Structural Variants in Heterogeneous Stock Rats

Denghui Chen[1*], Khai-Minh H. Nguyen[2], Milad Mortazavi[2], Daniel Munro[2, 3], James Guevara[2], Katarina Cohen[2], Melissa Gymrek[4, 5], Jonathan L. Sebat[2, 6, 7], Abraham A. Palmer[2,6]

[1] Bioinformatics and System Biology Program, University of California San Diego, La Jolla, CA, USA; [2] Department of Psychiatry, University of California San Diego, La Jolla, CA, USA;
[3] Center for Immunity and Immunotherapies, Seattle Children's Research Institute, Seattle, WA, USA; [4] 4 Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA;
[5] Department of Medicine, University of California San Diego, La Jolla, CA, USA; [6] Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA;
[7] Department of Cellular and Molecular Medicine and Pediatrics, University of California San Diego, La Jolla, CA, USA

NIDA Center of Excellence for *Genetics, Genomics, and Epigenetics of Substance Use Disorders in Outbred Rats*

UC San Diego PALMER LAB Behavioral Genetics of Mice, Rats and Men

Introduction

Determining the causal genes with an associated locus identified by Genome-wide association studies (**GWAS**) is challenging. The causal alleles may not be single nucleotide polymorphisms (**SNPs**), and SNPs may not always effectively tag the causal variants, such as structural variants (**SVs**). Structural variants are hard to detect due to their nature of large genome alterations. To address this challenge, we are developing a SV calling pipeline using Pacific Biosciences (**PacBio**) high fidelity (**HiFi**) sequencing. Our goal is to characterize the SV landscape of heterogeneous stock rats to extend GWAS studies to include SVs in the investigation of genetic basis behind addiction-related behaviors.

Animals

The N/NIH heterogeneous stock (**HS**) outbred rats were derived in 1984 by intercrossing 8 inbred rat strains and have been maintained as an outbred population for ~100 generations. To capture both common variants, inherited from the founders, and novel variants, new mutations accumulated during the outbreeding process, the animals used in this study include the 8 inbred HS founders and 87 outbred HS rats from generation 73 to 80.

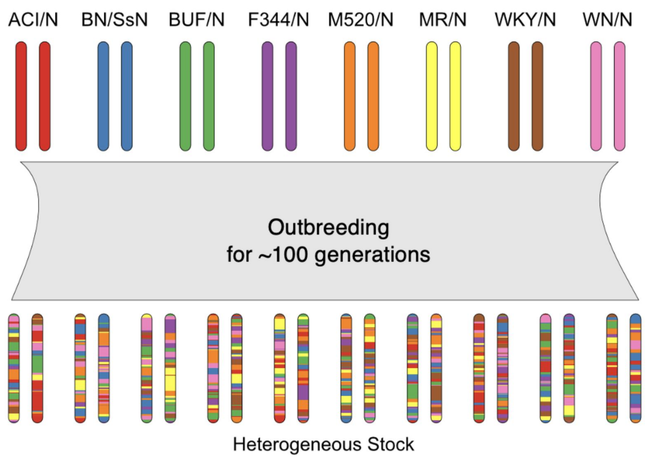

***Figure 1**: HS rats intercross strategy*

Sequencing and SV calling

We sequenced 8 inbred HS founders to 41.1x mean PacBio HiFi coverage and 87 outbred HS rats to 10.5x (**Fig. 3**). The sequences were passed into the SV calling pipeline shown in **Figure 2** to discover SVs. After filtering out SVs with < 2 supporting reads and SVs with ≥ 50% reciprocal overlap with known tandem repeats (**TR**) region on the rat genome, 110,987 high-quality SVs were retained (**Table 1**). The SV joint-calling pipeline were able to discover majority of the SVs in the population at ~10x coverage PacBio HiFi reads (**Fig. 3**).

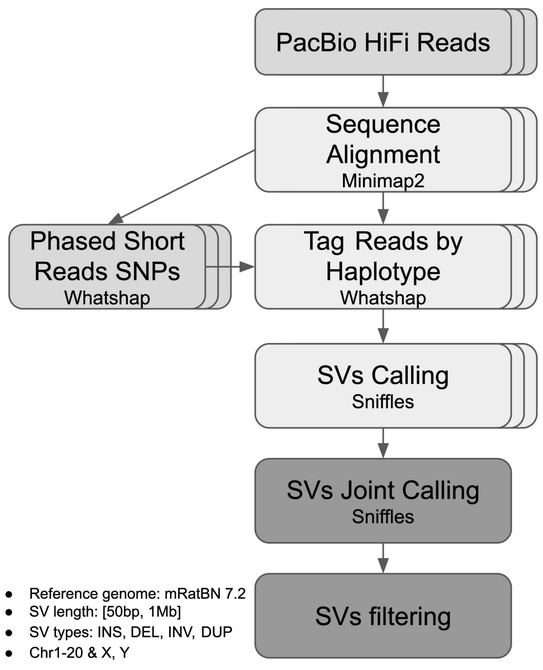

***Figure 2**: SV calling pipeline flowchart*

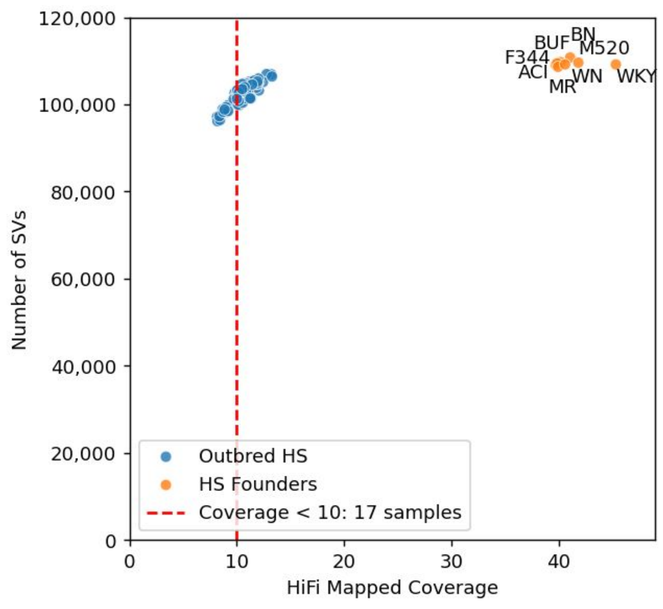

***Figure 3**: Number of SVs vs. HiFi coverage*

SV filtering

SV type	Raw SV count	Supporting reads ≥ 2	< 50% reciprocal overlap with TR	All filters
Total	136,876	136,834	111,022	110,987
INS	69,823	69,805	56,483	56,468
DEL	65,044	65,023	52,584	52,567
INV	763	760	763	769
DUP	1,246	1,246	1,192	1,192

***Table 1**: Number of SVs under different SV types and filters*

Acknowledgements

Funding: 1U01DA051234, P50DA037844; Sequencing: UCSD IGM, UCSD SCRM; Computing resources: SDSC TSCC

SV length distribution

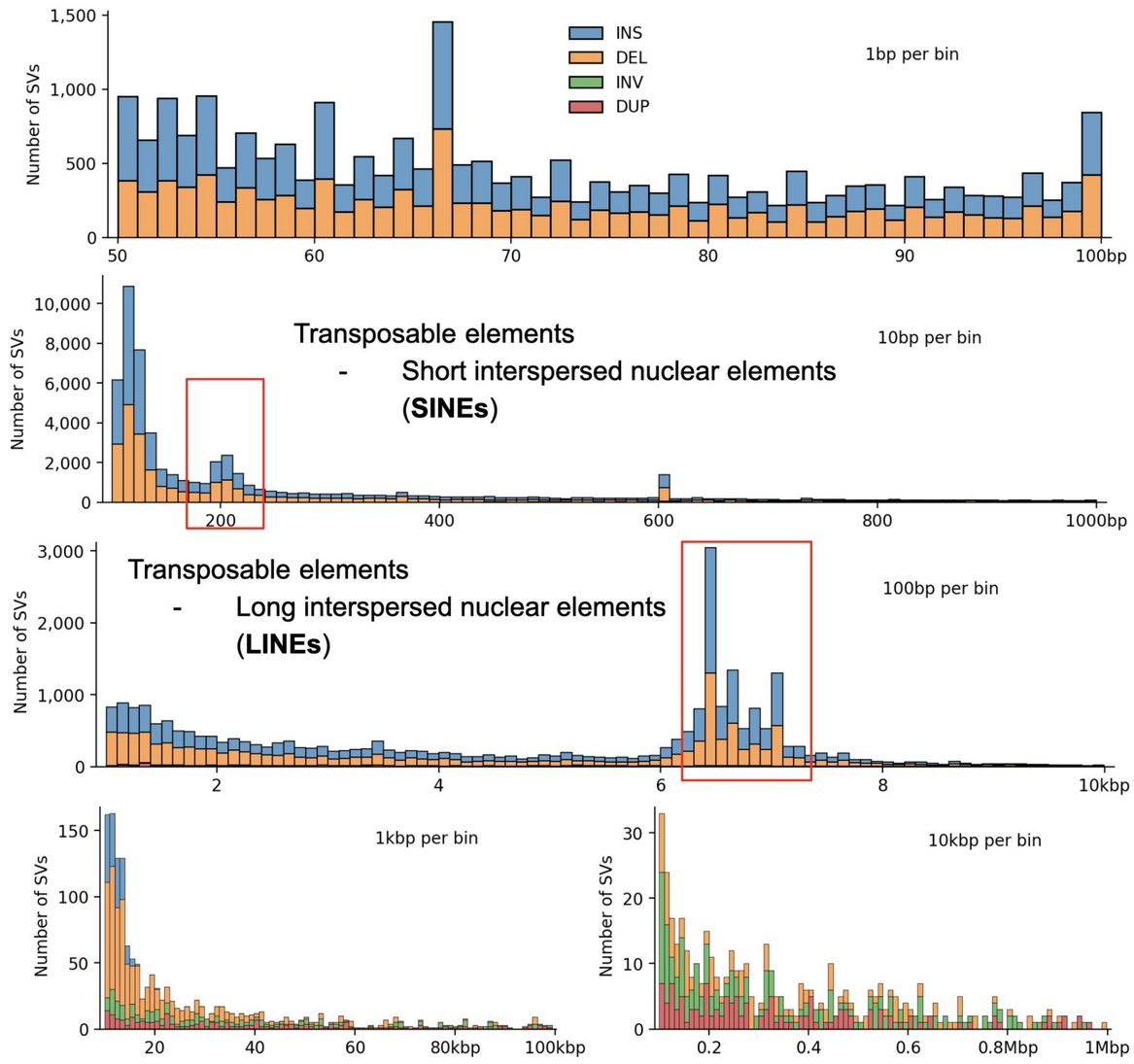

***Figure 4**: SV length distribution*

SV evaluation

To estimate the accuracy of the obtained SVs, we evaluated each common SV region on autosome by comparing founder haplotypes tagged using SVs and using existing SNPs data.

A buffer zone was added to both sides of an SV region when computing founder haplotypes from SNPs data. As the buffer zone size increases, more SV regions are tagged using SNPs data and included in the founder haplotype comparison. The founder haplotype concordance between tagged using SVs and SNPs peaks at 88.46% concordance for 75,116 SV regions with an 800 bp buffer on each side (**Fig 5**).

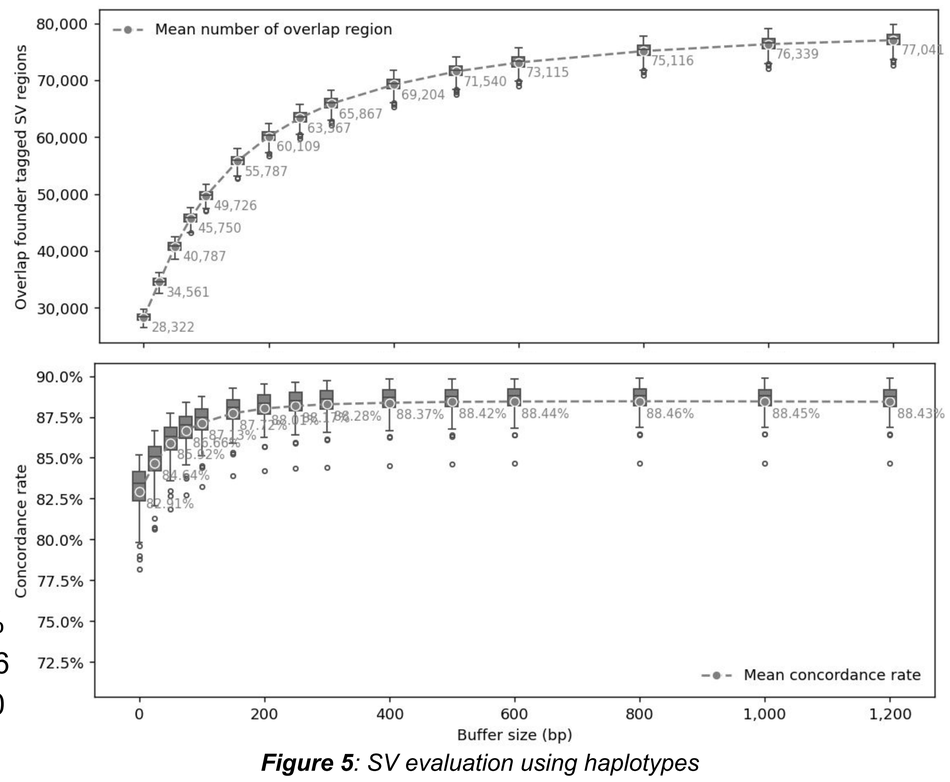

***Figure 5**: SV evaluation using haplotypes*

Summary

We developed a SV joint calling pipeline based on PacBio HiFi reads and discovered 110,987 structural variants in HS rats population. When comparing founder haplotypes tagged by SVs to by SNPs, the SV callset achieved 88.46% concordance.

Discussion

By imputing the discovered SVs into the outbred HS rats population, we will be able to identify associations between SVs and a variety of addiction-related behavioral traits in more than 20,000 outbred HS rats that have been extensively genotyped and phenotyped with the goal of revealing novel genetic mechanisms underlying drug abuse and other human diseases.