# Retroviral diversity of laboratory and wild mice M. musculus domesticus

**Stefanie Hartmann[1], Jens Mayer[2], Camila Mazzoni[3], Alex D Greenwood[4]**
[1]University of Potsdam, Potsdam, Germany
[2]University of Saarland, Saarbrücken, Germany
[3]Berlin Center for Genomics in Biodiversity Research, Berlin, Germany
[4]Leibniz-Institute for Zoo and Wildlife Research, Berlin, Germany

## Motivation and Objectives

The sensitivity and cost-effectiveness of Next Generation Sequencing (NGS) technologies has transformed almost every biological discipline, allowing to address questions whose answers seemed out of reach just a few years ago. One such area is the inventory and analysis of endogenous retroviruses of non-human origin. PCR-amplification and low-throughput Sanger-sequencing generally detects only retroviral variants that predominate in a given species or population, and rare sequence variants cannot easily be detected. NGS technologies, in contrast, allow to survey diversity and distribution of retroviruses among individuals, populations and species. Our poster will describe an analysis of targeted murine retroviral sequences from five mouse samples.

## Methods

Regions of approximately 400 bp that are conserved in most MLVs and in all known XMRV, PreXMRV-1, and PreXMRV-2 sequences were amplified. Amplicon clones were sequenced using the GS FLX Titanium platform, generating approximately 162,000 reads; thousands for each retroviral region. Although efficient algorithms for computing large multiple sequence alignments and phylogenetic trees exist, the visualization and interpretation of such large trees is technically and conceptually difficult. Furthermore, the fine-scale resolution of relationships within and between clades that phylogenetic trees provide may often not even be necessary. Instead of studying the retroviral diversity using standard phylogenetic analyses, we employed a clustering approach: The sequence reads, together with reference sequences, were clustered using the Markov Clustering algorithm as implemented in Tribe-MCL (Enright, 2002). The resulting clusters were then used to inventory retroviral sequences in the mouse samples. For sequences of selected clusters we also computed Maximum Likelihood phylogenies using RAxML (Stamatakis, 2006).

## Results and Discussion

For questions that are generally answered in the context of a phylogeny, massive amounts of sequence data often are a curse as much as a blessing. We will discuss clustering as an alternative to phylogenetic inference for the analysis and visualization of NGS technology-generated sequence data. Specifically, we characterized and will describe the distribution of mouse gamma retroviruses Xmv (xenotropic), Pmv (polytropic), and Mpmv (modified polytropic) in the three inbred laboratory mouse strains and two wild-caught mice. We also examined the distribution of Xenotropic Murine Leukemia Virus-related virus (XMRV), which is a laboratory recombinant of the two precursors PreXMRV-1 and PreXMRV-2 that have been reported to have very different distributions in mice. In addition, phylogenetic trees were computed for clusters containing XMRV and/or PreXMRV sequences.

## References

Enright AJ, Van Dongen S, Ouzounis CA. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**(7), 1575-1584.

Stamatakis A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21), 2688-90.