Submitter Name: Milad Mortazavi
Submitter email: smmortazavi@health.ucsd.edu
PI Name (if different): Abraham A. Palmer
PI email (if different): aapalmer@health.ucsd.edu

Polymorphic SNPs, short tandem repeats and structural variants cause differential gene expression among inbred C57BL/6 and C57BL/10 substrains

Milad Mortazavi[1], Yangsu Ren[1], Shubham Saini[2], Danny Antaki[1,3], Celine L St Pierre[4], April Williams[5], Abhishek Sohni[6], Miles Wilkinson[6,7], Melissa Gymrek[2,7], Jonathan Sebat[1,3,7], and Abraham A. Palmer[1,7]

[1]Department of Psychiatry, University of California San Diego; [2]Department of Computer Science and Engineering, University of California San Diego; [3]Department of Cellular and Molecular Medicine and Pediatrics, University of California San Diego; [4]Department of Genetics, Washington University; [5]Salk Institute for Biological Studies, [6]Department of Obstetrics, Gynecology and Reproductive Sciences, University of California San Diego; [7]Institute for Genomic Medicine, University of California San Diego

C57BL/6J is the most widely used inbred mouse strain and is the basis for the mouse reference genome. In addition to C57BL/6J, there are several closely related C57BL/6 and C57BL/10 substrains. Numerous phenotypic differences have been reported among these substrains, which are presumed to be due to the accumulation of new mutations. We performed whole genome sequencing and RNA-sequencing in 9 C57BL/6 and 5 C57BL/10 substrains. We identified 352,631 SNPs, 109,096 indels, 150,344 short tandem repeats (STRs), 3,425 structural variants (SVs) and 2,826 differentially expressed genes (DEGenes) among these 14 strains. 312,981 SNPs (89%) perfectly differentiated the B6 and B10 lineages and were clustered into 28 short segments that appear to be due to introgressed haplotypes rather than new mutations. However, these introgressed regions contained only 13% of the DEGenes. Outside of these introgressed regions, we identified numerous additional mutations including 53 SVs, protein-truncating SNPs and frameshifting indels that were strongly associated with DEGenes. The remaining DEGenes could not be definitively attributed to any specific variant. Crosses among substrains are called Reduced Complexity Crosses (RCCs), and represent a powerful mapping strategy that has been employed for a number of drug abuse relevant traits. Our results provide a catalog of variants and DEGenes that will greatly enhance the use of RCCs for forward genetics. Our results also identify numerous naturally occurring DEGenes that can be used for reverse genetic studies. More generally, our results illustrate how introgression and mutational processes give rise to differences among substrains.