Expressed sequence tags as a tool for phylogenetic analysis of placental mammal evolution.

Kullberg, Morgan; Hallström, Björn; Arnason, Ulfur; Janke, Axel

Published in:
PLoS ONE

DOI:
10.1371/journal.pone.0000775

2007

Citation for published version (APA):
Expressed Sequence Tags as a Tool for Phylogenetic Analysis of Placental Mammal Evolution

Morgan Kullberg*, Björn Hallström, Ulfur Arnason, Axel Janke

Department of Cell and Organism Biology, Division of Evolutionary Molecular Systematics, University of Lund, Lund, Sweden

Background. We investigate the usefulness of expressed sequence tags, ESTs, for establishing divergences within the tree of placental mammals. This is done on the example of the established relationships among primates (human), lagomorphs (rabbit), rodents (rat and mouse), artiodactyls (cow), carnivorans (dog) and proboscideans (elephant). Methodology/Principal Findings. We have produced 2000 ESTs (1.2 mega bases) from a marsupial mouse and characterized the data for their use in phylogenetic analysis. The sequences were used to identify putative orthologous sequences from whole genome projects. Although most ESTs stem from single sequence reads, the frequency of potential sequencing errors was found to be lower than allelic variation. Most of the sequences represented slowly evolving housekeeping-type genes, with an average amino acid distance of 6.6% between human and mouse. Positive Darwinian selection was identified at only a few single sites. Phylogenetic analyses of the EST data yielded trees that were consistent with those established from whole genome projects. Conclusions. The general quality of EST sequences and the general absence of positive selection in these sequences make ESTs an attractive tool for phylogenetic analysis. The EST approach allows, at reasonable costs, a fast extension of data sampling from species outside the genome projects.


INTRODUCTION

In 1992 Novacek [1] presented a widely known hypothesis for the phylogenetic tree of placental mammals based on a synthesis of morphological and molecular findings. At that time only limited amounts of sequence data were available, a circumstance that rendered many ordinal relationships unresolved. During an initial stage phylogenetic analyses of sequence data were generally based on single genes or parts of genes [2–4]. This changed gradually and during the 1990’s sequences of complete mitochondrial (mt) genomes became a common tool in phylogenetic analyses [e.g. [5,6]]. The combined sequences of all mt protein-coding genes yield alignment lengths of about 10–12 kbp, i.e. about 10-times the sequence amounts commonly used in the 1980s. However, in the absence of a closely related outgroup these analyses could not conclusively establish the direction of evolution in the placental tree. This limitation was amended by the first marsupial mt genome sequence, that of the opossum, Didelphis virginiana [7]. The marsupial rooting of the placental tree placed Rodentia (mouse, rat) as the sister group to remaining orders. This position of Rodentia was upheld in the great majority of the following mammalian mitogenomic (mtg) analyses, i.e. phylogenetic analyses based on the protein-coding genes of complete mt genomes [e.g. [8–12]]. However recent mtg studies joined rodents and primates on a common branch [e.g. [13]]. Thus, relationships within some basal parts of the placental tree remained equivocal, even in phylogenetic analyses of complete mt genomes. As some of these analyses demonstrated [9,12], the basal position of the rodents in the mtg tree of placental mammals was sensitive to the sampling of other basal taxa and to the analytical approaches applied.

In 2001 Murphy et al. [14] presented phylogenetic results that challenged some parts of the placental mtg tree. The study was based on both mt data and directly PCR amplified introns of nuclear genes. The contribution of individual taxa to the complete data set differed somewhat and the alignment of the nuclear sequences showed considerable numbers of gaps and ambiguous sites. This was particularly noticeable in three of the nuclear sequences (> 50% of the nuclear data) in which the amino acid distance between human and the mouse ranged between 20% and 40%, a circumstance that may adversely affect the aligning of homologous sites. Similarly the concatenation of genes showing great evolutionary rate variation may affect the estimation of model parameters such as the gamma distribution parameter, a [15,16]. The main parts of this nuclear gene tree [14] have nevertheless been supported in later studies based on far more comprehensive alignments [e.g. 17] and genome level characters like retroposon insertion and indel differences [18–20]. One of the main differences between this nuclear gene tree and previous mtg findings was that monophyletic Rodentia grouped with Lagomorpha, thereby supporting the morphological Glires hypothesis. Together with Primates, Dermoptera and Scandentia, Glires formed the superordinal clade Euarchontoglires. The sister group to the Euarchontoglires, called Laurasiatheria, included Artiodactyla, Carnivora and Perissodactyla among other orders. Euarchontaglires and Laurasiatheria are commonly joined in the Boreoeutheria.

The problems related to resolving basal placental relationships were again underlined in a recent study based on the sequences of eight housekeeping genes that were established by cDNA approaches from 22 placental mammals and three marsupials [21]. The total length of the alignment was 6 kb and all genes had similar evolutionary rates. Inconsistent with the results of Murphy...
et al. [14] the analyses favored a tree with Glires in a basal position relative to Primates rather than joining Primates and Glires on a common branch. Furthermore, and despite an extended and more uniform sequence representation of each individual taxon as compared to the study of Murphy et al. [14], the position of the root of the placental tree was not conclusively established when rate heterogeneity models were applied [21].

In this study we have selected the Boreoeutheria group to examine the utility ESTs for phylogenetic analyses, as a novel approach to economically obtain large amounts of protein-coding sequences. The procedure rests upon the random retrieval of ESTs from a cDNA library, which represents all expressed genes in a cell at a given time [22]. The ensuing database search allows subsequent complementation with orthologous sequences from species that are of interest to the phylogenetic study. This approach has hitherto been applied in only a limited number of phylogenetic studies that have addressed deep relationships among eukaryotes, plants, arthropods and mammalians [23–27].

Currently, genome projects of some 20 placental and marsupial mammals are in different stages of completion. Sequence data from these projects have allowed resolution of several ordinal and superordinal placental relationships [17] with which the results from the EST based phylogenetic analysis can be compared. In order to promote the identification of the root of the placental tree we have, along with the production of placental sequences established the homologous sequences from an Australian marsupial, the fat-tailed dunnart, Sminthopsis crassicaudata. With an upper palentological limit of about 125 million years before present, MYBP, for the divergence between marsupial and placental mammals [28] the inclusion of Sminthopsis constitutes a definite advantage in determining the root of the tree of placental mammals.

RESULTS

More than 1,200,000 nt sequences representing about 2000 EST sequences were retrieved from the Sminthopsis tissue culture cells (fibroblasts). About 1600 EST sequences with a minimum length of ≈400 bp were collected for further evaluation. After excluding vector and mt sequences, 854 individual nuclear cDNA sequences and contigs remained for the complementary database search. Orthology search against the human mRNA RefSeq database identified 455 protein-coding sequences with E-values <10−15 that were subsequently aligned. A list of the accession numbers of the putative 455 human orthologous mRNA sequences is provided in the Table S1. Several un-translated sequences were identified during the search. These sequences were not included in the study as it focuses on protein-coding genes. 344 of the 455 human mRNA transcripts could be classified according to the PANTHER classification system, while 109 sequences remained unclassified. Table 1 shows the classification for those gene classes that had more than five members.

Of these 455 sequences a total of 161 sequences were represented by seven placental species (elephant, mouse, rat, rabbit, human, cow and dog). This alignment (named maxspe) that maximized the mammalian representation for all sequences had a length of 77,328 nt (25,776 aa). A second alignment that maximized the number of sequences by allowing some sequences to be missing was also constructed. This alignment, which included 326 sequences (164,466 nt or 54,822 aa) from the eight species, is referred to as the maxgen alignment. Genomes with a low current sequencing coverage such as those of the elephant and the rabbit were allowed to lack 25% of the genes. In a few cases one or two sequences of cetferungulates (cow or dog) and/or rodents (mouse or rat) were allowed to be missing in the maxgen alignment.

<table>
<thead>
<tr>
<th>Function</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein biosynthesis</td>
<td>71</td>
</tr>
<tr>
<td>Transport related</td>
<td>35</td>
</tr>
<tr>
<td>mRNA Transcription</td>
<td>27</td>
</tr>
<tr>
<td>Cell structure</td>
<td>22</td>
</tr>
<tr>
<td>Proteolysis</td>
<td>21</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>17</td>
</tr>
<tr>
<td>Protein folding</td>
<td>16</td>
</tr>
<tr>
<td>mRNA splicing</td>
<td>13</td>
</tr>
<tr>
<td>Protein phosphorylation</td>
<td>12</td>
</tr>
<tr>
<td>Nucleotide, nt and na metabolism</td>
<td>10</td>
</tr>
<tr>
<td>DNA replication</td>
<td>10</td>
</tr>
<tr>
<td>Stress response</td>
<td>8</td>
</tr>
<tr>
<td>Cell motility</td>
<td>8</td>
</tr>
<tr>
<td>tRNA metabolism</td>
<td>8</td>
</tr>
<tr>
<td>Mitosis</td>
<td>8</td>
</tr>
<tr>
<td>Protein modification</td>
<td>7</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>7</td>
</tr>
<tr>
<td>Tricarboxylic acid pathway</td>
<td>6</td>
</tr>
<tr>
<td>Cell adhesion</td>
<td>6</td>
</tr>
<tr>
<td>Immunity and defense</td>
<td>6</td>
</tr>
<tr>
<td>Protein complex assembly</td>
<td>6</td>
</tr>
<tr>
<td>Protein glycosylation</td>
<td>6</td>
</tr>
<tr>
<td>Cell communication</td>
<td>6</td>
</tr>
<tr>
<td>Developmental processes</td>
<td>6</td>
</tr>
<tr>
<td>Intracellular protein traffic</td>
<td>6</td>
</tr>
<tr>
<td>Unclassified</td>
<td>109</td>
</tr>
</tbody>
</table>

NOTE–Only the most common classes according to the PANTHER classification with >5 identified homologues are shown.
doi:10.1371/journal.pone.0000775.0001

The chicken was not represented in about 33% of the alignments for both maxspe and maxgen and was therefore excluded from all analysis based on single genes. The general properties of the two datasets are given in Table 2.

The length of the individual and trimmed alignments excluding gaps varied from 126 to 1167 nt, with an average of 505 nt (Figure 1). The genetic distances between human and mouse ranged from 0% to 20% (mean = 4.5±4.5) for aa sequences (Figure 2), 0%–18% (mean = 3.2±2.9) for first and second codon positions (12cdp), and 3%–24% (mean = 11.0±3.3) for all codon positions (123cdp). Alignments with zero aa distance between human and mouse or human and cow were excluded from the

<table>
<thead>
<tr>
<th>Data set</th>
<th>Length nt</th>
<th>Gaps</th>
<th>Distance human-mouse</th>
<th>Constant sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>123cdp</td>
<td>12cdp</td>
<td>aa</td>
<td>123cdp</td>
</tr>
<tr>
<td>maxgen</td>
<td>164466</td>
<td>54%</td>
<td>0.111</td>
<td>0.033</td>
</tr>
<tr>
<td>maxspe</td>
<td>77328</td>
<td>6%</td>
<td>0.113</td>
<td>0.036</td>
</tr>
</tbody>
</table>

NOTE–The observed percentages and distances (substitutions per site) are shown.
doi:10.1371/journal.pone.0000775.0002
analysis. Additional properties such as the number of gaps and number of constant sites in the different alignments are shown in Table 2.

Estimates of potential sequencing errors in Sminthopsis ESTs indicated an error rate of approximately 0.01% and allelic variation of about 0.02%. Further evidence that sequence differences had been correctly classified as allelic variation rested on the observation that the sequence differences occurred always at silent 3rd codon positions. Most of the differences constituted frequent naturally occurring C-T transitions. A potential error rate of 0.01% was also recorded in 102,232 nt of mt ESTs with a 10-fold coverage of about 10,000 nt of overlapping mt protein-coding sites. Comparison between the EST data and the mt genome of another Sminthopsis individual showed 134 differences (0.1%). This value is within the expected sequence variation of mt sequences of different individuals. The results suggest that sequence differences related to sequencing errors are less frequent than natural allelic variation, although the statistics behind these differences is limited due to the low total numbers of differences. The findings suggest that potential sequencing errors in the EST sequencing study are at a level that effects the current phylogenetic analyses far less than allelic variation.

The aa distances within the two alignments, maxgen and maxspe, are shown in Table 3. Distances between the outgroup and the ingroup taxa differed by < 10%, indicating a limited difference in evolutionary rates among the ingroup species. There is a notable difference between the marsupial and placental distances relative to the chicken, indicating a faster evolution in the placentals. Among the placental mammals the sequences of Glires and the elephant appear to evolve faster than those of Homo and the cetferungulates. A chi-2 test as implemented in TREE-PUZZLE

![Figure 1. Distribution of the alignment lengths after trimming.](doi:10.1371/journal.pone.0000775.g001)

![Figure 2. Distribution of the pair wise aa distances between human and mouse.](doi:10.1371/journal.pone.0000775.g002)

<table>
<thead>
<tr>
<th>Chicken</th>
<th>Sminthopsis</th>
<th>Elephant</th>
<th>Dog</th>
<th>Cow</th>
<th>Human</th>
<th>Rabbit</th>
<th>Mouse</th>
<th>Rat</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.108</td>
<td>0.126</td>
<td>0.113</td>
<td>0.116</td>
<td>0.111</td>
<td>0.125</td>
<td>0.117</td>
<td>0.118</td>
<td></td>
</tr>
<tr>
<td>0.135</td>
<td>0.162</td>
<td>0.142</td>
<td>0.146</td>
<td>0.139</td>
<td>0.161</td>
<td>0.147</td>
<td>0.149</td>
<td></td>
</tr>
<tr>
<td>0.125</td>
<td>0.103</td>
<td>0.092</td>
<td>0.093</td>
<td>0.090</td>
<td>0.103</td>
<td>0.100</td>
<td>0.100</td>
<td></td>
</tr>
<tr>
<td>0.160</td>
<td>0.126</td>
<td>0.110</td>
<td>0.112</td>
<td>0.108</td>
<td>0.126</td>
<td>0.122</td>
<td>0.123</td>
<td></td>
</tr>
<tr>
<td>0.141</td>
<td>0.108</td>
<td>0.051</td>
<td>0.056</td>
<td>0.050</td>
<td>0.064</td>
<td>0.067</td>
<td>0.068</td>
<td></td>
</tr>
<tr>
<td>0.184</td>
<td>0.133</td>
<td>0.056</td>
<td>0.062</td>
<td>0.054</td>
<td>0.073</td>
<td>0.075</td>
<td>0.077</td>
<td></td>
</tr>
<tr>
<td>0.125</td>
<td>0.096</td>
<td>0.054</td>
<td>0.037</td>
<td>0.036</td>
<td>0.047</td>
<td>0.053</td>
<td>0.054</td>
<td></td>
</tr>
<tr>
<td>0.158</td>
<td>0.114</td>
<td>0.059</td>
<td>0.039</td>
<td>0.038</td>
<td>0.051</td>
<td>0.058</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>0.128</td>
<td>0.100</td>
<td>0.059</td>
<td>0.039</td>
<td>0.040</td>
<td>0.052</td>
<td>0.057</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>0.162</td>
<td>0.120</td>
<td>0.065</td>
<td>0.042</td>
<td>0.042</td>
<td>0.057</td>
<td>0.063</td>
<td>0.064</td>
<td></td>
</tr>
<tr>
<td>0.124</td>
<td>0.096</td>
<td>0.053</td>
<td>0.038</td>
<td>0.043</td>
<td>0.046</td>
<td>0.050</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>0.157</td>
<td>0.113</td>
<td>0.058</td>
<td>0.040</td>
<td>0.046</td>
<td>0.050</td>
<td>0.064</td>
<td>0.055</td>
<td></td>
</tr>
<tr>
<td>0.135</td>
<td>0.108</td>
<td>0.066</td>
<td>0.050</td>
<td>0.056</td>
<td>0.048</td>
<td>0.060</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td>0.176</td>
<td>0.132</td>
<td>0.074</td>
<td>0.054</td>
<td>0.061</td>
<td>0.052</td>
<td>0.066</td>
<td>0.070</td>
<td></td>
</tr>
<tr>
<td>0.132</td>
<td>0.106</td>
<td>0.072</td>
<td>0.058</td>
<td>0.063</td>
<td>0.055</td>
<td>0.048</td>
<td>0.020</td>
<td></td>
</tr>
<tr>
<td>0.170</td>
<td>0.129</td>
<td>0.082</td>
<td>0.064</td>
<td>0.069</td>
<td>0.060</td>
<td>0.071</td>
<td>0.071</td>
<td></td>
</tr>
<tr>
<td>0.133</td>
<td>0.108</td>
<td>0.074</td>
<td>0.060</td>
<td>0.064</td>
<td>0.056</td>
<td>0.066</td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td>0.171</td>
<td>0.132</td>
<td>0.084</td>
<td>0.066</td>
<td>0.071</td>
<td>0.062</td>
<td>0.071</td>
<td>0.024</td>
<td></td>
</tr>
</tbody>
</table>

Note—The values show the observed and JTT+4I+I distances (top to bottom). Above diagonal the maxgen and below maxspe alignment.

![Table 3. Pairwise aa distances between species.](doi:10.1371/journal.pone.0000775.t003)
showed that composition of the aa as well as 1st and 2nd codon positions (12cdp) is homogeneous among the mammalian species. However, nt composition was not homogeneous when the same test was applied to all three codon positions.

Figure 3 shows the Bayesian tree based on the nt sequences of the maxspe dataset. Posterior probability values were 1.00 for all nodes in the tree. ML rate heterogeneity bootstrap support from the maxspe data was moderate to high (70–100%) for the aa and cdp123 but low for cdp12 (39% for Boreoeutheria). When the chicken, the elephant, and the rabbit were excluded from the alignment there was a 0.85 posterior probability for the Euarchontoglires clade. Unpartitioned ML analyses including all species resulted in the same topology as in Figure 3, but rodents fell basal when the chicken, the elephant, and the rabbit were excluded. NJ and MP analysis generally placed rodents as sister group to all other placentals, regardless of the taxon sampling.

Likelihood ratio tests showed that partitioning of the data significantly improved the fit to the evolutionary model. The largest impact was seen from partitioning according to codon positions, where the increase in the logL values was several thousand. Partitioning according the genetic distances increased the logL values by a few hundred logL values, which was still a statistically significant improvement.

In order to further investigate the support for the best tree shown in Figure 3 the ΔlogL/S.E. ratio and the pSH values were calculated for alternative placental trees, Figure 4. The alternative trees represented the 2nd and 3rd best ML alternatives according to PROTML, an alternative tree based on housekeeping genes [21], the MP and NJ aa tree and the mtg tree [9]. Different evolutionary models were used to calculate the likelihood and pSH values for the maxgen and maxspe alignments on the basis of aa, 12cdp and 123cdp nt sequences (Table 4). All analyses identified tree-1 as the best tree, but alternative topologies could not be statistically rejected by all datasets, except in the case of tree-4.

The 161 genes of the maxspe alignment were individually analyzed in order to estimate the proportion of genes that supported each of the six alternative topologies shown in Figure 4. The logL for each of these topologies was calculated using TREE-PUZZLE. If the ΔlogL/S.E. ratio value between alternative trees exceeded 0.5 the topology was recorded. If the value was less than 0.5, the support was regarded as inconclusive. Although the cut-off ΔlogL/S.E. ratio of 0.5 was arbitrarily chosen, ΔlogL/S.E. ratios of this level indicate some support for the best topology over alternatives. Table 5 summarizes the support for alternative trees as recorded for individual genes applying the different analytical approaches. Most genes that contained enough phylogenetic information to distinguish between alternative topologies supported tree-1. The strongest phylogenetic support came from single gene analyses of the 123cdp alignments. However, the majority of the single genes does not carry sufficient phylogenetic information to distinguish between the alternative trees.

The nt and aa sequences of single genes from the maxspe dataset were analyzed with respect to their support for internal branches by calculating and comparing the ML values of different trees using PHYML and quartet puzzling (QP) as implemented in TREE-PUZZLE. The CONSENSE program (PHYLIP) was used

![Figure 3. Phylogenetic relationships based on partitioned Bayesian inference of the maxgen alignment (164466 nt) and a GTR+4I+I model of sequence evolution. The partitioning are according to four evolutionary rate classes and codon positions. The model parameters are estimated separate for every partitioning. doi:10.1371/journal.pone.0000775.g003](image)

![Figure 4. Relationships between species in the tested alternative topologies. doi:10.1371/journal.pone.0000775.g004](image)

**Table 4. ML support for the different phylogenetic hypothesis.**

<table>
<thead>
<tr>
<th></th>
<th>tree-1</th>
<th>tree-2</th>
<th>tree-3</th>
<th>tree-4</th>
<th>tree-5</th>
<th>tree-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>maxgen</td>
<td>JTT+4I+I</td>
<td>250468.7</td>
<td>2.11</td>
<td>1.63</td>
<td>3.03</td>
</tr>
<tr>
<td></td>
<td>pSH</td>
<td>1.00</td>
<td>0.313</td>
<td>0.417</td>
<td>0.010</td>
<td>0.020</td>
</tr>
<tr>
<td>maxspe</td>
<td>JTT+4I+I</td>
<td>[127791.9]</td>
<td>0.96</td>
<td>0.67</td>
<td>0.25</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>pSH</td>
<td>1.00</td>
<td>0.670</td>
<td>0.673</td>
<td>0.090</td>
<td>0.052</td>
</tr>
<tr>
<td>12cdp</td>
<td>maxgen</td>
<td>GTR+4I+I</td>
<td>[331163.9]</td>
<td>2.59</td>
<td>2.53</td>
<td>3.77</td>
</tr>
<tr>
<td></td>
<td>pSH</td>
<td>1.00</td>
<td>0.216</td>
<td>0.248</td>
<td>0.000</td>
<td>0.033</td>
</tr>
<tr>
<td>maxspe</td>
<td>GTR+4I+I</td>
<td>[152577.9]</td>
<td>1.43</td>
<td>1.79</td>
<td>2.23</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>pSH</td>
<td>1.00</td>
<td>0.488</td>
<td>0.404</td>
<td>0.072</td>
<td>0.203</td>
</tr>
<tr>
<td>123cdp</td>
<td>maxgen</td>
<td>GTR+4I+I</td>
<td>[668754.0]</td>
<td>6.84</td>
<td>6.03</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td>pSH</td>
<td>1.00</td>
<td>0.001</td>
<td>0.006</td>
<td>0.000</td>
<td>0.151</td>
</tr>
<tr>
<td>maxspe</td>
<td>GTR+4I+I</td>
<td>[326172.5]</td>
<td>4.33</td>
<td>5.35</td>
<td>4.47</td>
<td>1.93</td>
</tr>
<tr>
<td></td>
<td>pSH</td>
<td>1.00</td>
<td>0.056</td>
<td>0.069</td>
<td>0.000</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Note–The support is expressed as the ΔlogL/S.E. ratio and the pSH value. The -logL value of the best tree is shown in square brackets. doi:10.1371/journal.pone.0000775.t004


Depending on the mode of analysis the support for different clade selected internal branches according to the ML and QP analyses. Consensus tree. Table 6 shows the number of genes that favored to summarize the gene-trees by calculating a majority rule. Analyses of the concatenated dataset and the concatenated dataset confirmed that most maxgen 4/28521 of the other mammalian species studied. The effect of taxon sampling and the choice of outgroup species on the phylogenetic reconstruction were evaluated by analyzing the data after excluding one or more species at the time. Inclusion of the chicken as an extra outgroup besides Sminthopsis has no effect on the analysis, but when only the chicken was used as outgroup tree-1 and tree-5 became nearly indistinguishable. Likewise exclusion of rabbit led to the promotion of a basal position of rodents as in trees 4, 5 and 6. Euarchontaglires remained unresolved when the Sminthopsis was included in the analysis of single genes. Significant LRT at 1%.

### Table 5. Support of three alternative topologies by ML analysis of single genes.

<table>
<thead>
<tr>
<th>Topology</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tree/Species/Clade</td>
</tr>
<tr>
<td></td>
<td>aa+4I+1</td>
</tr>
<tr>
<td>tree-1</td>
<td>25</td>
</tr>
<tr>
<td>tree-2</td>
<td>3</td>
</tr>
<tr>
<td>tree-3</td>
<td>14</td>
</tr>
<tr>
<td>tree-4</td>
<td>4</td>
</tr>
<tr>
<td>tree-5</td>
<td>12</td>
</tr>
<tr>
<td>tree-6</td>
<td>11</td>
</tr>
<tr>
<td>unresolved</td>
<td>92</td>
</tr>
</tbody>
</table>

### Table 6. Numbers of genes that support a certain clade.

<table>
<thead>
<tr>
<th>Clade</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>aa+4I+1</td>
</tr>
<tr>
<td>Boreothia</td>
<td>13/55</td>
</tr>
<tr>
<td>Euarch.glir</td>
<td>7/49</td>
</tr>
<tr>
<td>Glires</td>
<td>17/61</td>
</tr>
<tr>
<td>Rodentia</td>
<td>111/147</td>
</tr>
<tr>
<td>Cetferungula</td>
<td>23/74</td>
</tr>
</tbody>
</table>

### Table 7. Sequences showing signs of positive selection on at least one branch under branch-site models. Significant LRT at 1%.}

<table>
<thead>
<tr>
<th>Gene function/name</th>
<th>Branch/Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene accession</td>
<td>Number of codons under pos. selection/number of positions</td>
</tr>
<tr>
<td>NM_001064 elephant</td>
<td>2/204 transketolase</td>
</tr>
<tr>
<td>NM_003670 human</td>
<td>1/132 basic helix-loop-helix</td>
</tr>
<tr>
<td>NM_005094 cow</td>
<td>1/86 solute carrier</td>
</tr>
<tr>
<td>NM_005866 rabbit</td>
<td>1/126 opioid receptor</td>
</tr>
<tr>
<td>NM_015150 Sminthopsis</td>
<td>2/76 lipid raft link</td>
</tr>
<tr>
<td>NM_012225 rabbit</td>
<td>2/163 nucleotide binding protein</td>
</tr>
<tr>
<td>NM_000546 rabbit</td>
<td>3/123 tumor protein p53</td>
</tr>
<tr>
<td>NM_004559 elephant</td>
<td>5/243 Y box binding protein</td>
</tr>
<tr>
<td>NM_001428 rat</td>
<td>1/284 p53-binding protein</td>
</tr>
<tr>
<td>NM_006739 rabbit</td>
<td>1/216 minichromosome maintenance complex</td>
</tr>
<tr>
<td>NM_018622 rat</td>
<td>11/207 presenilin associated</td>
</tr>
<tr>
<td>NM_000989 cow</td>
<td>4/114 ribosomal protein L30</td>
</tr>
<tr>
<td>NM_024092 elephant</td>
<td>2/182 transmembrane protein</td>
</tr>
<tr>
<td>NM_006449 dog</td>
<td>2/200 CDC42 effector protein</td>
</tr>
<tr>
<td>NM_018285 dog</td>
<td>2/172 ribonucleoprotein</td>
</tr>
<tr>
<td>NM_005530 cow</td>
<td>2/232 isocitrate dehydrogenase</td>
</tr>
<tr>
<td>NM_015292 cow</td>
<td>1/265 family with sequence similarity</td>
</tr>
<tr>
<td>NM_004135 cow</td>
<td>1/199 isocitrate dehydrogenase</td>
</tr>
<tr>
<td>NM_018361 rodentia</td>
<td>2/108 lyosphosphatic acid acyltransferase</td>
</tr>
<tr>
<td>NM_006530 rabbit</td>
<td>2/218 YETS domain containing 4</td>
</tr>
<tr>
<td>NM_002306 rabbit</td>
<td>2/182 galactoside-binding</td>
</tr>
<tr>
<td>NM_002306 mouse</td>
<td>1/182 galactoside-binding</td>
</tr>
<tr>
<td>NM_014300 rat</td>
<td>1/163 SEC11 homolog A</td>
</tr>
<tr>
<td>Concatenated Chicken</td>
<td>55/28521</td>
</tr>
<tr>
<td>Concatenated Sminthopsis</td>
<td>8/28521</td>
</tr>
<tr>
<td>Concatenated Elephant</td>
<td>40/28521</td>
</tr>
<tr>
<td>Concatenated Cow</td>
<td>7/28521</td>
</tr>
<tr>
<td>Concatenated Dog</td>
<td>3/28521</td>
</tr>
<tr>
<td>Concatenated Human</td>
<td>1/28521</td>
</tr>
<tr>
<td>Concatenated Rabbit</td>
<td>5/28521</td>
</tr>
<tr>
<td>Concatenated Mouse</td>
<td>1/28521</td>
</tr>
<tr>
<td>Concatenated Rat</td>
<td>4/28521</td>
</tr>
<tr>
<td>Concatenated Glires</td>
<td>0/28521</td>
</tr>
<tr>
<td>Concatenated Euarchontaglires</td>
<td>0/28521</td>
</tr>
<tr>
<td>Concatenated Boreothia</td>
<td>0/28521</td>
</tr>
<tr>
<td>Concatenated Placentalia</td>
<td>9/28521</td>
</tr>
</tbody>
</table>

To summarize the gene-trees by calculating a majority rule consensus tree. Table 6 shows the number of genes that favored selected internal branches according to the ML and QP analyses. Depending on the mode of analysis the support for different clade varied. The only branch receiving strong support was Rodentia, represented by the relatively closely related mouse and rat; all other branches were weakly supported, notably by the QP method.

The effect of taxon sampling and the choice of outgroup species on the phylogenetic reconstruction were evaluated by analyzing the data after excluding one or more species at the time. Inclusion of the chicken as an extra outgroup besides Sminthopsis has no effect on the analysis, but when only the chicken was used as outgroup tree-1 and tree-5 became nearly indistinguishable. Likewise exclusion of rabbit led to the promotion of a basal position of rodents as in trees 4, 5 and 6. Euarchontaglires remained unresolved when the Sminthopsis was included in the analysis of single genes. Significant LRT at 1%.

**NOTE** - The number of genes that support a fixed topology with a logL/S.E. ratio equal or better than 0.5 are shown. The resolution was regarded as unresolved when the jlogL/S.E. ratio was smaller than 0.5.

<table>
<thead>
<tr>
<th>Topology</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>tree-1</td>
<td>25</td>
</tr>
<tr>
<td>tree-2</td>
<td>3</td>
</tr>
<tr>
<td>tree-3</td>
<td>14</td>
</tr>
<tr>
<td>tree-4</td>
<td>4</td>
</tr>
<tr>
<td>tree-5</td>
<td>12</td>
</tr>
<tr>
<td>tree-6</td>
<td>11</td>
</tr>
<tr>
<td>unresolved</td>
<td>92</td>
</tr>
</tbody>
</table>
DISCUSSION

Phylogenetic studies of deeper mammalian relationships, such as those of placental orders, are to a great extent based on sequence analysis of protein-coding genes. Compared to the protein-coding regions the non-coding regions of the genomes evolve in general much faster and are often rearranged and randomized by multiple substitutions. Hitherto, nuclear gene sequences used in phylogenetic analyses have commonly come from genomic PCR-amplifications of single exons or from intron-less protein-coding genes [e.g. [14,29,30]]. The application of cDNA approaches circumvents this limitation as it allows the production of complete coding sequences from a variety of genes [21]. Sequencing of ESTs is one method to economically produce large amounts of protein-coding sequences for phylogenetic purposes.

About 45% of the EST sequences obtained in the current study constituted multiple nuclear sequences and about 7% were mt encoded. About 1/3rd of the nuclear sequences and contigs had putative homologues among the selected species and could be used for phylogenetic analysis. Most of the ESTs produced in the current study represented genes that can be classified as housekeeping genes [31–33]. Among mammalian orders the aa distances of housekeeping genes are generally limited to only a few percent, making them an ideal choice for phylogenetic analysis due to the facility with which correct alignments can be established. The limited distances among homologous housekeeping genes contrasts in this respect to nuclear genes such as vWF (von Willebrand factor), IRBP (interphotoreceptor retinoid binding protein) and the BRCA1 (breast and ovarian cancer susceptibility protein 1) sequence, which are commonly included in phylogenetic analyses. These three sequences have aa distance of 20–45% among mammalian orders. In comparison to these three sequences the effect of randomization in housekeeping genes can be considered as being limited. While it can be argued that the conserved nature of the housekeeping genes reduces the phylogenetic content of each single gene, this is compensated by the large amount of different EST sequences that can be produced from each individual taxon. Another advantage of applying cDNA sequences is that the risk of including pseudogenes is low. Although there are reports suggesting that some pseudogenes are transcribed [34], it is conceivable that pseudogenic sequences, if present, are rare among the large number of genic EST sequences. Thus, even if a few pseudogenes might occur among the ESTs they would be expected to have little or no influence on the phylogenetic outcome. By applying search against mRNA databases such as RefSeq the potential inclusion of pseudogenic sequences is also counteracted.

One particular difficulty in any phylogenetic study that utilizes nuclear encoded genes is the establishment of their orthology. When PCR based approaches are used to amplify sequences from genomic DNA [14,29,30] or cDNA [21] the orthology can only be assumed by similarity criteria. Criteria that use phylogenetic information [25] are generally preferred, but these depend on a known tree, the very subject that is under study. Synteny is another powerful criterion that has been used to define orthology [34]. Determination of orthology by the way of synteny analysis could not be achieved in the current study since it requires that the sequence of almost the whole genome is available, which is not the case for Sminthopsis. For this reason a number of precautions were undertaken in the current study in order to minimize the risk of including paralogous or pseudogenic sequences. This included reciprocal BLAST searches between human and the other species, with a cutoff E-value of 10−13 in order to ensure orthology between species. Also the use of a high quality databases such as RefSeq promotes the inclusion of known functional genes rather than pseudogenes in the analysis. In addition, after finding a possible human homologue, its full-length cDNA sequence instead of the shorter EST sequences was used for reciprocal searches in other species to further increase the chance of identifying orthologous genes in other species. The rigorous approach applied aims to maximize the probability of including only orthologous genes in the analyses.

There have been concerns about the quality of EST data for phylogenetic analyses [35], because most of the individual sequences are based on single reads. This study showed, however, that sequence differences due to sequencing errors of ESTs are at a level similar to that of allelic variation or even lower. The potential effect of errors of this kind can therefore be considered as negligible for the current phylogenetic results. The phylogenetic analyses appeared stable with respect to the assumption of evolutionary neutrality. The search for positive selection identified only few genes with single sites that had an ω value >1 in one or more branches. Other studies [36] have identified even fewer incidents of positive selection, when a pairwise method [37] was applied to compare ω between distant groups from different animal classes. The pairwise approach of that study probably identified fewer candidate genes, because the evaluation did not rely on phylogenetic information. Thus, selection may have been active on single branches in the past without the signal being recognized. The discrepancy between the number of sites identified by the analysis of the single genes and the concatenated musgn dataset may be related to more robust statistics when larger numbers of codons are involved in the analysis.

Compared to the total number of characters the low number of codons and the few genes that appear to be under selection appear to have no practical effect on the phylogenetic reconstruction. Only a few branches are affected by positive selection and the selection was not specific for a particular group, species or branch. If the phylogeny is incorrect, however, e.g. rodents or Glires as sister to the remaining species as in tree-4 to tree-6, fewer sites should actually be selected for, because in suboptimal phylogenies the number of sites under selection becomes over-estimated [38].

Phylogenetic examination of the eight placental taxa that are represented by nearly complete genomes joined primates and Glires as sister groups composing the clade Euarchontoglires. This clade is sister to Boreoeutheria as represented by the cow and the dog. The analyses favored a sister group relationship between the elephant and all other placentalts included. The same general topology was also found using sequences from the ENCODE consortium [18,39]. Other studies [14,19,40] have favored the same topology using shorter alignments but a more comprehensive taxon sampling. These results challenge previous studies using EST [24], large-scale genomic data [41–43], mtgs [9] and the analysis of few housekeeping genes [21].

Basal mammalian divergences have proven difficult to resolve, despite the use of large amount of nuclear sequences. This may reflect the potentially narrow temporal window within which the divergences took place, leaving only a low number of phylogenetically informative sites. While this constitutes a limitation that is common to all phylogenetic analyses the impact that taxon sampling may have on the tree is striking. Thus, in the current study Rodents tended to become the sister group to other placentalts when the rabbit was not included. The same tendency, which may be the effect of long branch attraction [44], has been observed earlier is studies with a limited taxon sampling [41–43]. An indication of a long branch attraction between rodents and the outgroup is that tree-5 was favored by MP and NJ analyses. Especially MP is known to be sensitive to evolutionary rate differences among the taxa and long branch attraction [44].
Among placental mammals the rodents may behave as a so-called rouge taxon, i.e. a lineages that tends to skew phylogenetic analyses [43]. This effect can be overcome by excluding taxa of this kind [46] or by including less deviating taxa for compensation. In the current study the attraction between rodents and the outgroup could be compensated by the inclusion of lagomorph sequences and complex model of sequence evolution. Other studies have questioned the use of long branch attraction phenomenon as an explanation to a basal position of rodents, because this particular topology received strong support from slowly evolving genes, while fast evolving genes supported Euarchontoglires [43]. The authors concluded that this contradicts the expectation from the effect of long branch attraction. The tendency of ML analysis to join fast (mouse) and slow (human) evolving taxa has instead been coined “long branch repulsion” or “opposite branch attraction” [43]. It is not clear how effects of this kind may have affected the current results, because it is not obvious why the rabbit would promote the opposite branch attraction phenomenon. The shift in the topology favored by ML when the rabbit is excluded form the analysis is more easily explained by long branch attraction between rodents and the outgroup. Further evidence for long branch attraction comes from the observation that ML cannot distinguish between tree-1 and tree-5 when only chicken is used as an outgroup. It appears that in these analyses the rodents are dragged towards the root of the placental tree. This illustrates the importance of choosing a not too distant outgroup and justifies the establishment of the marsupial EST sequences for this study.

Assumptions about evolutionary models have a major impact on the recovered phylogeny. A parameter rich model naturally fits the data better than a simpler model [47]. Dividing the data into partitions increases the number of parameters and thereby the fitness of the model. In order to create evolutionary models that are more realistic Kjer and Honeycutt [40] partitioned the sites in a mtg analysis of eutherian relationship into classes according to their relative MP consistency index. This resulted in a mtg phylogeny with strong support for the Atlantogonata hypothesis, which has not being supported by non-partitioned mtg analysis [9]. The performance of different partition strategies for concatenated data has been studied for mollusk sequence data. Partitioning data according to codon positions, and to a lesser extent also by genes, improves the fitness of models to the data [16]. In a large and variable dataset such as in this EST study, one can expect that extensive rate heterogeneity that may not be correctly accounted for by using non-partitioned analyses.

We tried to account for among site rate variation by partitioning the data into codon positions and also after the evolutionary rate of the genes. Obviously, partitioning according to single genes was impossible due to the number of genes included in this study. The size of each partition needs also to be reasonable large in order to get an accurate estimations of each parameter. We therefore divided the data according to distance classes and codon positions. Our approach to account for the among site rate variation satisfy the desire for a realistic model and still keeps the analysis on a computational acceptable level.

As mentioned above a limited taxon sampling may lead to an incorrect topology due to assuming models that are not consistent with the evolution. This could explain the results of some previous studies [24,41,42]. The strong attraction of rodent to the outgroup disappears for our data when they are partitioned and analyzed under Bayesian approaches.

While complete genomes are the ultimate data sets for resolving phylogenetic and evolutionary issues of different kinds [e.g. [28,41]], the costs of producing these data sets are still at a level that that precludes a dense taxonomic sampling among higher organisms. There is therefore a need to establish methods that at reasonable costs allow the production of sequence data that can be of general interest for phylogenetic studies. Producing EST sequences is such a method that will gain more attention in the future.

MATERIALS AND METHODS

Total RNA was isolated from fibroblast cell culture (EAEC number: SC 11) of Smimthopus crassicaudata (fat-tailed dunnart) using the acid phenol guanidium thiocyanate method, GTC-method [48]. Enriched mRNA was reverse transcribed, size fractionated, and cloned, yielding a total of 2000 cDNAs that were sequenced by QiAGEN (Germany) (Accession number EV533153-EV534021). The retrieved sequences were analyzed for identical or overlapping sequences from different clones using the Sequencher version 4.6 (Gene Codes) software. Contigs were assembled from overlapping genes. The sequencing error rate and the proportion of allelic variation were estimated by comparing more than 500 nucleotides (nt) from ESTs that were represented by two or more clones. NT differences were recorded as allelic variation when at least two different nt occurred at the same site, with each type being represented in at least two sequences. Other differences were counted as potential sequencing errors. As evident with this approach allelic variation becomes automatically under-estimated and sequencing errors overestimated unless comprehensive sequence coverage exists. Furthermore, the sequence error rate was estimated from comparing mt EST sequences among themselves and to the complete mt genome from another individual (Accession number NC_007631).

Individual sequences and contigs were used to search for homologous sequences in the mrna database using blastn as implemented in the EST-e-mate v1.0 program package (in-house application, code available from authors Hallstro¨m and Janke). In short, EST-e-mate blast the marsupial ESTs against the NCBI human mRNA RefSeq database [49]. The human sequences with the lowest E-value (expect value) were chosen as a template for searching for homologous sequences from the corresponding RefSeq database [49] of chicken (Gallus gallus), mouse (Mus musculus), rat (Rattus norvegicus), cow (Bos taurus) and dog (Canis familiaris) as at 30.Apr.2006. Rabbit (Oryctolagus cuniculus) and elephant (Loxodonta africana) were retrieved from ENSEMBL [50] 1.eahr.2007. Species which represent very close relatives (e.g. chimpanzee to human) were not included in the analysis. The elephant was chosen, because it shows the slowest evolutionary rate among the non-boreoeutherians [17]. Sequence hits with E-values above 10^-15 were excluded from further analysis. The program EST-e-mate utilizes ClustalW [51] for aligning the sequences, while keeping the reading frame intact with sequence alignments trimmed relative to the shortest sequence. Gaps and columns with ambiguous characters were removed. Potentially faulty alignments, i.e. alignments in which any taxon pair had amino acid (aa) distance value >0.6 were inspected further. All alignments were manually inspected using the Se-Al v2.0a11 software [52] and analyzed individually or as concatenated files. The mRNAs were functionally classified using the PANTHER classification system [53–54].

Sequence data were analyzed using the TREE-PUZZLE [55], PHYLIP [56], MOLPHY [57], MrBayes v3.1.2 [58], PAML3.15 [59], TREEFINDER [60], PHYML [61] or PAUP* [62] program packages. For the concatenated data the best-fitted model for nt sequence evolution and parameters were determined applying MODELTEST version 3.7 [63] and PROTTTEST version 1.3 [64]. The JTT model [65] of amino acid (aa) sequence evolution
and the GTR model [66] of nt evolution were used for distance and likelihood analyses. All phylogenetic analyses were computed assuming a gamma model of rate heterogeneity [67] with four classes of variable sites and one class of invariable site (A+F).

When a program did not allow for invariable sites, eight classes of variable sites and one class of invariable site (A+F) were used. For the merged alignment Bayesian analysis were conducted running two simultaneously analysis with MrBayes applying one cold and three heated chains for 10,000,000 MCMC (Markov chain Monte Carlo) generations, discarding the first 1,000,000 generations as burnin. To compensate for the rate heterogeneity in the data we divided the alignment into twelve partitions, each with its own individual GTR matrix, gamma distribution, proportion of invariable sites and base frequencies. The four main partitions were according to the observed (0–5%, 5–10%, 10–15%, >15%) aa distances between human and mouse. These four partitions were further divided according to codon positions. For aa and cdpl2 the dataset were divided only into four partitions according to aa distances. Bayesian analyses were made on the TITAN cluster of the Bioportal [68].

Analyses for potential selection were made for the concatenated sequences and single genes using codeml (PAML3.15) by estimating the non-synonymous (dN), synonymous (dS) substitution rates and ω (dN/dS) for one branch at a time. This approach corresponds to the branch-site model. In this model the branch of interest (foreground) can have sites with an ω-value larger than one and all other branches (background) are restricted to ω-values below or equal to one [69]. A Bayes empirical Bayes (BEB) procedure [70] was used to identify the sites evolving under potential positive selection. The codon frequencies were estimated from the data (CodonFreq = 3). All alignment columns containing ambiguities and gaps were excluded during the PAML analysis.

**SUPPORTING INFORMATION**

**Table S1** Accession numbers of 455 the human genes homologous to the ESTs.

Found at: doi:10.1371/journal.pone.0000775.s001 (0.12 MB DOC)

**ACKNOWLEDGMENTS**

The authors would like to express their gratitude to Maria Nilsson for comments on the manuscript, the Bioportal at University of Oslo for making computer resources public available and two anonymous reviewers for comments on an earlier version of the manuscript.

**Author Contributions**

Conceived and designed the experiments: AJ MK.Performed the experiments: MK. Analyzed the data: MK BH. Contributed reagents/materials/analysis tools: BH. Wrote the paper: UA AJ MK.

**REFERENCES**

15. DeBry RW (2003) Identifying conflicting signal in a multigene analysis reveals the GTR model of nt evolution were used for distance and likelihood analyses. All phylogenetic analyses were computed assuming a gamma model of rate heterogeneity [67] with four classes of variable sites and one class of invariable site (A+F).

When a program did not allow for invariable sites, eight classes of variable sites and one class of invariable site (A+F) were used. For the merged alignment Bayesian analysis were conducted running two simultaneously analysis with MrBayes applying one cold and three heated chains for 10,000,000 MCMC (Markov chain Monte Carlo) generations, discarding the first 1,000,000 generations as burnin. To compensate for the rate heterogeneity in the data we divided the alignment into twelve partitions, each with its own individual GTR matrix, gamma distribution, proportion of invariable sites and base frequencies. The four main partitions were according to the observed (0–5%, 5–10%, 10–15%, >15%) aa distances between human and mouse. These four partitions were further divided according to codon positions. For aa and cdpl2 the dataset were divided only into four partitions according to aa distances. Bayesian analyses were made on the TITAN cluster of the Bioportal [68].

Analyses for potential selection were made for the concatenated sequences and single genes using codeml (PAML3.15) by estimating the non-synonymous (dN), synonymous (dS) substitution rates and ω (dN/dS) for one branch at a time. This approach corresponds to the branch-site model. In this model the branch of interest (foreground) can have sites with an ω-value larger than one and all other branches (background) are restricted to ω-values below or equal to one [69]. A Bayes empirical Bayes (BEB) procedure [70] was used to identify the sites evolving under potential positive selection. The codon frequencies were estimated from the data (CodonFreq = 3). All alignment columns containing ambiguities and gaps were excluded during the PAML analysis.

**SUPPORTING INFORMATION**

**Table S1** Accession numbers of 455 the human genes homologous to the ESTs.

Found at: doi:10.1371/journal.pone.0000775.s001 (0.12 MB DOC)

**ACKNOWLEDGMENTS**

The authors would like to express their gratitude to Maria Nilsson for comments on the manuscript, the Bioportal at University of Oslo for making computer resources public available and two anonymous reviewers for comments on an earlier version of the manuscript.

**Author Contributions**

Conceived and designed the experiments: AJ MK. Performed the experiments: MK. Analyzed the data: MK BH. Contributed reagents/materials/analysis tools: BH. Wrote the paper: UA AJ MK.