Importance of polymorphic SNPs, short tandem repeats and structural variants for 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 St. Pierre[4], April Williams[5], Jonathan Sebat[1,3,6], Melissa Gymrek[2,6], and Abraham A. Palmer[1,6]

[1]Department of Psychiatry, University of California San Diego, La Jolla, CA

[2]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA

[3]Department of Cellular and Molecular Medicine and Pediatrics, University of California San Diego, La Jolla, CA

[4]Department of Genetics, Washington University School of Medicine, St. Louis MO

[5]Salk Institute for Biological Studies, La Jolla, CA

[6]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA

Corresponding Author:

Abraham A. Palmer (aap@ucsd.edu)

Keywords: mouse, animal models, inbred, C57BL, substrain

Abstract

C57BL/6J is the most widely used inbred mouse strain and is the basis for the mouse reference genome. In addition to C57BL/6J, several other C57BL/6 and C57BL/10 substrains exist. Previous studies have documented extensive phenotypic and genetic differences among these substrains, which are presumed to be due to the accumulation of new mutations. These differences can be used for genome wide association studies. They can also have unintended consequences for reproducibility when substrain differences are not properly accounted for. In this paper, we performed genomic sequencing and RNA-sequencing in the hippocampus of 9 C57BL/6 and 5 C57BL/10 substrains. We identified 985,329 SNPs, 150,344 Short Tandem Repeats (STR) and 896 Structural Variants (SV), out of which 330,178 SNPs and 14,367 STRs differentiated the C57BL/6 and C57BL/10 groups. We found several regions that contained dense polymorphisms. We also identified 578 differentially expressed genes for C57BL/6 substrains and 37 differentially expressed genes for C57BL/10 substrains (FDR < 0.01). We then identified nearby SNPs, STRs and SVs that matched the gene expression patterns. In so doing, we identified SVs in coding regions of *Wdfy1, Ide, Fgfbp3* and *Btaf1* that explain the expression patterns observed. We replicated several previously reported gene expression differences between substrains (*Nnt, Gabra2*) as well as many novel gene expression differences (e.g. *Kcnc2*). Our results illustrate the impact of new mutations on gene expression among these substrains and provides a resource for future mapping studies.

1. Introduction

Since Clarence C. Little generated the C57BL/6 inbred strain almost a century ago, the C57BL/6J has become the most commonly used inbred mouse strain. The C57BL/6 and C57BL/10 substrains were separated after they were already considered inbred (1, 2). The popularity of C57BL inbred mice led to the establishment of many substrains (defined as >20 generations of separation from the parent colony). Among the C57BL/6 branch, the two predominant lineages are based on C57BL/6J (from The Jackson Laboratory; **JAX**) and C57BL/6N (from the National Institutes of Health; **NIH** (3, 4)). Subsequently, several additional substrains were derived from the JAX and the NIH branches.

Spontaneous mutations are expected to accumulate in any isolated breeding population. Assuming they are selectively neutral, genetic drift dictates that some new mutations will to be lost, others will maintain an intermediate frequency, and others will become fixed, replacing the ancestral allele (5). The number of generations, effective population size, as well as other factors, influence the number and fate of new mutations. Mutations can be categorized into several classes: Single Nucleotide Polymorphisms (**SNPs**), Short Tandem Repeats (**STRs**), and Structural Variations (**SVs**). SNPs and SVs that differentiate C57BL/6J and C57BL/6N substrains have been previously reported (6-9), whereas to our knowledge, genome wide STR polymorphisms have never been identified among C57BL substrains. STRs are highly variable elements that play a pivotal role in multiple genetic diseases, population genetics applications, and forensic casework. STRs are among the most polymorphic variants in the human genome (10). STRs play a key role in more than 30 Mendelian disorders and recent evidence has underscored their profound regulatory role and potential involvement in complex traits (11). SVs include deletions, duplications, insertions, inversions, and translocations and are considerably less common than SNP and STRs, but can have greater functional consequences since they can alter gene expression via changes in regulatory factors or coding exons (12).

In addition to previous reports of genetic differences among C57BL/6J and C57BL/6N, numerous studies have reported phenotypic differences among various C57BL/6- and C57BL/10-derived substrains. For C57BL/6 substrains, these differences include learning behavior (13), prepulse inhibition (14), anxiety and depression (15), fear conditioning (16-18), glucose tolerance (19), alcohol-related (20, 21), and response to drugs (22-24). For

C57BL/10 substrains, these differences include seizure traits (25) and response to drugs (26).

In an effort to survey mutations that have arisen in various C57BL substrains, we performed whole genome sequencing in a single male individual from 9 C57BL/6 and 5 C57BL/10 substrains (~30x per substrain) and called SNPs, STRs and SVs. In addition, to identify functional consequences of these mutations, we performed RNA-sequencing of the hippocampal transcriptome in 7-11 male mice from each substrains, which allowed us to identify genes that were differentially expressed. This approach had two advantages: first it provided a large number of phenotypes that may be caused by substrain specific mutations. Second, we assumed that the gene expression differences would often be caused by a cis-eQTL, making it possible to narrow the number of potentially causal mutations without requiring the creation of intercrosses. Given that each differentially expressed gene had a characteristic Strain Distribution Pattern (**SDP**) for SPNs, STRs and SVs, we sought to identify features that were within 1 Mb of the differentially expressed gene and had the same SDP. For Structural Variations (SVs) we focused on SVs that overlapped the differentially expressed gene (exons, transcription start sites, UTRs) and matched the SDP. Thus, in addition to cataloging accumulated mutations among the substrains, we also established phenotypic (gene expression) consequences of these mutations. Some of these gene expression differences have been previously shown to cause differences in more complex traits, others are likely to have as-of-yet unappreciated effects on biomedically important traits.

2. Materials and Methods

2.1 Mice

We obtained a panel of 14 C57BL substrains from four vendors. The panel included 9 C57BL/6 and 5 C57BL/10 substrains: C57BL/6J, C57BL/6NJ, C57BL/6ByJ, C57BL/6NTac, C57BL/6JBomTac, B6N-TyrC/BrdCrlCrl, C57BL/6NCrl, C57BL/6NHsd, C57BL/6JEiJ, C57BL/10J, C57BL/10ScCr, C57BL/10ScSnJ, C57BL/10SnJ, C57BL/10ScNHsd (**Table 1**). All of the substrains were bred in house at the University of Chicago for one generation before tissue was collected for sequencing and RNA-sequencing; this minimized effects on gene expression due to environmental differences

among the four vendors. All procedures were approved by the University of Chicago IACUC.

2.2 *Whole-genome sequencing (WGS) and data processing*

DNA from one male animal per substrain (n=14) was extracted from spleens using a standard "salting-out" protocol. Sequencing libraries were prepared using a TruSeq DNA LT kit, as per the manufacturer's instructions. The DNA was sequenced at Beckman Coulter at an average depth of 5X coverage per sample using an Illumina HiSeq 4000 (paired-end 125bp). Subsequently, additional sequencing data was generated for the same libraries by Novogen at an average depth of 30X coverage on an Illumina HiSeq XTen (paired-end 150bp); for a total of ~35X coverage per sample (**Table 1**). For technical reasons only the Novogene reads are used in this paper.

Table 1. Summary of sequencing runs including number of reads and coverage. All of the mice were bred in house for one generation before tissue was collected for sequencing and RNA-sequencing.

			DNA Sequencing				RNA Sequencing	
			Beckman		Novogene		UCSD	
Strain	Vendor	Strain ID	Reads	Coverage	Reads	Coverage	Average Reads	# Samples
C57BL/6NTac	Taconic	B6	15,507,036,875	5.64	96,765,063,000	35.19	13,459,345	8
C57BL/6NJ	JAX	#005304	15,204,121,000	5.53	96,364,702,800	35.04	15,767,477	9
C57BL/6NHsd	Harlan	#044	11,082,436,750	4.03	92,297,327,532	33.56	14,656,480	9
C57BL/6NCrl	Charles River	#027	13,203,215,375	4.80	87,049,155,000	31.65	19,408,390	8
C57BL/6JEiJ	JAX	#000924	16,686,632,375	6.07	117,271,713,036	42.64	16,334,723	8
C57BL/6JBomTac	Taconic	B6JBom	10,119,288,500	3.68	88,797,126,300	32.29	17,340,684	7
C57BL/6J	JAX	#000664	14,953,129,750	5.44	94,211,981,700	34.26	18,555,108	7
C57BL/6ByJ	JAX	#001139	17,127,176,375	6.23	84,730,839,600	30.81	15,552,591	7
B6N/TyrC/BrdCrlCrl	Charles River	#493	11,001,132,750	4.00	90,882,123,900	33.05	15,139,469	8
C57BL/10SnJ	JAX	#000666	14,898,322,625	5.42	82,193,982,600	29.89	14,012,862	7
C57BL/10ScSnJ	JAX	#000476	12,804,319,875	4.66	82,380,679,200	29.96	18,947,235	8
C57BL/10ScNHsd	Harlan	#046	14,512,802,500	5.28	93,773,651,400	34.10	12,453,407	11
C57BL/10ScCr	JAX	#003752	15,014,884,375	5.46	83,468,156,700	30.35	16,384,792	8
C57BL/10J	JAX	#000665	15,422,674,250	5.61	89,515,708,500	32.55	17,550,456	7

2.3 SNPs

Reads were mapped to the mm10 reference genome using BWA-mem (v.0.7.12.; (27)). Subsequent processing was carried out with SAMtools v.1.2 (28), Genome Analysis Toolkit (GATK) v.3.3 (29), and Picard Tools v.1.129, which consisted of the following steps: sorting and merging of the BAM files, indel realignment, removal of duplicate reads, and recalibration of base quality scores for each individual, called

1,035,308 total SNPs. For subsequent analyses SNPs were further filtered using PLINK for high missing rate (--geno 0.1) yielding a total of 985,329 high quality SNPs (30). For the dendrogram tree we further LD-pruned (--indep-pairwise 50 5 0.5) the SNP panel, yielding a total of 376,824 SNPs.

*2.4 Short Tandem Repeat (**STR**)*

After read alignment described in the previous section, we used the BAM files to run the software HipSTR in order to call STRs (31). STRs for the 14 substrains were jointly genotyped on a single node local server in batches of 500 STRs. BED file containing STR regions to genotype for the mm10 assembly was obtained from the official HipSTR github repository. HipSTR version v0.6 was called individually per STR with default parameters. Resulting VCF files from each batch were merged to create a genome-wide callset in VCF format using bcftools concat for a total of 150,344 polymorphic STRs.

2.5 *Structural Variation (**SV**)*

We called SVs with LUMPY and CNVnator (32, 33). LUMPY is based on a general probabilistic representation of an SV breakpoint that allows any number of alignment signals to be integrated into a single discovery process. A breakpoint is defined as a pair of bases that are adjacent in an experimentally sequenced ‘sample’ genome but not in the reference genome. To account for the varying level of genomic resolution inherent to different types of alignment evidence, a breakpoint is represented with a pair of probability distributions spanning the predicted breakpoint regions. The probability distributions reflect the relative uncertainty that a given position in the reference genome represents one end of the breakpoint. Lumpy uses discordant paired ends and split reads to identify breakpoints for deletions, duplications, inversions, translocations, and complex SVs. SVs were called according to the default parameters of LUMPY (v.0.2.13) and SVtyper (v.0.7.1) to genotype variants. Since all of our samples are from inbred mice, we only considered homozygous calls even though recent mutations that have not yet become fixed might truly be observed as heterozygotes. In addition, we only included SVs in the autosomes and filtered out calls with length larger than 1Mb. This resulted in identification of 175 deletions, 184 duplications and 6 inversions across the C57BL/6 strains.

In addition to LUMPY we identified Copy Number Variations (CNVs) with CNVnator (v.0.4.1) (32) with bin size 100bp. CNVnator identifies CNVs based on read depth analysis. It combines the standard mean-shift approach with additional refinements such as multiple-bandwidth partitioning and GC correction. CNVnator can call deletions and duplications in regions where LUMPY typically fails, including regions with segmental duplications and repetitive regions. We only retained homozygous calls or calls in tandem segmental duplication regions, since a valid SV can occur in those regions by changing the number of repeats and be identified as a heterozygous call in the CNVnator output. CNVnator identified 236 deletions and 295 duplication. We then merged the LUMPY and CNVnator calls in a set of 896 unique SVs (**Supplementary material**).

2.6 RNA-sequencing and data processing

Total RNA was extracted from hippocampal samples from each of the 14 substrains using Trizol reagent (Invitrogen, Carlsbad, CA). RNA was treated with DNase (Invitrogen) and purified using RNeasy columns (Qiagen, Hilden, Germany). RNA-sequencing library prep and sequencing was performed by the UC San Diego Sequencing Core using Illumina TruSeq prep and Illumina HiSeq 4000 machine (single-end 50bp) (**Table 1**). Reads were mapped to mouse reference transcriptome (mm10) using the splice-aware alignment software HiSat2, and counts were normalized using HTSeq. We removed lowly expressed genes (CPM < 2) and two low quality samples, leaving us with gene expression data for 16,718 genes across 117 samples for the 14 substrains. To identify differentially expressed genes between C57BL/6 and C57BL/10 substrains we performed a two-factor t-test using the *t.test* function in R. To identify genes that were differentially expressed within C57BL/6 or within C57BL/10 substrains, we performed analysis of variance using the *anova* function in R separately for the C57BL/6 and C57BL/10 substrains. We further calculated the false-discovery rate (FDR) for each gene using the *p.adjust* function in R, which implements the Benjamini-Yekutieli procedure.

2.7 Construction of dendrograms

We generated dendrograms for each genomic element using SNPs, STRs and RNA-sequencing data (**Figure 1**). For the SNPs (**Figure 1b**), we first generated an identity-by-

state (**IBS**) matrix using the 376,824 SNPs in Plink (30), followed by plotting the dendrogram using the *hclust* function in R. The proximity of each substrain to one another represent genetic similarity. For the STRs (**Figure 1c**), we first generated a distance matrix for each pair-wise comparison between every sample across the 150,344 STRs using the *dist* function in R, then plotted the dendrogram using the *hclust* function. For the RNA-sequencing data, because we assumed that most of the genes would not be differentially expressed, we performed ANOVA for strains across all substrains and generated dendrograms for several FDR thresholds (FDR<0.1, FDR<0.01, and FDR<0.001) (**Figure 1d**).

2.8 Construction of chromosomal heatmaps

We generated chromosomal heatmaps for genomic features (SNPs, STRs and gene expression) using the *CMplot* package in R. The heatmaps showed the number of features binned in 1Mb windows that differ between C57BL/6 and C57BL/10 substrains (**Figure 2**), or that differ for at least one out of the nine C57BL/6 substrains (**Figure 3**), or that differ for at least one out of the five C57BL/10 substrains (**Figure 4**).

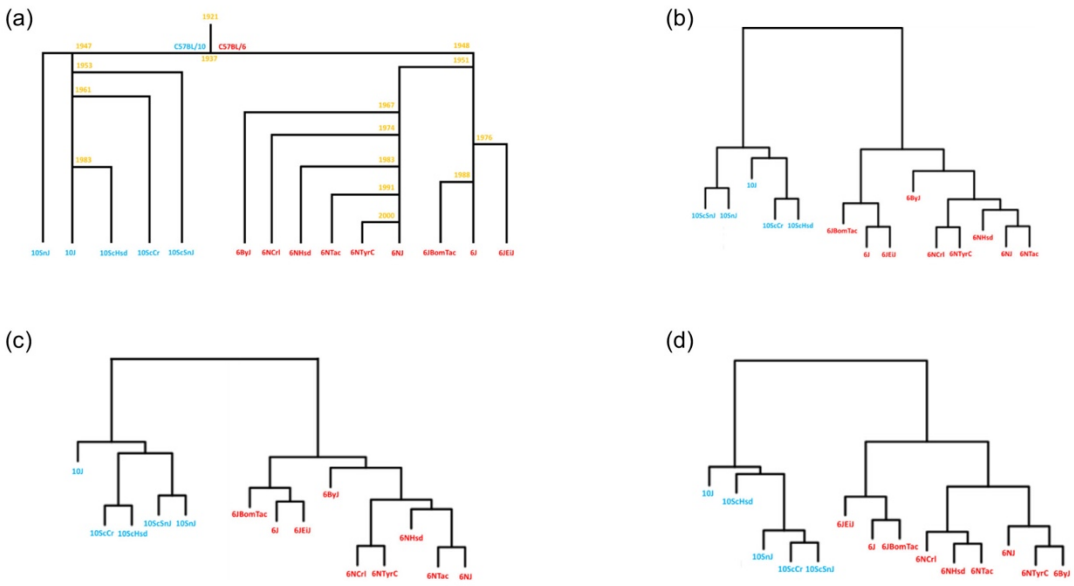


Figure 1. **Comparison of C57BL substrains chronological tree to dendrograms based on SNPs, STRs, and gene expression.** (a) Chronological tree of the derivation of C57BL/6 and C57BL/10 substrains since 1921. C57BL/6 substrains are shown in red and C57BL/10 substrains in blue, the year a strain was created is shown in yellow. (b) Dendrogram generated from IBS relatedness matrix using the 376,824 pruned SNPs. (c) Dendrogram generated from relatedness matrix using 150,344 STRs. (d) Dendrogram

generated from relatedness matrix using 641 differentially expressed transcripts from RNA-sequencing (FDR <0.001). For figure 'a', the distance along the x-axis is not intended to convey any meaning; however, for figures 'b'-'d' the distance between substrains reflects similarity.

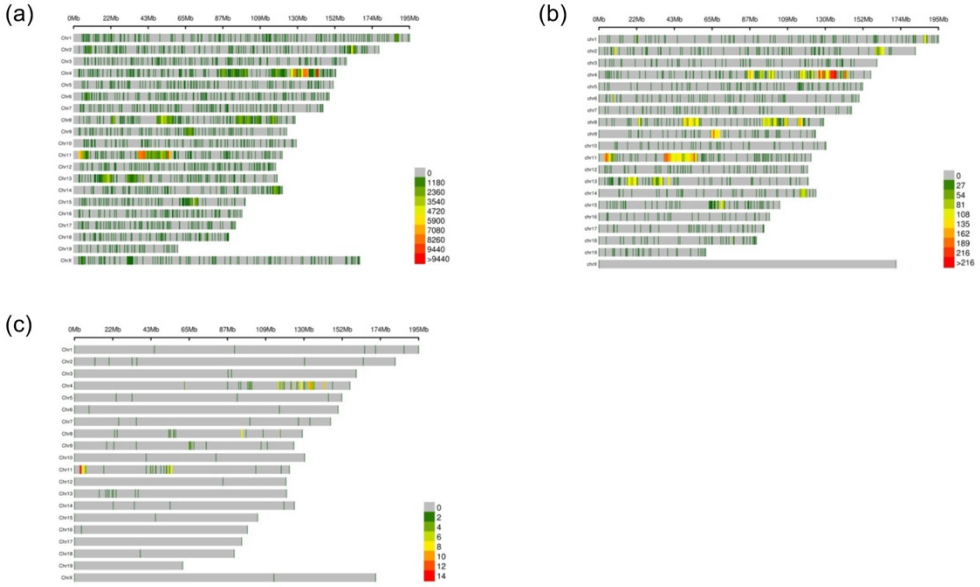


Figure 2. Chromosomal heatmap of genomic features and gene expression that differ between C57BL/6 and C57BL/10 substrains. Chromosomal heatmap of SNP (a) and STR (b) locations that are shared within C57BL/6 substains (n=9) and C57BL/10 substrains (n=5) but differ between the two groups. (c) shows the locations of genes that are differentially expressed between the C57BL/6 and C57BL/10 substrains (FDR<0.1).

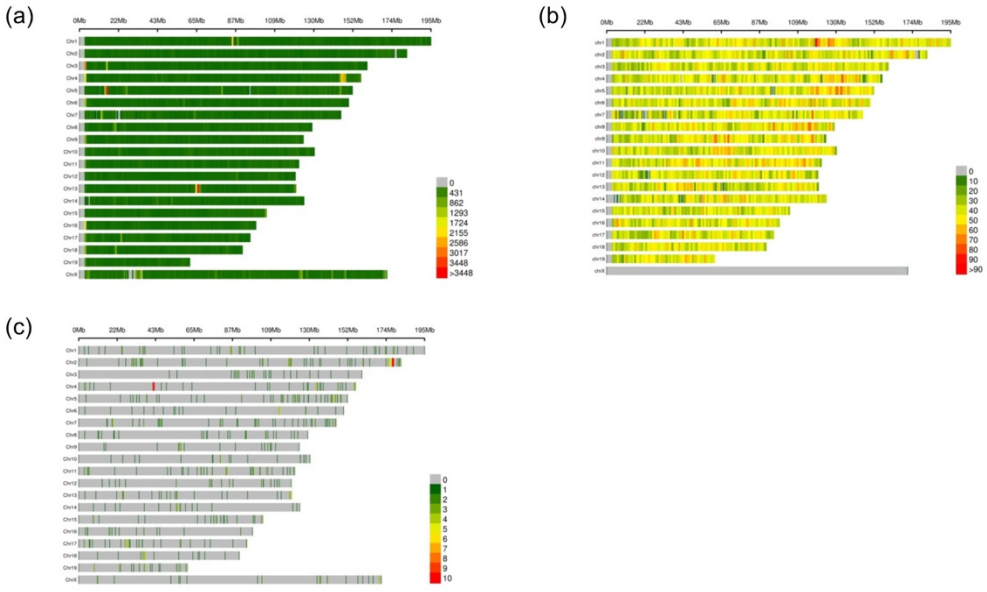


Figure 3. Chromosomal heatmap of genomic features and gene expression that differ between C57BL/6 substrain. Chromosomal heatmap of SNP (a) and STR (b) locations that have at least two unique

genotypes among the C57BL/6 substrains. (c) shows the locations of genes that are differentially expressed among the C57BL/6 substrains (FDR<0.01).

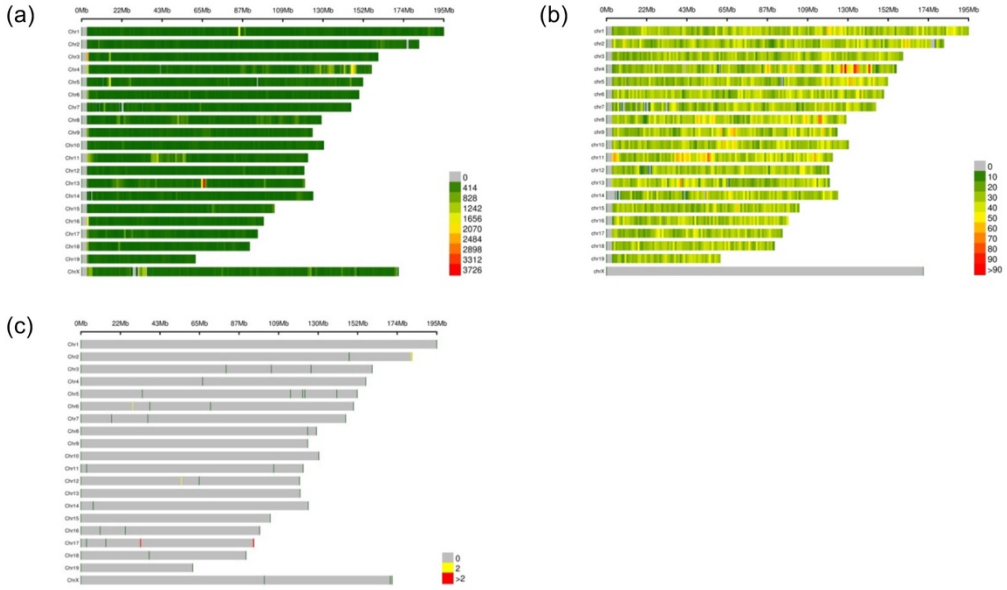


Figure 4. Chromosomal heatmap of genomic features and gene expression that differ between C57BL/10 substrain. Chromosomal heatmap of SNP (a) and STR (b) locations that have at least two unique genotypes among the C57BL/10 substrains. (c) shows the locations of genes that are differentially expressed among the C57BL/10 substrains (FDR<0.01).

3. Results

3.1 Genomic differences between and within C57BL/6 and C57BL/10 substrains

Joint variant calling identified 1,035,308 SNPs, 150,344 STRs and 896 SVs (SV were only called in the C57B/6 substrains). Using the RNA-sequencing data from 117 samples we identified 16,718 expressed genes (**Supplementary material**). To identify differences between all of the C57BL/6 and C57BL/10 strains, we first found genomic features (SNPs, STRs) that were identical among all C57BL/6 substrains (n=9) and among all C57BL/10 substrains (n=5), and then identified the subset of those genomic features that were different between the two groups. For SNPs and STRs, we found 330,178 and 14,367 genomic features that differentiated the C57BL/6 and C57BL/10 groups (**Figures 2a and 2b**). For gene expression we performed a two-factor t-test to identify genes that were differentially expressed (FDR<0.1) between the C57BL/6 substrains and C57BL/10 substrains, which yielded 225 genes (**Figure 2c**). Plotting the data revealed three clusters

of genomic features that differentiated the two groups; the first one on Chromosome 4 at about 130Mb and the other two on Chromosome 11 at about 5Mb and 43Mb.

We next identified polymorphisms (SNPs, STRs) within C57BL/6 and C57BL/10 substrains. We found 446,985 SNPs and 99,935 STRs that were polymorphic among C57BL/6 substrains (**Figures 3a, 3b**), and 419,376 SNPs and 76,524 STRs that were polymorphic among the C57BL/10 substrains (**Figures 4a, 4b**). To identify differentially expressed genes we performed ANOVA for the nine C57BL/6 substrains and five C57BL/10 substrains separately, which yielded 587 and 37 genes (FDR<0.01; **Figure 3c and 4c**). Examination of the heatmaps in **Figure 2** suggested that the distal portion of Chromosome 4 that we identified when comparing C57BL/6 and C57BL/10 was also polymorphic within both C57BL/6 and C57BL/10. In addition, a region at about 85 Mb on Chromosome 1 showed dense polymorphisms in both the C57BL/6 and C57BL/10 lineages.

3.2 Support for historical trees using dendrograms

We obtained information regarding the historical relationships among inbred mouse strains and used this data to draw a tree (**Figure 1a**) (34-36). We then checked to see if this historical information was consistent with dendrograms generated from empirical data. Specifically, we generated dendrograms using SNPs (**Figure 1b**) and STRs (**Figure 1c**). Finally, we used the gene expression data to generate a dendrogram. Given that most of the genes are not differentially expressed and therefore are not helpful for resolving relationships among the substrains, we performed an analysis of variance (ANOVA) for each gene using the between subject factor strain and removed genes that were not significantly differentially expressed among the 14 substrains using several significance thresholds: FDR≤1, FDR <0.1, FDR <0.01 and FDR <0.001. We found that 641 genes were differentially expressed at the most stringent threshold (FDR<0.001), this threshold produced a dendrogram that best matched the historical-, SNP- and STR-based dendrograms (**Figure 1d**). All dendrograms accurately separated the two predominant branches (C57BL/6 and C57BL/10).

3.3 Differential expression and genomic features that match the SDP

For the differential expression analyses, we performed ANOVA for C57BL/6 and C57BL/10 substrains separately. This resulted in 578 differentially expressed genes with FDR<0.01 among the C57BL/6 substrains and 37 genes with FDR<0.01 among the C57BL/10 substrains. We then examined the cis-region (+/- 1Mb) up and down stream of each differentially expressed gene to identify SNP and STR features with SDP that exactly matched the SDP of differential expression. We removed pseudo and predicted genes. For the C57BL/6 substrains, we found 551 differentially expressed genes with one or more features that exactly matched the SDP. Similarly, for the C57BL/10 substrains we found 35 differentially expressed genes with features that matched the expression patterns. Several examples are described in the next paragraphs.

Structural Variations in C57BL/6 substrains where identified by LUMPY and CNVnator (32, 33). LUMPY identified 175 deletions, 184 duplications and 6 inversions, and CNVnator identified 236 deletions and 295 duplications within the C57BL/6 substrains. We filtered the calls with length larger than 1Mb. For CNVnator, we filtered calls with length less than 1kb for deletions and 5 kb for duplications to reduce the number of false positives. We also filtered the calls in the gap regions of the reference genome (mm10). Since we are interested in the genomic differences among these substrains, we filtered the calls that are present in all the substrains and show no variations among them. After merging these calls, we identified 896 SVs. Of these 896 calls, 34 deletions and 10 duplications overlapped between LUMPY and CNVnator calls (>50% reciprocal overlap), however we used all calls, not just the ones that showed overlap between the two methods. The intersection of these 896 SVs with coding regions, including exons, Transcription Start Sites (TSS) and UTRs, were identified. We found 451 unique SV-feature intersections, which include 211 unique SVs and 609 unique genes. 40 of the SV-feature combinations involve a differentially expressed gene identified by RNAseq data.

We computed the correlation of the SV presence patterns among the substrains with expression patters by defining an SV signal as a binary vector indicating existence of an SV in a substrain, and an expression signal indicating the normalized (range 0 to 1) expression of the gene intersecting with the SV. The median of gene expressions among samples in each substrain was chosen as a representative of that substrain. A high

absolute correlation between an SV and gene expression indicates that the presence of the SV predicts gene expression. To establish an empirical significance threshold for these correlations, we performed a permutation test in which we obtained correlation coefficients for the 40 SV-gene expression pairs in which the SDP for SV status was permuted. **Figure 5** shows the thresholds established by the permutation null distribution and the distribution of correlation coefficients in our data. For FDR=0.05 and FDR=0.1 we performed the Benjamin-Hochberg procedure to control for false discovery rate. **Table 2** lists the genes and their corresponding SVs for FDR=0.05.

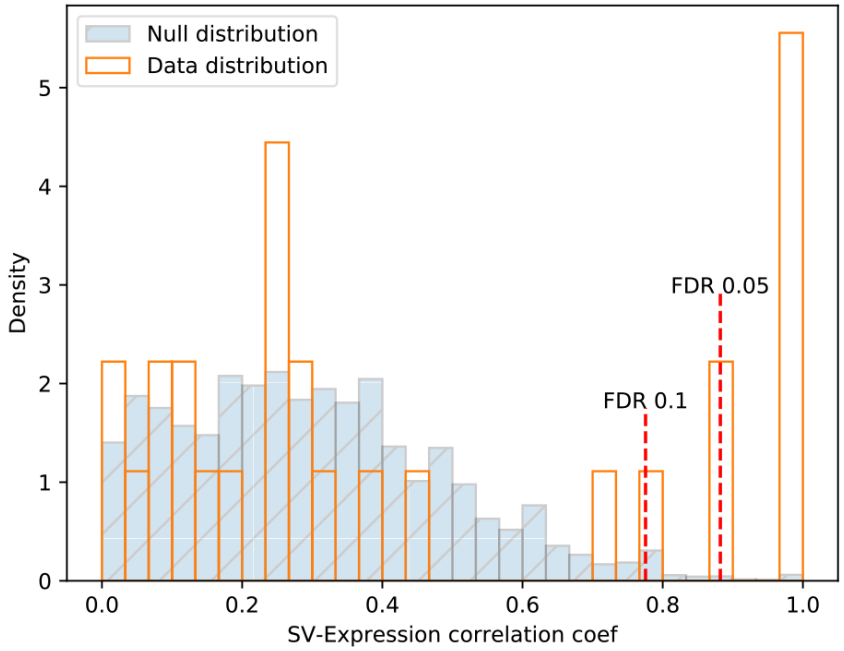


Figure 5. Correlation coefficient distribution for our data and the null hypothesis derived from a permutation test. The SV signal is a binary vector with length 9, showing the existence of an SV in each substrain, and the expression signal is a normalized vector, with components varying between zero and one, showing the variation of differentially expressed genes among substrains. The correlation coefficient is defined as the correlation between the SV signal and the expression signal. The dashed red lines show the significance thresholds for FDR=0.05 and 0.1, obtained by Benjamin-Hochberg procedure.

Table 2. Structural variations identified by LUMPY and CNVnator which have significant correlations between the SV and gene expression patterns. Strain genotypes code for substrains which have the structural variation. '1' for having and '0' for not having the SV. The order of the binary numbers from left to right corresponds to: C57BL/6J, C57BL/6NJ, C57BL/6ByJ, C57BL/6JEiJ, C57BL/6NTac, C57BL/6NCrl, C57BL/6NHsd, B6N-TyrC/BrdCrlCrl, C57BL/6JBomTac. Correlation coefficient is calculated from the strain

genotype signal and the normalized gene expression. FDR=0.05 is chosen for identifying significance based on the Benjamin-Hochberg procedure.

Gene	Region	SV type	Caller	Strain genotype	Corr Coef
Wdfy1	chr1:79715500-79729900	DEL	CNVnator	0;1;1;1;1;1;1;1;1	0.9855
Ide	chr19:37235100-37379500	DUP	CNVnator	1;0;0;0;0;0;0;0;0	0.9996
Fgfbp3	chr19:36911400-36971900	DUP	CNVnator	1;0;0;0;0;0;0;0;0	0.9898
Btaf1	chr19:36911400-36971900	DUP	CNVnator	1;0;0;0;0;0;0;0;0	0.8821

Wdfy1

Wdfy1 is a highly differentially expressed gene (**Figure 6**). The expression level of this gene in C57BL/6J is approximately half of the lever observed in the other 8 C57BL/6 substrains. This gene was previously reported to be differentially expressed between C57BL/6J and C57BL/6NCrl, and was listed as one of the candidate genes associated with reduction of alcohol preference in C57BL/6NCrl (38).

We identified a copy number variation in all substrains except C57BL/6J (**Table 2**) in a segmental duplication region. **Figure 7** shows coverage plots near *Wdfy1* on chromosome 1. A duplication is present in C57BL/6J which replicates three exons in the middle of *Wdfy1*. These extra exons do not exist in the other substrains. We hypothesize that this duplication in C57BL/6J causes a frameshift in *Wdfy1* and activates the nonsense mediated decay which reduces the expression of *Wdfy1* in C57BL/6J.

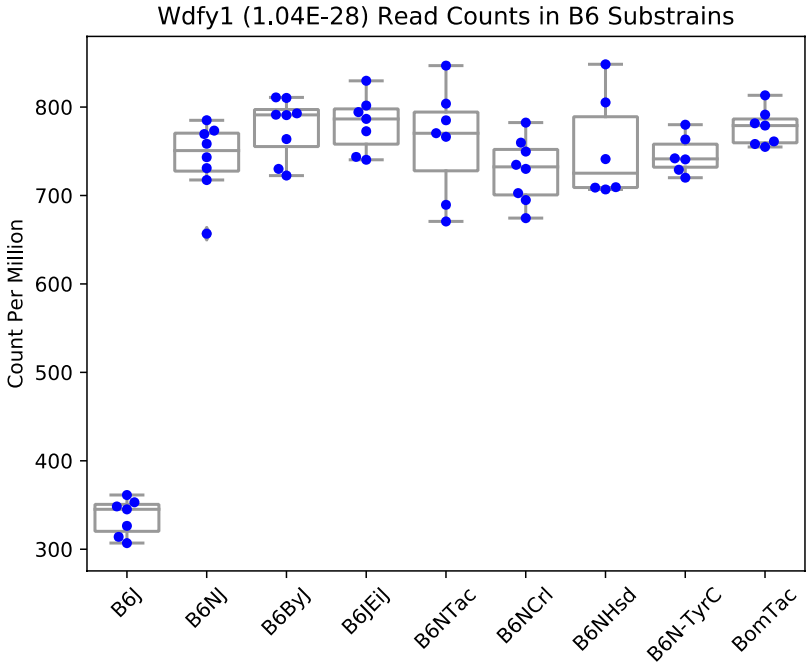


Figure 6. Gene expression results for *Wdfy1*. Gene expression in counts per million (CPM) for each of the C57BL/6 groups (n=7-11) shows a two-fold reduction in expression of *Wdfy1* in hippocampus of B6J.

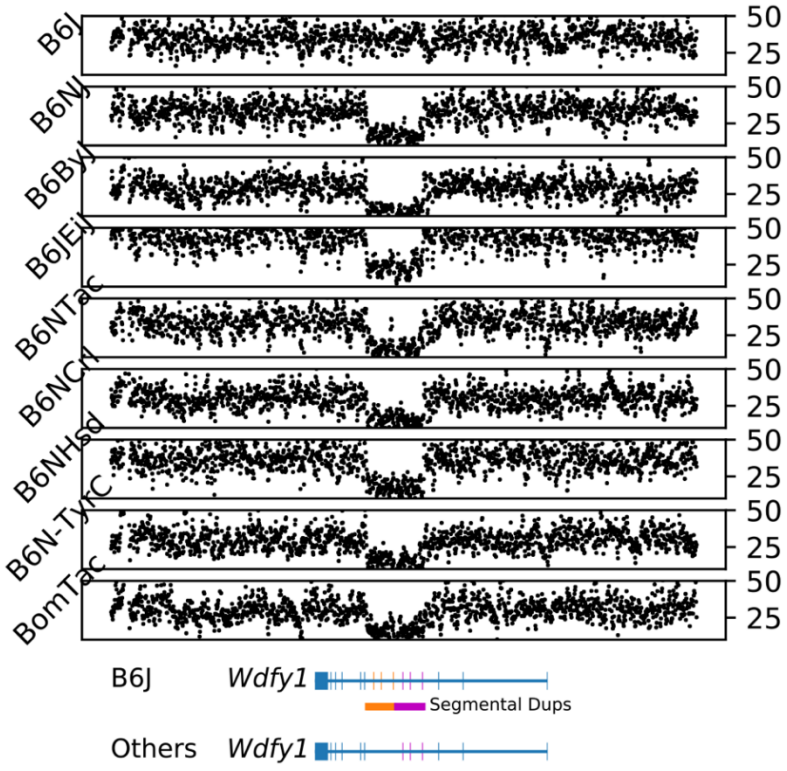


Figure 7. Read depth data near *Wdfy1*. Read coverages from illumina sequencing results show a copy number variation for all the substrains other than B6J at a segmental duplication region (indicated by the

orange and magenta bars). This shows that B6J (the reference genome) has a tandem duplication region which is missing in the other C57Bl/6 substrains that we studied. We hypothesize that this variation causes a frameshift and activates the nonsense mediated decay which reduces the expression of *Wdfy1* in the hippocampus.

Ide, Fgfbp3, Btaf1

RNAseq results in hippocampus show significant increases in expression of three nearby genes: *Ide, Fgfbp3* and *Btaf1* in C57BL/6J substrain (**Figure 8**). SV-expression correlation analysis linked two CNV calls (chr19:36911400-36971900 and chr19:37235100-37379500; **Table 2**) and one LUMPY call (chr19:36911370-37379559; **Figure 9**) to the expression pattern in *Ide, Fgfbp3* and *Btaf1*. The two CNV calls, and expression variation for *Ide* and *Fgfbp3* have been previously reported for C57BL/6J (39). *Ide* and *Fgfbp3* are completely within the borders of the two CNV regions, however *Btaf1* is partially inside the first CNV region. **Figure 8** shows that the increased expression of *Btaf1* is less than the increase for the other two genes. LUMPY also calls a duplication at the same location (**Figure 9**), however, the coverage plot shows that the two CNV calls are correct and LUMPY has incorrectly merged the two breakends, the beginning of the first CNV call and the end of the second.

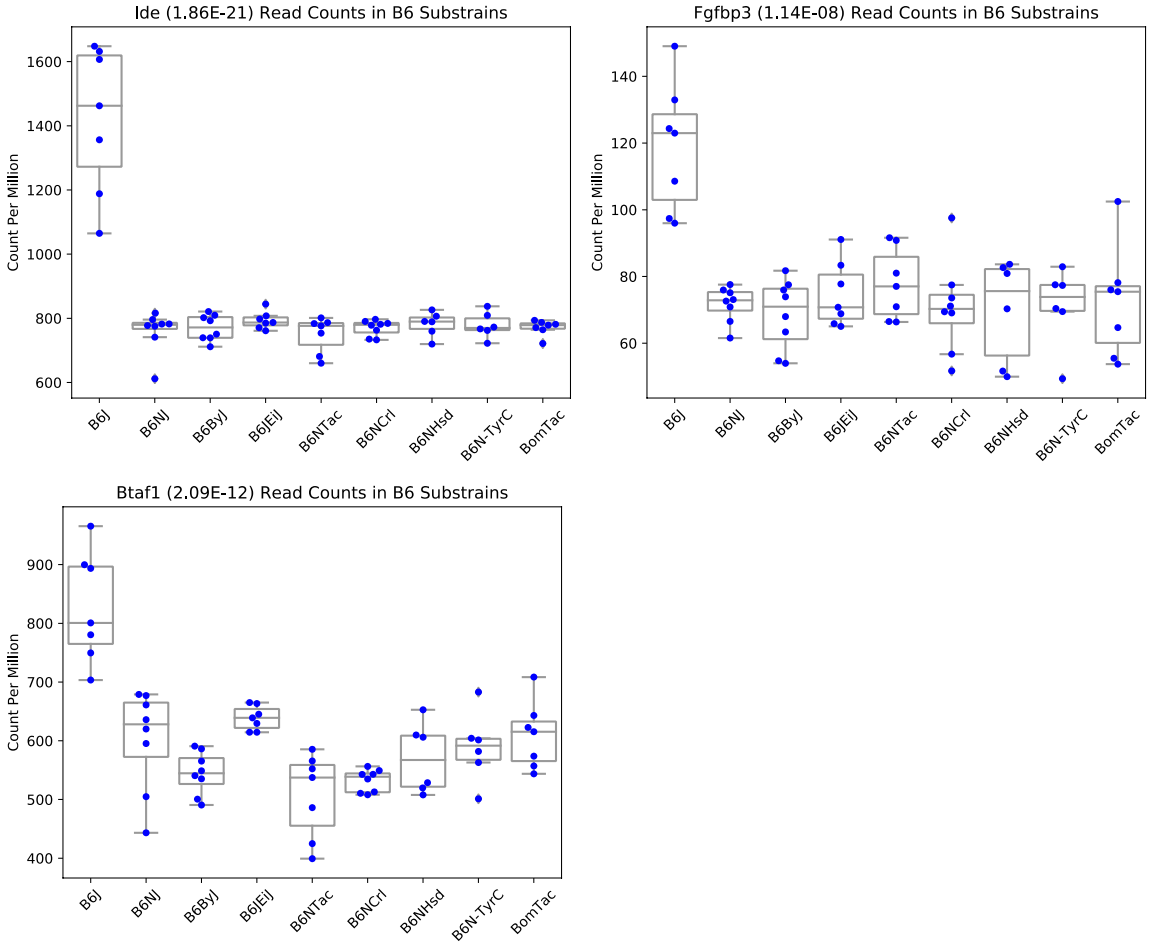


Figure 8. Gene expression results for *Ide, Fgfbp3 and Btaf1*. RNAseq results in hippocampus show significant increase in expression of *Ide, Fgfbp3* and *Btaf1* in B6J.

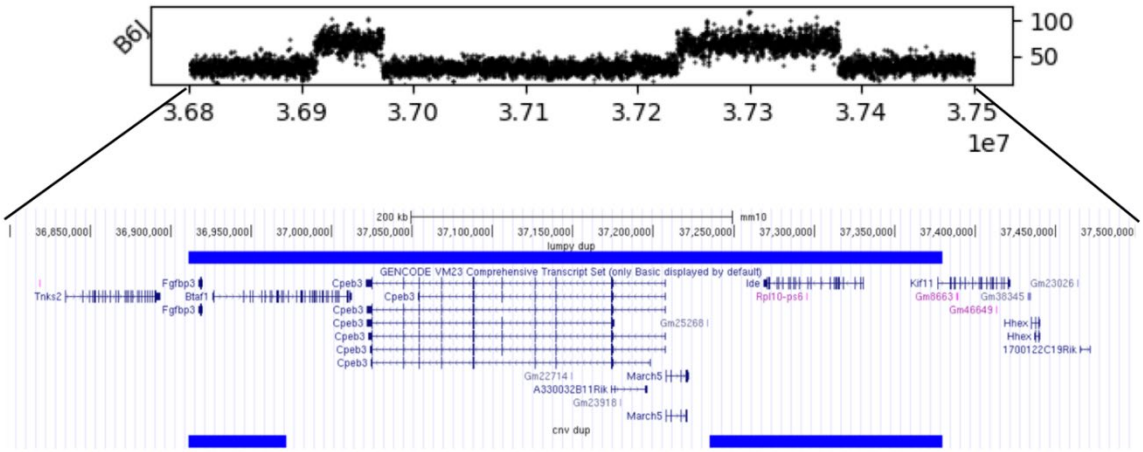


Figure 9. CNV calls in chromosome 19 overlapping *Ide, Fgfbp3* and *Btaf1*. Two CNV calls covering *Ide* and *Fgfbp3* completely and *Btaf1* partially are detected by CNVnator.

Nnt

Another significantly differentially expressed gene was *Nnt* (**Figures 10 & Table S2**). Once again, C57BL/6J was different from all other C57BL/6 substrains, however, in this case C57BL/6J showed approximately two-fold lower expression. This gene has been previously reported to show lower expression in the C57BL/6J substrain due to a unique multi-exon deletion that was previously detected by quantitative RT-PCR (19). However, we found no difference in the number of reads in this region in C57BL/6J compared to the other C57BL/6 substrains, which suggests that the previously reported deletion from exons 7 to exon 11 did not exist in the individual we sequenced, inconsistent with a previous report (19). However, we did find 49 genomic features that matched the SDP and thus could explain this gene expression difference.

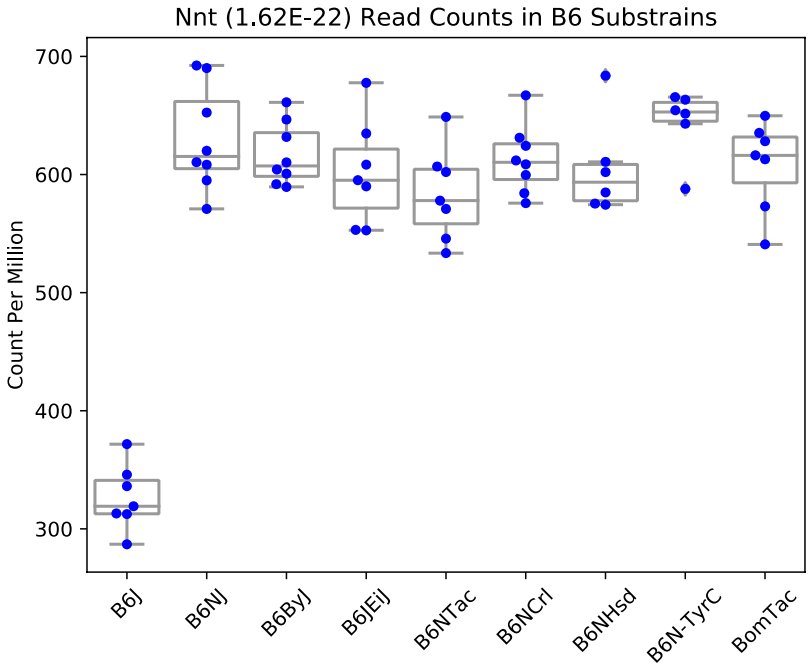


Figure 10. Gene expression results for *Nnt*. Boxplots for the gene expression in CPM are shown for each of the C57BL/6 groups (n=7-11) that underwent RNA-sequencing.

Gabra2

Another highly differentially expressed gene was *Gabra2* (**Figures 11 & Table S3**). C57BL/6J showed approximately half the expression level compared to all the other C57BL/6 substrains. There were 23 genomic features that matched the SDP and thus could explain the gene expression difference. The gene expression was consistent with previous findings that C57BL/6J mice have reduced *Gabra2* expression levels (40); in

that prior study, an intronic indel adjacent to a splice acceptor site was identified at Chr5:71,014,638. Repairing that SNP restored normal expression (40), demonstrating causality. We identified the same indel and also found that it matched the SDP.

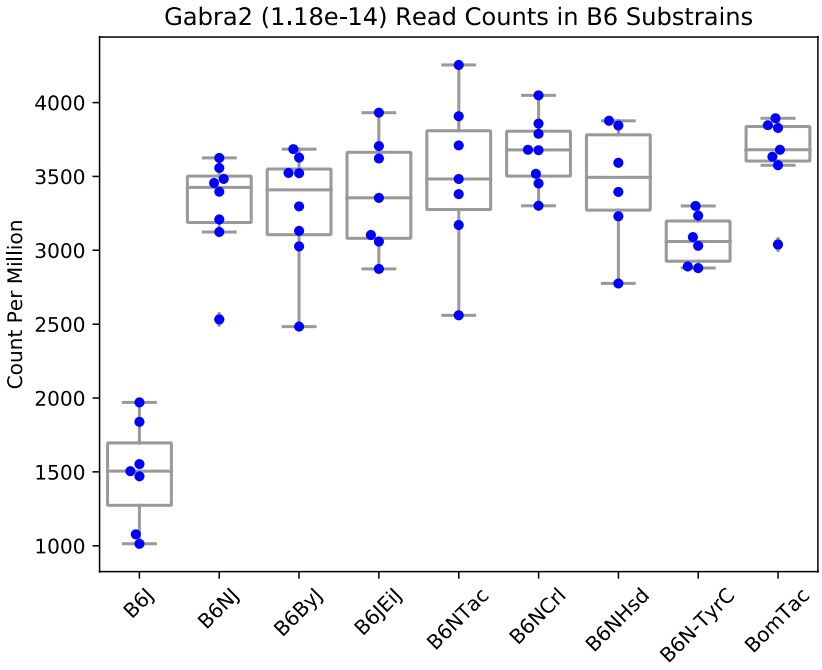


Figure 11. Gene expression results for *Gabra2*. Boxplots for the gene expression in CPM are shown for each of the C57BL/6 groups (n=7-11) that underwent RNA-sequencing.

Kcnc2

Another highly differentially expressed gene was *Kcnc2* (**Figures 12 & Table S4**). C57BL/6JEiJ showed approximately half the expression level compared the other C57BL/6 substrains. There were 19 genomic features that matched the SDP and thus could explain the gene expression difference. To our knowledge, this gene expression difference has not been previously reported. This represents one of the many novel gene expression differences we identified in the present study.

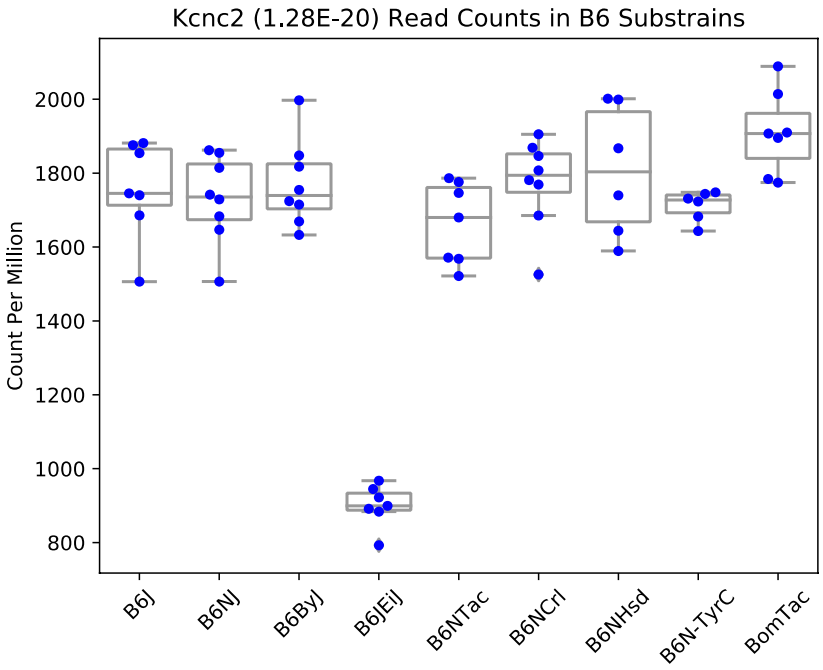


Figure 12. Gene expression results for *Kcnc2* with potentially causative genomic features. Boxplots for the gene expression in CPM are shown for each of the C57BL/6 groups (n=7-11) that underwent RNA-sequencing.

4. Discussion

Here we have performed a large scale multi-omics analysis of 14 commonly used C57BL substrains. We identified 985,329 SNPs, 150,344 STRs and 896 SVs, out of which 330,178 SNPs and 14,367 STRs differentiated the C57BL/6 and C57BL/10 groups. In order to demonstrate functional consequences of these polymorphisms, we examined gene expression differences, which provided a large number of traits. We identified 578 differentially expressed genes for C57BL/6 substrains and 37 differentially expressed genes for C57BL/10 substrains (FDR < 0.01). We then identified nearby polymorphisms that might be causally related to the gene expression differences based on their SDP. Previous studies had identified some of these differences, but many others were novel.

Dendrograms based on empirical data (SNPs, STRs, and gene expression) closely matched the chronological tree that was derived from historical records (**Figure 1**). Whereas our initial expectation was that polymorphisms would be due to new mutations, we identified evidence suggesting that a few regions were not completely inbred prior to the separation of C57BL/6 and C57BL/10 (Chromosomes 4 and 11; **Figure 2**). Thus, the

dendrograms reflect both new mutations and incomplete fixation. Furthermore, the heatmaps for only C57BL/6 or C57BL/10 identify a few regions that have been differentially fixed among these branches (Chromosome 1; **Figures 3 and 4**). The relatively uniform distribution of polymorphism across the rest of the genome are more consistent with new mutations.

To demonstrate that there were functional consequences to the polymorphisms, we examined gene expression. Gene expression was an attractive choice both because it offered a large number of traits, and because nearby polymorphisms could be associated with specific gene expression differences based on the SDP, thus avoiding the need to make crosses. We examined gene expression in the hippocampus because it can be quickly and reliably dissected, is relevant to many behavioral traits, and expresses a large number of genes. We found that 587 genes were differentially expressed (FDR<0.01) among the C57BL/6 substrains (**Figure 3c**) and 37 genes were differentially expressed (FDR<0.01) among the C57BL/10 substrains (**Figure 4c**). We identified the coding structural variations and defined a correlation coefficient between the SV signal and the corresponding gene expression. We performed a permutation test to find the null hypothesis distribution and the correlation coefficient corresponding to significant level for FDR=0.05 (**Figure 5**). We then identified significant coding SVs which match the expression pattern of their corresponding gene (**Table 2**). A CNV call within *Wdfy1* revealed that the segmental duplication present in the reference genome and C57BL/6J is missing in all the other substrains. We hypothesize that this duplication generates a frameshift in C57BL/6J and activates the nonsense mediated decay mechanism, and causes the expression of *Wdfy1* to decrease in C57BL/6J (**Figure 6, 7**). The increase in expression of *Ide, Fgfbp3* and *Btaf1* in C57BL/6J is linked to two duplication events in chromosome 19 which contain *Ide* and *Fgfbp3* completely and *Btaf1* partially (**Figure 8, 9**). These two CNV duplications and the increase in the expression of *Ide* and Fgfbp3 were previously observed in the C57BL/6J population, but not the expression increase in *Btaf1* (39). In greater than 95% of cases we were able to identify polymorphic genomic features that matched the SDP of the gene expression difference. Some of these had been previously reported, in some cases we independently identified the same causal

polymorphisms, however, in the case of *Nnt* our findings disputed an earlier report (**Figures 10-12**).

Interestingly, we identified a large number of heterozygous SNPs, which may seen counter-intuitive for an inbred strain (e.g. **Tables S1-S4**). In some cases, one individual was called as heterozygote while all others were homozygous; these could reflect mutations that have not yet reached fixation or genotyping errors. We also encountered a number of SNPs in which most of the strains were called as heterozygous but one or more were called homozygotes; these could reflect segmental duplications that are not accounted for in the current genome assembly, in that scenario, the homozygous calls would correspond to deletions or genotyping errors. Further evidence of polymorphisms that are segregating within individual substrains was provided by the gene expression data. For example, *Ide* was differentially expressed in C57BL/6J compared to all other C57BL/6 substrains, however, close inspection of the individual data points shows that several C57BL/6J individuals had intermediate expression levels. These individuals may have been heterozygous for the causal duplication. In contrast, *Nnt* did not show individuals with intermediate expression levels, suggesting that the causal alleles were fixed in all of the substrains. Because our sequence data represented only a single animal per substrain, we would have failed to capture a proportion of the segregating polymorphisms, which would prevent us from identifying all of the causal mutations.

Our results create a resource for future efforts to identify genes and causal polymorphisms that give rise to substrain differences. In the supplemental materials we have provided more than a million genomic features (SNPs, STRs and SVs), as well as gene expression differences in the hippocampus. A common study design has been to cross two phenotypically divergent inbred substrains, sometimes called near isogenic strains, and then map loci that influence a trait of interest. Once a locus has been identified, the data we provide will be extremely useful for identifying the causal polymorphism, which may be a coding difference or a regulatory difference. Another novel application would be to select two strains that are divergent for a coding difference, duplication, deletion or that differentially express a gene of interest and cross them to ascertain the gene's function. This approach is limited by the available polymorphisms.

However, since the cross would have few other polymorphisms, the results would be more readily interpretable.

Crossing near isogenic strains assumes that functional polymorphisms are randomly distributed across the genome. We observed a few chromosomal regions that had relatively dense polymorphisms (e.g. Chromosome 1 at about 85 Mb; **Figure 3a**), which will negatively impact the near isogenic approach. Such regions are likely to contain multiple putatively causative variants, making it more difficult to identify the causal variant, since the whole regions will often be inherited as a single unrecombined haplotype, thus erasing the benefits of using near isogenic substrains. Finally, although C57BL/6J is the most commonly use strain, the knock out mouse project (41) has instead used stem cells derived from a C57BL/6N substrain; our results can also inform possible complications that might arise from unintended crosses between these two strains.

One of the advantages of mice is the large number of inbred strains that are available. We have identified genomic features among a panel of C57BL substrains that can be used for future genetic studies. Whereas we assumed at the outset that polymorphisms among these strains would reflect new mutations, we have shown that incomplete inbreeding is another likely contributor. Future studies of mutational processes could mask the regions with dense polymorphisms, thus allowing for the study of mutagenic processes without interference from these regions.

Author Contributions

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements: MM, YR, CStP, AW and AAP were supported by P50DA037844. Additionally, YR was supported by T32MH018399 and AW was supported by T32MH020065.

REFERENCES

1. Festing M. (1979) *Inbred strains in biomedical research* (Oxford University Press, New York).
2. Lyon M, Searle A., Intenational Committee on Standardized Genetic Nomenclature for Mice (1989) *Genetic variants and strains of the laboratory mouse* (Oxford Univeristy Press).
3. Altman P, Kats, D., (1979) *Inbred and Gentically Defined Starins of Laboratory Animals, Part 1, Mouse and Rat* (Federation of American Society for Experimental Biology, Maryland).
4. Bailey D (1978) *Sources of Subline Divergence and their Relative Importance for Sublines of Six Major Inbred Strain of Mice* (Origins of Invred Mice, Academic Press, New York).
5. Reed C*, et al.* (2017) A Spontaneous Mutation in Taar1 Impacts Methamphetamine-Related Traits Exclusively in DBA/2 Mice from a Single Vendor. *Frontiers in pharmacology* 8:993.
6. Doran AG*, et al.* (2016) Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome biology* 17(1):167.
7. Quinlan AR*, et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome research* 20(5):623-635.
8. Simon MM*, et al.* (2013) A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome biology* 14(7).
9. Akinola LS*, et al.* (2019) C57BL/6 Substrain Differences in Pharmacological Effects after Acute and Repeated Nicotine Administration. *Brain sciences* 9(10).
10. Fan H & Chu JY (2007) A brief review of short tandem repeat mutation. *Genomics, proteomics & bioinformatics* 5(1):7-14.
11. Gatchel JR & Zoghbi HY (2005) Diseases of unstable repeat expansion: mechanisms and common principles. *Nature reviews. Genetics* 6(10):743-755.
12. Hurles ME, Dermitzakis ET, & Tyler-Smith C (2008) The functional impact of structural variation in humans. *Trends in genetics : TIG* 24(5):238-245.
13. Clapcote SJ & Roder JC (2004) Survey of embryonic stem cell line source strains in the water maze reveals superior reversal learning of 129S6/SvEvTac mice. *Behavioural brain research* 152(1):35-48.
14. Grottick AJ*, et al.* (2005) Neurotransmission- and cellular stress-related gene expression associated with prepulse inhibition in mice. *Brain research. Molecular brain research* 139(1):153-162.
15. Mayorga AJ & Lucki I (2001) Limitations on the use of the C57BL/6 mouse in the tail suspension test. *Psychopharmacology* 155(1):110-112.
16. Radulovic J, Kammermeier J, & Spiess J (1998) Generalization of fear responses in C57BL/6N mice subjected to one-trial foreground contextual fear conditioning. *Behav Brain Res* 95(2):179-189.
17. Siegmund A, Langnaese K, & Wotjak CT (2005) Differences in extinction of conditioned fear in C57BL/6 substrains are unrelated to expression of alpha-synuclein. *Behavioural brain research* 157(2):291-298.
18. Stiedl O*, et al.* (1999) Strain and substrain differences in context- and tone-dependent fear conditioning of inbred mice. *Behavioural brain research* 104(1-2):1-12.
19. Toye AA*, et al.* (2005) A genetic and physiological study of impaired glucose homeostasis control in C57BL/6J mice. *Diabetologia* 48(4):675-686.
20. Green ML*, et al.* (2007) Reprogramming of genetic networks during initiation of the Fetal Alcohol Syndrome. *Developmental dynamics : an official publication of the American Association of Anatomists* 236(2):613-631.
21. Khisti RT, Wolstenholme J, Shelton KL, & Miles MF (2006) Characterization of the ethanol-deprivation effect in substrains of C57BL/6 mice. *Alcohol* 40(2):119-126.
22. Diwan BA & Blackman KE (1980) Differential susceptibility of 3 sublines of C57BL/6 mice to the induction of colorectal tumors by 1,2-dimethylhydrazine. *Cancer letters* 9(2):111-115.

23. Roth DM, Swaney JS, Dalton ND, Gilpin EA, & Ross J, Jr. (2002) Impact of anesthesia on cardiac function during echocardiography in mice. *Am J Physiol Heart Circ Physiol* 282(6):H2134-2140.
24. Kumar V*, et al.* (2013) C57BL/6N mutation in cytoplasmic FMRP interacting protein 2 regulates cocaine response. *Science* 342(6165):1508-1512.
25. Kadiyala SB, Papandrea D, Herron BJ, & Ferland RJ (2014) Segregation of Seizure Traits in C57 Black Mouse Substrains Using the Repeated-Flurothyl Model. *Plos One* 9(3).
26. Markham BE, Kernodle S, Nemzek J, Wilkinson JE, & Sigler R (2015) Chronic Dosing with Membrane Sealant Poloxamer 188 NF Improves Respiratory Dysfunction in Dystrophic Mdx and Mdx/Utrophin(-/-) Mice. *Plos One* 10(8).
27. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
28. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
29. McKenna A*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20(9):1297-1303.
30. Purcell S*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81(3):559-575.
31. Willems T*, et al.* (2017) Genome-wide profiling of heritable and de novo STR variations. *Nature methods* 14(6):590-592.
32. Abyzov A, Urban AE, Snyder M, & Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research* 21(6):974-984.
33. Layer RM, Chiang C, Quinlan AR, & Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome biology* 15(6):R84.
34. Beck JA*, et al.* (2000) Genealogies of mouse inbred strains. *Nature genetics* 24(1):23-25.
35. Charles River Laboratories (2019) Available from: https://www.criver.com/.
36. Jackson Laboratory (2019) Available from: https://www.jax.org/.
37. Storey JD & Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 100(16):9440-9445.
38. Mulligan MK*, et al.* (2008) Alcohol trait and transcriptional genomic analysis of C57BL/6 substrains. *Genes, brain, and behavior* 7(6):677-689.
39. Watkins-Chow DE & Pavan WJ (2008) Genomic copy number and expression variation within the C57BL/6J inbred mouse strain. *Genome research* 18(1):60-66.
40. Mulligan MK*, et al.* (2019) Identification of a Functional Non-coding Variant in the GABA(A) Receptor alpha 2 Subunit of the C57BL/6J Mouse Reference Genome: Major Implications for Neuroscience Research. *Front Genet* 10.
41. Austin CP*, et al.* (2004) The knockout mouse project. *Nature genetics* 36(9):921-924.

5. Supplementary

Gene	Chromosome	Position	Category	B6J	B6NJ	B6ByJ	B6JEiJ	B6NTac	B6NCrl	B6NHsd	B6NTyrC	B6JBomTac
ide	19	36803402	INDEL	0/0	0/2	0/1	0/1	0/1	0/2	0/1	0/2	0/1
ide	19	36812936	SNP	TG	GG	GG	GG	GG	GG	GG	GG	GG
ide	19	36886323	INDEL	0/1	1/1	1/1	1/1	1/1	1/1	1/1	0/0	1/1
ide	19	36911371	SV	DUP(468188bp)	NO	NO	NO	NO	NO	NO	NO	NO
ide	19	37035314	SNP	AA	GA	GA	GA	GA	GA	GA	GA	GA
ide	19	37183862	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
ide	19	37204309	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
ide	19	37248620	INDEL	0/2	0/0	0/0	0/0	0/0	0/0	0/1	0/0	0/0
ide	19	37342195	INDEL	1/1	./.	./.	./.	./.	./.	./.	./.	./.
ide	19	37379512	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
ide	19	37460195	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
ide	19	37488595	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
ide	19	37530549	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
ide	19	37556701	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
ide	19	37565349	INDEL	0/1	1/1	1/1	1/2	0/2	0/2	0/0	0/0	2/2
ide	19	37645147	INDEL	0/2	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/0
ide	19	37786339	INDEL	0/0	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
ide	19	37791925	INDEL	0/2	1/1	1/1	0/1	1/1	1/1	1/1	1/1	1/1
ide	19	37791928	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791931	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791941	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791945	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791947	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791948	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791950	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791952	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791953	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791960	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791964	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37803861	INDEL	./.	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
ide	19	37817057	INDEL	0/1	0/0	0/2	0/2	1/1	0/2	0/0	0/2	0/0

Table S1. **Genomic features for the *Ide* gene that match the SDP of the expression difference.** SNPs, STRs, and SVs 500kb up/down-stream of the gene with SDP that matched the gene expression patterns are shown for all 9 C57BL/6 substrains.

Gene	Chromosome	Position	Category	B6J	B6NJ	B6ByJ	B6JEiJ	B6NTac	B6NCrl	B6NHsd	B6NTyrC	B6JBomTac
nnt	13	118847209	STR	0/2	0/1	0/0	0/0	0/1	0/0	0/1	0/1	0/0
nnt	13	118847256	INDEL	0/2	0/1	0/1	0/0	0/1	0/1	0/1	0/1	0/1
nnt	13	118859648	INDEL	0/2	0/0	0/0	0/0	1/2	0/0	0/0	0/0	0/0
nnt	13	118933907	SNP	GG	TT	TT	TT	GT	TT	TT	TT	TT
nnt	13	118957164	STR	0/1	0/0	0/0	2/2	0/0	0/0	0/0	0/0	0/0
nnt	13	119048487	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119067682	SNP	CA	AA	AA	AA	AA	AA	AA	AA	AA
nnt	13	119228909	STR	0/3	0/2	0/2	0/0	0/0	0/0	0/2	0/1	0/0
nnt	13	119392610	INDEL	0/1	0/0	0/0	0/0	0/2	1/1	0/0	0/0	0/0
nnt	13	119488749	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119489463	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119489765	SNP	GG	AG	AG	AG	AG	AG	AG	AG	AG
nnt	13	119595753	SNP	GC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119595994	SNP	CA	AA	AA	AA	AA	AA	AA	AA	AA
nnt	13	119596118	SNP	AA	GA	GA	GA	GA	GA	GA	GG	GA
nnt	13	119596370	SNP	CT	TT	TT	TT	TT	TT	TT	TT	TT
nnt	13	119596371	SNP	TG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119596676	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119598848	SNP	GT	TT	TT	TT	TT	TT	TT	TT	TT
nnt	13	119598892	SNP	GG	TG	TG	TG	TG	TG	TG	TG	TG
nnt	13	119599608	SNP	TT	AT	AT	AT	AT	AT	AT	AA	AT
nnt	13	119599716	SNP	GC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119599940	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
nnt	13	119600075	SNP	GT	TT	TT	TT	TT	TT	TT	TT	TT
nnt	13	119601965	SNP	GC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119602157	SNP	TG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119602739	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119602794	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119603540	INDEL	0/0	./.	./.	0/1	./.	1/1	0/1	1/1	1/1
nnt	13	119603562	SNP	CC	CT	CT	CT	CT	TT	CT	TT	CT
nnt	13	119603572	SNP	TT	TC	TC	TC	TC	TC	TC	CC	TC
nnt	13	119603588	SNP	CC	TC	TC	TC	TC	TC	TC	TC	TC
nnt	13	119603591	SNP	TT	AT	AT	AT	AT	AT	AT	AT	AT
nnt	13	119603604	SNP	CC	TC	TC	TC	TC	TC	TC	TC	TC
nnt	13	119603613	SNP	AA	CA	CA	CA	CA	CA	CA	CA	CA
nnt	13	119613570	SNP	CG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119613734	INDEL	0/2	2/3	0/1	1/3	1/2	1/3	1/2	2/3	1/3
nnt	13	119615372	INDEL	0/1	1/3	1/2	2/3	1/2	0/3	1/3	1/3	2/3
nnt	13	119624238	INDEL	1/1	./.	./.	./.	./.	./.	./.	1/2	./.
nnt	13	119685276	STR	0/0	2/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2
nnt	13	119685297	INDEL	./.	1/2	1/2	1/2	1/2	1/2	1/2	1/1	1/2
nnt	13	119685317	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
nnt	13	119685321	INDEL	0/0	0/2	0/2	0/2	0/1	0/2	0/1	0/1	0/2
nnt	13	119686806	STR	0/0	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
nnt	13	119686819	INDEL	0/0	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
nnt	13	119721309	INDEL	0/1	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0
nnt	13	119882818	SNP	GG	TG	TG	TG	TG	TG	TG	TG	TG
nnt	13	119882820	SNP	AA	TA	TA	TA	TA	TA	TA	TA	TA
nnt	13	119905572	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0

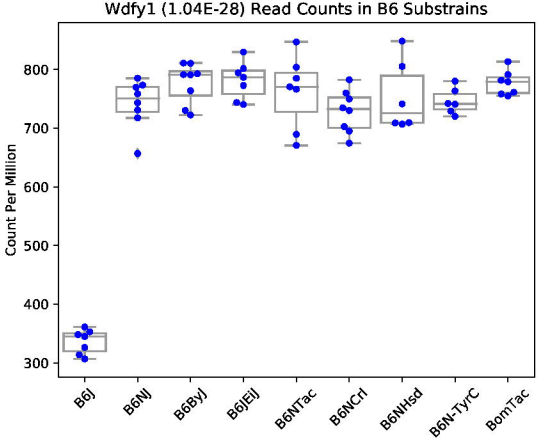
Table S2. **Genomic features for the *Nnt* gene that match the SDP of the expression difference.** SNPs, STRs, and SVs 500kb up/down-stream of the gene with SDP that matched the gene expression patterns are shown for all 9 C57BL/6 substrains.

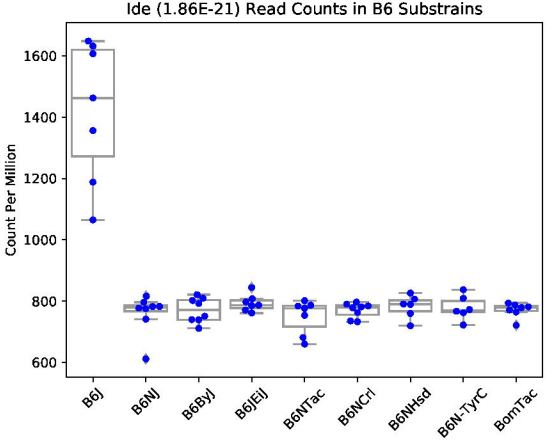
Gene	Chromosome	Position	Category	B6J	B6NJ	B6ByJ	B6JEiJ	B6NTac	B6NCrl	B6NHsd	B6NTyrC	B6JBomTac
gabra2	5	70509237	SNP	CC	GG	GG	GG	GG	GG	GG	GG	GG
gabra2	5	70514395	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
gabra2	5	70721116	INDEL	1/2	0/1	0/1	0/0	0/0	0/1	0/0	0/1	0/1
gabra2	5	70726237	INDEL	0/0	1/2	1/1	1/2	0/1	0/1	1/1	./.	1/2
gabra2	5	70737943	STR	2/2	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
gabra2	5	70774999	STR	0/2	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
gabra2	5	70786971	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
gabra2	5	70805456	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
gabra2	5	70925745	SNP	CT	CC	CC	TT	CC	CC	CC	TT	TT
gabra2	5	71004633	SNP	GT	TT	TT	TT	TT	TT	TT	TT	TT
gabra2	5	71014638	INDEL	0/0	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
gabra2	5	71091704	INDEL	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
gabra2	5	71126893	SNP	CT	TT	TT	TT	TT	TT	TT	TT	TT
gabra2	5	71157413	SNP	AC	CC	CC	CC	CC	CC	CC	CC	CC
gabra2	5	71215173	INDEL	0/1	./.	0/0	./.	./.	1/1	0/2	0/0	./.
gabra2	5	71218230	STR	0/2	0/0	0/0	0/0	0/1	0/0	0/1	0/0	0/0
gabra2	5	71274939	INDEL	1/2	0/2	0/0	0/0	0/0	0/1	0/1	0/1	0/1
gabra2	5	71342379	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
gabra2	5	71419173	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
gabra2	5	71427935	INDEL	0/2	0/0	0/0	0/0	0/1	0/0	0/0	0/0	0/1
gabra2	5	71447810	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
gabra2	5	71493034	STR	0/0	2/2	2/2	2/2	0/2	2/2	2/2	2/2	0/2
gabra2	5	71586254	INDEL	0/1	1/1	1/1	./.	1/1	1/1	1/1	./.	./.

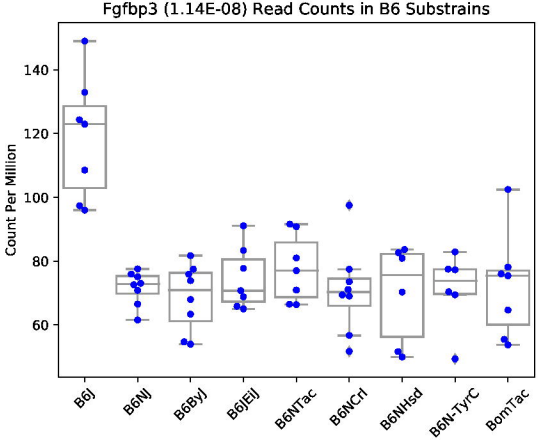
Table S3. **Genomic features for the *Gabra2* gene that match the SDP of the expression difference.** SNPs, STRs, and SVs 500kb up/down-stream of the gene with SDP that matched the gene expression patterns are shown for all 9 C57BL/6 substrains.

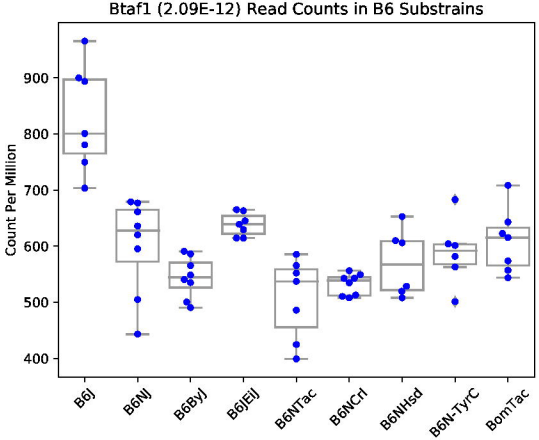
Gene	Chromosome	Position	Category	B6J	B6NJ	B6ByJ	B6JEiJ	B6NTac	B6NCrl	B6NHsd	B6NTyrC	B6JBomTac
kcnc2	10	111953664	INDEL	0/0	0/0	0/0	2/2	0/1	0/0	0/0	0/1	0/0
kcnc2	10	112013254	SNP	TT	TT	TT	GT	TT	TT	TT	TT	TT
kcnc2	10	112050694	SNP	GG	GG	GG	TT	GG	GG	GG	GG	GG
kcnc2	10	112052058	INDEL	0/1	0/1	0/1	0/0	0/1	0/1	0/1	0/1	0/1
kcnc2	10	112189298	SNP	TT	TT	TT	CC	TT	TT	TT	TT	TT
kcnc2	10	112201547	INDEL	0/1	0/1	0/1	./.	0/2	0/2	0/0	0/0	0/2
kcnc2	10	112381442	INDEL	0/2	1/2	0/0	1/1	0/2	0/0	0/1	1/2	0/0
kcnc2	10	112381448	INDEL	0/2	1/2	0/0	1/1	0/2	0/0	0/1	0/2	0/0
kcnc2	10	112385927	INDEL	0/0	0/1	0/0	0/2	0/1	0/1	0/1	0/1	0/0
kcnc2	10	112562168	INDEL	0/1	0/2	0/1	./.	0/1	0/1	0/0	0/1	0/0
kcnc2	10	112762924	SNP	GG	GG	GG	TT	GG	GG	GG	GG	GG
kcnc2	10	112783639	SNP	TT	TT	TT	GT	TT	TT	TT	TT	TT
kcnc2	10	112787091	STR	0/0	0/0	0/0	2/2	0/0	0/1	1/1	0/1	0/0
kcnc2	10	112819780	SNP	AA	AA	AA	CA	AA	AA	AA	AA	AA
kcnc2	10	112824134	SNP	AA	AA	AA	CA	AA	AA	AA	AA	AA
kcnc2	10	112834213	SNP	TT	TT	TT	CC	TT	TT	TT	TT	TT
kcnc2	10	112861890	INDEL	0/0	0/1	0/2	1/2	0/1	0/1	0/0	0/0	0/0
kcnc2	10	112898865	INDEL	0/0	0/1	0/1	1/2	0/0	1/1	0/0	0/2	1/1
kcnc2	10	112964630	INDEL	0/2	0/1	0/0	1/2	0/0	0/0	0/0	0/1	0/1

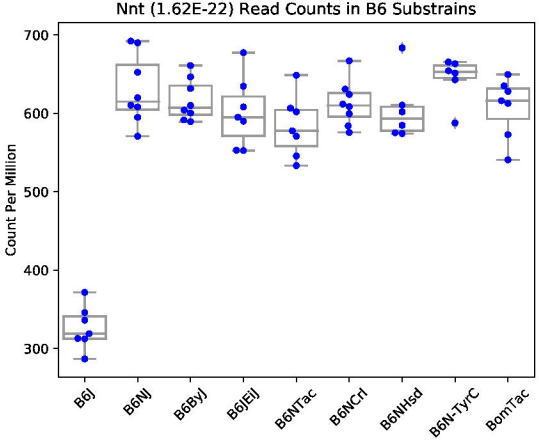
Table S4. **Genomic features for the *Kcnc2* gene that match the SDP of the expression difference.** SNPs, STRs, and SVs 500kb up/down-stream of the gene with SDP that matched the gene expression patterns are shown for all 9 C57BL/6 substrains.

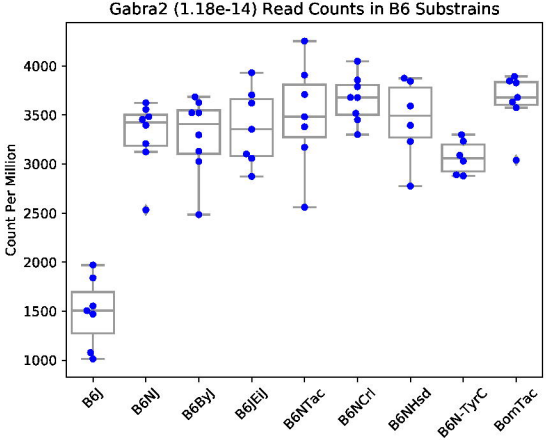

Wdfy1 (1.04E-28) Read Counts in B6 Substrains
Count Per Million
300
400
500
600
700
800
B6J
B6NJ
B6BYJ
B6JEIJ
B6NTac
B6NCrl
B6NHsd
B6N-TyrC
BomTac

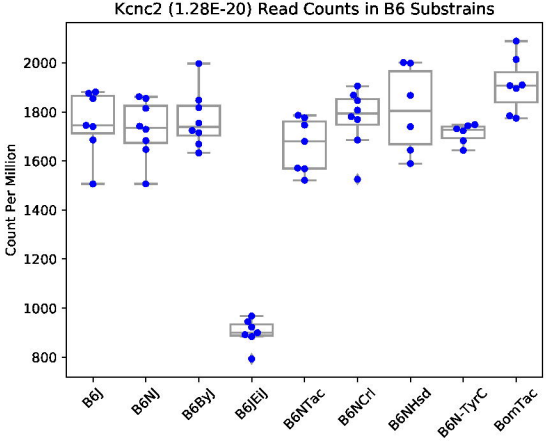

Ide (1.86E-21) Read Counts in B6 Substrains
Count Per Million
1600
1400
1200
1000
800
600
B6J
B6NJ
B6ByJ
B6JEiJ
B6NTac
B6NCrl
B6NHsd
B6N-TyrC
BomTac

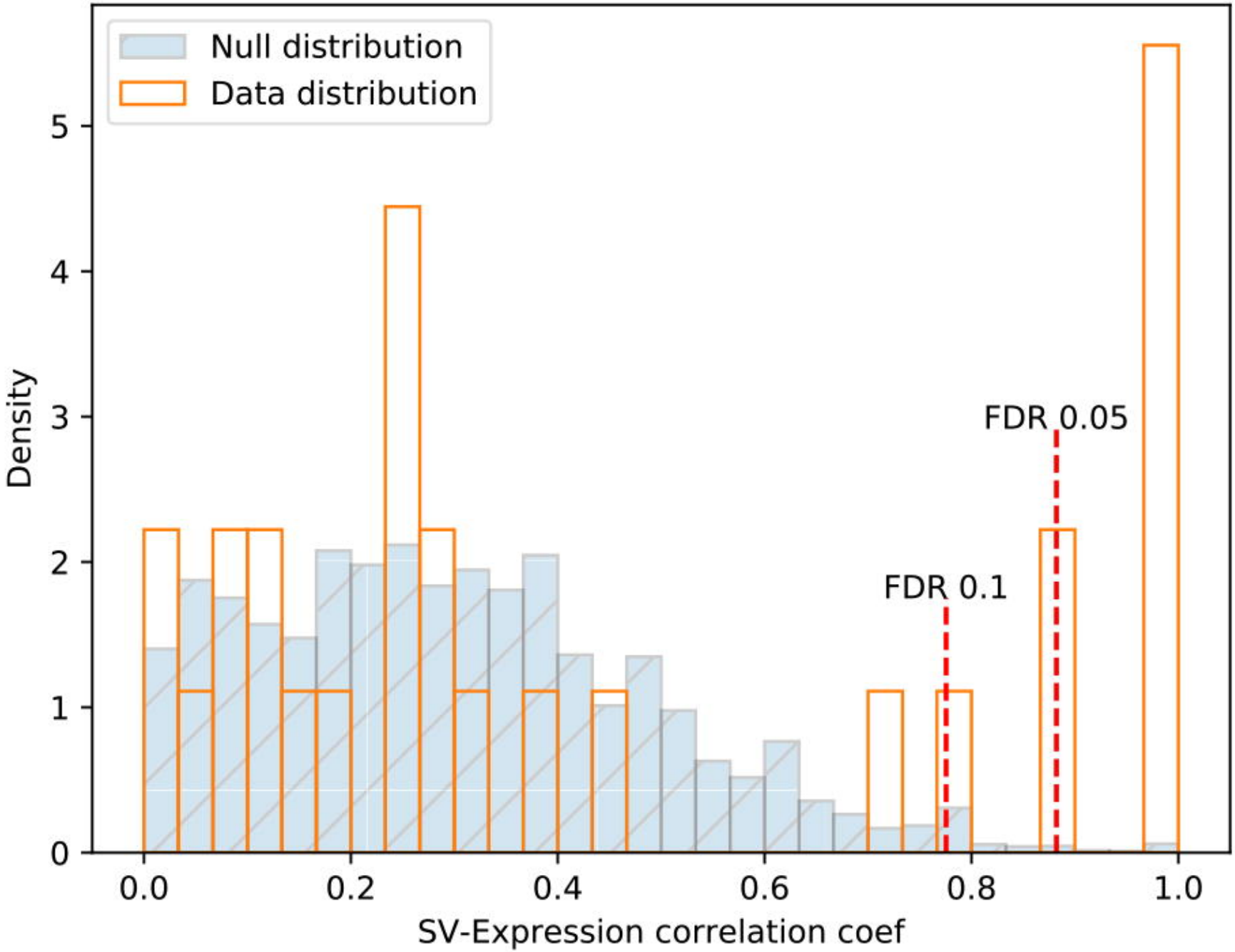

Fgfbp3 (1.14E-08) Read Counts in B6 Substrains
Count Per Million
140
120
100
80
60
B6J
B6NJ
B6BYJ
B6JEIJ
B6NTac
B6NCrl
B6NHsd
B6N-TyrC
BomTac

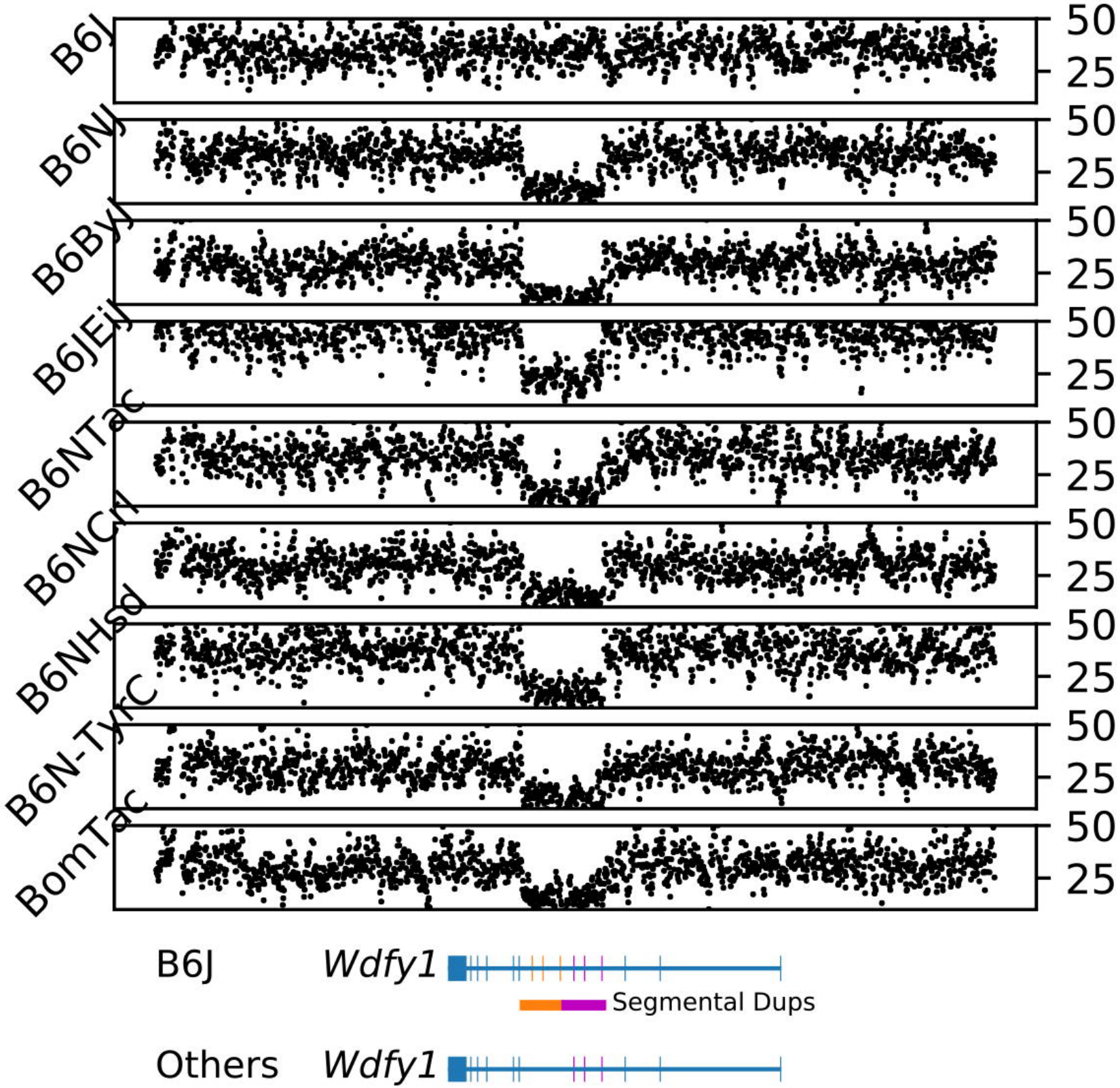

Btaf1 (2.09E-12) Read Counts in B6 Substrains
Count Per Million
900
800
700
600
500
400
B6J
B6NJ
B6BYJ
B6JEIJ
B6NTac
B6NCrl
B6NHsd
B6N-TyrC
BomTac

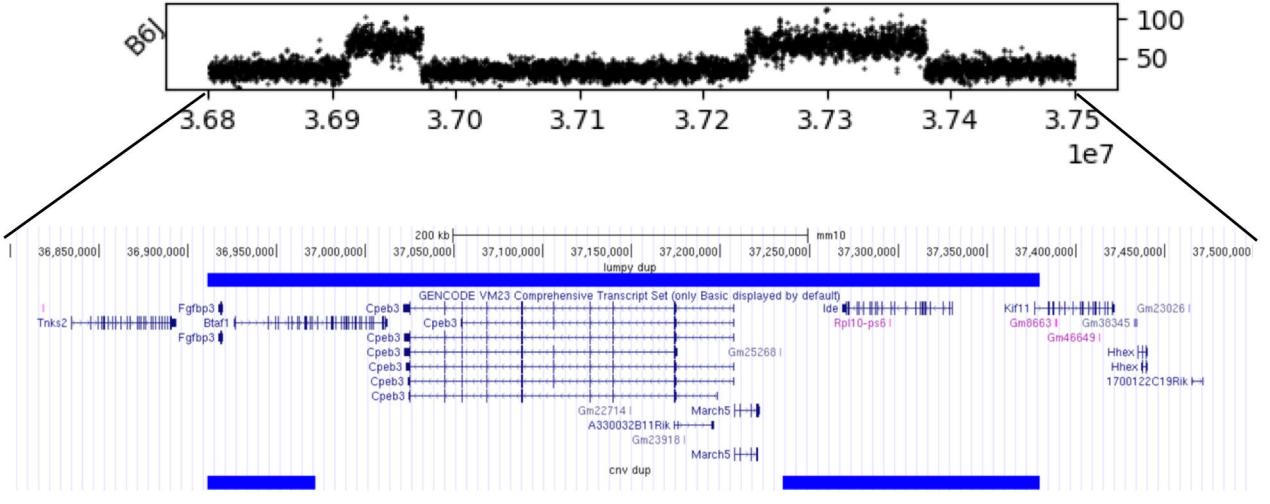

Nnt (1.62E-22) Read Counts in B6 Substrains
Count Per Million
700
600
500
400
300
B6J
B6NJ
B6ByJ
B6JEiJ
B6NTac
B6NCrl
B6NHsd
B6N-TyrC
BomTac


Gabra2 (1.18e-14) Read Counts in B6 Substrains
Count Per Million
4000
3500
3000
2500
2000
1500
1000
B6J
B6NJ
B6ByJ
B6JEiJ
B6NTac
B6NCrl
B6NHsd
B6N-TyrC
BomTac


Kcnc2 (1.28E-20) Read Counts in B6 Substrains
Count Per Million
2000
1800
1600
1400
1200
1000
800
B6J
B6NJ
B6ByJ
B6JEiJ
B6NTac
B6NCrl
B6NHsd
B6N-TyrC
BomTac


Null distribution
Data distribution
FDR 0.05
FDR 0.1
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
SV-Expression correlation coef


B6J
B6NJ
B6ByJ
B6JEiJ
B6NTac
B6NCrl
B6NHsd
B6N-TyrC
BomTac
50
25
B6J
Wdfy1
Segmental Dups
Others
Wdfy1


B6J
100
50
3.68
3.69
3.70
3.71
3.72
3.73
3.74
3.75
1e7
200 kb
mm10
36,850,000
36,900,000
36,950,000
37,000,000
37,050,000
37,100,000
37,150,000
37,200,000
37,250,000
37,300,000
37,350,000
37,400,000
37,450,000
37,500,000
lumpy dup
GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)
Tnks2
Fgfbp3
Btaf1
Fgfbp3
Cpeb3
Cpeb3
Cpeb3
Cpeb3
Cpeb3
Cpeb3
Cpeb3
Gm25268
Gm22714
March5
A330032B11Rik
Gm23918
March5
Ide
Rpl10-ps6
Kif11
Gm23026
Gm8663
Gm38345
Gm46649
Hhex
Hhex
1700122C19Rik
cnv dup

Gene	Chromosome	Position	Category	B6J	B6NJ	B6ByJ	B6JEiJ	B6NTac	B6NCrl	B6NHsd	B6NTyrC	B6JBomTac
ide	19	36803402	INDEL	0/0	0/2	0/1	0/1	0/1	0/2	0/1	0/2	0/1
ide	19	36812936	SNP	TG	GG	GG	GG	GG	GG	GG	GG	GG
ide	19	36886323	INDEL	0/1	1/1	1/1	1/1	1/1	1/1	1/1	0/0	1/1
ide	19	36911371	SV	DUP(468188bp)	NO	NO	NO	NO	NO	NO	NO	NO
ide	19	37035314	SNP	AA	GA	GA	GA	GA	GA	GA	GA	GA
ide	19	37183862	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
ide	19	37204309	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
ide	19	37248620	INDEL	0/2	0/0	0/0	0/0	0/0	0/0	0/1	0/0	0/0
ide	19	37342195	INDEL	1/1	./.	./.	./.	./.	./.	./.	./.	./.
ide	19	37379512	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
ide	19	37460195	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
ide	19	37488595	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
ide	19	37530549	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
ide	19	37556701	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
ide	19	37565349	INDEL	0/1	1/1	1/1	1/2	0/2	0/2	0/0	0/0	2/2
ide	19	37645147	INDEL	0/2	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/0
ide	19	37786339	INDEL	0/0	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
ide	19	37791925	INDEL	0/2	1/1	1/1	0/1	1/1	1/1	1/1	1/1	1/1
ide	19	37791928	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791931	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791941	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791945	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791947	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791948	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791950	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791952	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791953	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791960	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37791964	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
ide	19	37803861	INDEL	./.	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
ide	19	37817057	INDEL	0/1	0/0	0/2	0/2	1/1	0/2	0/0	0/2	0/0

Gene	Chromosome	Position	Category	B6J	B6NJ	B6ByJ	B6JEiJ	B6NTac	B6NCrl	B6NHsd	B6NTyrC	B6JBomTac
nnt	13	118847209	STR	0/2	0/1	0/0	0/0	0/1	0/0	0/1	0/1	0/0
nnt	13	118847256	INDEL	0/2	0/1	0/1	0/0	0/1	0/1	0/1	0/1	0/1
nnt	13	118859648	INDEL	0/2	0/0	0/0	0/0	1/2	0/0	0/0	0/0	0/0
nnt	13	118933907	SNP	GG	TT	TT	TT	GT	TT	TT	TT	TT
nnt	13	118957164	STR	0/1	0/0	0/0	2/2	0/0	0/0	0/0	0/0	0/0
nnt	13	119048487	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119067682	SNP	CA	AA	AA	AA	AA	AA	AA	AA	AA
nnt	13	119228909	STR	0/3	0/2	0/2	0/0	0/0	0/0	0/2	0/1	0/0
nnt	13	119392610	INDEL	0/1	0/0	0/0	0/0	0/2	1/1	0/0	0/0	0/0
nnt	13	119488749	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119489463	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119489765	SNP	GG	AG	AG	AG	AG	AG	AG	AG	AG
nnt	13	119595753	SNP	GC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119595994	SNP	CA	AA	AA	AA	AA	AA	AA	AA	AA
nnt	13	119596118	SNP	AA	GA	GA	GA	GA	GA	GA	GG	GA
nnt	13	119596370	SNP	CT	TT	TT	TT	TT	TT	TT	TT	TT
nnt	13	119596371	SNP	TG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119596676	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119598848	SNP	GT	TT	TT	TT	TT	TT	TT	TT	TT
nnt	13	119598892	SNP	GG	TG	TG	TG	TG	TG	TG	TG	TG
nnt	13	119599608	SNP	TT	AT	AT	AT	AT	AT	AT	AA	AT
nnt	13	119599716	SNP	GC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119599940	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
nnt	13	119600075	SNP	GT	TT	TT	TT	TT	TT	TT	TT	TT
nnt	13	119601965	SNP	GC	CC	CC	CC	CC	CC	CC	CC	CC
nnt	13	119602677	SNP	TG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119602739	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119602794	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119603540	INDEL	0/0	./.	./.	0/1	./.	1/1	0/1	1/1	1/1
nnt	13	119603562	SNP	CC	CT	CT	CT	CT	TT	CT	TT	CT
nnt	13	119603572	SNP	TT	TC	TC	TC	TC	TC	TC	CC	TC
nnt	13	119603588	SNP	CC	TC	TC	TC	TC	TC	TC	TC	TC
nnt	13	119603591	SNP	TT	AT	AT	AT	AT	AT	AT	AT	AT
nnt	13	119603604	SNP	CC	TC	TC	TC	TC	TC	TC	TC	TC
nnt	13	119603613	SNP	AA	CA	CA	CA	CA	CA	CA	CA	CA
nnt	13	119613570	SNP	CG	GG	GG	GG	GG	GG	GG	GG	GG
nnt	13	119613734	INDEL	0/2	2/3	0/1	1/3	1/2	1/3	1/2	2/3	1/3
nnt	13	119615372	INDEL	0/1	1/3	1/2	2/3	1/2	0/3	1/3	1/3	2/3
nnt	13	119624238	INDEL	1/1	./.	./.	./.	./.	./.	./.	1/2	./.
nnt	13	119685276	STR	0/0	2/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2
nnt	13	119685297	INDEL	./.	1/2	1/2	1/2	1/2	1/2	1/2	1/1	1/2
nnt	13	119685317	INDEL	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
nnt	13	119685321	INDEL	0/0	0/2	0/2	0/2	0/1	0/2	0/1	0/1	0/2
nnt	13	119686806	STR	0/0	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
nnt	13	119686819	INDEL	0/0	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
nnt	13	119721309	INDEL	0/1	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0
nnt	13	119882818	SNP	GG	TG	TG	TG	TG	TG	TG	TG	TG
nnt	13	119882820	SNP	AA	TA	TA	TA	TA	TA	TA	TA	TA
nnt	13	119905572	STR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0

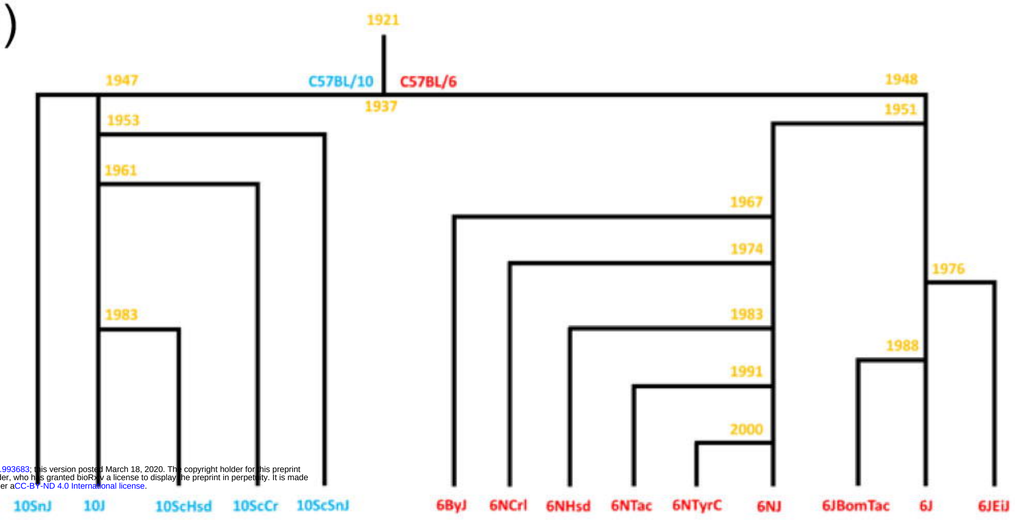
Gene	Chromosome	Position	Category	B6J	B6NJ	B6ByJ	B6JEiJ	B6NTac	B6NCrl	B6NHsd	B6NTyrC	B6JBomTac
gabra2	5	70509237	SNP	CC	GG	GG	GG	GG	GG	GG	GG	GG
gabra2	5	70514395	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
gabra2	5	70721116	INDEL	1/2	0/1	0/1	0/0	0/0	0/1	0/0	0/1	0/1
gabra2	5	70726237	INDEL	0/0	1/2	1/1	1/2	0/1	0/1	1/1	./.	1/2
gabra2	5	70737943	STR	2/2	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
gabra2	5	70774999	STR	0/2	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
gabra2	5	70786971	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
gabra2	5	70805456	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
gabra2	5	70925745	SNP	CT	CC	CC	TT	CC	CC	CC	TT	TT
gabra2	5	71004633	SNP	GT	TT	TT	TT	TT	TT	TT	TT	TT
gabra2	5	71014638	INDEL	0/0	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
gabra2	5	71091704	INDEL	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
gabra2	5	71126893	SNP	CT	TT	TT	TT	TT	TT	TT	TT	TT
gabra2	5	71157413	SNP	AC	CC	CC	CC	CC	CC	CC	CC	CC
gabra2	5	71215173	INDEL	0/1	./.	0/0	./.	./.	1/1	0/2	0/0	./.
gabra2	5	71218230	STR	0/2	0/0	0/0	0/0	0/1	0/0	0/1	0/0	0/0
gabra2	5	71274939	INDEL	1/2	0/2	0/0	0/0	0/0	0/1	0/1	0/1	0/1
gabra2	5	71342379	SNP	AG	GG	GG	GG	GG	GG	GG	GG	GG
gabra2	5	71419173	SNP	TC	CC	CC	CC	CC	CC	CC	CC	CC
gabra2	5	71427935	INDEL	0/2	0/0	0/0	0/0	0/1	0/0	0/0	0/0	0/1
gabra2	5	71447810	SNP	GA	AA	AA	AA	AA	AA	AA	AA	AA
gabra2	5	71493034	STR	0/0	2/2	2/2	2/2	0/2	2/2	2/2	2/2	0/2
gabra2	5	71586254	INDEL	0/1	1/1	1/1	./.	1/1	1/1	1/1	./.	./.

Gene	Chromosome	Position	Category	B6J	B6NJ	B6ByJ	B6JEiJ	B6NTac	B6NCrl	B6NHsd	B6NTyrC	B6JBomTac
kcnc2	10	111953664	INDEL	0/0	0/0	0/0	2/2	0/1	0/0	0/0	0/1	0/0
kcnc2	10	112013254	SNP	TT	TT	TT	GT	TT	TT	TT	TT	TT
kcnc2	10	112050694	SNP	GG	GG	GG	TT	GG	GG	GG	GG	GG
kcnc2	10	112052058	INDEL	0/1	0/1	0/1	0/0	0/1	0/1	0/1	0/1	0/1
kcnc2	10	112189298	SNP	TT	TT	TT	CC	TT	TT	TT	TT	TT
kcnc2	10	112201547	INDEL	0/1	0/1	0/1	./.	0/2	0/2	0/0	0/0	0/2
kcnc2	10	112381442	INDEL	0/2	1/2	0/0	1/1	0/2	0/0	0/1	1/2	0/0
kcnc2	10	112381448	INDEL	0/2	1/2	0/0	1/1	0/2	0/0	0/1	0/2	0/0
kcnc2	10	112385927	INDEL	0/0	0/1	0/0	0/2	0/1	0/1	0/1	0/1	0/0
kcnc2	10	112562168	INDEL	0/1	0/2	0/1	./.	0/1	0/1	0/0	0/1	0/0
kcnc2	10	112762924	SNP	GG	GG	GG	TT	GG	GG	GG	GG	GG
kcnc2	10	112783639	SNP	TT	TT	TT	GT	TT	TT	TT	TT	TT
kcnc2	10	112787091	STR	0/0	0/0	0/0	2/2	0/0	0/1	1/1	0/1	0/0
kcnc2	10	112819780	SNP	AA	AA	AA	CA	AA	AA	AA	AA	AA
kcnc2	10	112824134	SNP	AA	AA	AA	CA	AA	AA	AA	AA	AA
kcnc2	10	112834213	SNP	TT	TT	TT	CC	TT	TT	TT	TT	TT
kcnc2	10	112861890	INDEL	0/0	0/1	0/2	1/2	0/1	0/1	0/0	0/0	0/0
kcnc2	10	112898865	INDEL	0/0	0/1	0/1	1/2	0/0	1/1	0/0	0/2	1/1
kcnc2	10	112964630	INDEL	0/2	0/1	0/0	1/2	0/0	0/0	0/0	0/1	0/1

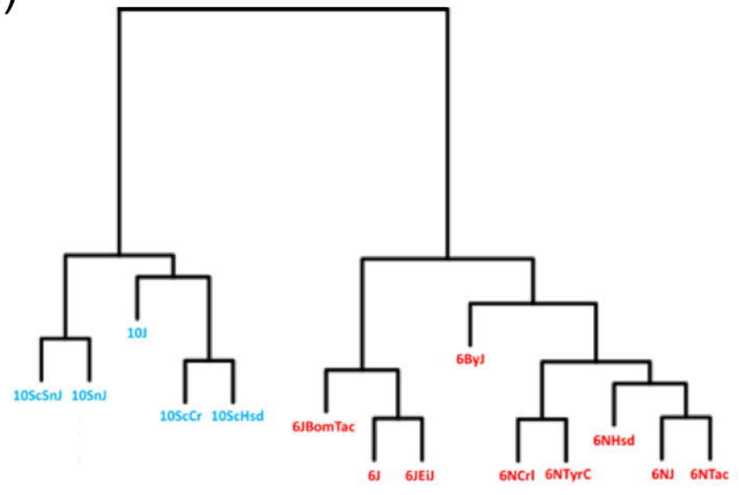
			DNA Sequencing				RNA Sequencing	
			Beckman		Novogene		UCSD	
Strain	Vendor	Strain ID	Reads	Coverage	Reads	Coverage	Average Reads	# Samples
C57BL/6NTac	Taconic	B6	15,507,036,875	5.64	96,765,063,000	35.19	13,459,345	8
C57BL/6NJ	JAX	#005304	15,204,121,000	5.53	96,364,702,800	35.04	15,767,477	9
C57BL/6NHsd	Harlan	#044	11,082,436,750	4.03	92,297,327,532	33.56	14,656,480	9
C57BL/6NCrl	Charles River	#027	13,203,215,375	4.80	87,049,155,000	31.65	19,408,390	8
C57BL/6JEiJ	JAX	#000924	16,686,632,375	6.07	117,271,713,036	42.64	16,334,723	8
C57BL/6JBomTac	Taconic	B6JBom	10,119,288,500	3.68	88,797,126,300	32.29	17,340,684	7
C57BL/6J	JAX	#000664	14,953,129,750	5.44	94,211,981,700	34.26	18,555,108	7
C57BL/6ByJ	JAX	#001139	17,127,176,375	6.23	84,730,839,600	30.81	15,552,591	7
B6N/TyrC/BrdCrlCrl	Charles River	#493	11,001,132,750	4.00	90,882,123,900	33.05	15,139,469	8
C57BL/10SnJ	JAX	#000666	14,898,322,625	5.42	82,193,982,600	29.89	14,012,862	7
C57BL/10ScSnJ	JAX	#000476	12,804,319,875	4.66	82,380,679,200	29.96	18,947,235	8
C57BL/10ScNHsd	Harlan	#046	14,512,802,500	5.28	93,773,651,400	34.10	12,453,407	11
C57BL/10ScCr	JAX	#003752	15,014,884,375	5.46	83,468,156,700	30.35	16,384,792	8
C57BL/10J	JAX	#000665	15,422,674,250	5.61	89,515,708,500	32.55	17,550,456	7

Gene	Region	SV type	Caller	Strain genotype	Corr Coef
Wdfy1	chr1:79715500-79729900	DEL	CNVnator	0;1;1;1;1;1;1;1;1	0.9855
Ide	chr19:37235100-37379500	DUP	CNVnator	1;0;0;0;0;0;0;0;0	0.9996
Fgfbp3	chr19:36911400-36971900	DUP	CNVnator	1;0;0;0;0;0;0;0;0	0.9898
Btaf1	chr19:36911400-36971900	DUP	CNVnator	1;0;0;0;0;0;0;0;0	0.8821

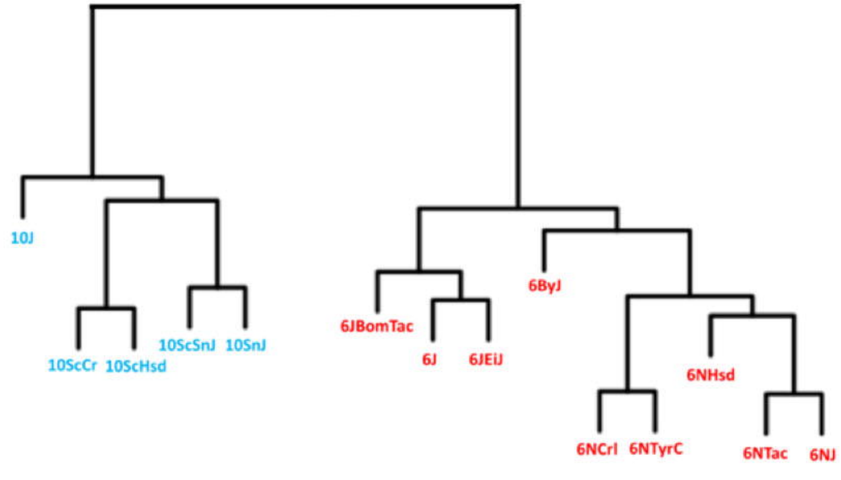
(a)



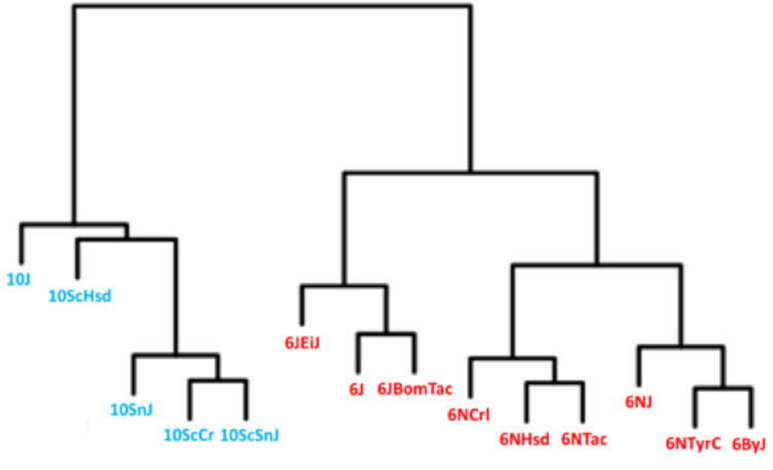
(b)



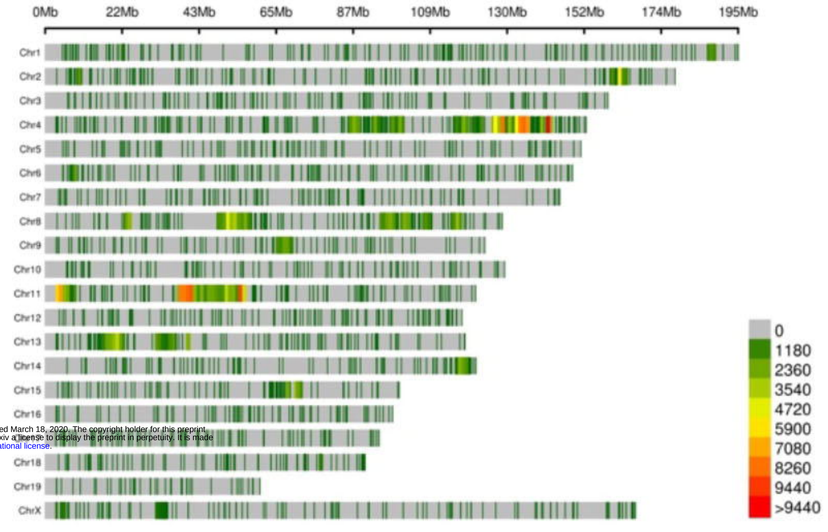
(c)



(d)

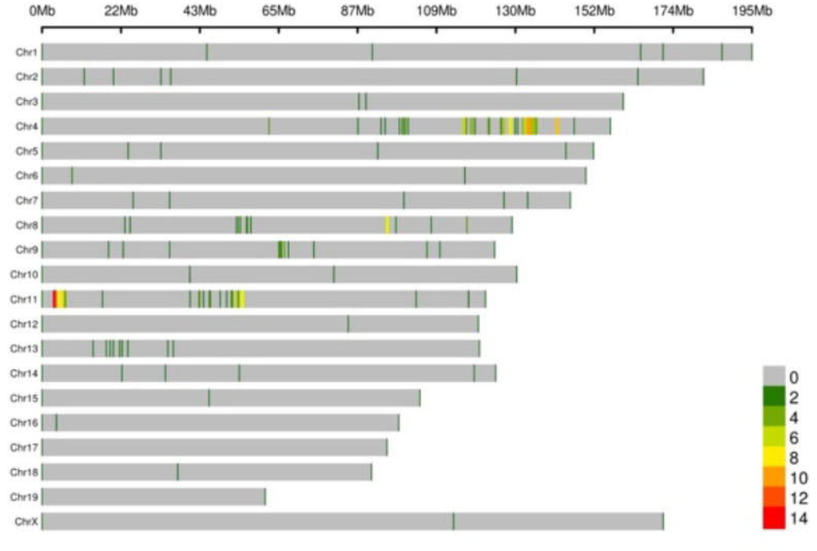


(a)

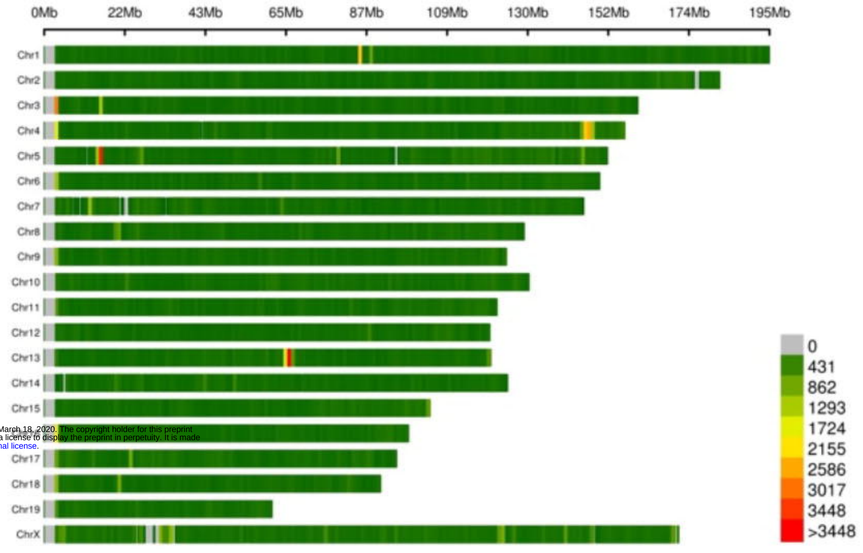


(b)

(c)

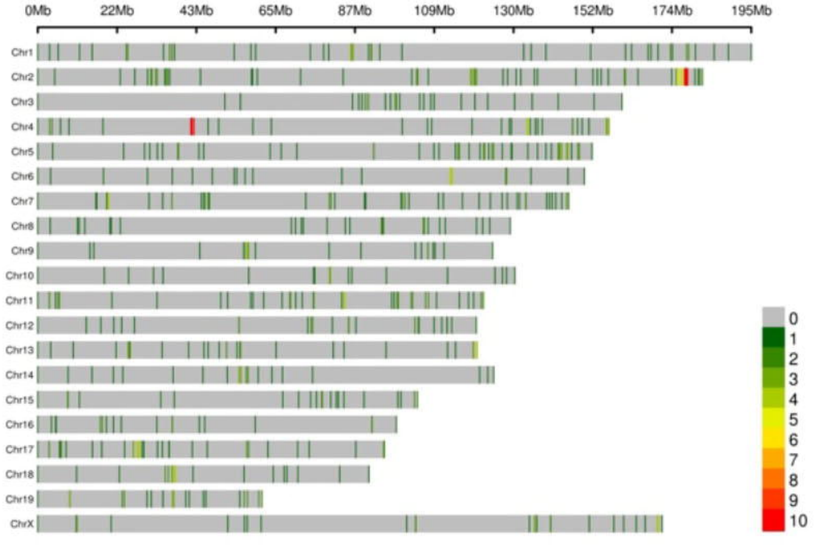


(a)

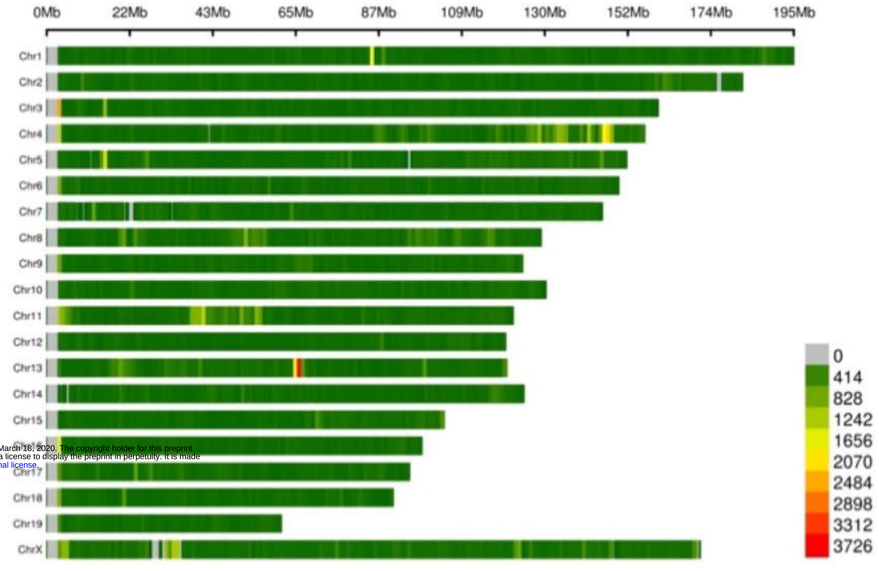


(b)

(c)



(a)



(b)

(c)