# Introduction

One of the goals of genetics is to better understand and interpret the patterns and processes of evolution. To investigate the evolutionary distances between species, genome-scale sequence alignment - a very powerful tool – can be used. However, alignments often favour hundreds of possible phylogenies. In order to successfully analyse and draw conclusions, one has to interpret this high dimensional data in a meaningful way, which is a challenging task to do. The interpretation needs to preserve all the important information, and discard some, to ease intelligibility. Efficient and accurate dimensionality reduction methods among competing phylogenies could significantly better diagnosis of current evolutionary models, and could potentially improve phylogenetic reconstitution methods. [1]

In the following report, some possible designs, and the implementation of those are going to be discussed. These designs aim for visualizing either the differences and commonalities of two possible phylogenetic trees, or the differences and commonalities of a whole scale of possible phylogenetic trees.

# Design

To have a brief overview of the data, a tree viewer software, FigTree was used. [2]

For sketching graphs, a photo-manipulating software, PhotoFiltre 7.2.1 was used. [3]

## Comparison of two phylogenies

1. In this case, there are only two trees, which are small. Therefore it is feasible to preserve all information, and simply visualize them on the same figure, highlighting the differences: the evolutionary distances and the branches (in state 0, orangutan is closer to gorilla than to siamang in terms of most recent common ancestry, while in state 23000, orangutan is closer to siamang). This can be visualized in numerous ways of which one is presented in the appendix (5. figure). Using coloured names makes it easy and quick to spot the differences between the two branches. The coloured branches and highlight each node with a grey dot further improve the visual.

   Additionally, after examining the complete dataset, the two trees can be labelled with the percentage of trees showing the same evolutionary path (having the same branches). This provides essential information to the viewer, on how frequent these paths are.

2. Looking at a phylogenetic tree of primates, one might be interested in how closely those species are related to us, humans? This question could emerge for various reasons: Is it possible and is it ethical to use them as model organisms? Is it possible to study behavioural or genetic patterns, are those parallel to human patterns?

   Therefore, for another way of visualizing the data a bar chart was chosen, showing the evolutionary distances from humans by species, by state. The states – again – could be labelled by the probabilities of the paths. This visual could be further improved by making it interactive:

---

[1] Wilgenbusch, James C., Wen Huang, and Kyle A. Gallivan. "Visualizing phylogenetic tree landscapes." *BMC bioinformatics* 18.1 (2017): 85.

[2] http://tree.bio.ed.ac.uk/software/figtree/ (14. 05. 2017.)

[3] http://www.photofiltre-studio.com/ (14. 05. 2017.)

The user could chose the species to which all the other species are compared in terms of evolutionary distance. (6. figure)[4]

3. An important difference between the two trees is in the branches. A third useful figure would visualize the common branches, and the distinct branches. Here, the distinct parts are blue and red, and the common part is purple (the "intersection" of the two trees, the colour is purple, because it is the mixture of blue and red). (7. figure)

### Entire scale of possible phylogenies

1. The first design for visualizing all of the states at once, attempts to catch the most possible information in only one plot. Therefore, the figure consists of two graphs: one for each possible branch order. But, instead of indicating only one possible branch-length, the graphs show the entire range of possible branch lengths within the dataset. The figures also indicate the percentage of trees within the dataset matching that specific branch order.
   This design captures almost all the information from the dataset, in a simple and easily understandable way. However, this method would work best with small trees. (8. figure)
2. One might be interested in the evolutionary distance between these species, and humans, for reasons mentioned above.
   a. This can be visualized on line graphs for each species separately, which are showing the evolutionary distance (y axis) depending on the state (x axis). The graphs designed below also show the average evolutionary distance, along with the minimum and the maximum (9. figure).
      This design is not only capturing changes over the states, but at the same time shows a basic statistical summary of it. Again, the design works best, if the number of species is relatively small, as the number of graphs increases with the number of species.
   b. Another, more compact way of using the idea of line graphs is to plot them in the same coordinate system. This visual is more convenient in terms of comparison, although comes at a cost of small data loss (showing the minimum and maximum values would make the plot too messy) (10. figure).

   These visuals could be further improved by making them interactive: The user could chose the species to which all the other species are compared in terms of evolutionary distance.
3. The previous two designs aimed to compress data, to show some information – that could be of interest – and hide the rest, reducing the tremendous data into one or few charts. This last design's goal is to keep all the graphs, and try to navigate through it, rather than just compressing them.
   Once the tree is opened, the possible branch orders are displayed on the screen, with average branch lengths, the percentage of the trees with that particular branch order would also be displayed (11. figure). Once a tree is clicked (12. figure), the first state that exhibits that branch order is displayed, and the user can scroll through all the trees that exhibit that branch order (13. figure).

## Implementation

From "Comparison of two phylogenies" the second design, and from "Entire scale of possible phylogenies" also the second design was implemented. For the implementation, LiClipse 3.3.0 with python 3.6.1, and R Studio 1.0.143 with R 3.4.0 was used.

For details concerning the logic of the code, see the screencast (Screencast).

[4] Sketch of bar chart made with: https://www.onlinecharttool.com/ (14. 05. 2017.)

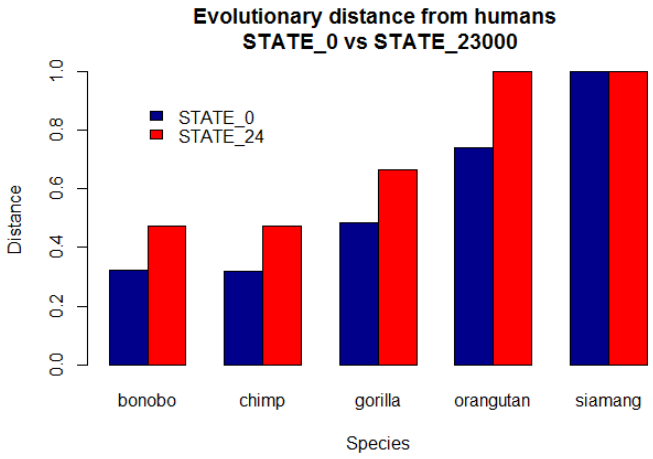
The evolutionary distances were extracted from the nexus file along with the states, and written into a text file, so working with the data in R would be easy. The data from the generated text file was loaded in R Studio, and manipulated such that the evolutionary distances were scaled to one, and relative to humans. Relativizing to any other species is easily manageable, by changing 'human' in the following piece of code (1. figure) to the primate of interest.

```
#get evolutionary distances for the primates relative to humans
for(i in (1:ncol(stl))){
  stl[,i] <- ifelse(stl[,i]>=stl$human,stl[,i],stl$human)
}
```

*1. figure – relativizing of evolutionary distances to humans; change the highlighted piece of code to the primate of interest, to get values relative to that primate.*

## Comparison of two phylogenies – Design 2

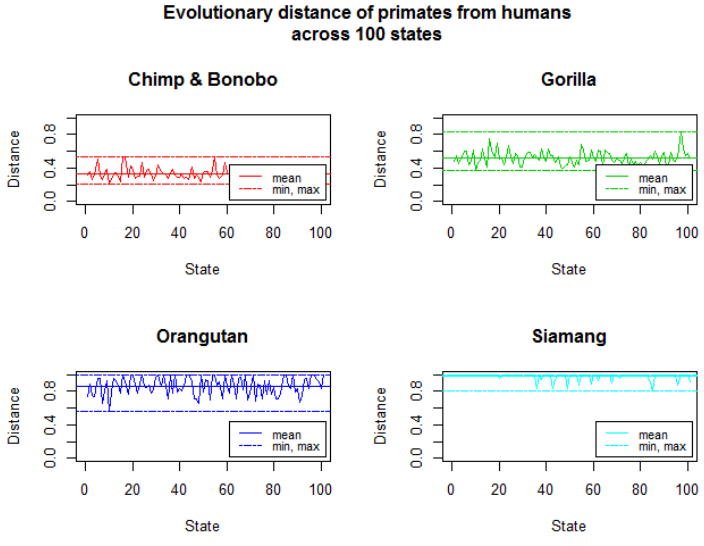
After pre-processing the dataset, coupled boxplots were generated, that show the information in a manner described in the design section previously.




*2. figure – design 1.2 Evolutionary distance form humans, comparison between STATE_0 and STATE_23000; Distance values scaled to one.*

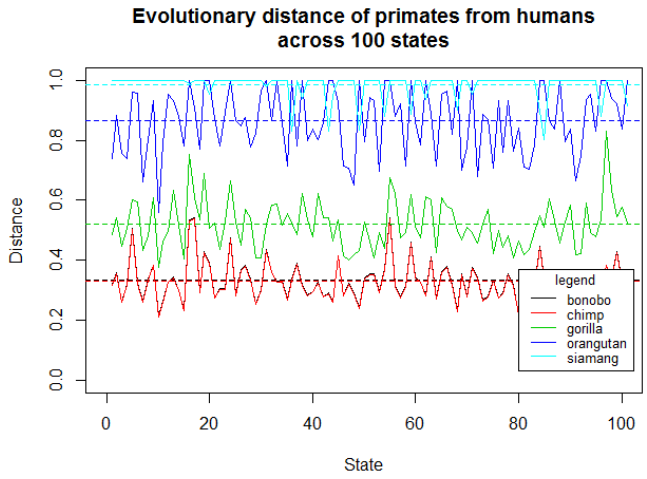
## Entire scale of possible phylogenies – Design 2, a) & b)

After processing the dataset a little further, the following line graphs were generated. The legends can be placed as needed due to the *locator()* function, to make sure, the graphs are not covered. (3. figure)




*3. figure – design 2.2.a Evolutionary distance of primates from humans across 100 states*

Another, more compact way of using the idea of line graphs is to plot them in the same coordinate system. This visual is more convenient in terms of comparison, although comes at a cost of small data loss. Again, with the *locator()* function, the legend can be placed manually.




*4. figure – design 2.2.b Evolutionary distance of primates form humans across 100 states*

## Insight

In design 2.2.b (4. figure), it can be seen that the graphs of siamang and orangutan cross each other multiple times. The variance changes from species to species as well. On the plot, we can compare the mean values of species.

Depending on the variance and the extremes (max, min), in some cases, it might be a good idea to use other methods to describe the population average, than the mean (e. g. median, quartiles /as speciation is most likely to be linked to a range of time, rather than a strict time point/ …etc.). Therefore, the fluctuation of evolutionary distance across species, across states could be subject of statistical analysis, to capture the characteristics of a set of trees better.

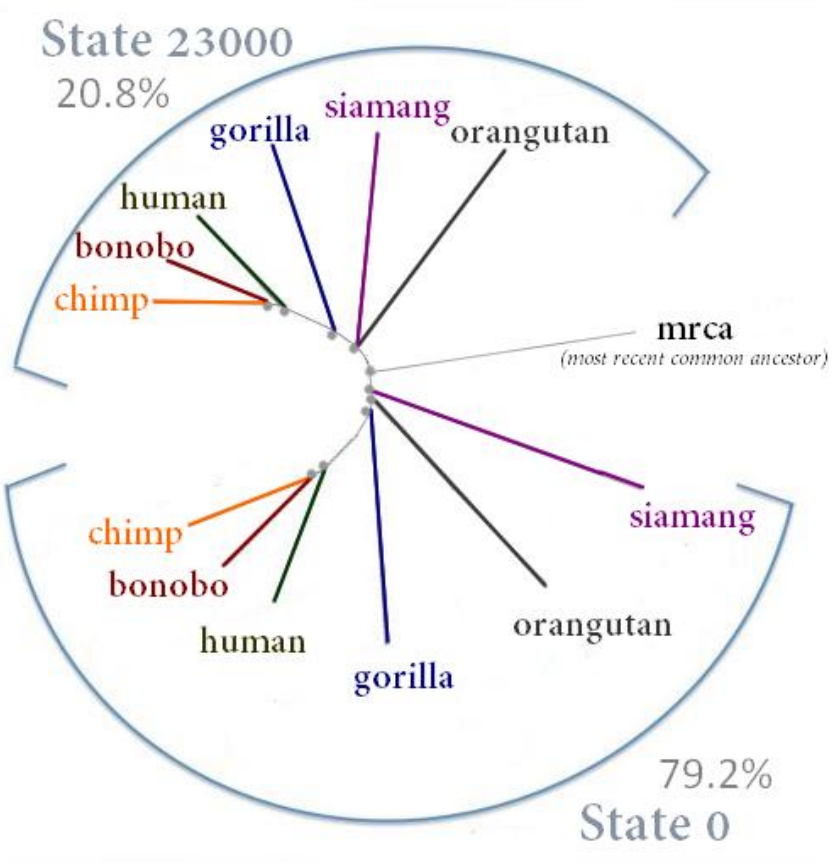
To decide what the most probable absolute branch order is, the mean values could be examined. However, as on the graph above can be seen, some graphs can cross each other multiple times. Thus it could also be interesting to look at the total area enclosed by these cross-overs, and compare that to the total area enclosed by the graphs and their mean lines. (15. figure)

Such data could be used for evaluation of the method the phylogenies were built; the residual errors' clear trend (14. figure) could indicate bias or other problems concerning the phylogeny-building method.

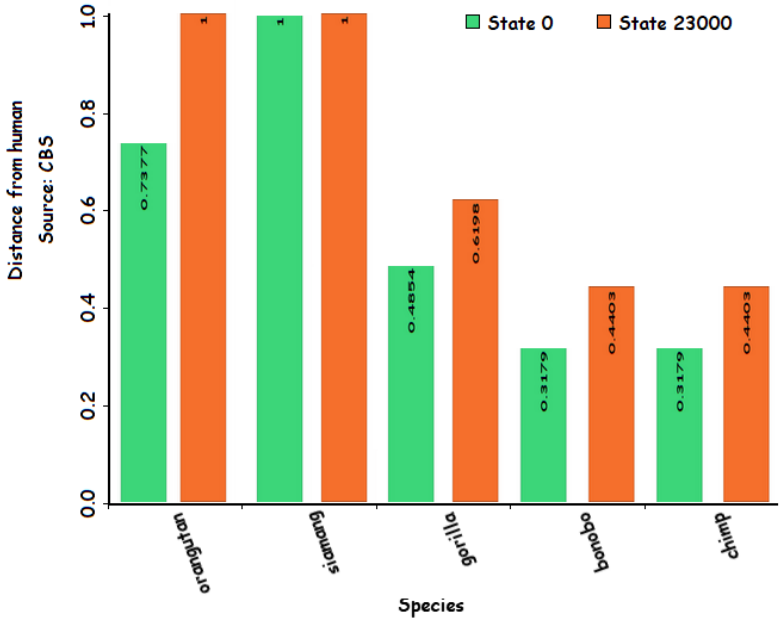
## Screencast

To introduce the code that was used for implementation of the above designs, a screencast was recorded. The demonstration is available at the following link: https://youtu.be/OUr611BY4Zc

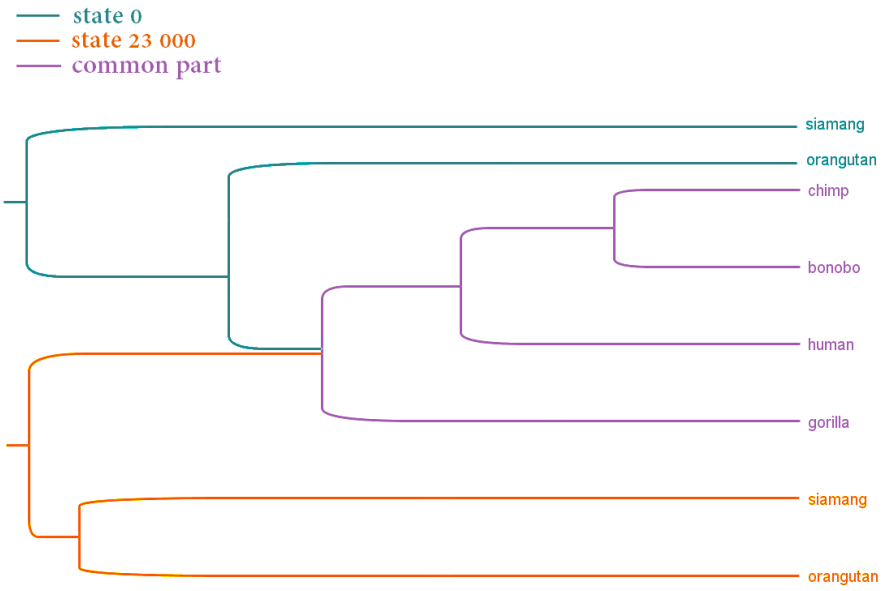
# Appendix



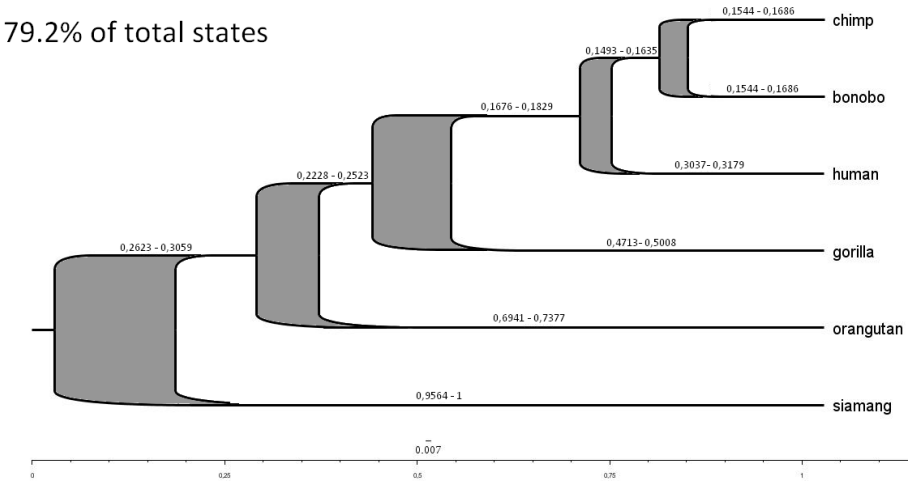

*5. figure*

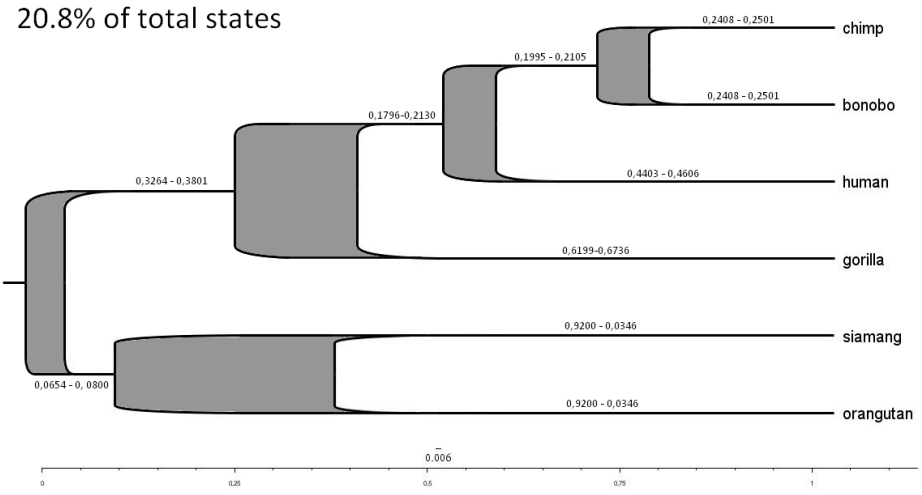



*6. figure*

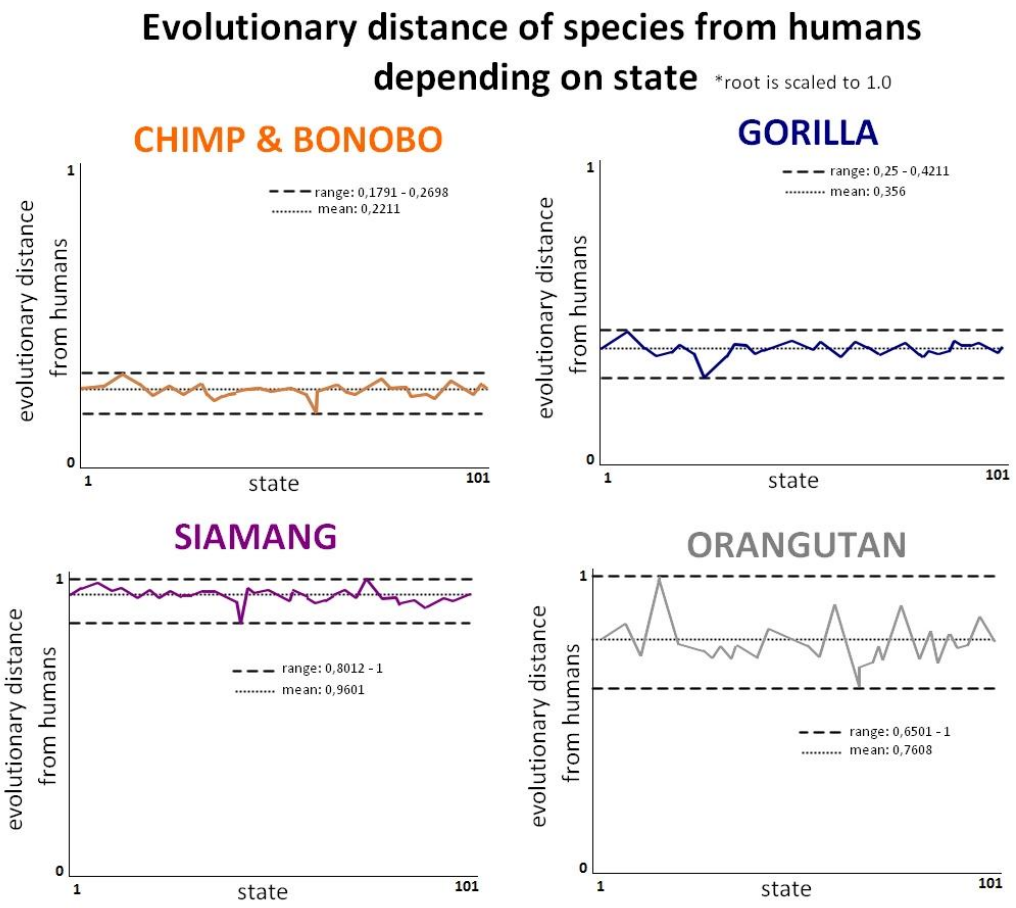



*7. figure*

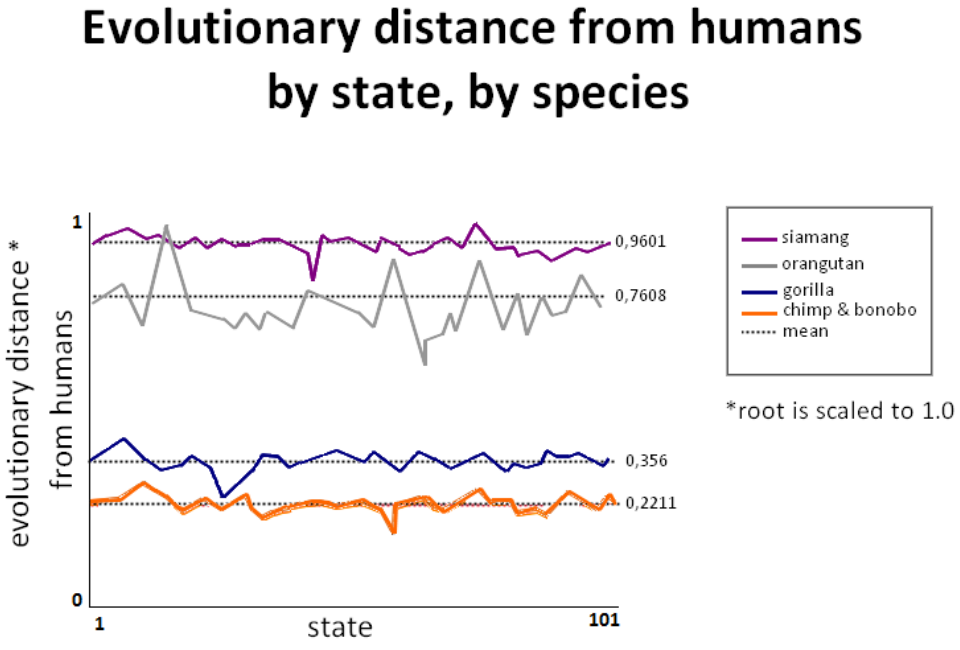


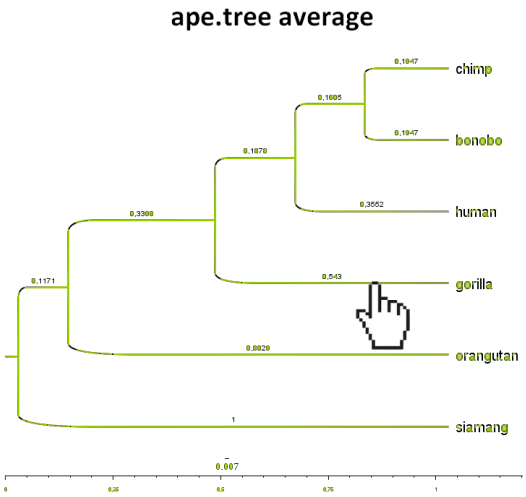

*8. figure*



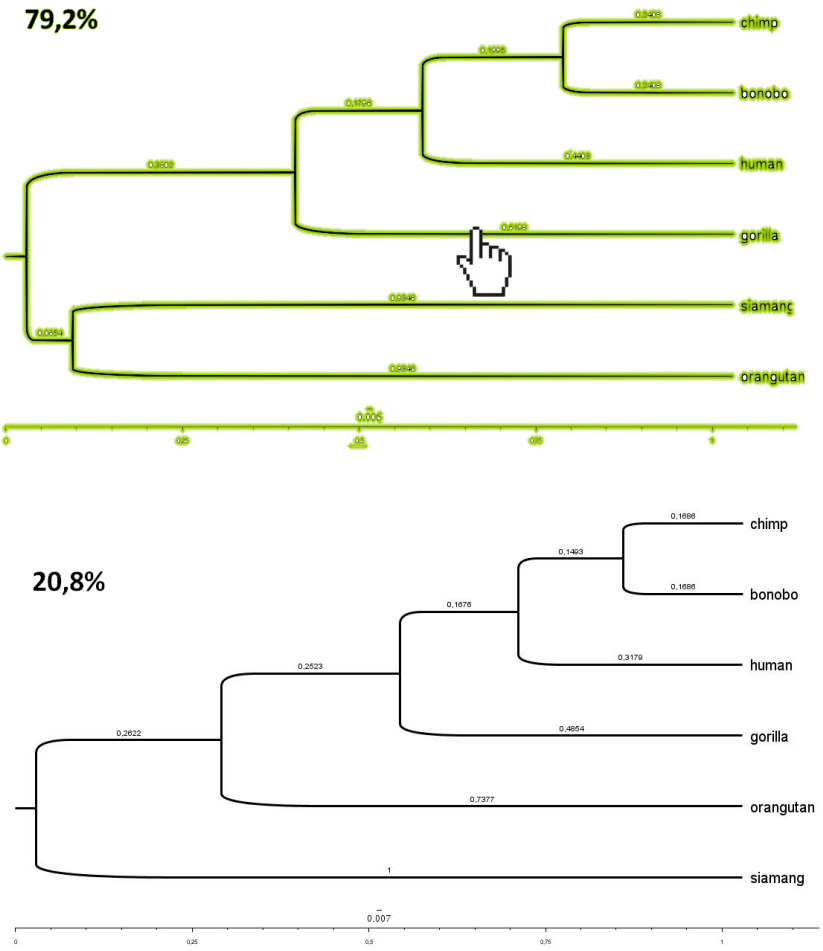

*9. figure*



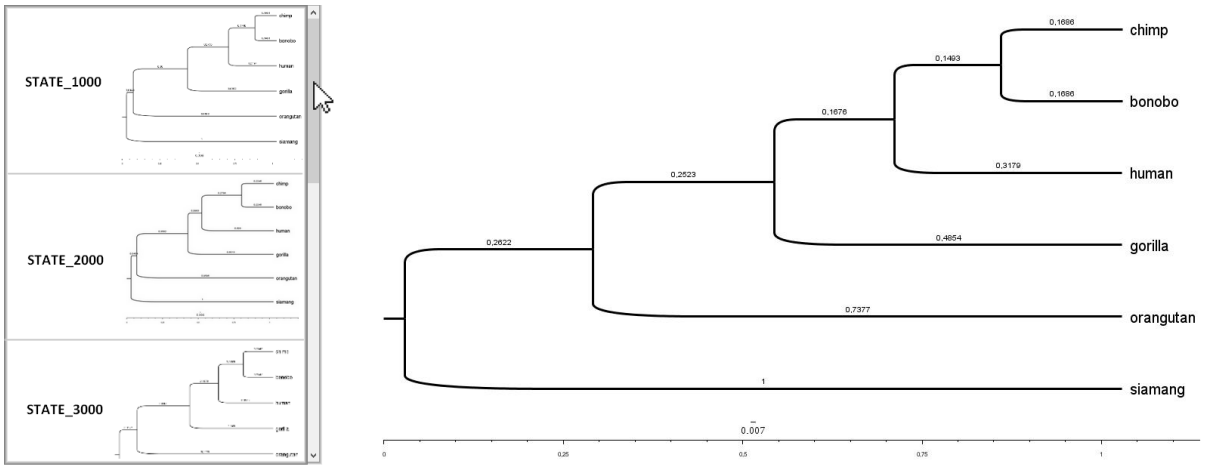

*10. figure*



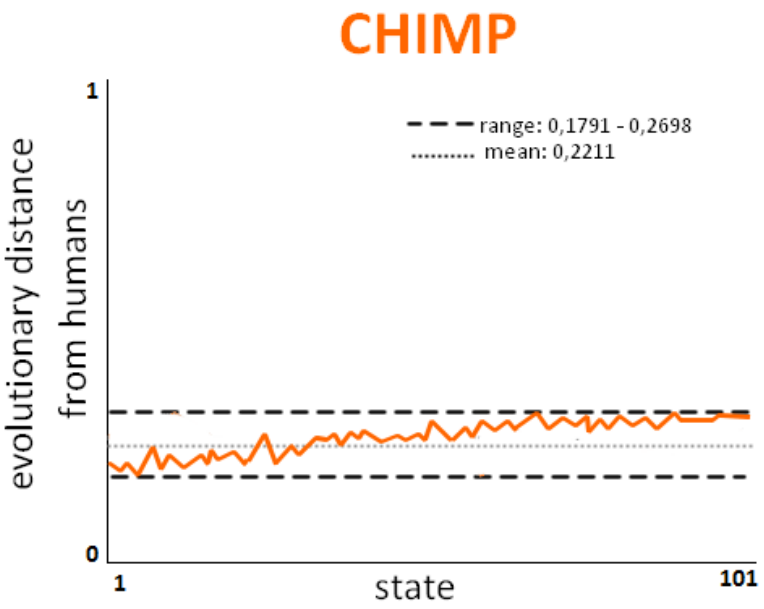

*11. figure*



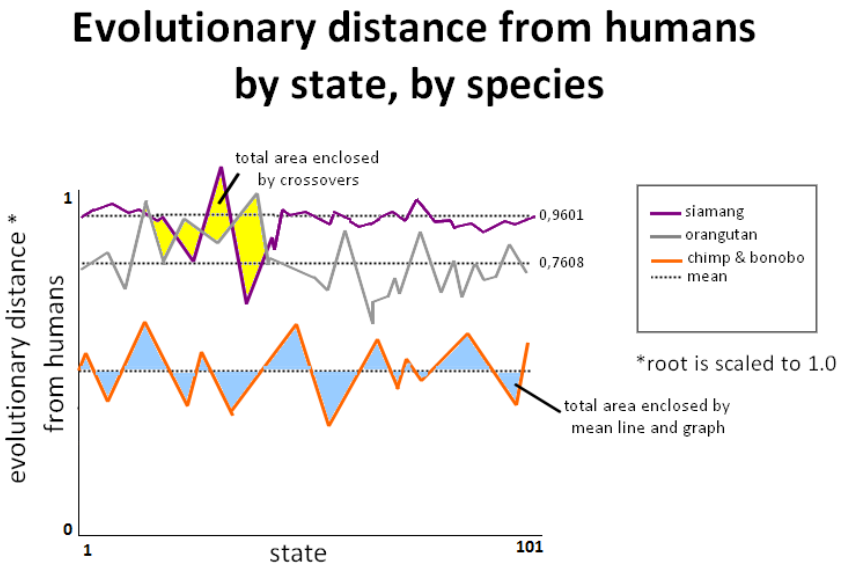

*12. figure*


STATE_1000
STATE_2000
STATE_3000
0,1686
chimp
0,1493
0,1686
bonobo
0,1678
0,3179
human
0,2523
0,4864
gorilla
0,2622
0,7377
orangutan
1
siamang
0,007
0
0,25
0,5
0,75
1

*13. figure*




*14. figure*




*15. figure*