The risk of developing complex diseases or disorders and the future approaches for treating or preventing them may benefit from high-throughput, computational, or bioinformatics based approaches. For example, computational approaches, such as used in genome wide association studies, exome sequencing, proteomics, and microarray analyses, have shown great promise in recent years. Related advances in biotechnology have facilitated the identification of genotypes that may be factors involved in the heritability of complex genetic diseases (Yonan et al., 2003). For example, specific genotypes can be associated with a probabilistic value of susceptibility relative to the gene(s) they influence and thus correlated with a disease phenotype (Li et al., 2005; Newton-Cheh et al., 2009; Klein et al., 2012).

Due to a lack of knowledge about the specific mechanisms by which multiple genetic factors may influence complex diseases, pharmacotherapies are often aimed at managing symptoms or laboratory values, and are thus reactionary and not curative. Often, the approach to complex disease management involves attempting environmental changes, such as can be conveyed through patient education or lifestyle modification, to reduce susceptibility in addition to pharmacotherapy (Estruch et al., 2013). A major current goal of biomedical research is therefore to better characterize the genetic factors that may contribute to developing complex diseases. The fact that the genetic environment influences susceptibility to complex disease implicates the structural or functional relationships between some or all members of a disease associated gene network in the development of the disease (Li et al., 2005). This relationship might be a direct physical interaction between the protein products of the genes, parallel functionality in metabolic pathways, or co-localization of protein products in a certain cell or tissue type (Li et al., 2005). These data are not easily elucidated using an experimental approach focused on a single gene or pathway and require a broader systems-based methodology. These types of relationships may be reflected in the evolutionary conservation of genes or gene groups among organisms with and without susceptibility to a given disease (Thornton & DeSalle, 2000; Watson et al., 2014). Mapping the evolutionary patterns of gene conservation or co-evolution associated with a complex disease may identify previously unknown clusters of genes or functional pathways that have impact on a disease process.

Simultaneous Analysis

Phylogenetic analyses infer potential evolutionary relationships based on similarities implying common descent from shared ancestry and are performed on data sets consisting of

physical, functional, or molecular representations (Swiderski et al., 1998). Genomic analyses typically construct the analytic matrix using nucleotide or amino-acid sequences from different individuals or species (termed "taxa"; singular "taxon"). Classically, the resulting data are presented as trees where the branching points (termed "nodes") give rise to hierarchical groupings of more similar taxa (akin to leaves on a branch). These trees can be used to explore potential patterns of divergence from a common ancestor as well as the degree of difference among taxa included in the tree. This degree of difference is usually described as an evolutionary "distance" that can be inferred multiple ways, but typically represents a measure of evolutionary change (based upon sequence differences) or an amount of time since divergence likely occurred (Zharkikh, 1994; Hedges et al., 2006).

However, like experiments focused on a single gene or pathway, an isolated phylogenetic analysis may not capture important features of co-evolution or conservation of gene clusters impacting complex disease processes. Additionally, reliance on phylogenetic trees of individual genes may not fully address the potential for genetic changes such as lateral gene transfer, reversion of mutations, or recombination events (Dagan, 2011; Layeghifard et al., 2013).

To account for multiple evolutionary patterns represented by multiple genes, data matrices can be combined into a single phylogenetic analysis through a "simultaneous analysis" (SA) approach (Nixon & Carpenter, 1996; Gatesy et al., 1999; Rokas et al., 2003). In SA, individual data blocks (e.g., a sequence matrix for a particular gene; referred to as a "partition") are systematically combined to enable higher-order analyses that transcend data derived from analysis of an individual partition. Frequently, SA values are derived by applying arithmetic operations on other (already determined) SA values, so the workflow tends to follow a stepwise pattern. Previous studies have shown that SA techniques strengthen the overall support for the

evolutionary patterns represented by trees determined by single partition phylogenetic analyses (Baker et al., 1998).

In this study, a previous automated SA approach (Automated Simultaneous Analysis Phylogenetics; ASAP (Sarkar et al., 2008)) was refined to collect and analyze disease genes based on: (1) the degree of corroboration between partitions; and (2) the support for an overall consensus tree modeling a putative evolutionary relationship common to all partitions, using maximum parsimony analysis (Fitch, 1971). The final phase then generates a phylogenetic network based on the Robinson-Foulds tree similarity metric (Robinson & Foulds, 1981).

Alzheimer Disease

Alzheimer Disease (AD), a complex neurodegenerative disorder, is the most common form of dementia, accounting for between 70 – 90% of diagnosed cases of dementia (Ferri et al., 2005; Mayeux & Stern, 2012). Worldwide, more than 24 million people are estimated to have AD, with estimates exceeding 80 million to be affected over the next 30 years (Ballard et al., 2011) The pathognomonic histological finding that originally defined AD is the presence amyloid plaques in cortical brain tissues (Hardy & Selkoe, 2002; Nelson et al., 2009; 2012). The plaques arise from the production, and eventual extracellular precipitation, of fibrillar aggregates of amyloid-β peptides causing disruption of the neural architecture and induction of inflammation (Hardy & Selkoe, 2002). Clinically, AD is characterized by progressive memory loss and cognitive decline, leading to general functional impairment (Blacker et al., 1994; Dubois et al., 2007; Querfurth & LaFerla, 2010; Karran et al., 2011; Nelson et al., 2009; 2012). However, the precise etiology of AD remains elusive. Amyloid plaques have been shown to differ widely in manifestation with regards to protein composition, structural characteristics, and prevalence

(Silverman et al., 1999; Lu et al., 2013). Despite advances in predicting the presence of the disease based on symptoms and diagnostic imaging, a definitive diagnosis of AD can only be made after an autopsy of the affected brain after death (Blacker et al., 1994; Dubois et al., 2007). As such, researchers are faced with significant difficulties both in studying the progression of the disease as well as identifying potential therapeutic techniques to prevent or treat the disease.

Further compounding the difficulty in identifying causes of AD is the fact that the majority of cases do not follow a well-characterized pattern of heritability, even though susceptibility to AD is widely considered to have a genetic basis (Sjögren et al., 1952; Silverman et al., 1999; McIlroy et al., 2000; Ramanan et al., 2013). Familial forms of AD have been identified resulting from single or double gene lesions leading to increased amyloid plaque burden, but these account for less than 5% of the total cases of AD (Tanzi, 2012). Over the course of decades of study, a wide range of human genes have been linked to susceptibility at various stages of life, suggesting sporadic AD has a multifactorial, complex genetic component (Tanzi, 2012). Thus, AD is a perfect candidate disease for testing the potential utility of SA techniques.

The overall goal of this study was to develop an automated method for an SA phylogenetic analysis and to use it to construct an evolutionary history for a set of genes that may be correlated with AD. The co-evolution was examined for nine genes previously identified as associated with AD susceptibility using a SA technique mediated with a series of scripts written in the Ruby scripting language. The resulting phylogenetic tree provides the first description of the co-evolution of genes that may impact the development and pathogenesis of AD. The resulting phylogenetic network further highlights a potentially important role of oxidative stress genes in the evolution of the AD gene network. The developed technique provides a framework

for an automated approach to study the co-evolution of gene sets associated with a complex disease using a robust phylogenetic methodology.

Materials and Methods

The automated collection and simultaneous phylogenetic analysis process was developed using a sequential set of Ruby (Matsumoto, n.d.) scripts (entitled and referred to henceforth as "ASAP2") that made use of the Bioruby gem (Goto et al., 2010) as well as the following freely available genomic or phylogenetic analysis tools: BLAST+ (Basic Local Alignment Search Tool (Camacho et al., 2009)), MUSCLE (MUltiple Sequence Comparison by Log-Expectation (Edgar, 2004)), and TNT (Tree analysis using New Technology (Goloboff et al., 2000)) . The overall workflow for ASAP2 is shown in **Figure 1**.

Alzheimer Disease Gene Sequences

Nucleotide sequences for genes implicated as contributing to a higher risk for Alzheimer Disease (AD) in humans were manually identified using the data associated with the "Alzheimer Disease; AD" entry in Online Mendelian Inheritance in Man (OMIM) (OMIM ID #104300) (McKusick, 2007). This yielded ten discrete genes shown to be related to AD, which were loaded into ASAP2. For the purposes of this study, nonspecific chromosomal regions that encompass numerous genes or the noncoding regions between them were not included.

Identifying Potentially Related Disease Genes Based on Sequence Similarity

From the initial set of human AD disease gene sequences, ASAP2 performed two types of sequence searches from within the non-redundant (nr) protein database maintained at the United

States National Library of Medicine's National Center for Biotechnology Information (NCBI) using NCBI BLAST+ (Camacho et al., 2009). First, a PSI-BLAST (Position-Specific Iterated BLAST) search (which searches for similar sequences using an iterative profiling approach (Altschul et al., 1997)) was done for each gene sequence. Second, a recursive search was initiated with a BLASTx (which uses a translated nucleotide sequence query to perform a protein search) search of the nr protein database for the gene of interest and the best results used to iteratively search for additional protein sequences using BLASTp (which searches for protein sequences using an amino-acid query) until no additional sequences were found. An expect value (E-value) of 1.0×10^{-256} was used as the criterion for inclusion of results for both the PSI-BLAST and the recursive BLAST algorithm. Candidate data partition matrices for each gene were then constructed based on the combination of PSI-BLAST and recursive BLAST results. Corresponding nucleotide sequences were determined based on information in the DBSOURCE metadata field that links a given protein sequence to its coding nucleotide sequence.

After candidate data partitions were assembled, ASAP2 culled taxa and sequences from each data partition that were not uniformly represented for each gene (i.e., a sequence for a given species must be present in each data partition for that species to be retained for further analysis). Additionally, if any species was represented more than once in any partition, ASAP2 only kept the first (most similar according to BLAST) protein sequence for that species. ASAP2 then assembled the resulting data partitions into FASTA files, aligned them using the default parameters of MUSCLE, and formatted them into TNT-compatible data matrix files. The generated FASTA files and TNT data matrix files are available as Supplementary Data.

Phylogenetic Analyses

ASAP2 used TNT to conduct the maximum parsimony phylogenetic analyses of each data partition. Trees were constructed using tree bisection and reconnection (TBR) rearrangement, finding optimal scores 20 times followed by 10 cycles of tree-drifting. Subsequently, group support values were determined by counting the minimum number of steps needed to lose each group by TBR rearrangement (Goloboff & Farris, 2001). The TNT analysis included individual plotting of apomorphies and synapomorphies, bootstrap resampling, and calculation of both the relative and absolute Bremer support values at each branch.

ASAP2 then generated a SA consensus tree using TNT by creating an interleaved matrix of the data partitions. The interleaved matrix was built by concatenating each aligned data partition (minus headers and metadata) sequentially, separated by line breaks, into a single TNT data file. This data file was then interpreted by TNT as if the sequences for each species were concatenated in the order of the data partitions in the interleaved matrix. The tree building routine was the same as used for analyzing the individual data partitions, except 30 cycles of tree-drifting were used.

The Partitioned Bremer Support (PBS; also known as Partitioned Branch Support) at each node in the SA consensus tree was used as the primary criterion for the evaluation of each data partition. The PBS value is defined as "the minimum number of character steps for [a] partition on the shortest topologies for the combined data set that do not contain that node, minus the minimum number of character steps for that partition on the shortest topologies for the combined data set that do contain that node" (Gatesy et al., 1999). Therefore, a specific PBS value can be interpreted as a measurement of how well the data from a particular partition either support (represented by positive values) or refute (negative values) a particular node on the consensus tree. Branch Support (BS) values, defined as "the minimum number of character steps for that

data set on the shortest topologies that do not contain that node, minus the minimum number of character steps for that data set on the shortest topologies that contain that node" were used as the second criterion for evaluation of the SA consensus tree (Gatesy et al., 1999). After determining PBS values across all tree nodes on the consensus tree for each data partition, the BS was determined for each node on the consensus tree by the sum of all PBS values for that particular node. A positive BS score indicates that the overall combined set of data partitions support the topology at that node rather than refute it. ASAP2 uses a slightly modified version of a previously developed TNT script to calculate the PBS values (Peña et al., 2006). Modifications were made to the original TNT script were to facilitate automated data input and processing of output as required by ASAP2 without altering the tree building routines, and minimizing the text-based front end displayed to the user.

The Hidden Branch Support (HBS) for a particular node on the consensus tree was computed as the difference between the BS value at that node in the consensus tree and the sum of the BS values for that node from each data partition. The magnitude of an HBS value of a given node in the consensus tree was used as the final criterion for determining the overall strength of supporting or refuting the topology at the node.

Finally, a phylogenetic network was generated from the consensus analyses for each data partition using the Robinson-Foulds (RF) metric to quantify the distance between each pair of trees (Robinson & Foulds, 1981). This was implemented using a previously written TNT script (Goloboff, n.d.) that was modified to fit within the automated workflow of ASAP2. All calculations and parameters in the script were unchanged from the original version. To transform *RF* values onto a scale where larger values corresponded to more similarity (conventionally,

higher RF values indicate greater dissimilarity based on a normalized count of symmetric differences between trees), the following calculation was used:

$$RF' = \frac{1}{e^{RF}}$$

Cytoscape (Smoot et al., 2011) was then used to visualize the network relationship among the gene trees based on the RF' values as normalized edge weights using a force-directed layout.

Results

ASAP2

ASAP2 was developed as a set of Ruby scripts and is available at GitHub under the GNU General Public License (https://github.com/JDRomano2/ASAP2). The script guides the process of performing a SA from an initial set of Genbank identifiers. By the end of the analysis, ASAP2 produces files containing the data partitions, E-value tables, FASTA files of the final data partitions (both unaligned and aligned), TNT data matrices, and all TNT output, including log files and parenthetically-notated tree files. The ASAP2 data workflow is illustrated in **Figure 1**.

Gathering Uniform Taxonomic Distribution of AD Genes

Ten genes associated with Alzheimer Disease susceptibility were initially selected through and OMIM for analysis using ASAP2. Due to incompatibility issues with BLASTx, one gene (PAX-interacting protein 1 [PAXIP1]; GI:530387259) was removed from the analysis. In brief, because PAXIP1 contains six BRCT (BRCA C terminus) domains that are homologous to many sequences in GenBank, BLASTx quit at each attempt due to memory overflow. The nine remaining genes used for the remainder of the study are listed in **Table 1.**

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.230v1 | CC-BY 4.0 Open Access | received: 1 Feb 2014, published: 1 Feb 2014

The combined PSI-BLAST and recursive BLAST results for each gene included in this study resulted in nine data partitions representing 34 unique species (including *Homo sapiens*; **Table 2**). If the BLAST analyses resulted in any species being represented more than once in a data partition, only the first sequence (the one most similar to the query sequence) was kept. The protein sequences identified, along with the corresponding source nucleotide sequences, using this process are provided in **Supplementary Table 1**.

Simultaneous Analysis

All phylogenetic analyses were rooted to *Dasypus novemcinctus* (nine-banded armadillo), which was determined to be the furthest diverged from humans using TimeTree (Hedges et al., 2006). Individual maximum parsimony trees for each nucleotide and protein data partition are shown in **Figure 2** and **Figure 3**, respectively. Consensus SA trees based on the combination of the nine data partitions are shown in **Figure 4** and **Figure 5** (nucleotide and protein tree, respectively). Computed Branch Score (BS) values are shown on the consensus trees, and corresponding Partitioned Bremer Support (PBS) values are listed in **Table 3** and **Table 4**, respectively.

Individual trees for the respective nucleotide and protein data partitions yielded an evolutionary lineage for each individual gene, but empirical comparison of partition trees did not show coherent patterns. However, the SA trees did show a distinct branching pattern, with no more than two branches emerging at any single node. Furthermore, while some PBS values were negative (indicating that the data in a specific partition was not congruent with the consensus tree at that branch), all the BS values on the protein SA tree were positive. The nucleotide SA tree had

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.230v1 | CC-BY 4.0 Open Access | received: 1 Feb 2014, published: 1 Feb 2014

positive BS scores at each node with no polytomies, suggesting that the genes selected for this study supported all the internal branches in the protein simultaneous analysis tree.

While the topologic organization of the SA trees generally followed canonical patterns of mammalian evolution there were some notable exceptions that received high levels of statistical support. In the SA nucleotide tree, most primates were grouped together into the monophyletic clade rooted at node 13, with the exception of *Macaca mulatta* (rhesus macaque), *Callithrix* jacchus (common marmoset) and *Otolemur garnettii* (northern greater galago) that each occurred distally from all other primates (**Figure 4**). In the SA protein tree, primates were divided into two distinct clades: (1) a monophyletic clade rooted at node 27, or (2) a paraphyletic clade rooted at node 8 that also included *Sus scrofa* (wild boar) and *Jaculus jaculus* (lesser Egyptian jerboa) (**Figure 5**).

Comparison of Trees Using the Robinson-Foulds Metric

The Robinson-Foulds (RF) metric was used to quantify the similarity between the generated trees. The pairwise comparisons between each of the nucleotide and protein data partitions are shown in **Table 5** and **Table 6**, respectively. Additionally, the RF (and RF') distances for respective nucleotide and protein trees for a given partition as well as for the SA trees are shown in **Table 7**. The RF' distances were used as input into CytoScape to visualize the relative relationship between nucleotide and protein sequences for a given gene based on shared evolutionary history, shown in **Figure 6**. The resulting phylogenetic networks showed a tight clustering of MPO, A2M, NOS3, SORL1, and PLAU evolutionary patterns.

Discussion

The use of ASAP2 enabled the generation of the first integrated phylogeny of Alzheimer Disease associated genes. The results are robust and generally consistent with accepted patterns of taxonomic evolution. Examination of the resultant phylogenetic network also identified a clustering of evolution patterns among oxidative stress related genes associated with the development of AD. As the results suggest, SA techniques may have utility in development of large-scale studies that aim to model the evolutionary development, transmission, and interaction of disease associated gene sets.

ASAP2 Function

ASAP2 consolidates the application of SA techniques into a single pipeline of Ruby scripts designed to expose higher-order quantitative relationships between genes not visible through more traditional single-gene based analyses. Implementing SA techniques often requires a significant amount of manual data curation that is both labor- and time-intensive. ASAP2 was designed as a flexible automated tool that performs these tasks with minimal intervention beyond entering the initial GenBank identifiers. ASAP2 thus supports the ability to do large-scale phylogenetic analyses in a tractable manner. The data structures produced by ASAP2 were intentionally designed to be user-readable and manually editable during a given analysis. This supports the ability to adjust subsequent analyses based on results generated at any point along the analysis pipeline.

The original Perl version of ASAP (Sarkar et al., 2008) required a prior file containing sequences that was then aligned using MUSCLE and the SA subsequently executed using PAUP* (Wilgenbusch & Swofford, 2003). ASAP also allowed for the inclusion of pre-aligned or morphological data. By contrast, ASAP2 was developed in Ruby, uses MUSCLE based

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.230v1 | CC-BY 4.0 Open Access | received: 1 Feb 2014, published: 1 Feb 2014

alignment with the SA analyses done in TNT (which is freely licensable, unlike PAUP*). Additionally, ASAP2 was specifically designed to work exclusively with molecular data available from GenBank/GenPept, requiring only that the user provide an initial set of Accession numbers.

In this study, the utility of ASAP2 was demonstrated by performing analyses on a discrete set of pre-identified disease associated genes. However, the script may also be used for a myriad of large-scale multi-gene phylogenetic investigations. For example, one could use ASAP2 to study whole genomes with the goal to identify essential, evolutionarily conserved genes in groups of species (Rokas et al., 2003; Klein et al., 2012).

Putative Orthologue Sequence Identification

Based on an initial OMIM query for Alzheimer Disease, orthologues for 34 species were identified across nine disease genes. In addition to the recursive BLAST based approaches implemented by ASAP2, there are specific orthologue resources that could have also been searched to identify orthologous sequences for each of the nine disease genes. For example, inParanoid (Ostlund et al., 2010) and OrthoMCL (Li et al., 2003) had eight and 13 species spanning the nine genes, respectively. Interestingly, ASAP2 and inParanoid only found putative orthologues from mammalian species, while OrthoMCL data included several non-mammalian species, including *Danio rerio* (zebrafish), *Takifugu rubripes* (tiger blowfish), *Tetraodon nigroviridis* (spotted green pufferfish), and *Gallus gallus* (chicken). Additionally, both inParanoid and OrthoMCL identified the species *Canis familiaris* (dog) and *Equus caballus* (horse), while ASAP2 did not. The differences in orthologue identification may be due to the conservative filtering parameters used for BLAST queries in ASAP2 that were tuned to ensure a high degree

of similarity between sequences and to minimize the possibility of random homologies (as implicated by using an E-value of 1.0×10^{-256}). Neither inParanoid nor OrthoMCL identified the same set of additional species across all nine genes that were the focus of this study. ASAP2 does allow for the inclusion or removal of sequences to increase or reduce the taxonomic diversity of a given analysis immediately following the BLAST analyses; however, since no additional taxa were identified uniformly across the nine genes of interest by either OrthoMCL or inParanoid, no such modification of taxon diversity was performed in this study. Additionally, future studies may benefit from starting with a wider empirical set of genes or with parameters for the recursive BLAST strategy that are tuned to higher E-values that could lead to greater taxonomic diversity.

Phylogenetic Analysis

The TNT analyses used by ASAP2 were optimized to only include the most unambiguous groupings. As such, the TNT scripts produce fewer trees, but the likelihood of the trees reflecting evolutionary history is correspondingly more reliable. The final consensus tree represents a likely model of evolutionary transmission of the group of Alzheimer Disease genes studied, and the partitioned Bremer support values indicate the degree to which each gene fits the predicted pattern of evolution. The partitioned Bremer values may also be used to identify genes or species in a study that did not (for one reason or another) follow a similar pattern of transmission as the others. Topologically, the SA protein tree in this study exhibited a small number of groupings that differ from the accepted model of mammalian evolution, notably the separation of primates into two distinct clades. On the SA nucleotide tree, the paraphyletic grouping of some primates also merits scrutiny since this suggests that the genes included in this study deviate from taxonomically accepted evolution. It is important to note that the aggregate PBS values for these

different nodes are low and may be subject to topologic changes with the addition of more partitions. However, these "alternative" placements of certain primate species in the SA tree might also be explained by a reversion to an ancestral state for a particular disease gene. In this instance, the "state" being referred to would be patterns of interaction between the disease genes included in the study. This type of deviation from taxonomic evolution represents potential evolutionary divergence of the Alzheimer Disease gene network within isolated species. The presence of these types of alternate evolutionary patterns point to potential differential susceptibility of species to the development of AD. For example, the APBB2 and APP PBS values at node 13 in the nucleotide SA tree are significantly higher than for other partitions – 993 and 986, respectively (compared to an average value of 131.4 ± 620.1). These values suggest a potential interaction (based on a strongly corroborated evolutionary history) between the protein products of APBB2 and APP in primates. Building on the known interaction between APBB2 and APP in *H. sapiens*, exploration of the polymorphisms in these genes in *M. mulatta* and *O. garnettii* may elucidate the potential for differences in functional interactions. Such further exploration of these types of findings, especially relative to critical synapomorphic characters, could therefore yield valuable data regarding the evolutionarily important functional or potentially interacting sites for a given disease gene.

The individual data partition protein trees had a high incidence of polytomy, which is when more than two species branch off of a single node. This is generally considered uninformative in determining ancestry, as there are not enough data to determine whether species branching off of the same node are more or less closely related. However, these observations highlight the evolutionary conservation of fundamental protein sequences over many related organisms (Alexander et al., 2007; Kaneko et al., 1995; Pardossi-Piquard et al., 2005; Liu et al.,

2007; Nikolaev et al., 2009). APP, one of the central genes in Alzheimer Disease research, displays the most drastic examples of polytomy, with 17 branches underneath one node alone. This reinforces previous studies showing a high degree of conservation of the APP gene family over time (Freir et al., 2011; Yang et al., 2011; Manczak & Reddy, 2013; Coulson et al., 2000; Tharp & Sarkar, 2013).

While the protein phylogenies demonstrate conservation of structure across multiple species, the nucleotide sequences generate trees allowing a more precise elucidation of ancestry. Since nucleotide sequences can have differences that do not affect protein structure or function due to the degeneracy of the genetic code, rates of change in nucleotide sequences are more closely tied to evolutionary time (Brown, 2002; Bejerano et al., 2004; McKusick, 2007; Lehmann & Libchaber, 2008). Among the individual partition nucleotide trees, on the APBB2 tree has an one occurrence of more than two branches rooted at a single parent node. The branch generated at this node contains four species of very closely related great apes (*N. leucogenys*, *G. gorilla gorilla*, *P. troglodytes*, and *P. paniscus*). This suggests that the nucleotide sequences corresponding to APBB2 in each of these species are so similar that a more descriptive phylogenetic relationship between them cannot be determined, which underscores the fact that APBB2 is highly conserved among closely related species.

Determination of the distance between individual trees prior to constructing a consensus tree can help to preliminarily identify clustering patterns among specific genes prior to constructing a consensus tree (Vilella et al., 2009). Additionally, once a consensus is reached, these distances can be used to explain the strength of the support for the SA tree and generate representations of the gene network (Degnan et al., 2009). While multiple methods may be used to evaluate the distance between trees consisting of the same set of taxa, this study used the

Robinson-Foulds (RF) distance (Robinson & Foulds, 1981). The RF distance between two trees is defined by the sum of the number of data partitions implied by one, but not both, of the trees. A variety of algorithms exist for computing RF distance (Bansal et al., 2010; Chaudhary et al., 2013), and an optimal method is usually selected on the basis of algorithmic complexity and worst-case running time (Goloboff & Farris, 2001; Pattengale et al., 2007; Chaudhary et al., 2012). In this study, a phylogenetic network was constructed based on tree topology similarity using RF (transformed to RF', which converts RF values onto a scale where higher values correspond to less similarity). On examination of the phylogenetic network for the Alzheimer Disease genes used in this study, a tight clustering of oxidative stress genes was observed with the gene for plasminogen activator (PLAU) and a member of the sortilin related receptor gene family (SORL1). While SORL1 has been found to have an important association with Alzheimer Disease and oxidative stress genes are involved in the unfolded protein response associated with increased amyloid formation, a relationship between these genes has not been shown before (Rogaeva et al., 2007; Haataja et al., 2008). This type of association is not observable using single pathway experiments or phylogenetic methods that do not incorporate an SA approach. Further investigation will be needed to understand the nature of this network clustering.

A final aspect of this study is that it further highlights the fact that choice of model organism is paramount for the study of complex disease. The relatively short lifespan of *M. musculus* and malleability of the murine genome has led to an explosion of experimental approaches centered on manipulation of genes thought to be involved in human disease (Bedell et al., 1997*a*; 1997*b*). However, especially with relation to complex diseases, alternative model organisms need to be considered (Ostrander, 2012). The recent increase in biological systems data and continued growth in bioinformatics methodologies for analyzing these data may allow

for the development of more data driven choices of model organisms for complex diseases. For example, based on the preliminary findings of this study of the shared evolution of a limited set of genes thought to influence AD susceptibility in humans, the SA consensus trees suggests that *Sus scrofa* (pig), *Jaculus jaculus* (jerboa), and *Mustela putorius furo* (weasel) may be more suitable model organisms than rodents..

Conclusion

Phylogenomic studies using Simultaneous Analysis techniques are positioned to become more commonplace as increasing amounts of genomic data are available across the spectrum of life and systematically available through resources such as GenBank. Here, an automated tool (ASAP2) is presented with the intent of enabling researchers to leverage these data to support studies that aim to unveil potentially significant relationships that may be embedded in co-evolution. The application of ASAP2 to a set of nine genes associated with Alzheimer Disease demonstrated a potentially important clustering of genes around components of oxidative stress pathways.

References

Alexander PA, He Y, Chen Y, Orban J & Bryan PN. 2007. The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci USA* 104:11963–11968.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.

Badano JL & Katsanis N. 2002. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet* 3:779–789.

Baker RH, Yu X & DeSalle R. 1998. Assessing the relative contribution of molecular and morphological characters in simultaneous analysis trees. *Mol Phylogenet Evol* 9:427–436.

Ballard C, Gauthier S, Corbett A, Brayne C, Aarsland D & Jones E. 2011. Alzheimer's disease. *Lancet* 377:1019–1031.

Bansal MS, Burleigh JG, Eulenstein O & Fernández-Baca D. 2010. Robinson-Foulds supertrees. *Algorithms Mol Biol* 5:18.

Bedell MA, Jenkins NA & Copeland NG. 1997*a*. Mouse models of human disease. Part I: techniques and resources for genetic analysis in mice. *Genes Dev* 11:1–10.

Bedell MA, Largaespada DA, Jenkins NA & Copeland NG. 1997*b*. Mouse models of human disease. Part II: recent progress and future directions. *Genes Dev* 11:11–43.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS & Haussler D. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–1325.

Blacker D, Albert MS, Bassett SS & Go R. 1994. Reliability and validity of NINCDS-ADRDA criteria for Alzheimer's disease: the National Institute of Mental Health Genetics Initiative. *Archives of ….*

Brown TA. 2002. *Genomes 2*. Garland Publishing.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K & Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Chaudhary R, Burleigh JG & Fernández-Baca D. 2012. Fast local search for unrooted Robinson-Foulds supertrees. *IEEE/ACM Trans Comput Biol Bioinform* 9:1004–1013.

Chaudhary R, Burleigh JG & Fernández-Baca D. 2013. Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms Mol Biol* 8:28.

Coulson EJ, Paliga K, Beyreuther K & Masters CL. 2000. What the evolution of the amyloid protein precursor supergene family tells us about its function. *Neurochem Int* 36:175–184.

Dagan T. 2011. Phylogenomic networks. *Trends Microbiol* 19:483–491.

Degnan JH, DeGiorgio M, Bryant D & Rosenberg NA. 2009. Properties of consensus methods for inferring species trees from gene trees. *Systematic Biology* 58:35–54.

Dubois B, Feldman HH, Jacova C & DeKosky ST. 2007. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet ….*

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

Estruch R, Ros E, Salas-Salvadó J, Covas M-I, Corella D, Arós F, Gómez-Gracia E, Ruiz-Gutiérrez V, Fiol M, Lapetra J, Lamuela-Raventos RM, Serra-Majem L, Pintó X, Basora J,

Muñoz MA, Sorlí JV, Martínez JA, Martínez-González MAPREDIMED Study Investigators. 2013. Primary prevention of cardiovascular disease with a Mediterranean diet. *N Engl J Med* 368:1279–1290.

Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, Hall K, Hasegawa K, Hendrie H, Huang Y, Jorm A, Mathers C, Menezes PR, Rimmer E, Scazufca MAlzheimer's Disease International. 2005. Global prevalence of dementia: a Delphi consensus study. *Lancet* 366:2112–2117.

Fitch WM. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology* 20:406–416.

Freir DB, Nicoll AJ, Klyubin I, Panico S, Mc Donald JM, Risse E, Asante EA, Farrow MA, Sessions RB, Saibil HR, Clarke AR, Rowan MJ, Walsh DM & Collinge J. 2011. Interaction between prion protein and toxic amyloid β assemblies can be therapeutically targeted at multiple sites. *Nat Commun* 2:336.

Gatesy J, O'Grady P & Baker RH. 1999. Corroboration among Data Sets in Simultaneous Analysis: Hidden Support for Phylogenetic Relationships among Higher Level Artiodactyl Taxa. *Cladistics* 15:271–313.

Goloboff P, Farris S & Nixon K. 2000. TNT (Tree analysis using New Technology).

Goloboff PA & Farris JS. 2001. Methods for Quick Consensus Estimation. *Cladistics* 17:S26–S34.

Goloboff PA ed. n.d. *TNT wiki*. Available at: http://tnt.insectmuseum.org/index.php/Scripts/RF_distances [Accessed December 12, 2013].

Goto N, Prins P, Nakao M, Bonnal R, Aerts J & Katayama T. 2010. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics* 26:2617–2619.

Haataja L, Gurlo T, Huang CJ & Butler PC. 2008. Islet amyloid in type 2 diabetes, and the toxic oligomer hypothesis. *Endocr Rev* 29:303–316.

Hardy J & Selkoe DJ. 2002. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* 297:353–356.

Hedges SB, Dudley J & Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22:2971–2972.

Kaneko I, Yamada N, Sakuraba Y, Kamenosono M & Tutumi S. 1995. Suppression of mitochondrial succinate dehydrogenase, a primary target of beta-amyloid, and its derivative racemized at Ser residue. *J Neurochem* 65:2585–2593.

Karran E, Mercken M & De Strooper B. 2011. The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nat Rev Drug Discov* 10:698–712.

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M & Tsui LC. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* 245:1073–1080.

Klein BA, Tenorio EL, Lazinski DW, Camilli A, Duncan MJ & Hu LT. 2012. Identification of essential genes of the periodontal pathogen Porphyromonas gingivalis. *BMC Genomics* 13:578.

Layeghifard M, Peres-Neto PR & Makarenkov V. 2013. Inferring explicit weighted consensus networks to represent alternative evolutionary histories. *BMC Evol Biol* 13:274.

Lehmann J & Libchaber A. 2008. Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. *RNA* 14:1264–1269.

Li L, Stoeckert CJ & Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189.

Li Y et al. 2005. Genetic association of the APP binding protein 2 gene (APBB2) with late onset Alzheimer disease. *Hum Mutat* 25:270–277.

Liu Q, Zerbinatti CV, Zhang J, Hoe H-S, Wang B, Cole SL, Herz J, Muglia L & Bu G. 2007. Amyloid precursor protein regulates brain apolipoprotein E and cholesterol metabolism through lipoprotein receptor LRP1. *Neuron* 56:66–78.

Lu J-X, Qiang W, Yau W-M, Schwieters CD, Meredith SC & Tycko R. 2013. Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. *Cell* 154:1257–1268.

Manczak M & Reddy PH. 2013. Abnormal interaction of oligomeric amyloid-β with phosphorylated tau: implications to synaptic dysfunction and neuronal damage. *J Alzheimers Dis* 36:285–295.

Matsumoto Y ed. n.d. *Ruby Programming Language*. Available at: https://www.ruby-lang.org/en/ [Accessed January 20, 2014].

Mayeux R & Stern Y. 2012. Epidemiology of Alzheimer disease. *Cold Spring Harb Perspect Med*; DOI: 10.1101/cshperspect.a006239.

McIlroy SP, Crawford VL, Dynan KB, McGleenon BM, Vahidassr MD, Lawson JT & Passmore AP. 2000. Butyrylcholinesterase K variant is genetically associated with late onset Alzheimer's disease in Northern Ireland. *J Med Genet* 37:182–185.

McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* 80:588–604.

Nelson PT et al. 2012. Correlation of Alzheimer disease neuropathologic changes with cognitive status: a review of the literature. *J Neuropathol Exp Neurol* 71:362–381.

Nelson PT, Braak H & Markesbery WR. 2009. Neuropathology and cognitive impairment in

Alzheimer disease: a complex but coherent relationship. *J Neuropathol Exp Neurol* 68:1–14.

Newton-Cheh C et al. 2009. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* 41:666–676.

Nikolaev A, McLaughlin T, O'Leary DDM & Tessier-Lavigne M. 2009. APP binds DR6 to trigger axon pruning and neuron death via distinct caspases. *Nature* 457:981–989.

Nixon KC & Carpenter JM. 1996. ON SIMULTANEOUS ANALYSIS. *Cladistics* 12:221–241.

Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O & Sonnhammer ELL. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 38:D196–D203.

Ostrander EA. 2012. Franklin H. Epstein Lecture. Both ends of the leash--the human links to good dogs with bad genes. *The New England journal of medicine* 367:636–646. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22894576&retmode=ref&cmd=prlinks.

Pardossi-Piquard R, Petit A, Kawarai T, Sunyach C, Alves da Costa C, Vincent B, Ring S, D'Adamio L, Shen J, Müller U, St George-Hyslop P & Checler F. 2005. Presenilin-dependent transcriptional control of the Abeta-degrading enzyme neprilysin by intracellular domains of betaAPP and APLP. *Neuron* 46:541–554.

Pattengale ND, Gottlieb EJ & Moret BME. 2007. Efficiently computing the Robinson-Foulds metric. *J Comput Biol* 14:724–735.

Peña C, Wahlberg N, Weingartner E, Kodandaramaiah U, Nylin S, Freitas AVL & Brower AVZ. 2006. Higher level phylogeny of Satyrinae butterflies (Lepidoptera: Nymphalidae) based on DNA sequence data. *Mol Phylogenet Evol* 40:29–49.

Querfurth HW & LaFerla FM. 2010. Alzheimer's disease. *N Engl J Med* 362:329–344.

Ramanan VK, Risacher SL, Nho K, Kim S, Swaminathan S, Shen L, Foroud TM, Hakonarson H, Huentelman MJ, Aisen PS, Petersen RC, Green RC, Jack CR, Koeppe RA, Jagust WJ, Weiner MW & Saykin AJ. 2013. APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study. *Mol Psychiatry*; DOI: 10.1038/mp.2013.19.

Rees DC, Williams TN & Gladwin MT. 2010. Sickle-cell disease. *Lancet* 376:2018–2031.

Robinson DF & Foulds LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*.

Rogaeva E et al. 2007. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. *Nat Genet* 39:168–177.

Rokas A, Williams BL, King N & Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.

Sarkar IN, Egan MG, Coruzzi G, Lee EK & DeSalle R. 2008. Automated simultaneous analysis phylogenetics (ASAP): an enabling tool for phlyogenomics. *BMC Bioinformatics* 9:103.

Sillén A, Forsell C, Lilius L, Axelman K, Björk BF, Onkamo P, Kere J, Winblad B & Graff C. 2006. Genome scan on Swedish Alzheimer's disease families. *Mol Psychiatry* 11:182–186.

Silverman JM, Smith CJ, Marin DB, Birstein S, Mare M, Mohs RC & Davis KL. 1999. Identifying families with likely genetic protective factors against Alzheimer disease. *Am J Hum Genet* 64:832–838.

Sjögren T, Sjogren H & LINDGREN AG. 1952. Morbus Alzheimer and morbus Pick; a genetic, clinical and patho-anatomical study. *Acta Psychiatr Neurol Scand Suppl* 82:1–152.

Smoot ME, Ono K, Ruscheinski J, Wang P-L & Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27:431–432.

Swiderski DL, Zelditch ML & Fink WL. 1998. Why morphometrics is not special: coding quantitative data for phylogenetic analysis. *Systematic Biology*.

Tanzi RE. 2012. The genetics of Alzheimer disease. *Cold Spring Harb Perspect Med*; DOI: 10.1101/cshperspect.a006296.

Tharp WG & Sarkar IN. 2013. Origins of amyloid-β. *BMC Genomics* 14:290.

Thornton JW & DeSalle R. 2000. Gene family evolution and homology: genomics meets phylogenetics. *Annu Rev Genomics Hum Genet* 1:41–73.

Tiraboschi P, Hansen LA, Thal LJ & Corey-Bloom J. 2004. The importance of neuritic plaques and tangles to the development and evolution of AD. *Neurology* 62:1984–1989.

Velinov M, Slaugenhaupt SA, Stoilov I, Scott CI, Gusella JF & Tsipouras P. 1994. The gene for achondroplasia maps to the telomeric region of chromosome 4p. *Nat Genet* 6:314–317.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R & Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19:327–335.

Watson JD, Baker TA, Bell SP, Gann A, Levine M & Losick R. 2014. *Molecular Biology of the Gene*. Benjamin Cummings.

Wilgenbusch JC & Swofford D. 2003. Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics* Chapter 6:Unit6.4.

Yang J, Ji Y, Mehta P, Bates KA, Sun Y & Wisniewski T. 2011. Blocking the apolipoprotein E/amyloid-β interaction reduces fibrillar vascular amyloid deposition and cerebral microhemorrhages in TgSwDI mice. *J Alzheimers Dis* 24:269–285.

Yonan AL, Alarcón M, Cheng R, Magnusson PKE, Spence SJ, Palmer AA, Grunn A, Juo S-HH,

Terwilliger JD, Liu J, Cantor RM, Geschwind DH & Gilliam TC. 2003. A genomewide screen of 345 families for autism-susceptibility loci. *Am J Hum Genet* 73:886–897.

Zharkikh A. 1994. Estimation of evolutionary distances between nucleotide sequences. *J Mol Evol* 39:315–329.

Table 1: The nine Alzheimer Disease genes used in the study.

Gene	Common Name	GenBank GI	GenPept GI
apbb2	Amyloid-β precursor protein-binding family B member 2	225007611	50083291
nos3	Nitric oxide synthase 3	231571328	231571329
plau	Urokinase-type plasminogen activator	222537757	4505863
sorl1	Sortilin-related receptor L	307611954	4507157
a2m	α-2-macroglobulin	66932946	62088808
blmh	Bleomycin hydrolase	530411126	194378004
mpo	Myeloperoxidase	4557758	4557759
ace	Angiotensin I coverting enzyme	295844836	295844837
app	Amyloid-β precursor protein	324021737	324021738

Table 2: The 34 species identified by ASAP2 using Alzheimer Disease gene queries.

Scientific Name	Vernacular Name	NCBI Taxon ID
Homo sapiens	Human	9606
Pan troglodytes	Chimpanzee	9598
Gorilla gorilla gorilla	Western lowland gorilla	9595
Nomascus leucogenys	Northern white-cheeked gibbon	61853
Macaca mulatta	Rhesus monkey	9544
Pan paniscus	Bonobo	9597
Papio Anubis	Olive baboon	9555
Callithrix jacchus	White-tufted-ear marmoset	9483
Saimiri boliviensis boliviensis	Bolivian squirrel monkey	39432
Mus musculus	House mouse	10090
Otolemur garnettii	Small-eared galago	30611
Pteropus alecto	Black flying fox	9402
Ovis aries	Sheep	9940
Cavia porcellus	Domestic guinea pig	10141
Sus scrofa	Pig	9823
Rattus norvegicus	Norway rat	10116
Bos Taurus	Cattle	9913
Oryctolagus cuniculus	Rabbit	9986
Ailuropoda melanoleuca	Giant panda	9646
Tupaia chinensis	Chinese tree shrew	246437
Felis catus	Domestic cat	9685
Cricetulus griseus	Chinese hamster	10029
Heterocephalus glaber	Naked mole-rat	10181
Ceratotherium simum simum	Southern white rhinoceros	73337
Orcinus orca	Killer whale	9733
Odobenus rosmarus divergens	Pacific walrus	9708
Dasypus novemcinctus	Nine-banded armadillo	9361
Chinchilla lanigera	Long-tailed chinchilla	34839
Ictidomys tridecemlineatus	Thirteen-lined ground squirrel	43179
Trichechus manatus latirostris	Florida manatee	127582
Mustela putorius furo	Domestic ferret	9669
Condylura cristata	Star-nosed mole	143302
Octodon degus	Degu	10160
Jaculus jaculus	Lesser Egyptian jerboa	51337

Table 3: Nucleotide PBS values for each internal node on nucleotide simultaneous analysis tree for each data partition. Node numbers are ASAP2 assigned labels, reported next to BS score on SA tree (Figure 4).

Node	apbb2	nos3	plau	sorl1	a2m	blmh	mpo	ace	app
1	1929	-188	788	-2136	-111	775	54	134	-305
2	-457	-173	-21	-604	-197	493	-49	1634	-68
3	360	-175	-46	-739	-203	874	-45	128	-142
4	360	-175	-46	-739	-203	874	-45	128	-142
5	-2	-52	844	-505	181	-435	-101	368	-205
6	-2	-52	844	-505	181	-435	-101	368	-205
7	771	-9	-5	-1414	-231	897	-34	-25	67
8	-1748	-209	1283	1781	31	-538	-200	347	-267
9	360	-175	-46	-739	-203	874	-45	128	-142
10	771	-9	-5	-1414	-231	897	-34	-25	67
11	-46	-197	1889	-772	103	-529	-141	370	-436
12	830	-42	27	-1570	-138	-1040	510	339	1607
13	993	-110	152	-343	179	-1005	57	274	986
14	133	-14	21	76	38	-1942	366	223	1116
15	950	627	-1229	1429	-568	554	-365	452	-191
16	133	-14	21	76	38	-1942	366	223	1116
17	133	-14	21	76	38	-1942	366	223	1116
18	138	1203	23	80	39	-1989	368	211	1094
19	360	-175	-46	-739	-203	874	-45	128	-142
20	771	-9	-5	-1414	-231	897	-34	-25	67
21	771	-9	-5	-1414	-231	897	-34	-25	67
22	771	-9	-5	-1414	-231	897	-34	-25	67
23	1216	-158	757	-1448	-176	-69	19	21	-47
24	268	300	21	-552	-35	-1	5	-40	223
25	1086	-348	2094	-781	-267	-622	-173	183	-107
26	360	-175	-46	-739	-203	874	-45	128	-142
27	360	-175	-46	-739	-203	874	-45	128	-142
28	36	-7	-35	-63	10	22	1	22	35
29	360	-175	-46	-739	-203	874	-45	128	-142
30	360	-175	-46	-739	-203	874	-45	128	-142
31	-109	1	45	-502	-60	522	152	140	-126

Table 4: Protein PBS values given for each internal node on protein simultaneous analysis tree for each data partition. Node numbers are ASAP2 assigned labels, reported next to BS score on SA tree (Figure 5).

Node	apbb2	nos3	plau	sorl1	a2m	blmh	mpo	ace	app
1	6	10	12.5	1	4.5	4	3.5	-13.5	3
2	265	0	-37	11	-131	-6	-10	-69	0
3	263	6	-28	16	-112	-8	-7	-56	19
4	0	13	3	-4	-1	0	3	2	0
5	261	-1	-38	11	-131	-7	-12	-62	0
6	273	6	-15	18	-89	1	-14	-49	5
7	2	1	0	3	-5	3	-1	1	1
8	0	3	10	2	14	2	7	-31	39
9	-3	-1	-25	-21	-61	-583	-13	581	186
10	27	1	5	6	16	2	8	15	-16
11	0	7	0	0	1	3	12	1	0
12	0	548	0	0	2	0	12	0	-18
13	1	3	0	13	4	0	-3	2	0
14	4	1	0	9	17	1	-1	12	2
15	2	3	2	5	7	-1	6	9	1
16	13	6	3	11	24	5	13	14	1
17	4	3	8	9	18	-2	7	6	0
18	3	1	-1	-6	-4	-2	-1	2	17
19	3	1	-1	-6	-4	-2	-1	2	17
20	-16	10	7	20	40	7	2	-27	-5
21	-16	10	7	20	40	7	2	-27	-5
22	-3	-1	-25	-21	-61	-583	-13	581	186
23	-16	10	7	20	40	7	2	-27	-5
24	-4	2	1	0	4	2	0	5	2
25	-21	5	5	16	12	1	-8	10	-2
26	30	4	14	-1	24	-909	160	571	176
27	-2	6	2	3	5	-3	-1	7	1
28	2	13	1	3	-4	8	15	-33	40
29	12	27	31	42	120	7	90	68	-334
30	2	13	1	3	-4	8	15	-33	40
31	-3	-1	-25	-21	-61	-583	-13	581	186

Table 5: RF and RF' distance between each pair of trees for nucleotide sequence data partitions. Unshaded cells show RF distances, and shaded cells show RF' distances.

	A2M	ACE	APBB2	APP	BLMH	MPO	NOS3	PLAU	SORL1
A2M	--	0.379	0.405	0.417	0.417	0.444	0.472	0.430	0.404
ACE	0.969	--	0.380	0.430	0.379	0.379	0.379	0.391	0.391
APBB2	0.905	0.968	--	0.380	0.392	0.405	0.380	0.392	0.405
APP	0.875	0.844	0.968	--	0.391	0.379	0.417	0.379	0.430
BLMH	0.875	0.969	0.937	0.938	--	0.391	0.430	0.404	0.391
MPO	0.812	0.969	0.905	0.969	0.938	--	0.404	0.444	0.404
NOS3	0.750	0.969	0.968	0.875	0.844	0.906	--	0.417	0.417
PLAU	0.844	0.938	0.937	0.969	0.906	0.812	0.875	--	0.404
SORL1	0.906	0.938	0.905	0.844	0.938	0.906	0.875	0.906	--

Table 6: RF and RF' distance between each pair of trees for protein sequence data partitions. Unshaded cells show RF distances, and shaded cells show RF' distances.

	A2M	ACE	APBB2	APP	BLMH	MPO	NOS3	PLAU	SORL1
A2M	--	0.492	0.438	0.421	0.472	0.634	0.553	0.745	0.590
ACE	0.709	--	0.401	0.402	0.423	0.456	0.449	0.463	0.424
APBB2	0.826	0.915	--	0.435	0.400	0.418	0.426	0.443	0.436
APP	0.864	0.911	0.833	--	0.384	0.402	0.408	0.386	0.402
BLMH	0.750	0.860	0.917	0.957	--	0.454	0.465	0.461	0.454
MPO	0.455	0.786	0.872	0.911	0.789	--	0.468	0.562	0.526
NOS3	0.592	0.800	0.854	0.897	0.765	0.760	--	0.499	0.487
PLAU	0.294	0.769	0.814	0.951	0.774	0.577	0.696	--	0.584
SORL1	0.527	0.857	0.830	0.911	0.789	0.643	0.720	0.538	--

Table 7: RF and RF' values between corresponding nucleotide and protein trees for each gene.

Gene	RF	RF'
APBB2	0.880	0.415
NOS3	0.593	0.553
PLAU	0.714	0.490
SORL1	0.900	0.407
A2M	0.559	0.572
BLMH	0.770	0.463
MPO	0.367	0.693
ACE	0.900	0.407
APP	0.837	0.433

Figure 1

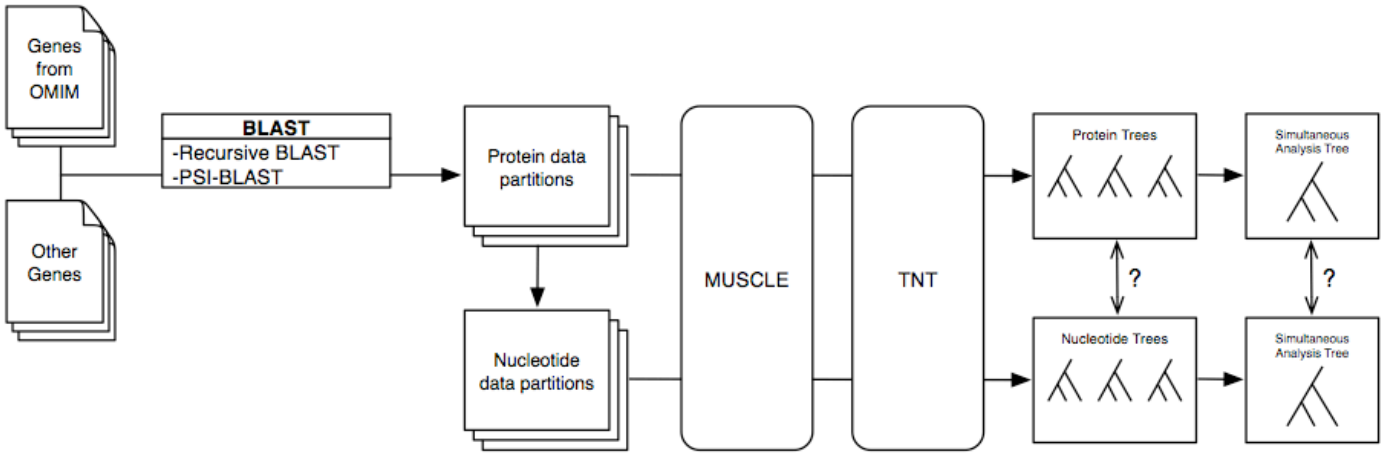
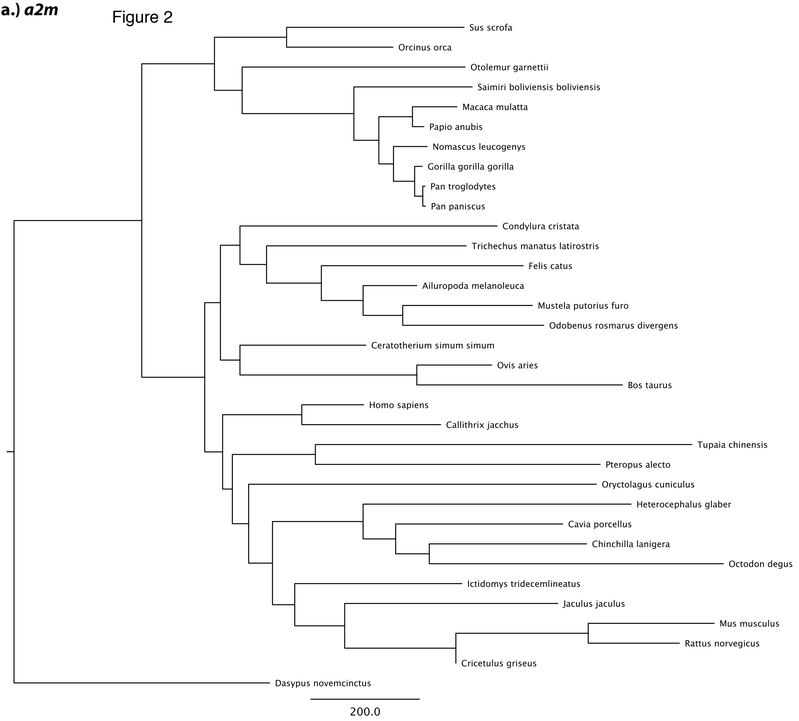
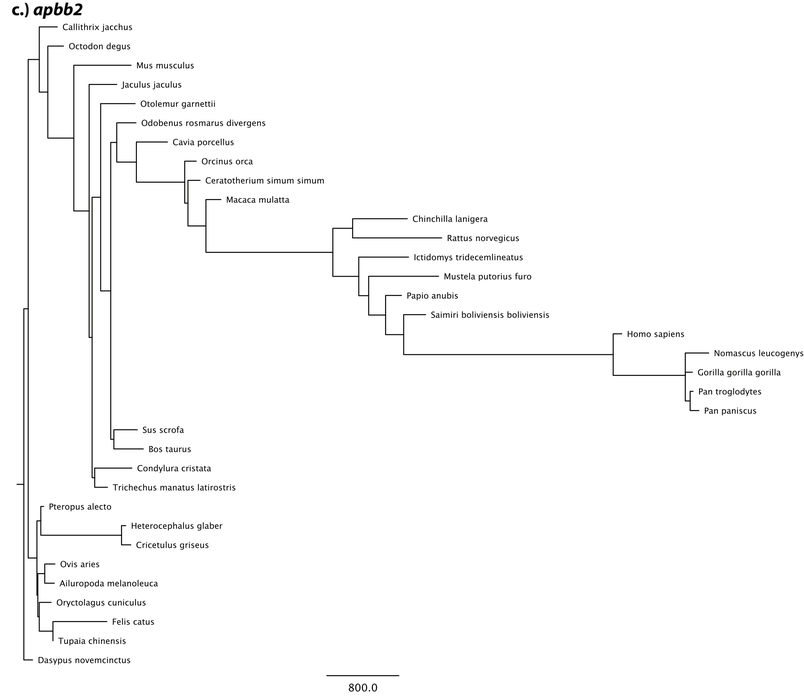
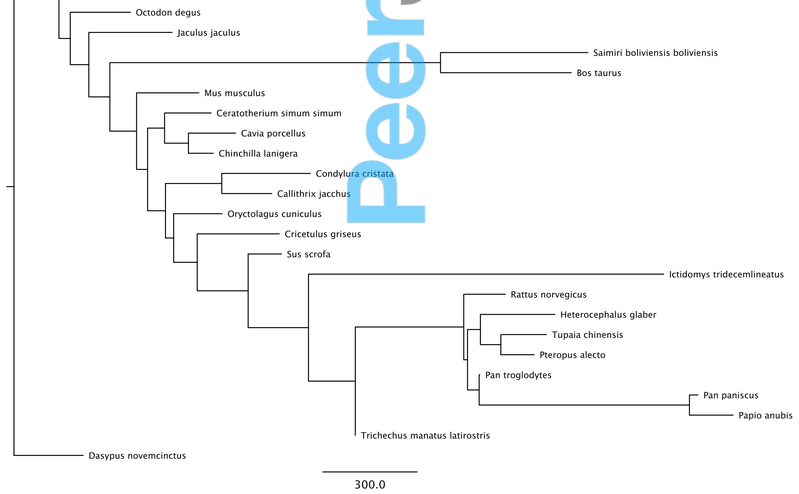
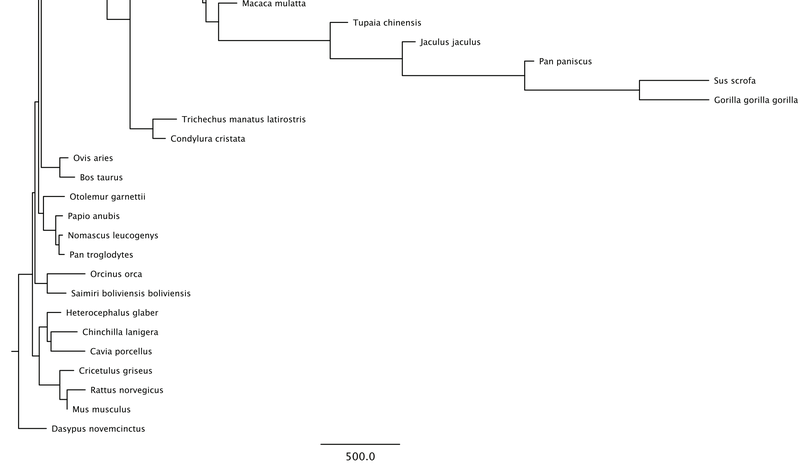


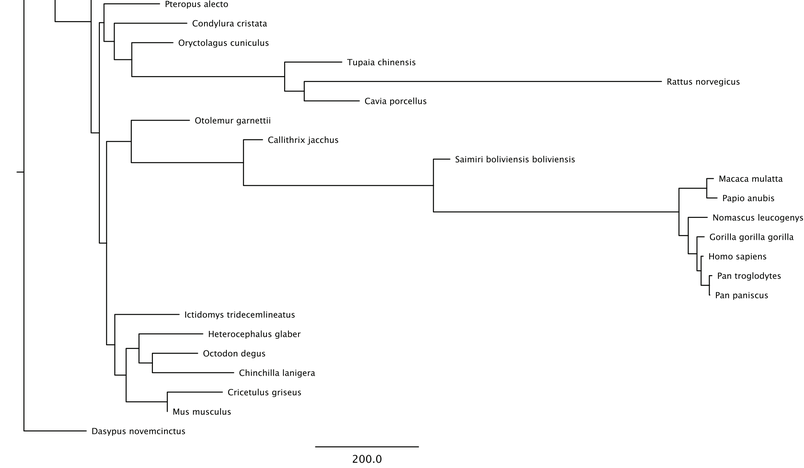
Figure 2

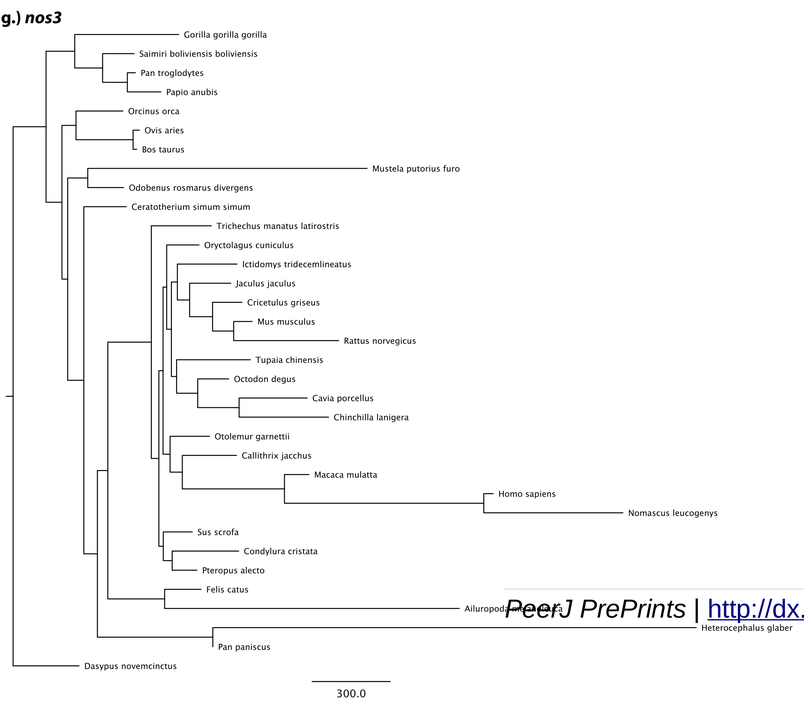


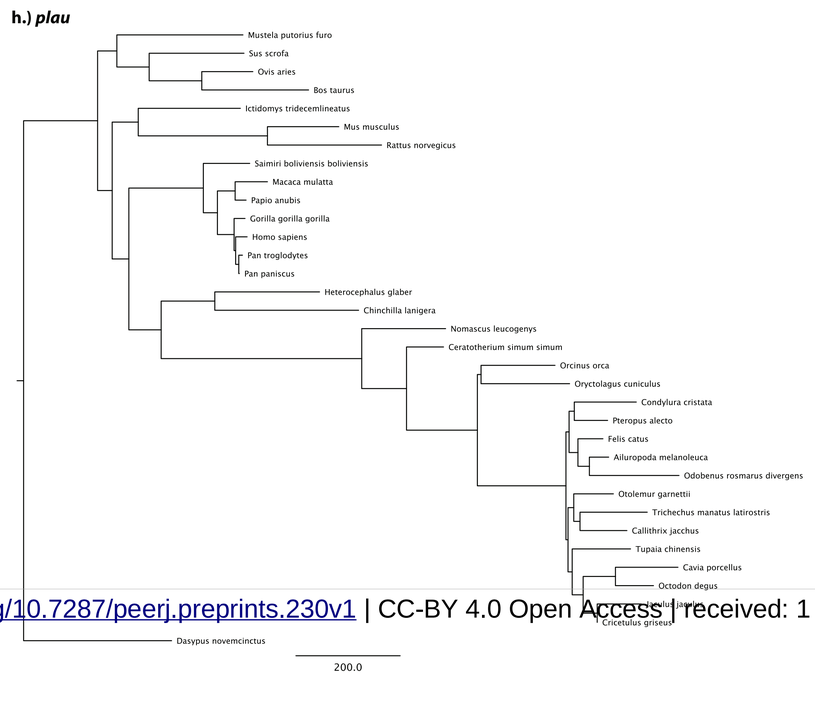












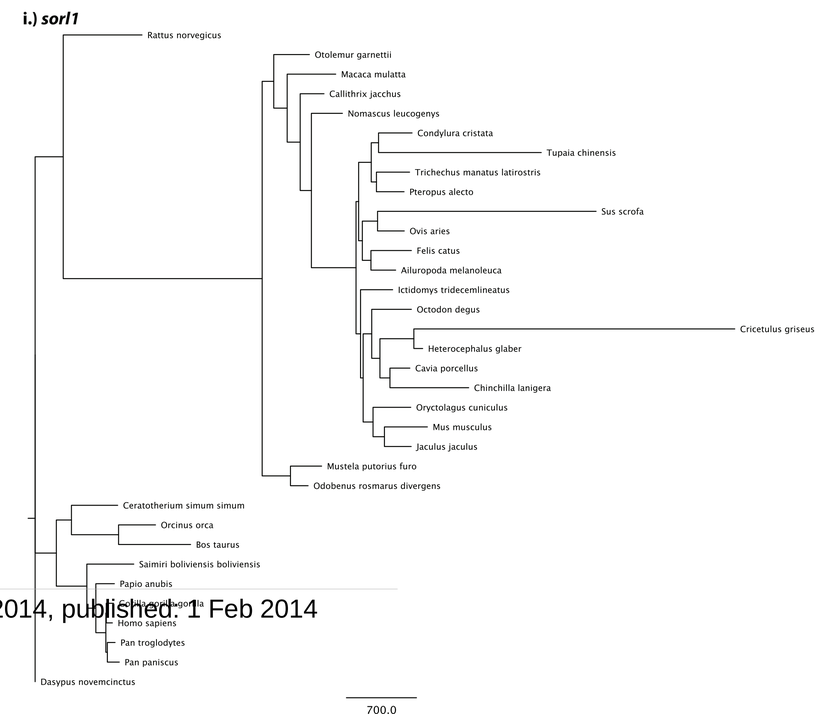
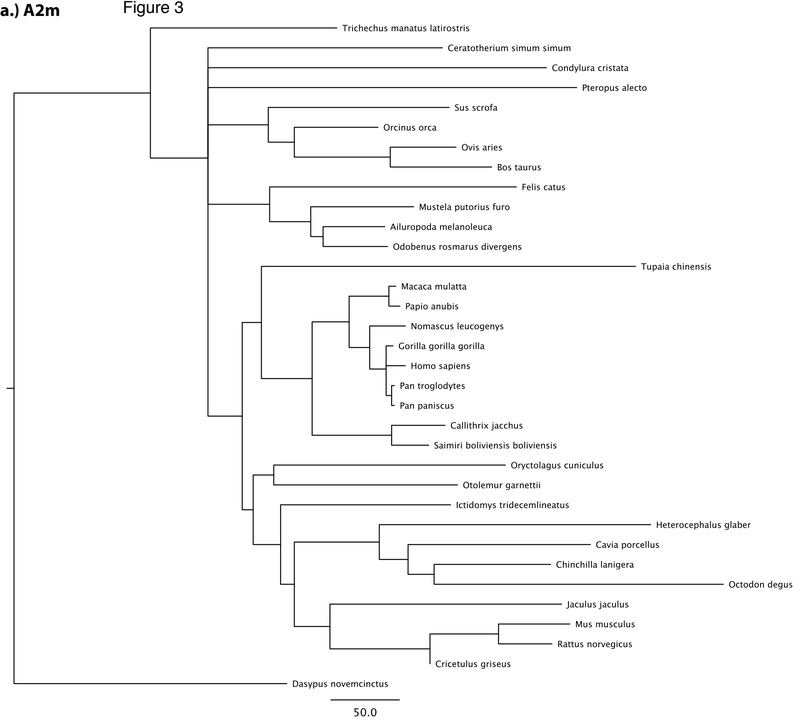
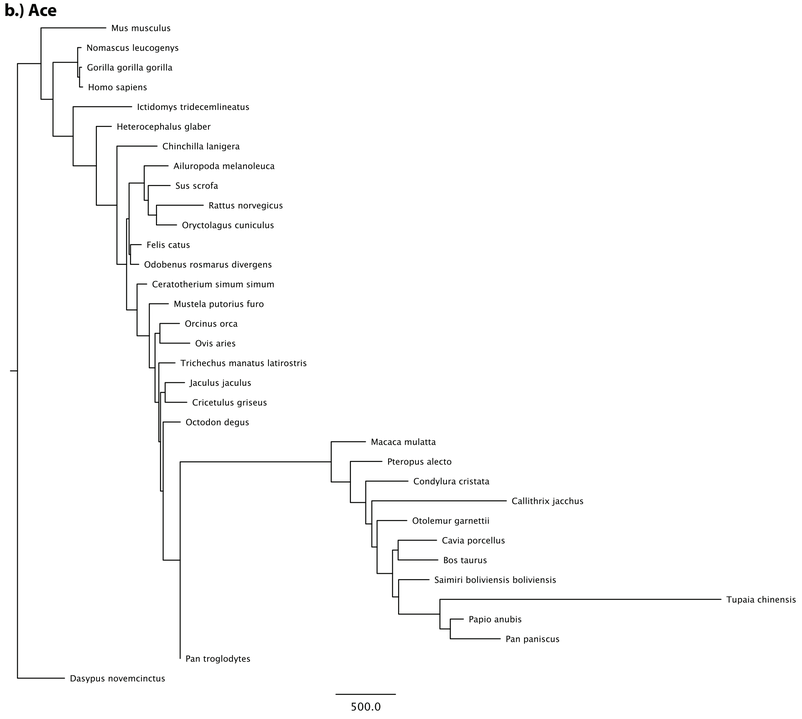
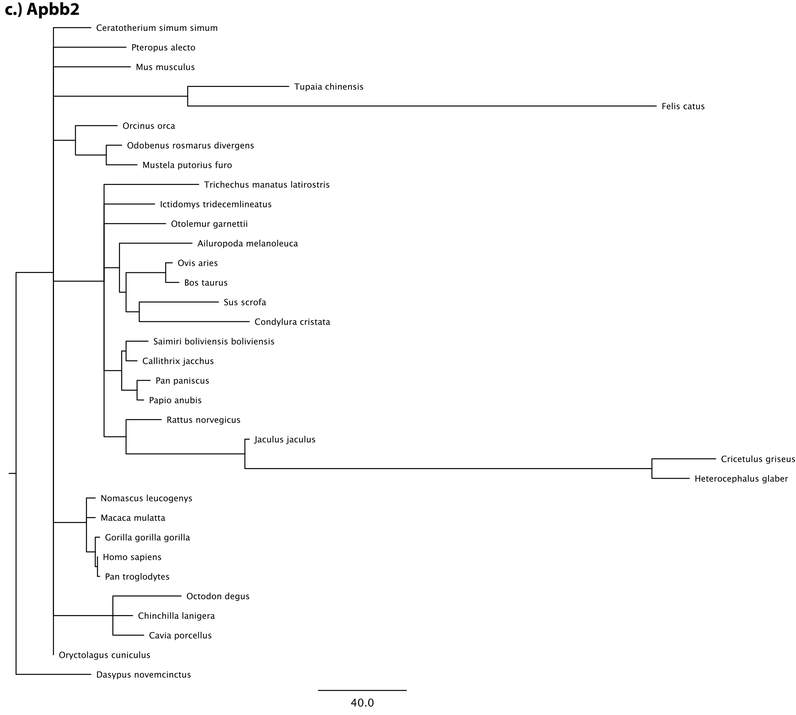
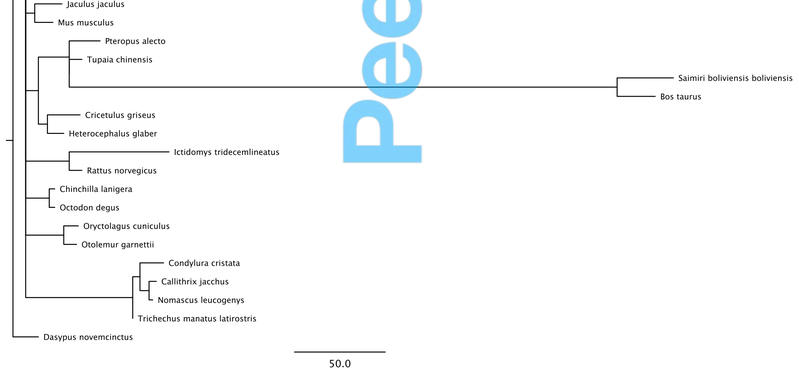


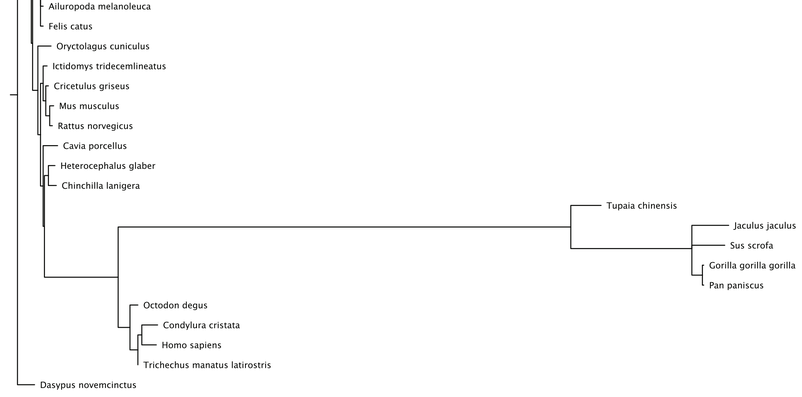
Figure 3

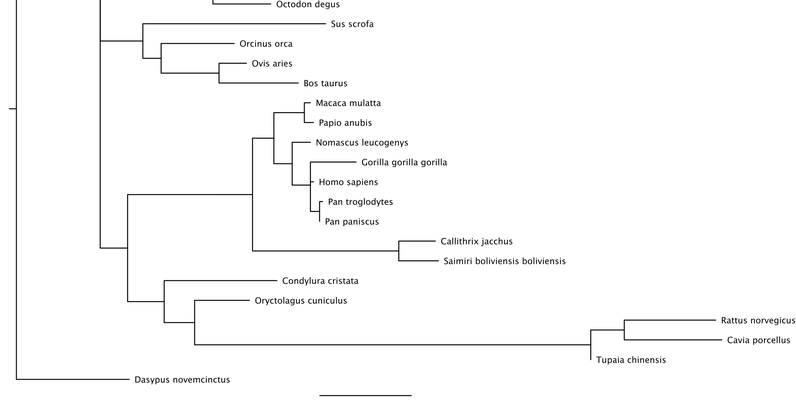


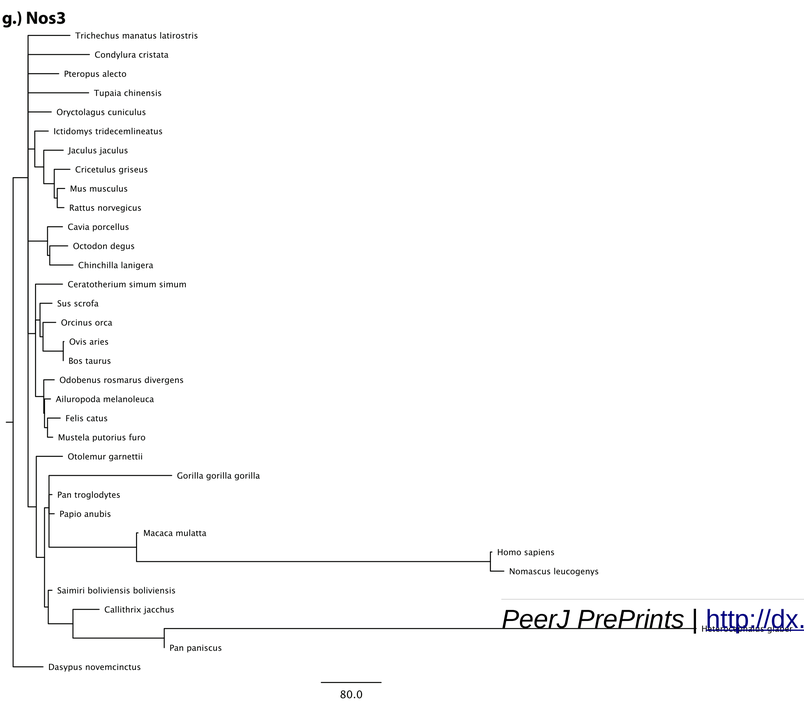


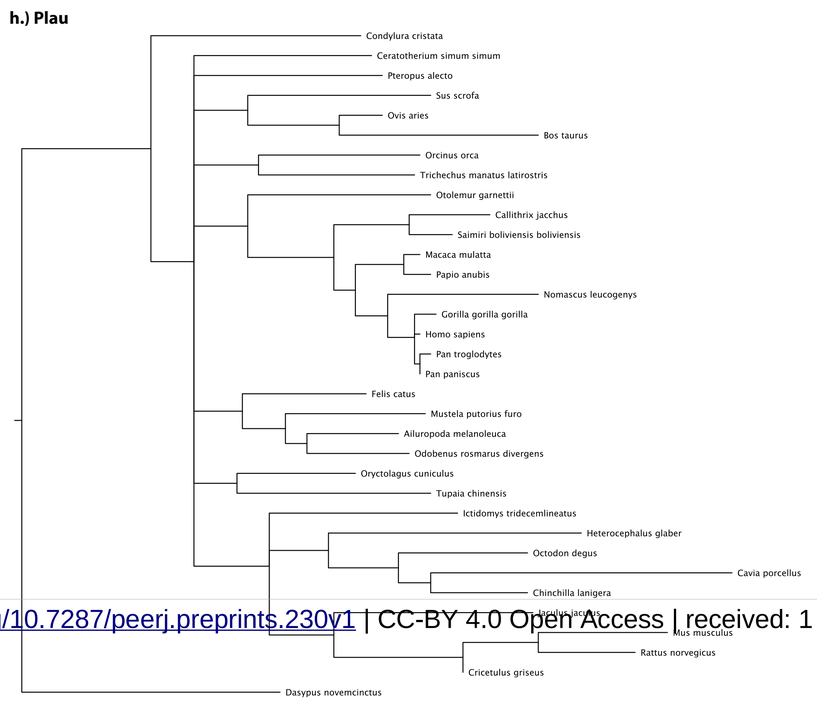


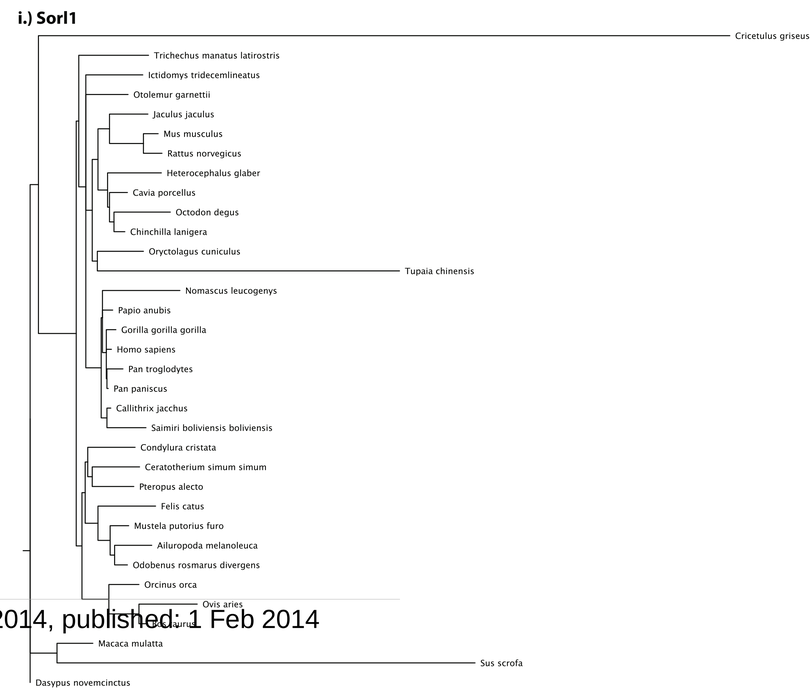






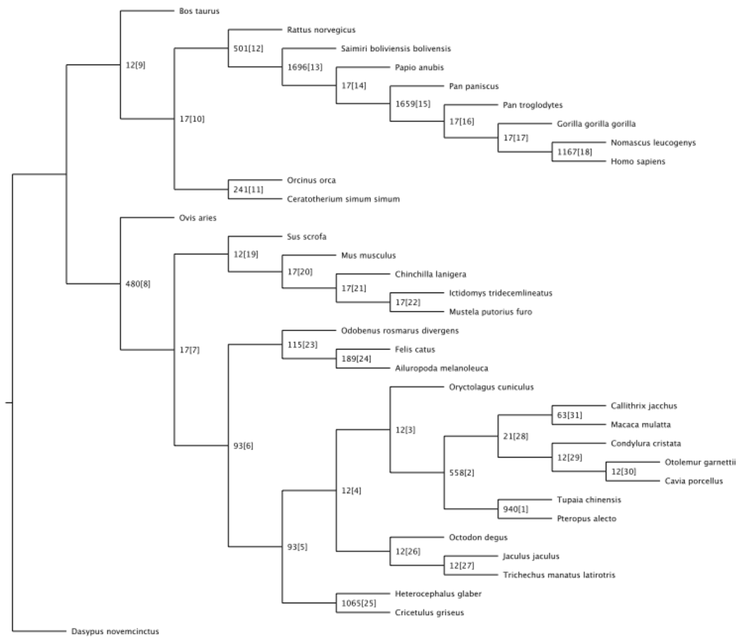






PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.230v1 | CC-BY 4.0 Open Access | received: 1 Feb 2014, published: 1 Feb 2014

Figure 4



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.230v1 | CC-BY 4.0 Open Access | received: 1 Feb 2014, published: 1 Feb 2014

Figure 5

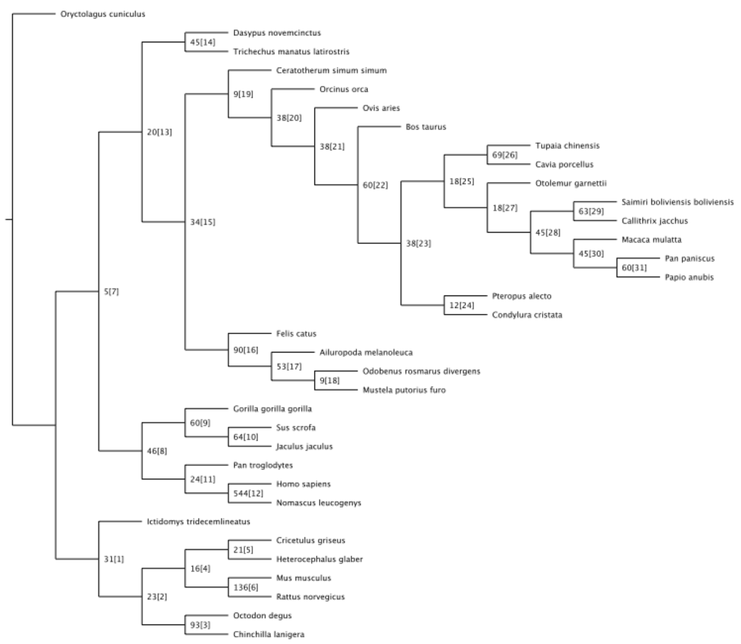


Figure 6a

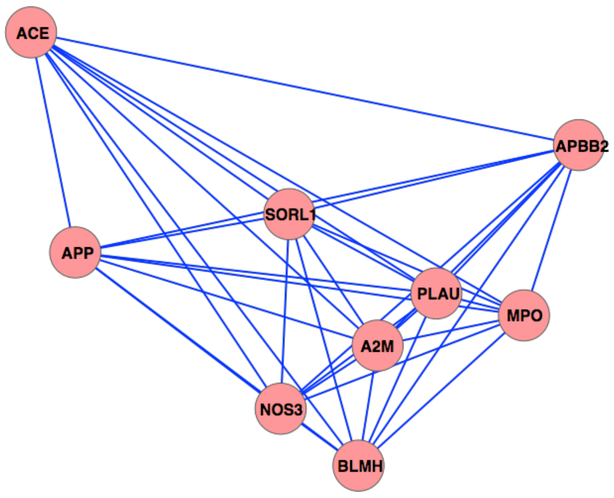


Figure 6b