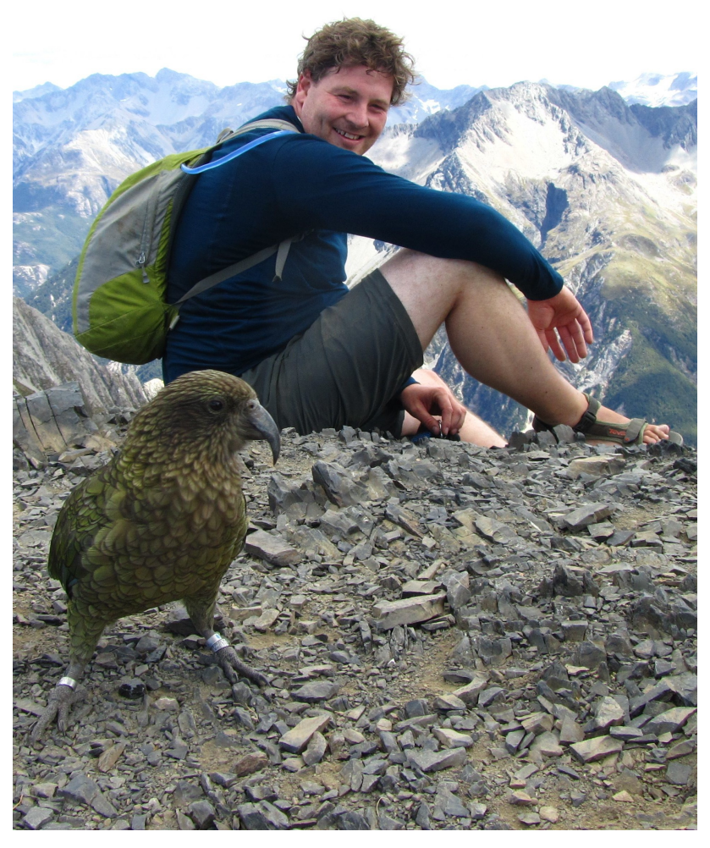
# Does your phylogenetic tree fit your molecular data?

THE UNIVERSITY OF AUCKLAND
FACULTY OF SCIENCE
Department of Mathematics

**Steffen Klaere**
s.klaere@auckland.ac.nz

**Jessica W. Leigh**
jleigh@maths.otago.ac.nz



## Motivation

In phylogenetic inference one looks among a set of models for the model(s) under which the given data are most likely to have occurred. The set of models is being constantly extended. Many methods to infer the most suitable model prior to tree inference have been proposed, and numerous methods dealing with testing the variability of inferred topologies have been introduced (see [1] for an excellent review on the area). However, it remains to find a good test for the event that the fit between model and data is absolutely poor. Or as Gatesy [2] put it: *Given the simplicity of most models, it is possible that model selection in modern systematics is analogous to an overweight man shopping in the petites department of a women's clothing store. A particular garment might fit the portly man best, but this does not imply a good overall fit.*

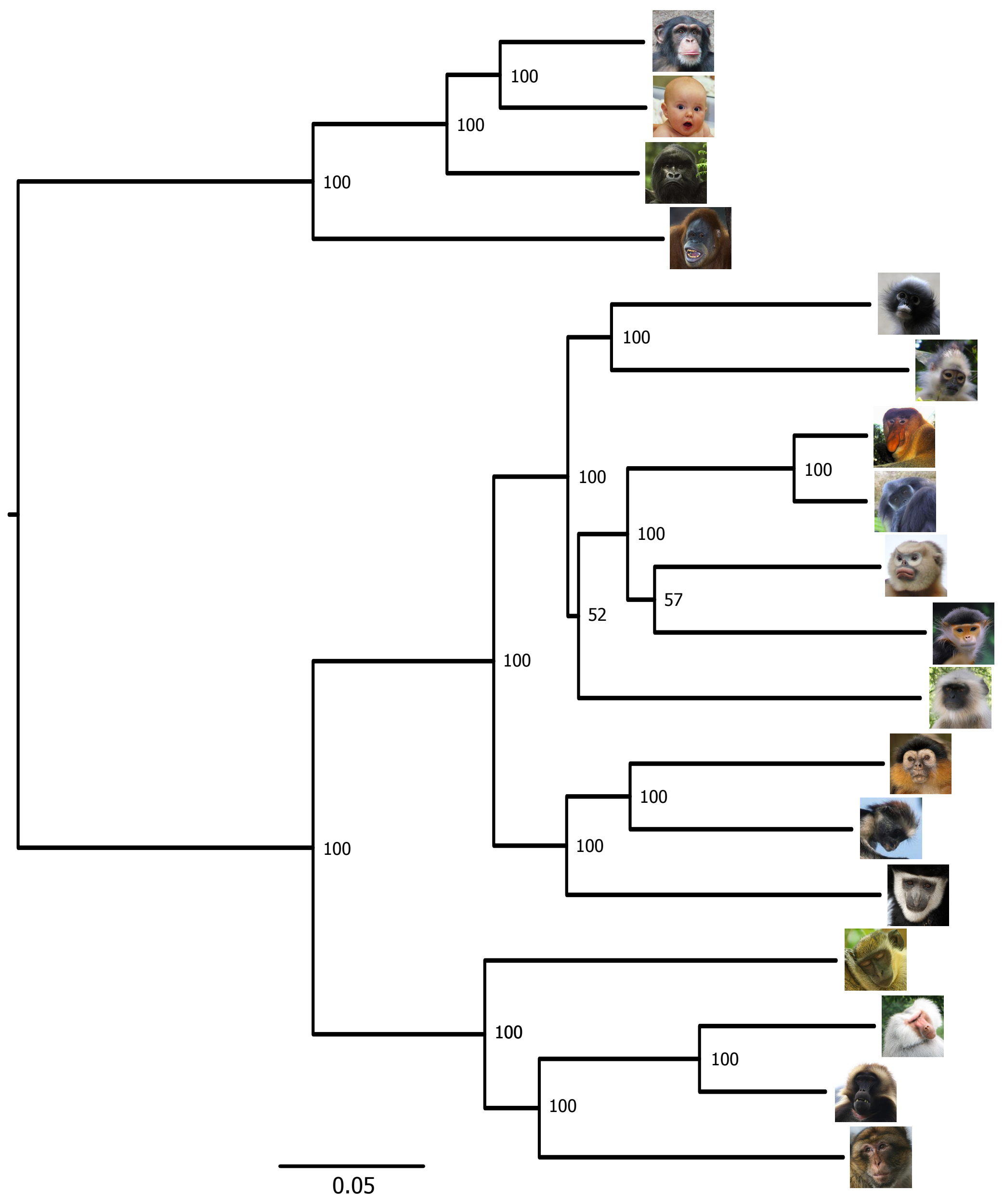
One way of investigating the fitness between model and data is to estimate the number of sites in an alignment which are less compatible with the inferred model. A few outlier detection methods have recently been proposed, using statistics like influence functions or noise reduction.

Here, we will graphically explore the data-to-model fitness by looking closer at the outlier identification step of Misfits [3].

## Data manipulation

We inferred a phylogenetic tree from the mtDNA genome alignment of 18 primate species (hominoids, cercopithecines, colobines). The alignment has been taken from [4]; we added the associated sequences for a Gorilla using Muscle to graft the sequence to the alignment. Gappy regions have been removed using GBlocks.

The tree is inferred under a GTR+Γ model in RAxML with 100 bootstrap samples.

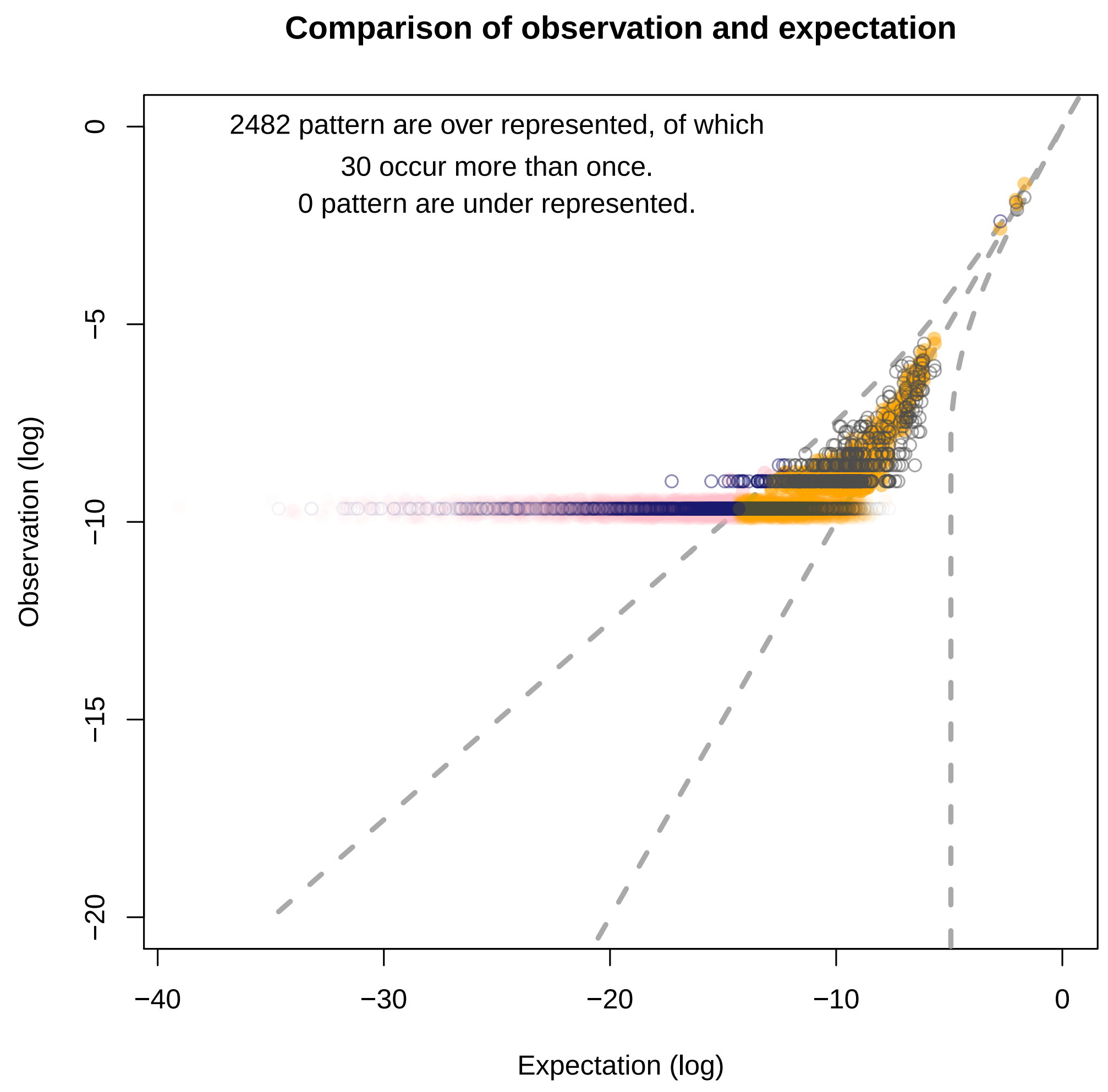


The alignment comprises 15,765 well-conserved sequence sites thus a well-chosen model should show an excellent fit to the data.

However, the authors of the study proclaimed to have identified hybridization in the clade of odd-nosed monkeys. So it also provides a little controversy.

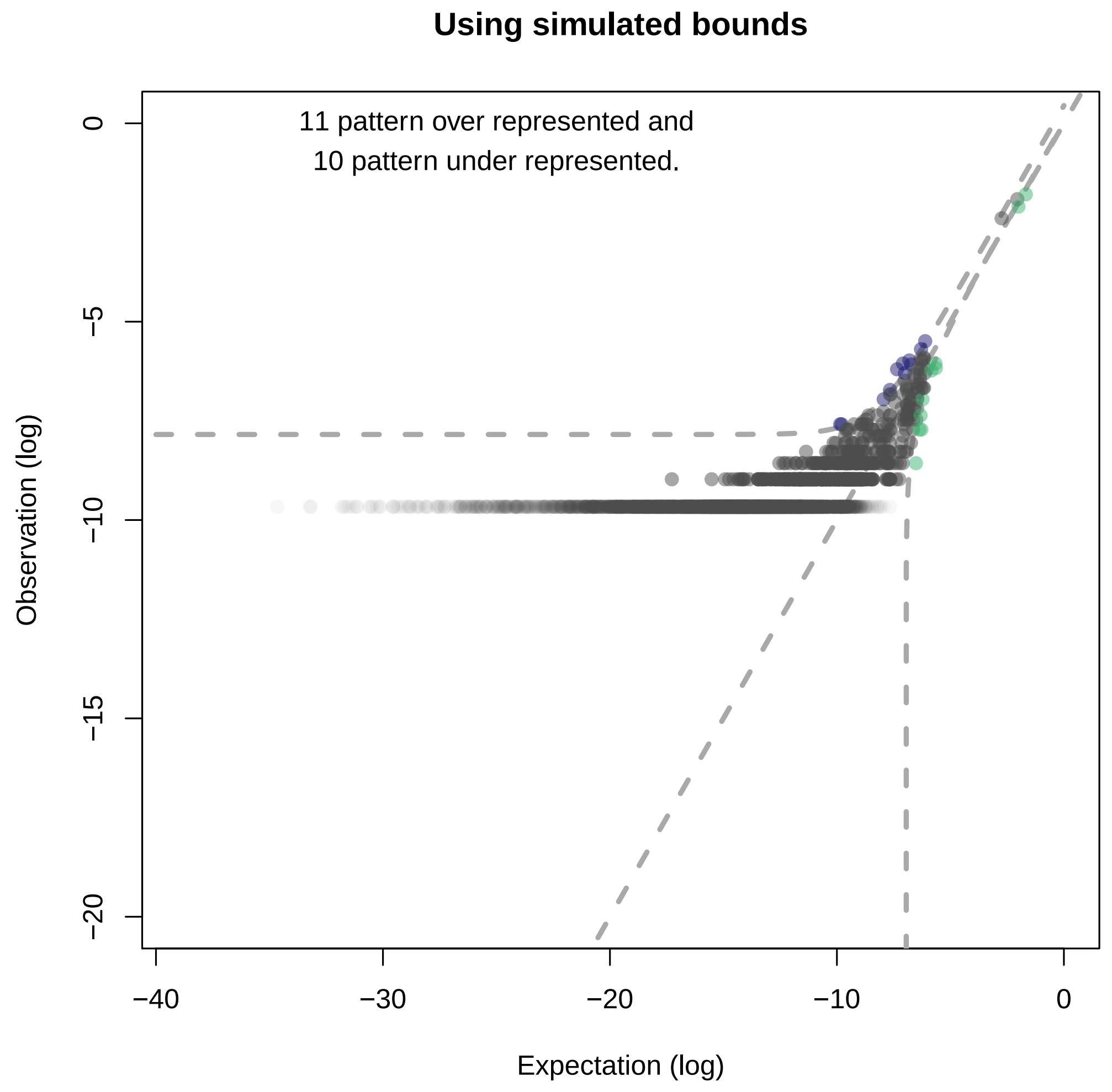
## Highlighting outliers using Misfits

We use Misfits [3] to visualize the relative fitness between data and model. The plots relate expected to observed site pattern frequencies on a log scale. Each dot indicates one of 5,087 distinct site patterns. The middle line indicates the case where observation equals expectation. The outer lines are upper and lower bounds computed using a simultaneous Gold confidence region. Blue dots indicate site patterns for which $f_s$ exceeds the upper bound of the confidence region. The plot indicates that about 50% of the site patterns are over represented in the Gold confidence region. These over represented site patterns account for about 25% of the alignment sites.



However, a simulated alignment (orange and pink dots in above plot) under the inferred model using P4 shows very similar behavior to the original data, thus suggesting that the bounds are too strict for the model.

To obtain more relaxed model bounds we generated 1,000 alignments under the inferred model, recorded for a range of possible site pattern frequencies the highest and the lowest loglikelihood score, and used a least squares approach to fit a confidence region to the data. Under this refined bound only 11 and 10 of the site patterns from the original alignment are over and under represented, respectively. These "extremal" site pattern comprise 2% and 29% of the alignment sites, respectively, although one does not need to replace all sites to improve the fit (dots are close to boundary).



The simulated bounds suggest a better fit between model and data. In particular, while the Gold confidence region selects its outliers primarily among site patterns of low occurrence, our simulated bound considers these a good fit. However, one can still improve on these bounds, e.g., by storing quantiles of the simulation runs instead of maximum and minimum.

## References

### References

[1] Kelchner SA, Thomas MA. Model use in phylogenetics: Nine key questions. *Trends Ecol Evol* 22(2):87–94, 2007.

[2] Gatesy J. A tenth crucial question regarding model use in phylogenetics. *Trends Ecol Evol* 22(10):509–10, 2007.

[3] Nguyen MAT, Klaere S, von Haeseler A. MISFITS: Evaluation the Goodness of Fit between a Phylogenetic Model and an Alignment. *Mol Biol Evol* 28(1): 143–52, 2011.

[4] Roos C, Zinner D, Kubatko LS, Schwarz C, Yang M, Meyer M, Nash SD, Xing J, Batzer MA, Brameier M,Leendertz FH, Ziegler T, Perwitasari-Farajallah D, Nadler T, Walter L, Osterholz M. Nuclear versus mitochondrial DNA: evidence for hybridization in colobine monkeys. *BMC Evol Biol* 11:77, 2011.