Exploring residual heterozygosity in inbred rat strains: How much, where, and why?

Farnaz Salehi[1], Andrea Guarracino[1], Denghui Chen[3], Flavia Villani[1], David G. Ashbrook[1], Vincenza Colonna[1], Abraham Palmer[3], Robert W. Williams[1], Hao Chen[2], Erik Garrison[1]

[1]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA. [2] Department of Pharmacology, Addiction Science, and Toxicology, University of Tennessee Health Science Center, Memphis, TN, USA. [3] Department of Psychiatry, University of California San Diego, San Diego, CA, USA

Inbred strains are expected to achieve 98.7% homozygosity after 20 filial (F) generations of matings and 99.98% after 40, yet long-read genome assemblies of inbred rats even after F100 reveal many heterozygous regions, raising concerns about genetic stability, breeding errors, and possible assembly and alignment challenges. We are interested in resolving sources of residual heterozygosity and possible mechanisms of its maintenance. We are now generating PacBio HiFi sequences at coverages from 40–52X for strains belonging to three groups of male samples: 1. common inbred strains (*n* = 10, mean F>100); 2. the HXB/BXH RI family (*n* = 4, average F98), 2. the FXLE/LEXF RI family (*n* = 1, F27). Our initial HiFi assemblies contain primary and alternate contigs totalling ~3.27 Gbp per strain. We hypothesize that the additional 0.5 Gbp per assembly is contributed by alternate haplotypes and regions of residual heterozygosity. To investigate this, we applied a multi-assembly pangenome variant calling approach to extract heterozygous small variants from the alignments of the alternate haplotypes against the primary and reference genome. We combined primary and alternate assemblies with GRCr8 to create strain-specific pangenome graphs with the PanGenome Graph Builder. Heterozygosity was defined from graph-based variant calls generated with the *variation graph toolkit*. On average we detect 2.3 million heterozygous SNPs per assembly, equivalent to ~0.07%. Values range from 0.06% in BN—the strain used for GRCr8 reference—to 0.11% in WKY/NCrl. The F27 sample of FXLE12 was expected to have 0.13% residual heterozygosity and we observed 0.08% with a heterozygous locus on Chr 9. The strain with highest heterozygosity —WKY/NCrl—appears to have been contaminated by admixture affecting almost all chromosomes. Regions within 5 Mb of the centromere have unusually high levels of heterozygosity on Chrs 2, 10, 11, 12, and X. For example, Chr 12 is heterozygous in many strains and includes *Brca2* and two large 5s rRNA clusters. Heterozygosity is intermediate on Chrs 1, 3, 4, 5, 7, and 8, and very low on the shorter chromosomes—Chrs 9, and 13 through 19. We are now resolving to what extent heterozygosity in these regions is driven by selective pressure, *de novo* mutations, rapid mutation in satellite repeats, or residual assembly, alignment, and variant calling problems. Once technical confounders have been eliminated we will be able to dissect potential causes underlying heterozygosity and the maintenance of genetic diversity in inbred strains.