Initial effort in generating a rat pangenome

*Flavia Villani[1], Andrea Guarracino[2], Jun Huang[3], David Ashbrook[1], Robert W. Williams[1], Vincenza Colonna[1,4], Hao Chen[3], and Erik Garrison[1]

*fvillan1@uthsc.edu

[1] *Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, TN, United States*
[2] *Genomics Research Centre, Human Technopole, Viale Rita Levi-Montalcini 1, Milan, 20157, Italy*
[3] *Department of Pharmacology, Addiction Science and Toxicology, University of Tennessee Health Science Center, Memphis, TN, United States.*
[4] *Institute of Genetics and Biophysics "Adriano Buzzati-Traverso", National Research Council, Naples, Italy.*

A pangenome can contain the full genomic information of a species. Pangenome graphs provide a compact representation of the mutual alignment of collections of genomes. In these graphs, nodes represent sequences in the pangenome, and paths describe genomes as walks through the graph. The members of the HXB/BXH family of recombinant inbred strains of rats have been used in many genetic mapping studies of physiological and behavior traits. It is part of the hybrid rat diversity panel, which is used in several ongoing large-scale genetic mapping studies on substance abuse related traits and phenome-wide association analysis. Current genomic studies using this family assume a single linear reference genome, making it difficult to observe sequences diverging from the reference, therefore limiting the accuracy and completeness of analyses. We sequenced all 30 members of the HXB family using linked-read libraries and built a pangenome graph with the PanGenome Graph Builder (PGGB) (*Garrison et al., 2021*) to study genetic variation. The pangenome enhanced the discovery of complex variants not seen by traditional genomics methods and provided calls with good precision and sensitivity. The ratio of the number of transitions to the number of transversions (Ts/Tv) in the pangenomic calls is 2.2, which is slightly higher than figures from LongRanger (2.1), this potentially reflects a slight enrichment of complex variants in the pangenomic set. In masked regions *(Gonzalez et al., 2021)* (i.e. regions that do not contain SINE, ALUs, LINE, LTR, and other simplex and complex DNA repeats), precision and sensitivity of vg (Hickey et al., 2020) calls are 90% and 84%, respectively, and for SNPs these figures rise to 94% and 80%. Overall we were able to reproduce data on known genetic variants and capture novel variations, which are being validated using Sanger sequencing. In summary, we demonstrate that pangenomes can be accurately built from linked-reads and that the pangenome produced by short linked reads can be informative.