Insights on the origin, fate, and functional consequences of *de novo* ERV mutations in the Collaborative Cross

Anwica Kashfeen[1,2], Paul A Cotney[2], James Xenakis[2],
Fernando Pardo Manuel de Villena[2,3], Leonard McMillan[1]

Department of Computer Science[1], Department of Genetics[2], Lineberger Comprehensive Cancer Center[3],
University of North Carolina at Chapel Hill, NC, USA

Endogenous Retroviruses (ERVs) represent a large fraction of mammalian genomes. In mouse, ERVs are the most active of Class I retroelements. We have recently characterized the genomic location, segregation status, and subtypes of thousands of ERVs in the Collaborative Cross (CC) population. The significance of this work derives in part from the fact that the CC is a popular platform for research and because its structure allows for genetic mapping, to make strong inferences about the origin of each *de novo* ERV, characterize the functional consequences of segregating ERVs, and study the fate of *de novo* mutations. As part of this effort, we identified 366 *de novo* ERVs in 92 mice from 78 CC strains. This sample set is particularly useful to estimate mutation rates for different ERV types, study the functional effect and evolutionary significance of recent mutations, and potentially identify active ERVs in the germline of the CC. *De novo* ERVs differ from ERVs that are fixed and segregating in the CC, in several key aspects. For example, when ERVs are classified into types based on the LTR sequence, the frequency of *de novo* ERV types is significantly different from those that are fixed and segregating. Specifically, ERVB7_1-LTR_MM and RLTR4_MM account for nearly half of all *de novos* (182 out of 366) and are highly overrepresented in this class (23X and 6X, respectively). Furthermore, *de novo* ERVs are significantly overrepresented ($P<0.001$) in exons and introns compared to older ERVs. This observation confirms that many *de novo* ERVs are deleterious and that insertions in exons and introns are subject to purifying selection. Finally, the number of *de novo* ERVs varies in a strain-specific manner among CC strains. This indicates that there is some type of genetic control for the mutation rate. We have characterized the fate of *de novo* ERVs using both a deep pedigree in CC027/GeniUnc and a set of independent whole genome sequence datasets from 32 CC strains. This latter set of Most-Common Recent Ancestors (MRCAs) is particularly useful because it is derived from more recent CC samples (2021 versus 2016), and each is a pool of DNAs from the entire set of breeders for that strain used at that date. Using the MRCAs, we are able to ascertain the fates of all 180 *de novos* discovered in 32 CC strains with MRCA datasets. We conclude that 90 ERVs were lost, 35 were still segregating, and 55 were fixed. The observed fixation rate (31%) is artificially high due to the small effective population sizes of the CC and the fact that CC strains underwent a severe bottleneck just prior to 2021. We used the 55 fixed *de novo* ERVs to explore whether the 26 located in introns have an effect on expression of the corresponding gene and found that two ERVs (8%) lead to a highly significant reduction in expression of *Prkca1* in CC004/TauUnc and *Gmpr* in CC061/GeniUnc. In the case of *Prkca1,* expression is completely abolished making CC004/TauUnc a null for this gene. In a complementary analysis we used the *de novo* ERVs to determine the founder that contributed active ERVs, define, count and estimate the activity of subtypes, and identify the minimal set of potentially active ERV progenitors (fixed and segregating ERVs identified in the CC). We conclude that every CC founder contributed a set of active ERVs and no founder contributed all (i.e., many *de novo* ERVs are absent from the GRCm38 mouse reference genome). There are dozens of active subtypes and their activities vary significantly based on the number of *de novo* explained. Ancestry analysis allows us to identify and drastically reduce the set of most likely active ERVs. These active ERVs, causing *de novo* insertions are an important source of private mutations in the CC. Some of these private ERVs are causally responsible for the emergence of phenotypic outliers among CC strains.