AARHUS UNIVERSITY

Faculty of Natural Sciences

Department of Molecular Biology and Genetics

Martyna Stankevičiūtė

The relationship between evolutionary constraint and genetic diversity across primate genomes

Master's Thesis

Master's degree programme in Bioinformatics

Supervisor: **Juraj Bergman**

Tenure-track Assistant Professor,

Department of Molecular Biology and Genetics - BiRC,

Aarhus University

Section for Bioinformatics and Computational Biology - BiRC

Aarhus 2026

Contents

Acknowledgments

I want to express my gratitude to all people who help me along the during this study. Thank you to all the amazing people from BiRC at Aarhus Univeristy, the professors and fellow students. A huge thank you for people who maintain the local HPC cluster “GenomeDK” which i used for all the computational tasks. I want to thank my closest 3 friends who i met during my studies: Antonia, Laurids and Sebastian - for always reminding to take a break and boosting the mood. I am grateful for my family and Mykolas for always waiting for me to comeback home. Finally, my biggest thank you goes to the people behind the Primate Diversity Panel: for greatest help in the most important moments of the thesis to Bjarke Pedersen and for the best supervision and constant support to my supervisor Juraj Bergman.

Abstract

Genomic regions under strong purifying selection are expected to harbor less standing genetic variation. Testing this relationship systematically across the primate order requires both a high-resolution map of genomic constraint and diversity estimates spanning a phylogenetically broad sample of species.

This study combines a 49-species whole-genome alignment with heterozygosity estimates from 219 non-human primate species to ask whether evolutionary constraint, quantified at 100 kb resolution, predicts patterns of genetic diversity across the primate genome. PhastCons and PhyloP conservation scores were derived from the whole-genome alignment using the PHAST package and combined with per-window heterozygosity estimates, calculated as heterozygous sites per callable site. Heterozygosity was systematically reduced in more conserved genomic windows, a negative relationship observed consistently across all 219 species individually despite differences in baseline diversity. Linear mixed models confirmed this relationship even after accounting for species-specific diversity, which explained 27% of total variance independently of local constraint. The X chromosome showed lower diversity than autosomes beyond its elevated conservation.

In parallel, a phylogeny-aware DNA language model GPN-star was trained, providing a foundation for future inference of per-site conservation scores using a non-parametric approach.

Together, the results demonstrate that purifying selection leaves a consistent and recoverable imprint on primate genetic diversity across 100 kb genomic windows, detected within all 219 species regardless of their baseline genomic diversity.