A Murine Database of Structural Variants Enables the Genetic Architecture of a Spontaneous Murine Lymphoma to be Characterized

Wenlong Ren[1], Zhuoqing Fang[1], Egor Dolzhenko[2,] Christopher T. Saunders[2], Zhuanfen Cheng[1], Victoria Popic[3] and Gary Peltz[1]

[1]Department of Anesthesia, Pain and Perioperative Medicine, Stanford University School of Medicine, Stanford CA 94305; [2]Pacific Biosciences, Menlo Park, CA; and [3]Broad Institute of MIT and Harvard, Cambridge, MA

W.R., V.P., Z.F, Z.C. and G.P. declare no conflict of interest. C.T. and E.D. are employees and shareholders of Pacific Biosciences.

[*]Address correspondence to: gpeltz@stanford.edu

Abstract

A more complete map of the pattern of genetic variation among inbred mouse strains is essential for characterizing the genetic architecture of the many available mouse genetic models of important biomedical traits. Although structural variants (SVs) are a major component of genetic variation, they have not been adequately characterized among inbred strains due to methodological limitations. To address this, we generated high-quality long-read sequencing data for 40 inbred strains; and designed a pipeline to optimally identify and validate different types of SVs. This generated a database for 40 inbred strains with 573,191SVs, which included 10,815 duplications and 2,115 inversions, that also has 70 million SNPs and 7.5 million insertions/deletions. Analysis of this SV database led to the discovery of a novel bi-genic model for susceptibility to a B cell lymphoma that spontaneously develops in SJL mice, which was initially described 55 years ago. The first genetic factor is a previously identified endogenous retrovirus encoded protein that stimulates CD4 T cells to produce the cytokines required for lymphoma growth. The second genetic factor is a newly found deletion SV, which ablates a protein whose promotes B lymphoma development in SJL mice. Characterizing the genetic architecture of SJL lymphoma susceptibility could provide new insight into the pathogenesis of a human lymphoma that has similarities with this murine lymphoma.

Abstract word count: 215 words

Abbreviations: DEL, deletions; DUP, duplications; GNN, graph neural network; HL, Hodgkin Lymphoma; HRS, Hodgkin Reed Sternberg; IGV, integrated genome viewer; INS, insertions; INV, inversions; KO, knockout; NGS, next generation sequencing; LRS, long read sequencing; RAG2, Recombination activating gene 2 protein; SRS, short read sequencing; SV, structural variants; TADs, topologically associating domains; TB, terabases; TR, tandem repeats; UTR, untranslated region.

The laboratory mouse has been the premier model organism for biomedical research, and the large number of phenotypically well-characterized inbred strains has enabled genetic factors for important biomedical traits to be identified using murine genetic models [1-3]. However, mouse genetic discovery is critically dependent upon having a complete map of genetic variation among these strains. While SNPs and small indels have been extensively characterized in mouse strains [4-6], the lack of a comprehensive database of structural variation has limited our ability to fully analyze and interpret the mouse genome. Prior efforts to characterize murine structural variants (SVs) (i.e., genomic alterations >50 bp in size) have included only a few strains [7] and relied on short-read sequencing (SRS) [8,9], which has a limited ability to detect SVs in repetitive regions of the genome. Long-read sequencing (LRS) platforms, which produce reads of length ≥20kb, have improved our ability to identify SVs, especially in difficult genomic regions [10-12]. LRS has doubled the estimated number of SVs in the human genome versus prior SRS estimates [10]. Our prior analysis of six strains with LRS revealed that: (i) SVs are very abundant (4.8 per gene), which indicates that they are likely to impact genetic traits; and (ii) as in human studies, the SVs previously identified using SRS[13] accounted for only 25% of those identified by our LRS analysis [14]. A recent analysis of SVs in 14 inbred strains used LRS [15]; but this analysis primarily reported only deletions and insertions. Sequencing and alignment artifacts, along with a heavy reliance on heuristics limit the ability of many existing programs to accurately identify additional types of SVs. We found that characterization of duplications or inversions was particularly problematic, even when murine LRS was analyzed [14]. To comprehensively characterize SVs across the mouse genome, we sequenced 40 inbred mouse strains using high-accuracy PacBio HiFi long reads. Simulations were used to evaluate the performance of several SV detection methods for identifying different types and sizes of murine SVs. In addition to state-of-the-art alignment [16,17,18] and assembly-based [19] heuristic methods, we also evaluated a recently developed deep-learning method (Cue [20]) for SV detection. Based upon these simulation results, we designed a custom pipeline to characterize a broader set of SVs from these 40 strains, which included deletions (DEL), insertions (INS), duplications (DUP), and inversions (INV) of varying size. This approach generated a comprehensive database of 573,191SVs among 40 inbred strains that includes a significant number of novel DUPs, INVs and large INS.

The utility of this SV database was demonstrated by identifying a genetic susceptibility factor for an unusual lymphoma that spontaneously appears in SJL mice [21-23]. These tumors originate in B cell germinal centers and are of interest because some of their features resemble those seen in one type of human non-Hodgkins lymphoma [24]. SJL lymphomas contain activated T cells, which produce the cytokines required for lymphoma propagation *in vitro* [25]. One susceptibility factor was identified as an endogenous retrovirus (***Mtv29***) in the SJL genome that encodes a tumor associated antigen (**vSAg29**) [26,27], which stimulates a subset of CD4 T cells [28] to produce the cytokines required for lymphoma development [29]. However, a second genetic susceptibility factor also must contribute because: (i) a strain (MA/My) that expresses vSAg29 does not develop (or has a very low incidence of) lymphoma [30]; and (ii) analyses of SJL intercross progeny indicated that one autosomal dominant genetic factor is present in multiple other strains, which suppresses lymphoma development [31,32]. Although this tumor suppressor had not been identified in the 55 years since this lymphoma was described in 1969, our AI mouse genetic discovery pipeline [33] identified this second genetic factor as an SJL-unique SV that ablates a tumor suppressor.

Results

Genomic sequencing and SNP/INDEL identification. Genomic sequencing was performed using a PacBio Revio instrument with the HiFi system, which achieves a median read accuracy reaching 99.9% [34], to generate LRS from 40 inbred strains. A total of 3.54 TB of sequence was generated, with an average of 88.5 GB (30x coverage) per strain (**Table S1**). Since they are commonly used in genetic models, we separately report on SNPs, Indels and SVs present the 35 classic inbred strains and those present in all 39 sequenced strains, which includes the four wild derived strains (CAST, Spret, MOLF, WSB). Using DeepVariant [35], 70,051,144 SNP sites and 7,540,144 sites with insertions or deletions (INDELs) were identified in the 39 strains (vs the C57BL/6 reference genome GRCm39). Consistent with our previous finding that the four wild-derived strains had patterns of genetic variation that were distinct from the 35 commonly used classical inbred strains [36], most of the SNPs (64.7M or 92%) and INDELs (6.8 M or 91%) were present in the 4 wild-derived strains, which also had most of the SNP or INDEL alleles that were present in only one strain. There were 21,331,225 SNP and 2,290,861 INDEL sites in the 35 classical inbred strains; the alleles in 5,286,543 SNPs and 696,031 INDELs were only present in the classic inbred strains; and 4.7M SNPs and 0.52M INDELs had alleles that were present in only a single classic inbred strain (**Figs. S1-S4**). To ensure that SNP alleles were correctly identified, we examined the output of our AI mouse genetic analysis pipeline [33] using the new SNP and INDEL databases. The previously identified causative genetic factors for four traits [14,33,37,38] (two of which were caused by SNP alleles and two were caused by INDELs) were identified by the AI using the new SNP and INDEL databases (**Figs. S5-S6**). These results indicate that the SNP and INDEL alleles in the new database are congruent with those previously identified.

Assessing SV identification programs. Since SV analysis programs vary in their ability to identify different types of SVs [39,40], we performed an extensive set of simulations to examine the ability of five programs (Cue [20], Sawfish [18], Sniffles2 [16], PBSV [17], and Dipcall [19]) to detect SVs that were artificially inserted into the mouse genomic sequence. We separately assessed their ability to identify small (<1 kb) and large (1-100 kb) DELs, INSs, INVs and DUPs (**Table 1**). As described in Supplemental Note 1, the simulation results were used to design the SV pipeline used to assemble this SV database (**Fig. 1**). Whenever possible, we used a consensus-based SV identification strategy. However, when only a single caller achieved high recall in the simulations (i.e., large INSs), or when low agreement was observed among the programs when the actual data was analyzed (i.e., INVs) (**Figs. 2, S7**), the consensus strategy was replaced with one where the individual predictions obtained from one or more high recall tools were selected if they were validated by another program (VaPoR [41]) or by visual inspection. Overall, our results indicate that DEL and INS can be reliably identified, but improved methods for identification and validation of DUPs and INVs are needed. Nevertheless, an increased number of SVs, which includes DUPs and INVs, were identified by this pipeline.

Characterization of SVs. A total of 210,926 small (50 to 1000 bp) DEL SVs were identified in the 39 inbred strains (vs the C57BL/6 reference), and 80,195 of them were present in the 35 classical inbred strains (**Fig. 1**). There were 47,984 and 21,666 large (1 to 100 KB) DELs identified in the 39 or 35 inbred strains, respectively (**Figs. 2, S7).** Like the SNP alleles, most small (61%) and many large (53%) DELs are only present in the four wild-derived strains. There were 263,186 and 95,622 small INS SVs in the 39 or 35 inbred strains, respectively (**Figs. 2A, S7).** Dipcall identified 38,165 or 23,379 VaPoR-validated large INS in all 39 strains or in the 35 classical inbred strains, respectively. Consistent with the simulation results, 94% (38,165 out of 40,460) of the large INS identified by Dipcall were validated by VaPoR, and most large INS that could not be validated had sequence complexity that precluded VaPoR analysis (i.e., rated as not assessable). The 846 (or 275) small INVs and 1,269 (or 420) large INVs were validated by visual inspection using the integrated genome viewer (IGV) (**Fig. S8**). In addition, the 9,943 (or 5,420) small DUP and 872 (or 598) large DUP in all 39 (or 35 classical) inbred strains were also validated using VaPoR. Most of the small and large DUPs identified by the analysis programs could not be assessed by VaPoR (**Fig. 2B, S7**). Among the 573,191SVs identified, the number of INS (n=301,351) and DEL (n=258,910) was far greater than DUPs (n=10,815) or INVs (2,115) (**Fig. 3A**) in the 39 strains analyzed. Spret, CAST/Ei and MOLF had the largest total number of SVs and of strain-unique SVs (**Fig. 3B-C**). Three features of this analysis were notable. (i) Most small DUPs (54%) were within INS, while only 4.0% of the large DUPs were within INS (Figs. 2B, **S7**). (ii) Only a minority of the DUPs identified by Cue or PBSV could be validated by VaPoR (Figs. 2B, S7B); visual inspection confirmed that sequence complexity made it difficult to assess these SVs. (iii) Most SV alleles were shared by three or fewer inbred strains (**Figs. 3D, S9**).

The variant effect predictor (VEP) program [42] was used to analyze SV impact. Manual inspection of the high impact (protein coding) small INS (performed using the integrated genome viewer, IGV) revealed that 99.8% were true positives (2 false positives out of 801 analyzed), and 99.3% of the small DELs were true positives (5 false positive out of 705). Also, >99% of the large INS (1 false positive out of 518) and 98.6% of large DELs (4 false positives out of 293) were true positives (**Table S3**). To facilitate genetic discovery the genes with high impact SVs present in the 35 classical inbred strains are provided for small (n=705 in 654 unique genes) and large (n=293 in 270 unique genes) DELs, for small (n=801 SVs in 765 unique genes) and large (n=518 in 487 unique genes) INS, for all INVs, and for high impact DUPs in **supplemental data files 1-4.** In summary, the 2,305 high impact SVs with alleles in the 35 classical inbred strains provides a set of highly curated genetic variants that could impact a substantial number of biomedical traits.

Although we have only a limited ability to interpret the impact of SVs located within intergenic and noncoding regions, chromatin is compartmentalized into topologically associating domains (**TADs**), which are megabase-sized genomic segments that are separated by boundary regions [43,44]. TADs provide a regulatory scaffold for gene expression; they are linked with variation in gene expression because their structure facilitates enhancer-promoter interactions; and they insulate regions from the effect of other regulators [45]. TADs are created by the binding of a DNA sequence-specific transcription factor (CCCTC binding factor or **CTCF**) to its consensus binding element. A multi-subunit protein (cohesin) then binds to CTCF to form the 3D loop-like structures that alter gene transcription within a domain. Of note, we found 1,877 DELs that contain a CTCF recognition sequence (CCGCGNGGNGGCAG) among the 35 inbred strains (**Supplemental Data File 5**). Since CTCF binding is critical for TAD formation, some of the 1,877 SVs that delete a CTCF binding site could significantly alter chromatin structure; and by this mechanism could affect gene expression patterns and genetic traits.

Identification of the 2nd lymphoma susceptibility factor. Based upon the hypothesis that SJL mice lack a tumor suppressor, the AI mouse genetic pipeline [33] was used to identify lymphoma-associated (MeSH Term: D008223) genes with high impact SV alleles uniquely present in SJL mice (i.e., absent in the 34 other classic inbred strains). The AI identified a 1641 bp SJL-unique deletion in *high mobility group A1b* (*Hmga1b),* which ablated the exon encoding the entire Hmga1b protein, as the candidate genetic factor (**Fig. 4).** As discussed in *supplementary note 2*, Hmga1b was the only gene with an SJL-specific high impact DEL that was directly associated with lymphoma. HMGA family members are low molecular weight proteins that bind to AT-rich regions in nuclear chromatin where they exert positive or negative effects on gene expression by enabling other transcription factors to bind at nearby sites [46,47]. Two murine genes encode nearly identical HMGA1 proteins [48]. *Hmga1b* on chromosome 11 encodes a 107 amino acid protein. *Hmga1* on chromosome 17 generates two predominant mRNAs that produce: a 96 amino acid protein (whose sequence is identical to Hmga1b except 11 amino acids are deleted; or a 107 amino acid protein whose sequence is identical to Hgma1b, which arises by differential splicing (**Fig. 5A**). Analysis of *Hmga1* or *Hmga1b* mRNAs in SJL liver and spleen tissue by RT-PCR amplification indicated that the SJL mRNAs are identical to those present in other strains. In contrast, the level of expression of *Hmga1 (or Hmga1b)*-derived mRNAs in SJL thymus are greatly reduced relative to those in the thymus of other strains (**Fig. 5B**). Because the 3' UTRs of *Hmga1*- and *Hmga1b*-encoded mRNAs have a sequence difference at a corresponding site, transcript sequencing was used to identify the gene(s) that encoded these mRNAs in different C57BL/6 tissues. The mRNAs in spleen, liver, bone marrow, kidney and lymph nodes were *Hmga1* encoded, while the thymic mRNAs were *Hmga1b* encoded (**Figs. 5C, S10**). Hence, *Hmga1b* mRNA expression predominates in a tissue where lymphocyte development occurs. HMGA protein family members are strongly associated with leukemia and lymphoma in mice [49] and humans [47], which explains why reduced HGMA protein function in the SJL thymus contributes to lymphoma susceptibility (discussed below).

Discussion

This dataset represents the most comprehensive and carefully performed analysis of SVs in the mouse genome. High quality LRS with a high level of genome coverage for 40 inbred strains, and recently developed state of the art programs (Sawfish [18], Sniffles2 [16], CUE [20]) that were adapted for LRS analysis were used with earlier programs (PBSV, Dipcall) for SV identification. These programs were selected because simulations indicated that they performed optimally for identification of a certain type/size of SV. Large INS and all DUPs were validated by a separate program, and all INVs were individually validated by manual observation. Our results are consistent with prior observations that no single SV calling algorithm was optimal for detection of the different sizes and types of SVs, and that there can be a high level of divergence when the results of simulated and actual data are compared [40]. Our data also demonstrates that there is a critical need for improved methods for analysis of DUPs and INVs, which probably will require machine learning based programs. Nevertheless, since murine DUPs and INVs were particularly hard to identify, even when LRS was used [14]; the INV and DUP, along with the INS and DELs identified here could facilitate many genetic discoveries. DUPs and INVs are already known to contribute to Autism [50,51] and to impact brain function [50]. Recent analyses have indicated that segmental duplications occupy ~7% of the human genome [52]. One study limitation is that the currently used SV callers were designed and evaluated using human genomic sequences. Since mouse-specific features of the inbred strains' genomic sequences could be confounding, the existing programs may need to be further refined for mouse genomes. Another study limitation is that tandem repeats (**TRs**) – a type of SV with multiple repeats of short DNA sequence motifs, which are associated with multiple human diseases [53,54] - were not covered here; but they will be analyzed in a subsequent paper.

The discovery of the second genetic susceptibility factor for SJL lymphoma was facilitated by this murine SV database and the AI genetic discovery pipeline. A novel bi-genic model explains why B cell lymphomas uniquely develop in SJL mice, and this model is consistent with all available data about its pathogenesis. The first genetic factor is a protein (vSAg29) encoded by an endogenous retrovirus in SJL mice [26,27] that stimulates CD4 T cells to produce cytokines [28] required for B cell lymphoma development [29]. The second genetic factor is an SJL-unique SV generates an *Hmga1b* KO. This SJL-unique genetic factor was predicted by murine intercross experiments, which indicated that a genetic factor present in multiple other strains suppressed lymphoma development [31,32]. The *Hmga1b* deletion SV allele is also present in 3 wild derived strains (Cast/Ei, Spret, MOLF); and the endogenous retrovirus is present in another inbred strain (MaMy); but these strains do not spontaneously develop lymphomas. Hence, a unique combination of two genetic factors (vsAg29 and the *Hmga1b* deletion SV) is required to produce lymphomas in SJL mice. Hopefully, the genetic architecture of other murine genetic models of important biomedical traits will be uncovered using this SV database.

How could a SJL SV allele that ablates Hmga1b promote lymphoma development? While Hmga1 and Hmga1b have virtually identical protein sequences, our data demonstrates that only *Hmga1b* mRNAs are expressed in the thymus. A series of *in vitro* studies demonstrated that Hmga1 represses the expression of the Recombination activating gene 2 protein **(**RAG2) endonuclease [55], which plays a key role in lymphocyte development; but increased RAG expression can cause DNA damage and an increased risk for oncologic transformation [56]. *Hmga1* knockout (KO) mice have increased RAG2 activity in their spleens [55]. An *Hmga1* KO altered T cell development, increased B cell development, and caused the mice to develop B cell lymphomas and other hematopoietic malignancies [57,58]. SJL spleen and lymph node tissues have an increased number of germinal centers with IL-21 producing T follicular helper (TfH) cells, an increased level of IL-21 production; and SJL lymphoma development is IL-21 dependent [29]. Hence, the HMGA tumor suppressor function is (at least partly) mediated through repression of RAG2; loss of this repressor function in the SJL thymus will alter T cell development and could induce DNA translocations in other tissues. The abnormalities in thymic T cell development along with VsAg29-induced T cell proliferation explains why SJL (and *Hmga1* KO) mice have an increase in IL-21 producing TfH cells in their germinal centers, which leads to the development of a population of B cells with a high level of abnormal IgH rearrangements. By this mechanism, the activity of an expanded and abnormal population of TfH cells generates an IL-21 dependent B cell lymphoma in SJL mice [55,58].

Understanding the genetic architecture of SJL lymphoma susceptibility could provide new insight into the pathogenesis of certain types of human lymphomas. For example, SJL lymphomas have transcriptomic and phenotypic similarities with one type of non-Hodgkin human lymphoma: angioimmunoblastic T-cell lymphoma (AITL) [24,29]. Like SJL lymphomas, AITL develops late in life, is driven by IL-21-producing TfH cells [59], and it sometimes develops into a B cell lymphoma [60]. Analogous to the role of vSAg29 and the *Hmga1b* deletion allele in driving lymphomas in SJL mice, Epstein Barr virus (EBV) is detected in 66-86% of AITL patients [61,62] and mutations in genes encoding epigenetic modifiers have frequently been detected in AITL patients [24]. Also, IL-21 induces the expression of EBV latent membrane protein 1 (LMP1) in human B cells [63], which has been shown to provide survival signals for B cells [64] and can rescue transformed cells from apoptosis [65]. Given the similarities in their pathogenesis, additional studies on SJL lymphomas could provide new information about how infectious agents and host genetic factors jointly contribute to the pathogenesis of AITL, and possibly other types of lymphomas.

Data availability: The SNP, INDEL and SV data will be available at the Mouse Phenome Database (GenomeMUSter https://mpd.jax.org/genotypes). The LRS data files, will be available at the NCBI SRA database http://www.ncbi.nlm.nih.gov/bioproject/788143

Acknowledgements: This work was supported by NIH awards (1R01DC021133 and 1 R24 OD035408) to GP; and NHGRI Award R01HG012467 to VP. The funder had no role in the writing of this paper.

Supplemental Data Files

Supplemental Data File 1. Lists of high impact DELs present in all 39 or in the 35 classical inbred strains. The chromosome, starting and ending position, size, gene symbol, strains with the variant allele, and the predicted effect for each DEL are shown.

Supplemental Data File 2. Lists of high impact INS present in all 39 or in the 35 classical inbred strains. The chromosome, starting and ending position, size, gene symbol, strains with the variant allele, and the predicted effect for each INS are shown.

Supplemental Data File 3. Lists of INVs present in all 39 or in the 35 classical inbred strains. The chromosome, starting and ending position, size, gene symbol, strains with the variant allele, and the predicted effect for each INV are shown.

Supplemental Data File 4. Lists of high impact DUPs present in all 39 or in the 35 classical inbred strains. The chromosome, starting and ending position, size, gene symbol, strains with the variant allele, and the predicted effect for each DUP are shown.

Supplemental Data File 5. List of DELs present in 35 classical inbred strains with CTCF binding sites. The chromosome, starting and ending position of the CTCF recognition element and its sequence, size, gene symbol, strains with the variant allele for each DEL, the CTCF score (the strength of the motif match based on the position weight matrix), p-value and Q-value (p-value that is adjusted for the false discovery rate) are shown. These were determined using the CTCF (v0.99.11) R package in Bioconductor [66]. All predicted sites have a *P*-value that is below the recommended cutoff (1 x 10^{-6}).

References

1 Wang, J., Liao, G., Usuka, J. & Peltz, G. Computational Genetics: From Mouse to Man? *Trends in Genetics* **21**, 526-532 (2005).

2 Zheng, M., Dill, D. & Peltz, G. A better prognosis for genetic association studies in mice. *Trends Genet* **28**, 62-69 (2012). https://doi.org:10.1016/j.tig.2011.10.006

3 Fang, Z. & Peltz, G. 21st century mouse genetics is again at an inflection point. *Laboratory Animal* **In Press** (2024).

4 Peltz, G. *et al.* Next-Generation Computational Genetic Analysis: Multiple Complement Alleles Control Survival After Candida Albicans Infection *Infection and Immunity* **79**, 4472-4479 (2011).

5 Arslan, A. *et al.* High Throughput Computational Mouse Genetic Analysis *BioRxiv* **https://www.biorxiv.org/content/10.1101/2020.09.01.278465v2** (2020).

6 Ball, R. L. *et al.* GenomeMUSter mouse genetic variation service enables multitrait, multipopulation data integration and analysis. *Genome Res* **34**, 145-159 (2024). https://doi.org:10.1101/gr.278157.123

7 Mortazavi, M. *et al.* SNPs, short tandem repeats, and structural variants are responsible for differential gene expression across C57BL/6 and C57BL/10 substrains. *Cell Genom* **2** (2022). https://doi.org:10.1016/j.xgen.2022.100102

8 Doran, A. G. *et al.* Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome Biol* **17**, 167 (2016). https://doi.org:10.1186/s13059-016-1024-y

9 Yalcin, B. *et al.* Sequence-based characterization of structural variation in the mouse genome. *Nature* **477**, 326-329 (2011). https://doi.org:10.1038/nature10432

10 van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet* **34**, 666-681 (2018). https://doi.org:10.1016/j.tig.2018.05.008

11 Mantere, T., Kersten, S. & Hoischen, A. Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in genetics* **10**, 426 (2019). https://doi.org:10.3389/fgene.2019.00426

12 Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat Rev Genet* (2020). https://doi.org:10.1038/s41576-020-0236-x

13 Arslan, A. *et al.* Analysis of Structural Variation Among Inbred Mouse Strains Identifies Genetic Factors for Autism-Related Traits. *BioRxiv* **https://www.biorxiv.org/content/10.1101/2021.02.18.431863v1** (2022). https://doi.org: https://doi.org/10.1101/2021.02.18.431863

14 Arslan, A. *et al.* Analysis of Structural Variation Among Inbred Mouse Strains. *BMC Genommics* **24**, 97-109 (2023). https://doi.org:10.1186/s12864-023-09197-5

15 Ferraj, A. *et al.* Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements. *Cell Genom* **3**, 100291 (2023). https://doi.org:10.1016/j.xgen.2023.100291

16 Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* (2024). https://doi.org:10.1038/s41587-023-02024-y

17 Topfer, A. *PacBio structural variant calling and analysis tools*, <https://github.com/PacificBiosciences/pbsv> (2023).

18 Saunders, C. T. *et al.* Sawfish: Improving long-read structural variant discovery and genotyping with local haplotype modeling. *BioRxiv* **https://doi.org/10.1101/2024.08.19.608674;** (2024). https://doi.org:https://doi.org/10.1101/2024.08.19.608674;

19 Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**, 595-597 (2018). https://doi.org:10.1038/s41592-018-0054-7

20 Popic, V. *et al.* Cue: a deep-learning framework for structural variant discovery and genotyping. *Nat Methods* **20**, 559-568 (2023). https://doi.org:10.1038/s41592-023-01799-x

21 Murphy, E. D. Transplantation behavior of Hodgkin's-like reticulum cell neoplasms of strain SJL-J mice and results of tumor reinoculation. *J Natl Cancer Inst* **42**, 797-807 (1969).

22 Ponzio, N. M., Brown, P. H. & Thorbecke, G. J. Host-tumor interactions in the SJL lymphoma model. *Int Rev Immunol* **1**, 273-301 (1986).

23 Bonavida, B. Intimate host-tumor interaction in the spontaneous reticulum cell sarcoma of SJL/J mice: is it an exceptional case? *Surv Immunol Res* **4**, 271-282 (1985). https://doi.org:10.1007/BF02918735

24 Mhaidly, R. *et al.* New preclinical models for angioimmunoblastic T-cell lymphoma: filling the GAP. *Oncogenesis* **9**, 73 (2020). https://doi.org:10.1038/s41389-020-00259-x

25 Lasky, J. L. & Thorbecke, G. J. Characterization and growth factor requirements of SJL lymphomas. II. Interleukin 5 dependence of the in vitro cell line, cRCS-X, and influence of other cytokines. *Eur J Immunol* **19**, 365-371 (1989). https://doi.org:10.1002/eji.1830190222

26 Thomas, R. M. *et al.* Regulation of mouse mammary tumor virus env transcriptional activator initiated mammary tumor virus superantigen transcripts in lymphomas of SJL/J mice: role of Ikaros, demethylation, and chromatin structural change in the transcriptional activation of mammary tumor virus superantigen. *J Immunol* **170**, 218-227 (2003).

27 Sen, N. *et al.* META-controlled env-initiated transcripts encoding superantigens of murine Mtv29 and Mtv7 and their possible role in B cell lymphomagenesis. *J Immunol* **166**, 5422-5429 (2001).

28 Tsiagbe, V. K. *et al.* Syngeneic response to SJL follicular center B cell lymphoma (reticular cell sarcoma) cells is primarily in V beta 16+ CD4+ T cells. *J Immunol* **150**, 5519-5528 (1993).

29 Jain, S. *et al.* IL-21-driven neoplasms in SJL mice mimic some key features of human angioimmunoblastic T-cell lymphoma. *Am J Pathol* **185**, 3102-3114 (2015). https://doi.org:10.1016/j.ajpath.2015.07.021

30 Zhang, D. J., D'Eustachio, P. & Thorbecke, G. J. The Mtv29 gene encoding endogenous lymphoma superantigen in SJL mice, mapped to proximal chromosome 6. *Immunogenetics* **46**, 163-166 (1997).

31 Bubbers, J. E. Single-gene abrogation of spontaneous pleomorphic SJL/J mouse reticulum cell sarcoma expression. *J Natl Cancer Inst* **71**, 795-799 (1983).

32 Bubbers, J. E. Identification and linkage analyses of a gene, Rcs-1, suppressing spontaneous SJL/J lymphoma expression. *J Natl Cancer Inst* **72**, 441-446 (1984).

33 Fang, Z. & Peltz, G. An Automated Multi-Modal Graph-Based Pipeline for Mouse Genetic Discovery. *Bioinformatics* **38**, 3385-3394 (2022). https://doi.org:10.1093/bioinformatics/btac356

34 *Revio system: reveal more with accurate long-read sequencing at scale*, <https://www.pacb.com/revio/> (2024).

35 Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983-987 (2018). https://doi.org:10.1038/nbt.4235

36 Wang, M., Fang, Z., Yoo, B., Bejerano, G. & Peltz, G. The Effect of Population Structure on Murine Genome-Wide Association Studies. *Frontiers in genetics* **12**, 745361 (2021). https://doi.org:10.3389/fgene.2021.745361

37 Wang, M., Fang, Z., Yoo, B., Bejarano, G. & Peltz, G. The Effect of Population Structure on Murine Genome-Wide Association Studies. *Frontiers in genetics* **In press** (2021).

38 Zheng, M. *et al.* The Role of Abcb5 Alleles in Susceptibility to Haloperidol-Induced Toxicity in Mice and Humans *PLoS Medicine* **12**, e1001782 (2015). https://doi.org:10.1371/journal.pmed.100172

39 Liu, Y. H., Luo, C., Golding, S. G., Ioffe, J. B. & Zhou, X. M. Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data. *Nature communications* **15**, 2447 (2024). https://doi.org:10.1038/s41467-024-46614-z

40 Liu, Z. *et al.* Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol* **23**, 68 (2022). https://doi.org:10.1186/s13059-022-02636-8

41 Zhao, X., Weber, A. M. & Mills, R. E. A recurrence-based approach for validating structural variation using long-read sequencing technology. *Gigascience* **6**, 1-9 (2017). https://doi.org:10.1093/gigascience/gix061

42 Hunt, S. E. *et al.* Annotating and prioritizing genomic variants using the Ensembl Variant Effect Predictor-A tutorial. *Hum Mutat* **43**, 986-997 (2022). https://doi.org:10.1002/humu.24298

43 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012). https://doi.org:10.1038/nature11082

44 de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499-506 (2013). https://doi.org:10.1038/nature12753

45 Ibrahim, D. M. & Mundlos, S. Three-dimensional chromatin in disease: What holds us together and what drives us apart? *Curr Opin Cell Biol* **64**, 1-9 (2020). https://doi.org:10.1016/j.ceb.2020.01.003

46 Sumter, T. F. *et al.* The High Mobility Group A1 (HMGA1) Transcriptome in Cancer and Development. *Curr Mol Med* **16**, 353-393 (2016). https://doi.org:10.2174/1566524016666160316152147

47 Akade, E. & Jalilian, S. The role of high mobility group AT-hook 1 in viral infections: Implications for cancer pathogenesis. *Int J Biochem Cell Biol* **169**, 106532 (2024). https://doi.org:10.1016/j.biocel.2024.106532

48 Johnson, K. R., Cook, S. A. & Davisson, M. T. Chromosomal localization of the murine gene and two related sequences encoding high-mobility-group I and Y proteins. *Genomics* **12**, 503-509 (1992). https://doi.org:10.1016/0888-7543(92)90441-t

49 De Martino, M., Esposito, F. & Fusco, A. Critical role of the high mobility group A proteins in hematological malignancies. *Hematol Oncol* **40**, 2-10 (2022). https://doi.org:10.1002/hon.2934

50 Mullegama, S. V. *et al.* MBD5 haploinsufficiency is associated with sleep disturbance and disrupts circadian pathways common to Smith-Magenis and fragile X syndromes. *Eur J Hum Genet* **23**, 781-789 (2015). https://doi.org:10.1038/ejhg.2014.200

51 Blumenthal, I. *et al.* Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and multiplex autism families. *Am J Hum Genet* **94**, 870-883 (2014). https://doi.org:10.1016/j.ajhg.2014.05.004

52 Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022). https://doi.org:10.1126/science.abj6965

53 Paulson, H. Repeat expansion diseases. *Handbook of clinical neurology* **147**, 105-123 (2018). https://doi.org:10.1016/B978-0-444-63233-3.00009-9

54 Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19**, 286-298 (2018). https://doi.org:10.1038/nrg.2017.115

55 Battista, S. *et al.* High-mobility-group A1 (HMGA1) proteins down-regulate the expression of the recombination activating gene 2 (RAG2). *Biochem J* **389**, 91-97 (2005). https://doi.org:10.1042/BJ20041607

56 Teng, G. *et al.* RAG Represents a Widespread Threat to the Lymphocyte Genome. *Cell* **162**, 751-765 (2015). https://doi.org:10.1016/j.cell.2015.07.009

57 Battista, S. *et al.* Loss of Hmga1 gene function affects embryonic stem cell lympho-hematopoietic differentiation. *FASEB J* **17**, 1496-1498 (2003). https://doi.org:10.1096/fj.02-0977fje

58 Fedele, M. *et al.* Haploinsufficiency of the Hmga1 gene causes cardiac hypertrophy and myelo-lymphoproliferative disorders in mice. *Cancer Res* **66**, 2536-2543 (2006). https://doi.org:10.1158/0008-5472.CAN-05-1889

59 Dupuis, J. *et al.* Expression of CXCL13 by neoplastic cells in angioimmunoblastic T-cell lymphoma (AITL): a new diagnostic marker providing evidence that AITL derives from follicular helper T cells. *Am J Surg Pathol* **30**, 490-494 (2006). https://doi.org:10.1097/00000478-200604000-00009

60 Willenbrock, K., Brauninger, A. & Hansmann, M. L. Frequent occurrence of B-cell lymphomas in angioimmunoblastic T-cell lymphoma and proliferation of Epstein-Barr virus-infected cells in early cases. *Br J Haematol* **138**, 733-739 (2007). https://doi.org:10.1111/j.1365-2141.2007.06725.x

61 Tokunaga, T. *et al.* Retrospective analysis of prognostic factors for angioimmunoblastic T-cell lymphoma: a multicenter cooperative study in Japan. *Blood* **119**, 2837-2843 (2012). https://doi.org:10.1182/blood-2011-08-374371

62 Kawano, R. *et al.* Epstein-Barr virus genome level, T-cell clonality and the prognosis of angioimmunoblastic T-cell lymphoma. *Haematologica* **90**, 1192-1196 (2005).

63 Kis, L. L. *et al.* IL-21 imposes a type II EBV gene expression on type III and type I B cells by the repression of C- and activation of LMP-1-promoter. *Proc Natl Acad Sci U S A* **107**, 872-877 (2010). https://doi.org:10.1073/pnas.0912920107

64 Kapatai, G. & Murray, P. Contribution of the Epstein Barr virus to the molecular pathogenesis of Hodgkin lymphoma. *J Clin Pathol* **60**, 1342-1349 (2007). https://doi.org:10.1136/jcp.2007.050146

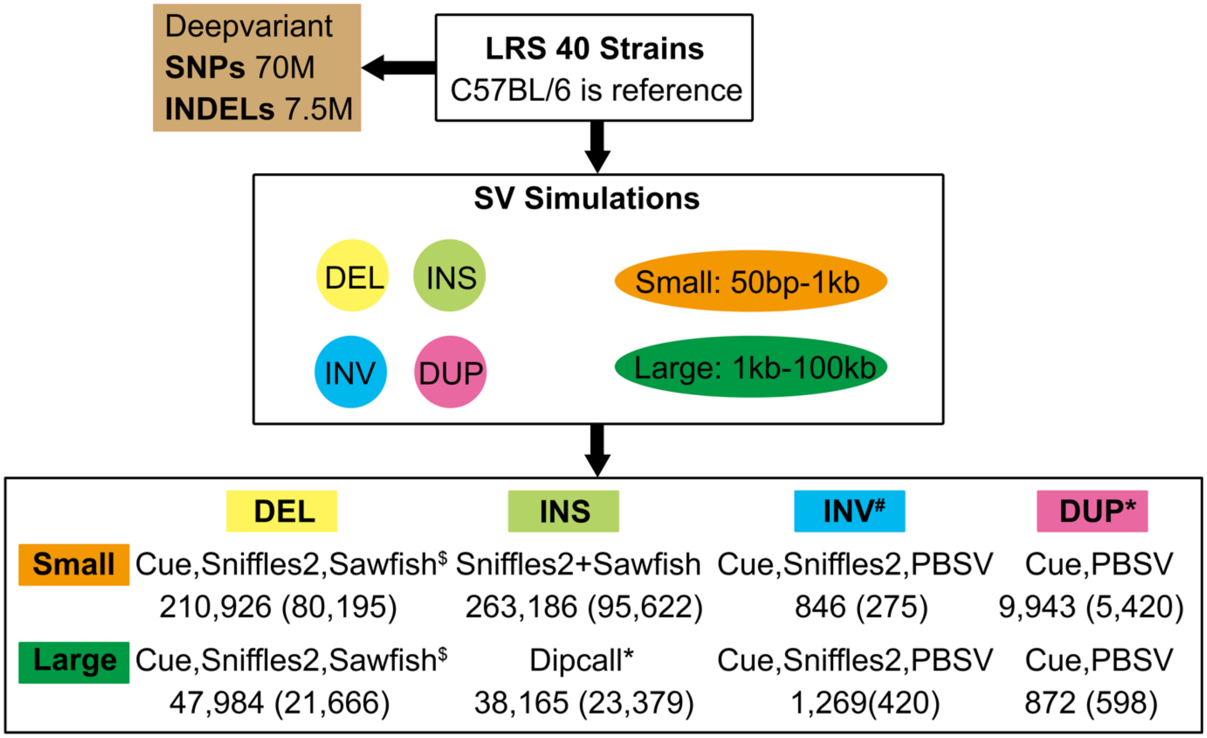
65 Mancao, C. & Hammerschmidt, W. Epstein-Barr virus latent membrane protein 2A is a B-cell receptor mimic and essential for B-cell survival. *Blood* **110**, 3715-3721 (2007). https://doi.org:10.1182/blood-2007-05-090142

66 Dozmorov, M. G. *et al.* CTCF: an R/bioconductor data package of human and mouse CTCF binding sites. *Bioinform Adv* **2**, vbac097 (2022). https://doi.org:10.1093/bioadv/vbac097

67 Ono, Y., Hamada, M. & Asai, K. PBSIM3: a simulator for all types of PacBio and ONT long reads. *NAR Genom Bioinform* **4**, lqac092 (2022). https://doi.org:10.1093/nargab/lqac092

Table 1. The results of simulations assessing the ability of Cue, Sniffles2, Sawfish, PBSV and Dipcall to identify various types of SVs are shown. For each simulation, 1000 SVs of a single type (DEL, INS, INV, DUP) with a restricted size range [small (50 - 1000 bp) or large (1 - 100 KB)] were artificially inserted at random positions into the GRCm39 reference sequence. The precision (P), recall (R) and F1 statistic were calculated based upon the ability of each program to detect each type of inserted SV. Cue, Sniffles2, Sawfish and PBSV analyzed Hifi LRS (30x coverage) that were aligned with minimap2. The simulations were performed using PBSIM3 [67]. The genomic sequences analyzed by Dipcall were assembled using HiFiasm. NA: not assessed. W* and $W^{\$}$: withdrawn for the reasons discussed in Supplemental Note 1.

		Cue	**Sniffles2**	**Sawfish**	**PBSV**	**Hifi-asm Dipcall**
		LRS alignment+DL	*LRS alignment*	*LRS alignment*	*LRS alignment*	*LRS assembly*
DEL (50-1000bp)	**P**	1	1	1	1	0.99
	R	0.994	0.999	1	1	0.702
	F1	0.996	0.9995	1	1	0.821
DEL (1-100kb)	**P**	1	0.991	0.989	0.997	1
	R	0.998	0.548	0.995	0.648	0.494
	F1	0.999	0.706	0.992	0.785	0.661
INS (50-1000bp)	**P**	NA	0.998	1	1	0.987
	R	NA	0.978	0.982	0.979	0.811
	F1	NA	0.988	0.991	0.989	0.89
INS (1-100kb)	**P**	NA	1	1	1	1
	R	NA	0.182	0.358	0.13	0.936
	F1	NA	0.308	0.527	0.23	0.967
INV (50-1000bp)	**P**	0.999	0.908	W*	0.967	NA
	R	0.909	0.434	W*	0.758	NA
	F1	0.952	0.587	W*	0.85	NA
INV (1-100kb)	**P**	0.999	0.994	W*	0.994	NA
	R	0.999	0.998	W*	0.901	NA
	F1	0.999	0.996	W*	0.9454	NA
DUP (50-1000bp)	**P**	0.997	0	$W^{\$}$	0.792	NA
	R	0.976	0	$W^{\$}$	0.795	NA
	F1	0.987	0	$W^{\$}$	0.793	NA
DUP (1-100kb)	**P**	0.999	0.92	0.956	0.986	NA
	R	0.993	0.916	0.762	0.982	NA
	F1	0.996	0.918	0.848	0.984	NA



	DEL	**INS**	**INV[#]**	**DUP***
Small	Cue,Sniffles2,Sawfish[$] 210,926 (80,195)	Sniffles2+Sawfish 263,186 (95,622)	Cue,Sniffles2,PBSV 846 (275)	Cue,PBSV 9,943 (5,420)
Large	Cue,Sniffles2,Sawfish[$] 47,984 (21,666)	Dipcall* 38,165 (23,379)	Cue,Sniffles2,PBSV 1,269(420)	Cue,PBSV 872 (598)

[$] Any two of three [#] Any one + IGV inspection * Any one + VaPoR validation

Figure 1. Summary of the pipeline used to analyze the genomic sequences of 40 inbred mouse strains to produce the SNP, INDEL, and SV databases and their content. Long Range Sequencing of 40 inbred strains was performed, and the C57BL/6 sequence was used as the reference sequence. Deepvariant [35] was used to identify alleles for 70M SNPs and 7.5 million insertions-deletions in the 40 strains. SVs were separately analyzed based upon their size (small, 50 to 1000 bp; large, 1 to 100 KB) and type (DEL, INS, INV, DUP). The simulation results (Table 1) and the overlap among SVs identified by the programs were analyzed to produce the SV identification procedures shown in this figure. There were 210,926 small DELs, 47984 large DELs, and 263,186 small INS that were jointly identified by two of the three analysis programs (Cue, Sniffles2, Sawfish). Since none of these programs could reliably identify large INS, the genomic sequences were individually assembled and Dipcall was used to identify 38,165 large INS that were verified using the VaPoR program. To identify small and large INVs, any INV identified by Cue, Sniffles2, or PBSV that was manually verified using the IGV program was reported. The 9,943 small and 872 large DUPs identified by Cue or PBSV were validated by the VaPoR program. The numbers within parenthesis indicate the number of each type of SV present in the 35 commonly used inbred strains.

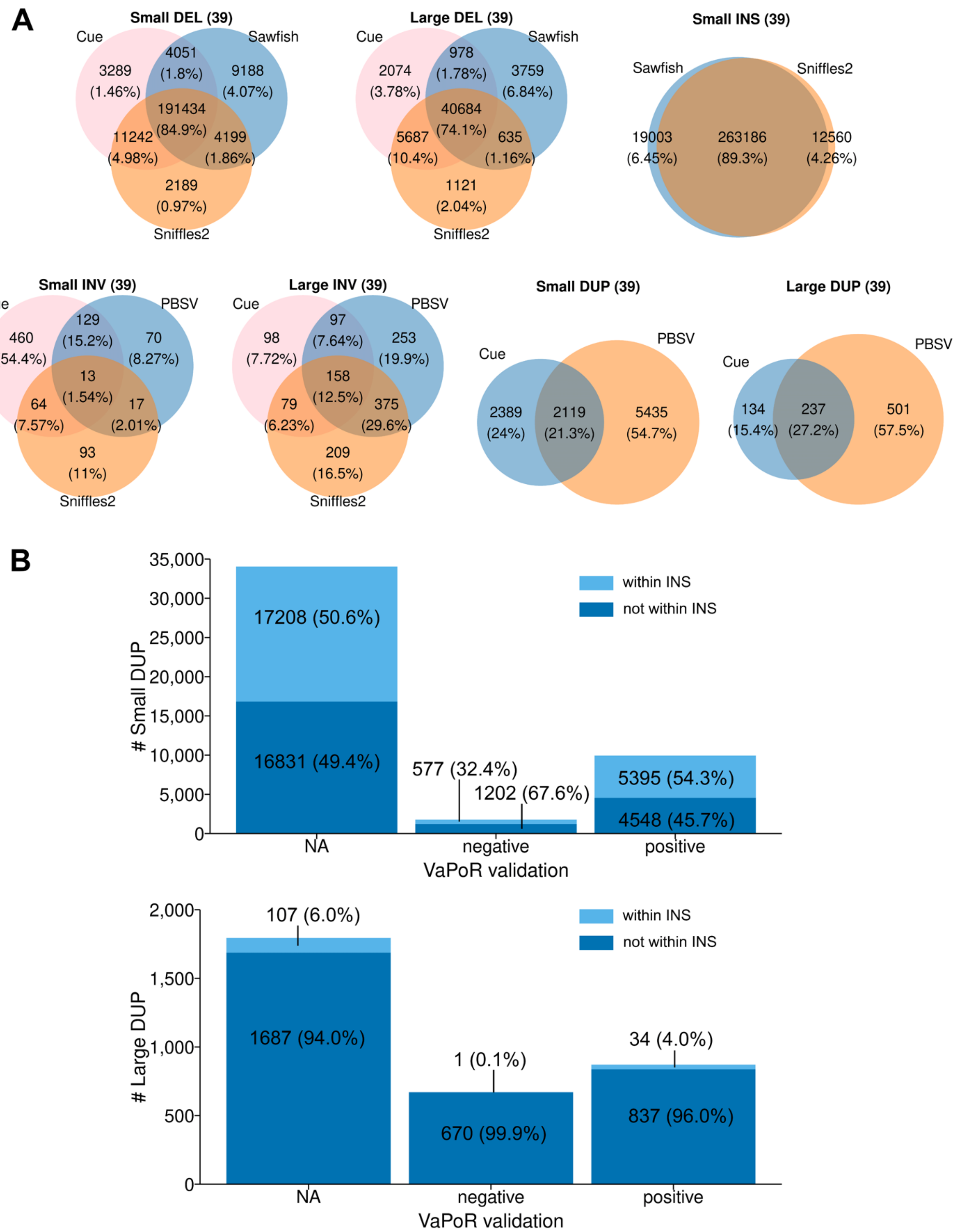


Figure 2. Assessing the performance of different SV identification programs. (**A**) SV identification programs have an acceptable level of agreement for identification of small (50 – 1000 bp) and large (1 kb – 500 kb) deletions (DELs), and for small insertions (INSs). The results

obtained from analysis of the genomic sequences all 39 inbred strains by each indicated program are shown in separate Venn diagrams. The number of SVs identified by an individual (or a combination of) programs are shown within the circles and the percentages are shown in parenthesis. In contrast, there was substantial divergence in the INVs identified by Cue, PBSV and Sniffles2. Because of this, the INVs identified by each program were verified by manual inspection using the IGV program. (**B**) The Vapor validation results for DUPs identified by Cue or PBSV. Most of the DUPs identified by Cue or PBSV could not be assessed by VaPoR (represented by NA). The DUPs that passed or failed the VaPoR validation are represented as positive or negative, respectively. We also examined whether DUPs were found within an identified INS: 54% of validated small DUPs were within an INS, but only 4.0% of validated large DUPs were within an INS.

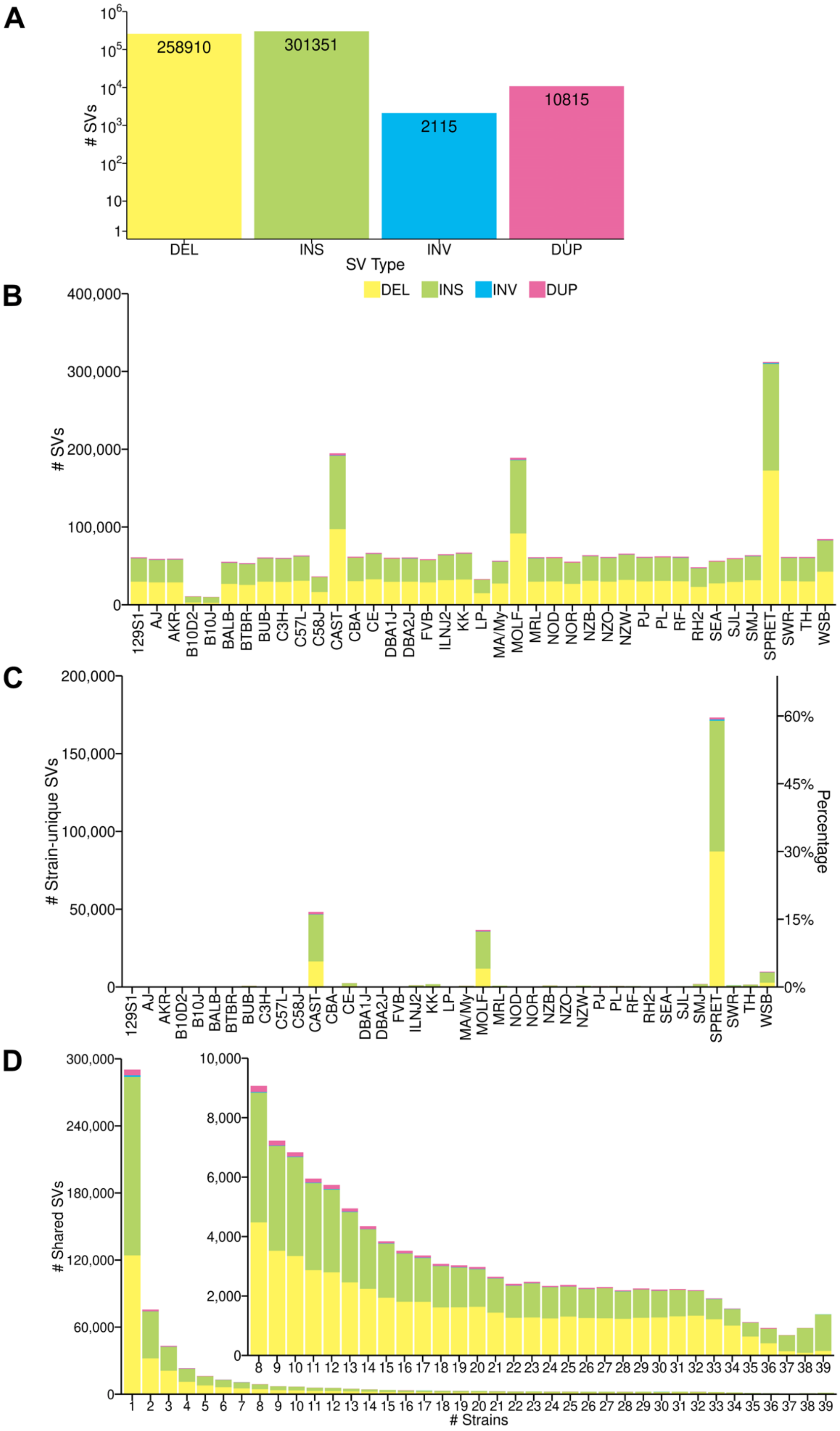
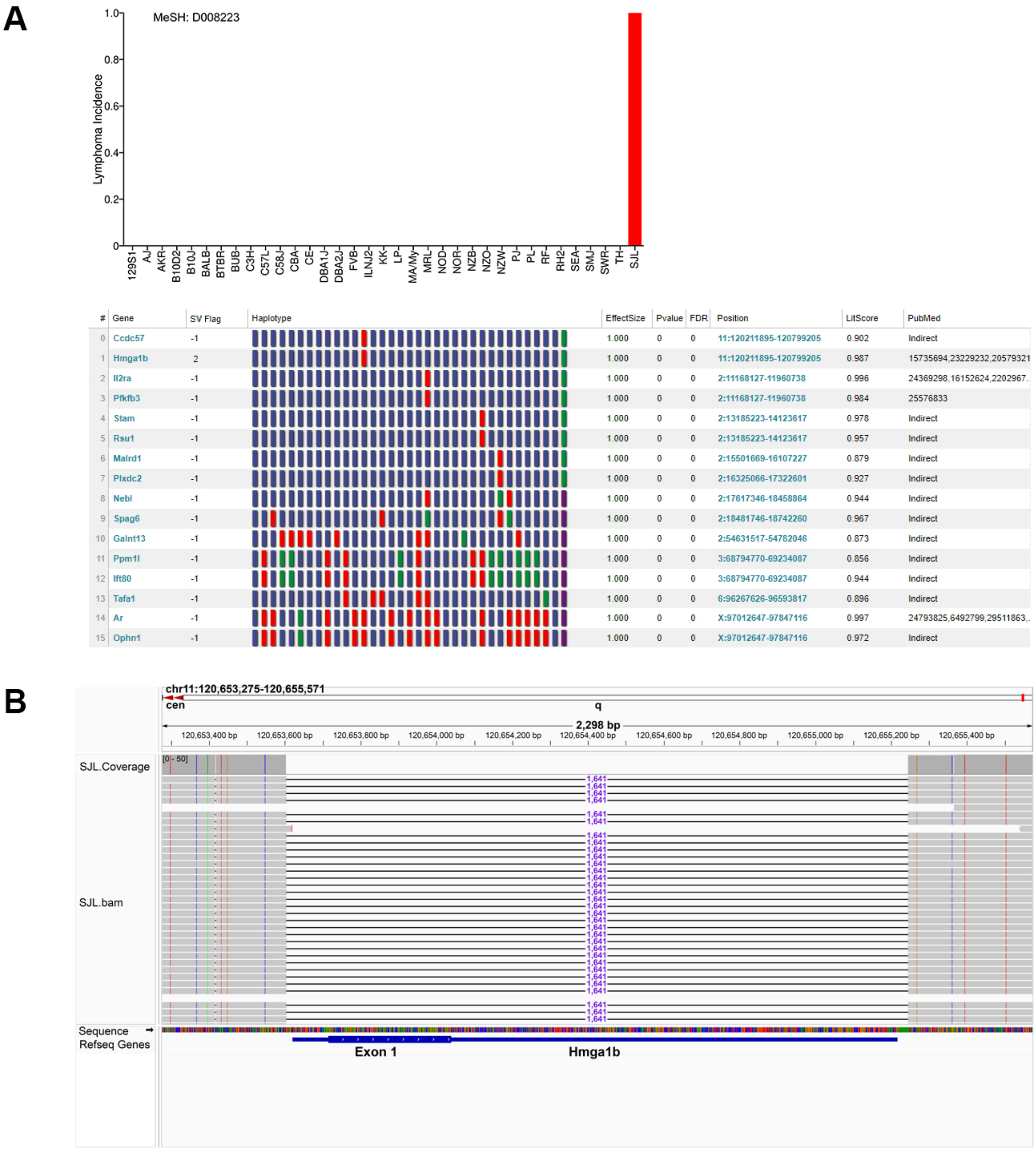

A
# SVs
10^6
10^5
10^4
10^3
10^2
10
1
258910
301351
2115
10815
DEL
INS
INV
DUP
SV Type
B
DEL
INS
INV
DUP
# SVs
400,000
300,000
200,000
100,000
0
129S1 AJ AKR B10D2 B10J BALB BTBR BUB C3H C57L C58J CAST CBA CE DBA1J DBA2J FVB ILNJ2 KK LP MA/My MOLF MRL NOD NOR NZB NZO NZW PJ PL RF RH2 SEA SJL SMJ SPRET SWR TH WSB
C
# Strain-unique SVs
200,000
150,000
100,000
50,000
0
Percentage
60%
45%
30%
15%
0%
129S1 AJ AKR B10D2 B10J BALB BTBR BUB C3H C57L C58J CAST CBA CE DBA1J DBA2J FVB ILNJ2 KK LP MA/My MOLF MRL NOD NOR NZB NZO NZW PJ PL RF RH2 SEA SJL SMJ SPRET SWR TH WSB
D
# Shared SVs
300,000
240,000
180,000
120,000
60,000
0
10,000
8,000
6,000
4,000
2,000
0
8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
# Strains

Figure 3. Characteristics or the SV alleles among the inbred strains. (**A**) The total number deletion (DEL), insertion (INS), inversion (INV) or duplication (DUP) SVs detected in all 39 strains are shown. (**B**) The number of each type of SV detected in each of the 39 strains are shown. Each type of SV is indicated by the color shown in A. (**C**) The number (left y-axis) and percentage (right y-axis) of strain-unique SVs detected in each of the 39 inbred strains is shown. Three wild derived strains (CAST, SPRET, MOLF) have most of the strain-unique SVs. (**D**) The number of strains with a shared SV allele are shown for all 39 inbred strains. Each type of SV is indicated by the color shown in A. Most of the minor SV alleles are shared by 1-3 strains. The inset graph shows the number of SV alleles shared by 8 or more strains.



#	Gene	SV Flag	Haplotype	EffectSize	Pvalue	FDR	Position	LitScore	PubMed
0	Ccdc57	-1		1.000	0	0	11:120211895-120799205	0.902	Indirect
1	Hmga1b	2		1.000	0	0	11:120211895-120799205	0.987	15735694,23229232,20579321
2	Il2ra	-1		1.000	0	0	2:11168127-11960738	0.996	24369298,16152624,2202967.
3	Pfkfb3	-1		1.000	0	0	2:11168127-11960738	0.984	25576833
4	Stam	-1		1.000	0	0	2:13185223-14123617	0.978	Indirect
5	Rsu1	-1		1.000	0	0	2:13185223-14123617	0.957	Indirect
6	Malrd1	-1		1.000	0	0	2:15501669-16107227	0.879	Indirect
7	Plxdc2	-1		1.000	0	0	2:16325066-17322601	0.927	Indirect
8	Nebl	-1		1.000	0	0	2:17617346-18458864	0.944	Indirect
9	Spag6	-1		1.000	0	0	2:18481746-18742260	0.967	Indirect
10	Galnt13	-1		1.000	0	0	2:54631517-54782046	0.873	Indirect
11	Ppm1l	-1		1.000	0	0	3:68794770-69234087	0.856	Indirect
12	Ift80	-1		1.000	0	0	3:68794770-69234087	0.944	Indirect
13	Tafa1	-1		1.000	0	0	6:96267626-96593817	0.896	Indirect
14	Ar	-1		1.000	0	0	X:97012647-97847116	0.997	24793825,6492799,29511863,.
15	Ophn1	-1		1.000	0	0	X:97012647-97847116	0.972	Indirect

Figure 4. The AI pipeline identifies a genetic factor for lymphoma susceptibility in SJL mice. (**A**) *Top panel:* Lymphoma was treated as a qualitative trait, which appears in SJL mice (incidence = 1), and not in the 34 other classical inbred strains (incidence = 0). *Bottom panel:* The AI pipeline performed a GWAS to identify haplotype blocks with allelic patterns that corresponded with lymphoma susceptibility in the 35 strains, and then selected those with haplotype blocks that contained a large deletion SV that was only present in SJL ((i.e., genetic effect size = 1 and genetic association p-value = 0). The 16 genes (indicated by gene symbol) meeting these

criteria are shown. Within the haplotype box: each block color represents a haplotype for one strain, strains with the same haplotype have the same color, and the blocks are shown in the same strain order as in panel A. The chromosome and the starting and ending position of each haplotype block are also shown. The LitScore represents the strength of the association of each gene with lymphoma as determined by the AI-mediated literature search (MeSH term: Lymphoma D008223). PubMed identification numbers are only provided for genes that have a direct link with the MeSH term. Otherwise, the AI indicates that the gene has an indirect association, which results from MeSH term relationships identified with other proteins that are associated with the gene candidate. The SV flag indicates the impact of the SV deletion as determined by VEP analysis: high impact, 2; and modifier -1. Only *Hmga1b* has a high impact deletion and is directly associated with lymphoma. (**B**) Among the 35 classical inbred strains, SJL mice uniquely have a 1641 bp deletion within *Hmga1b*. This large deletion (as visualized using the integrative genomics viewer) removes the *Hmga1b* exon that encodes the entire 107 amino acid protein.

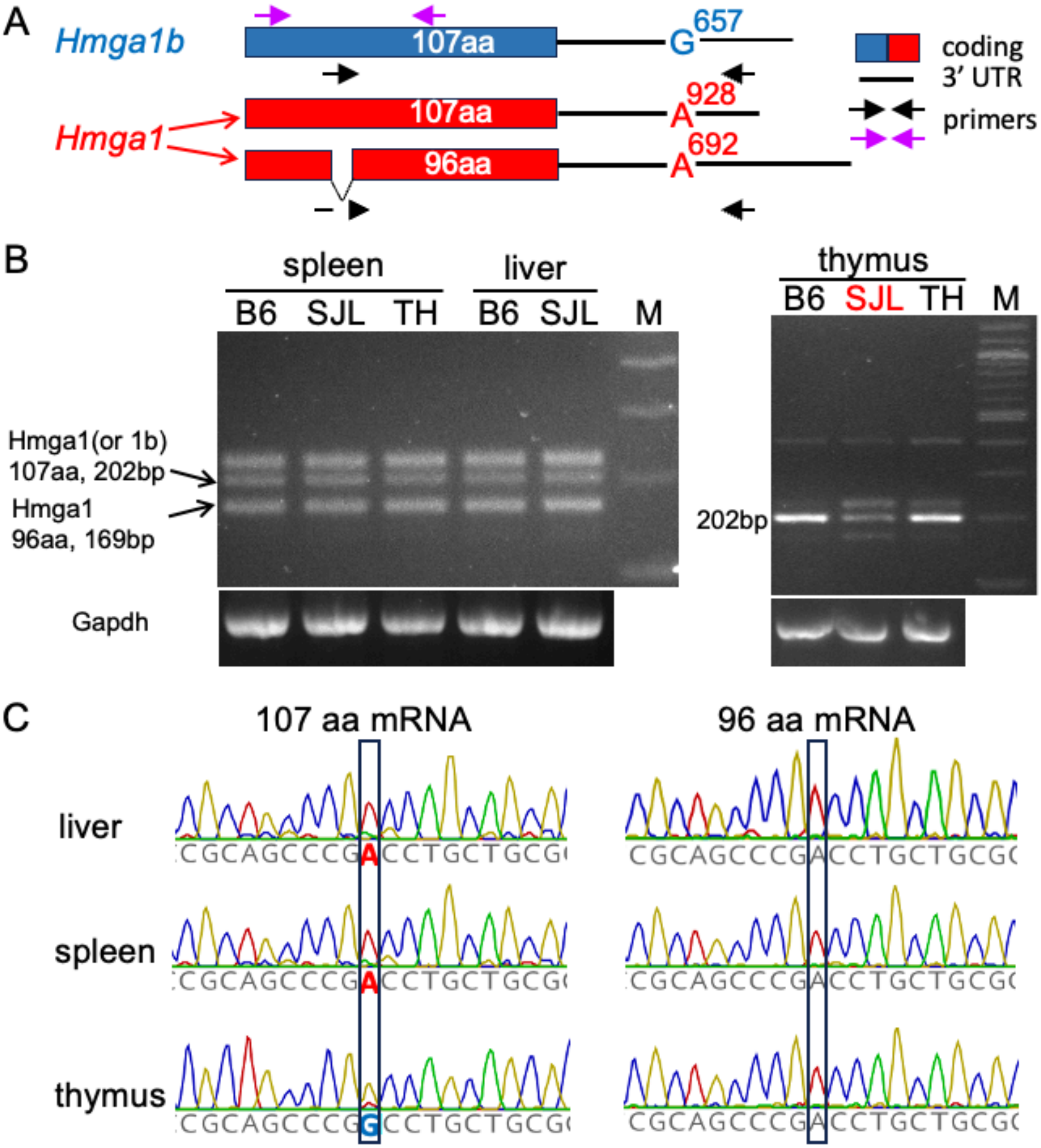


Figure 5. *Hmga1b* is selectively expressed in the thymus. (**A**) A diagram of the *Hmga1* and *Hmga1b* encoded mRNAs. *Hmga1b* encodes a 1626 bp mRNA, which generates a 107 amino acid protein. *Hmga1* generates two principal mRNAs: a 1045 bp mRNA that produces a 107 amino acid protein whose sequence is identical to that of Hmga1b; and a 1631 bp mRNA that produces a 96 amino acid protein whose sequence is identical to Hmga1b, except for the deletion of a 11 amino acid segment. The site of a sequence difference at corresponding sites in

the 3' UTR of the *Hmga1* (A) *Hmga1b* (G) mRNAs is shown. The purple arrows indicate the primers used for PCR amplification of the bands shown in B, and black arrows are the primers used to generate the amplicons used for sequencing in C. (**B**) RT-PCR was performed on spleen, liver and thymic tissues obtained from C57BL/6 (B6), TallyHo (TH) and SJL mice. The amplicons from the *Hmga1* (96 amino acid protein) and the 107 amino acid *Hmga1* (or *Hmga1b*)-encoded proteins are indicated by arrows. The pattern of amplicon expression in liver and spleen is similar in all three strains. In contrast, thymic tissue primarily expresses the 107 amino acid protein, SJL thymus has a much lower level of expression this protein. (**C**) The thymus only expresses *Hmga1b* mRNA. mRNA was prepared from spleen, liver and thymic tissue obtained from C57BL/6 mice. RT-PCR amplicons from *Hmga1* and *Hmga1b* mRNAs were prepared and sequenced. The 3' UTR of *Hmga1* mRNA has an A, while *Hmga1b* mRNA has a G at the boxed position. The results indicate that spleen and liver mRNAs are encoded by *Hmga1*, while thymic mRNAs are encoded by *Hmga1b*.

A Murine Database of Structural Variants Enables the Genetic Architecture of a Spontaneous Murine Lymphoma to be Characterized

Supplemental Note 1: Assessing SV identification programs. Since SV analysis programs vary in their ability to identify different types of SVs [1,2], we performed a set of targeted simulations to examine the ability of five programs (Cue [3], Sawfish [4], Sniffles2 [5], PBSV [6], and Dipcall [7]) to detect SVs that were artificially inserted into the mouse genomic sequence. We separately assessed their ability to identify small (<1 kb) and large (1-100 kb) DELs, INSs, INVs and DUPs (**Table 1**). This strategy measures the upper bound on recall for each program and SV category to identify those that systematically miss specific types of SVs. Multiple methods had high recall (>90%) for small and large DELs (Cue, Sniffles2, Sawfish and PBSV), large DUPs (Cue, PBSV), large INVs (Cue, Sniffles2 and PBSV), and small and large INS (Cue, PBSV). However, only Dipcall (which used individually assembled genomic sequences) could reliably detect large INS.

The simulation results were used to select the SV identification programs and database curation methods used to assemble the SV database (**Fig. 1**). A recent analysis of human LRS data revealed that SVs found by only one program are more likely to be false positives [1]. Therefore, a consensus-based approach was used for SV types where multiple programs exhibited high recall in the simulations, and where a high level of agreement was observed when the real data was analyzed by these programs. Also, whenever possible, we used a machine learning-based method (Cue) with a heuristic-based method (Cue, Sniffles2, PBSV), since programs that use similar heuristics may jointly produce false positives. For example, 85% of the small and 74% of DELs were commonly identified by Cue, Sawfish and Sniffles2, which indicates that these SVs were valid (**Figs. 2, S7**). However, if only a single caller achieved high recall in the simulations (i.e., large INSs), or when low agreement was observed among the programs when the actual data was analyzed (i.e., INVs) (Figs. 2, S7), the consensus strategy was replaced with one where the individual predictions obtained from one or more high recall tools were selected if they could be validated by a computational program or visual inspection. We used VaPoR [8], which is a program that autonomously validates SVs identified by analysis of LRS data, to identify the high-quality SVs. For example, many of the validated INVs were only identified by two of the three analysis programs utilized, and most small and large DUPs were only identified by one program (Figs. 2, S7). The low level of overlap supports our strategy of using different methods but only incorporating SVs, which were validated by another method, into the database. We are aware that the VaPoR validation requirement is likely to eliminate some true positive SVs, but we accept this risk to ensure that only true positive SVs were included in the database.

The simulation results for Sniffles2 and Sawfish for inversions and small DUPs were withheld due to performance issues that became apparent while evaluating the results produced by those programs. First, we examined why the simulations for recognition of small DUPs by Sniffles2 and Sawfish produced low values for precision, recall and F1 ($W^{\$}$). By changing the parameters [typeignore=true] used by the Truvari software [9] to calculate the precision, recall and F1 values, we found that small DUPs were incorrectly labelled by these programs as INS. Although we used IGV images to visually validate the high impact INS, it is possible that some DUPs were labelled as INS. Second, we also found a bug that was specifically present in the early version of Sawfish that was used in this study interfered with the recognition of INVs (W*).

Supplementary Note 2: Analysis of other genes identified by the mouse genetic AI pipeline. In addition to *Hmga1b*, only three other genes with SJL-specific large deletions were identified as having a potential direct association with lymphoma. (i) *Il2ra* was associated with lymphoma because *Il2ra* mRNA expression is a prognostic marker for Burkitt lymphoma [10], acute myelogenous leukemia [11], mycosis fungoides [12], diffuse large B cell lymphoma [13] and other cancers; and IL-2 is administered as an anti-cancer agent [14]. The absence of *Il2ra* could reduce T and NK cell mediated anti-cancer immunity, which potentially could facilitate B cell lymphoma development. However, the *Il2ra* SV is not high impact (it labelled as a modifier and does not affect the coding sequence), and will not have a major effect on *Il2ra* expression. (ii) Although *Pfkb2* was identified as having a direct association with lymphoma, the paper [15] identified by the AI was a false positive association with lymphoma. (iii) *Ccdc57* identified by the AI because it is within the same haplotype block as *Hmga1b*; but *Ccdc57* has no direct association with lymphoma. Ccdc57 is in the Human Protein Atlas, which contains lymphoma tissue. No other gene with SJL-specific high impact large deletion was directly associated with lymphoma.

Methods

Animal experiments. All animal experiments were performed according to protocols that were approved by the Stanford Institutional Animal Care and Use Committee. All mice were obtained from Jackson Labs, and the results are reported according to the ARRIVE guidelines [16].

DNA sequencing. Forty inbred strains (**Table S1**) were subject to LRS using the HiFi REVIO system (PacBIO). For thirty strains, mouse liver was obtained from mice purchased from the Jackson Laboratory, snap-frozen in liquid nitrogen and shipped on dry-ice to the DNA Technologies Core of the Genome Center, University of California Davis were high molecular DNA purification and REVIO sequencing was performed. The genomic DNA for ten strains (the bottom 10 listed in Table S1) were kindly provided Dr. Laura Reinholdt, Co-Director of the Mutant Mouse Resource and Research Center at the Jackson Laboratory (Bar Harbor, ME); and REVIO sequencing was also performed at the UC Davis Genome Center.

Simulations for assessing SV program performance. We generated 8 synthetic genomes using insilicoSV (https://github.com/PopicLab/insilicoSV), which inserted 1000 simulated SVs of each type and size into the GRCm39 reference sequence. For each synthetic genome, PacBio HiFi reads at 30x coverage using PBSIM3 [17] were simulated, and the simulated reads were aligned using minimap2 (v2.28) [18].

Identification of SNPs and INDELs. PacBio HIFI long read sequence (LRS) data were aligned to GRCm39 (mm39) reference genome using minimap2 (v2.28) and pbmm2 (v1.13.1) to generate the BAM files. SNP and INDEL alleles were identified for each strain using the GPU-based DeepVariant (v1.6.1) [19] and the alleles from all strains were merged using GLnexus (v1.4.1) [19].

Identification of SVs. To identify small and large deletions, Cue [3] and Sniffles2 (v2.3.2) [20] were used to perform SV calling based on the minimap2 alignment, and Sawfish (v0.10.0) [21] was used for SV calling based on the pbmm2 alignment. Then, SURVIVOR (v1.0.7) [22] was used to merge the results obtained by these three programs. The SVs commonly identified by at least two of these programs were included in the final dataset. Sniffles2 and Sawfish were used to identify small insertions; the small insertions jointly identify by both programs were merged using SURVIVOR and were incorporated into the final dataset. An assembly-based program Dipcall [7] was used to identify large insertions, and VaPoR (v1.0) [8] was used for their validation. The large INS with a VaPoR statistic GS⩾0.15 and QS ⩾ 0.1 were considered to be validated, and were incorporated into the final dataset. Cue, Sniffles2 (v2.3.2) and PBSV (v2.9.0) were used to identify INVs: the Cue and Sniffles2 analyses were based on minimap2 alignments, whereas the PBSV analysis was based on pbmm2 alignment. SURVIVOR was used to merge those identified by any of the three methods; these INVs were then inspected using the Integrative Genomics Viewer (IGV) (v2.17.4) [23] as described below; and visually-validated INVs were incorporated into the dataset. DUPs were identified by PBSV or Cue, and then validated by VaPoR as described above. In addition, the starting and ending positions of the identified DUPs were examined to determine if they were located within an identified insertion. Functional annotation was performed using the Ensembl Variant Effect Predictor (VEP) (v112.0)[24].

Visual inspection of high impact deletions and insertions, and all inversions was performed by examining IGV images. To do this, the IGV image script code was modified as follows:

```
“new
genome mm39
snapshotDirectory /pathway1/
load /pathway2/name.identify.vcf
load /pathway3/name.alignment.bam
goto chr*:start-end position
scrollToTop
preference SAM.COLOR_BY READ_STRAND
preference SAM.LINK_READS TRUE
preference SAM.LINK_TAG READNAME
preference SAM.GROUP_OPTION ZMW
preference SAM.SHOW_MISMATCHES TRUE
preference SAM.MAX_VISIBLE_RANGE 500
snapshot output.png
exit”
```

We found that altering three of the six display settings was important. (i) The 'preference SAM.COLOR_BY READ_STRAND' indicates that reads from different strands are displayed in different colors. (ii) The 'preference SAM.GROUP_OPTION ZMW' specifies the ZMW group option, which is optimal for analysis of PacBio long read sequence. (iii) The 'preference SAM.MAX_VISIBLE_RANGE 500' sets the maximum length of a displayed SV at 500 KB. The use of the alternate parameters improved the display of the inversions (Fig. S8). The predicted CTCF binding sites in the genomic sequences of the inbred strains were identified using the CTCF (v0.99.11) R package in Bioconductor [25] with the JASPAR 2022 database and the recommended *P*-value cutoff (1 x 10^{-6}). Of note, we used split-read alignments with strand flips when validating INVs by visual inspection, but this approach has limitations since it can only validate INVs where split-read alignments are available.

Analysis of Hmga1b and Hmga1 mRNAs. Total RNA was purified from mouse thymus, spleen and liver obtained from SJL, C57BL/6 and TallyHo mice; and from kidney and hind leg bone marrow of SJL and C57BL/6 mice using the TRIzol (Thermo Fisher) reagent with the Direct-zol RNA miniprep kit (Zymo Research). One half ug each of the total RNAs were reverse transcribed using the High-Capacity cDNA Reverse Transcription kit (Applied Biosystems) in a 20 µl volume. One µl each of the cDNAs were then subject to PCR using the GoTaq® G2 DNA polymerase master mix (Promega) and primers for *Hmga1b, Hmga1* or *Gapdh* (control) transcripts. *Hmga1b/Hm*ga1 transcript primers: Hmga-F1: AGCGAGTCGGGCTCAAAGTC and Hmga-R1: CGCCCTTATTCTTGCTTCCCTTT. Primers for either the 107aa transcript or the 96aa transcripts were: Hmg-107aa-F1, TGAGTCCTGGGACGGCGCT, Hmg-96aa-F, AGCAGCCTCCGAAAGAGCC, Hmga-R, GAATGCTCCCAGGACCCTCTA. Primers for the *Gapdh* transcript: Gapdh-F2: GTAGACAAAATGGTGAAGGTCGGT and Gapdh-R1, GGTCCAGGGTTTCTTACTCCTTG. The PCR amplified cDNA generated using Hmg-107aa-F1 and Hmga-R (the 107aa transcripts only) primers or the Hmg-96aa-F and Hmga-R (the 96aa transcripts only) primers were gel purified and then subject to Sanger sequencing.

References

1 Liu, Y. H., Luo, C., Golding, S. G., Ioffe, J. B. & Zhou, X. M. Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data. *Nature communications* **15**, 2447 (2024). https://doi.org:10.1038/s41467-024-46614-z

2 Liu, Z. *et al.* Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol* **23**, 68 (2022). https://doi.org:10.1186/s13059-022-02636-8

3 Popic, V. *et al.* Cue: a deep-learning framework for structural variant discovery and genotyping. *Nat Methods* **20**, 559-568 (2023). https://doi.org:10.1038/s41592-023-01799-x

4 Saunders, C. T. *et al.* Sawfish: Improving long-read structural variant discovery and genotyping with local haplotype modeling. *BioRxiv* **https://doi.org/10.1101/2024.08.19.608674;** (2024). https://doi.org:https://doi.org/10.1101/2024.08.19.608674;

5 Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* (2024). https://doi.org:10.1038/s41587-023-02024-y

6 Topfer, A. *PacBio structural variant calling and analysis tools*, <https://github.com/PacificBiosciences/pbsv> (2023).

7 Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**, 595-597 (2018). https://doi.org:10.1038/s41592-018-0054-7

8 Zhao, X., Weber, A. M. & Mills, R. E. A recurrence-based approach for validating structural variation using long-read sequencing technology. *Gigascience* **6**, 1-9 (2017). https://doi.org:10.1093/gigascience/gix061

9 English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol* **23**, 271 (2022). https://doi.org:10.1186/s13059-022-02840-6

10 Xu, Y. F., Wang, G. Y., Zhang, M. Y. & Yang, J. G. Hub genes and their key effects on prognosis of Burkitt lymphoma. *World J Clin Oncol* **14**, 357-372 (2023). https://doi.org:10.5306/wjco.v14.i10.357

11 Aoki, T. *et al.* High IL2RA/CD25 expression is a prognostic stem cell biomarker for pediatric acute myeloid leukemia without a core-binding factor. *Pediatr Blood Cancer* **71**, e30803 (2024). https://doi.org:10.1002/pbc.30803

12 Flechon, L. *et al.* Genomic profiling of mycosis fungoides identifies patients at high risk of disease progression. *Blood Adv* **8**, 3109-3119 (2024). https://doi.org:10.1182/bloodadvances.2023012125

13 Sadras, T. *et al.* Differential expression of MUC4, GPR110 and IL2RA defines two groups of CRLF2-rearranged acute lymphoblastic leukemia patients with distinct secondary lesions. *Cancer Lett* **408**, 92-101 (2017). https://doi.org:10.1016/j.canlet.2017.08.034

14 Yang, Y. & Lundqvist, A. Immunomodulatory Effects of IL-2 and IL-15; Implications for Cancer Immunotherapy. *Cancers (Basel)* **12** (2020). https://doi.org:10.3390/cancers12123586

15 Koiri, R. K., Trigun, S. K. & Mishra, L. Activation of p53 mediated glycolytic inhibition-oxidative stress-apoptosis pathway in Dalton's lymphoma by a ruthenium (II)-complex containing 4-carboxy N-ethylbenzamide. *Biochimie* **110**, 52-61 (2015). https://doi.org:10.1016/j.biochi.2014.12.021

16 Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M. & Altman, D. G. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* **8**, e1000412 (2010). https://doi.org:10.1371/journal.pbio.1000412

17 Ono, Y., Hamada, M. & Asai, K. PBSIM3: a simulator for all types of PacBio and ONT long reads. *NAR Genom Bioinform* **4**, lqac092 (2022). https://doi.org:10.1093/nargab/lqac092

18 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018). https://doi.org:10.1093/bioinformatics/bty191

19 Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983-987 (2018). https://doi.org:10.1038/nbt.4235

20 Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461-468 (2018). https://doi.org:10.1038/s41592-018-0001-7

21 *Sawfish*, <https://github.com/PacificBiosciences/sawfish> (2024).

22 Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature communications* **8**, 14061 (2017). https://doi.org:10.1038/ncomms14061

23 Robinson, J. T., Thorvaldsdottir, H., Turner, D. & Mesirov, J. P. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* **39** (2023). https://doi.org:10.1093/bioinformatics/btac830

24 Hunt, S. E. *et al.* Annotating and prioritizing genomic variants using the Ensembl Variant Effect Predictor-A tutorial. *Hum Mutat* **43**, 986-997 (2022). https://doi.org:10.1002/humu.24298

25 Dozmorov, M. G. *et al.* CTCF: an R/bioconductor data package of human and mouse CTCF binding sites. *Bioinform Adv* **2**, vbac097 (2022). https://doi.org:10.1093/bioadv/vbac097

26 Wang, M., Fang, Z., Yoo, B., Bejarano, G. & Peltz, G. The Effect of Population Structure on Murine Genome-Wide Association Studies. *Frontiers in genetics* **In press** (2021).

27 Zheng, M. *et al.* The Role of Abcb5 Alleles in Susceptibility to Haloperidol-Induced Toxicity in Mice and Humans *PLoS Medicine* **12**, e1001782 (2015). https://doi.org:10.1371/journal.pmed.100172

28 Arslan, A. *et al.* Analysis of Structural Variation Among Inbred Mouse Strains. *BMC Genommics* **24**, 97-109 (2023). https://doi.org:10.1186/s12864-023-09197-5

29 Fang, Z. & Peltz, G. An Automated Multi-Modal Graph-Based Pipeline for Mouse Genetic Discovery. *Bioinformatics* **38**, 3385-3394 (2022). https://doi.org:10.1093/bioinformatics/btac356

Table S1. Characteristics of the LRS data obtained from inbred 40 strains. The strain name, Jackson Lab #, sequence amount in gigabases (GB); and the mean, median and number of reads (in millions) for each strain are shown. The N50 read-length distribution statistic assesses read-length quality: it represents the length of the shortest read in the group of the longest sequences that together represent >50% of the nucleotides in that sequence set. Sequencing was performed with a PacBio Revio instrument using the HiFi system.

Strain	JAX Strain #	GB	# Reads (M)	Mean Read Length	Median Read Length	N50 Read length (bp)
A/J	#000646	85.12	4.69	18132	18061	20255
C57BL/6J	#000664	87.65	5.05	17359	17127	18965
C57BL/10J	#000665	92.54	5.76	16071	15825	17362
C57L/J	#000668	92.20	5.92	15583	15347	16863
CBA/J	#000656	88.19	5.12	17230	17062	19088
DBA/1J	#000670	109.58	6.58	16644	16300	17989
DBA/2J	#000671	106.62	6.03	17694	17341	19069
FVB/NJ	#001800	97.39	6.07	16039	15682	17481
KK.Cg-Ay/J	#002468	88.64	5.40	16426	16085	17674
LP/J	#000676	89.05	5.49	16210	16089	17873
NOD/ShiLtJ	#001976	87.21	4.92	17720	17496	19586
NOR/LtJ	#002050	93.00	5.77	16114	15819	17722
NZB/BINJ	#000684	87.23	5.27	16550	16390	18197
NZW/LacJ	#001058	90.83	6.10	14884	14629	15914
TALLYHO/JngJ	#005314	85.01	5.59	15211	14949	16694
AKR/J	#000648	94.11	5.63	16725	16477	18181
B10.D2-Hc1 H2d H2-T18c/nSnJ	#000463	98.45	5.89	16722	16546	18087
C3H/HeJ	#000659	87.20	5.53	15754	15717	17701
RF/J	#000682	92.80	5.49	16893	16648	18797
SJL/J	#000686	101.17	6.45	15679	15487	17057
MRL/MpJ	#000486	96.00	6.17	15566	15309	16825
NZO/HILtJ	#002105	88.20	5.03	17538	17314	19153
WSB/EiJ	#001145	97.26	6.03	16135	15869	17391
CAST/EiJ	#000928	79.22	5.94	13344	12413	15657
MOLF/EiJ	#000550	84.45	5.50	15345	14993	16945
SPRET/EiJ	#001146	82.31	4.86	16936	16748	18858
SWR/J	#000689	85.84	7.04	12194	10771	14418
129S1/SvImJ	#002448	84.49	5.16	16387	16181	17795
BALB/cJ	#000651	108.88	6.67	16333	16056	18050
BTBR T+ Itpr3tf/J	#002282	82.31	4.72	17433	17224	18906
BUB/BnJ	#000653	81.32	6.38	12741	12688	14612
MA/MyJ	#000677	87.84	6.22	14119	14041	15684
C58/J	#000669	79.26	7.21	10997	10550	13356
SM/J	#000687	46.31	4.70	9864	9138	11960
CE/J	#000657	74.83	6.10	12266	12004	13752
PL/J	#000680	70.32	5.96	11803	11505	13551
SEA/GnJ	#000644	90.93	7.14	12737	12637	14857
I/LnJ	#000674	91.93	6.65	13829	13785	15595
RHJ/LeJ	#001591	81.86	7.67	10668	10493	12523
P/J	#000679	92.88	5.80	16150	15778	17645
Total		**3540.45**	**233.69**	**15301**		
Mean		**88.51**	**5.84**			

Table S2. The functional consequences of SVs that were predicted by VEP. Their estimated severity (HIGH, MODERATE, LOW, MODIFIER) and the type of effect (consequences) was determined for SVs in all 39 strains or those in the 35 classical inbred strains.

Impact	Consequences	#SV 39 strains	#SV 35 strains
HIGH	transcript_ablation	2068	1079
	splice_acceptor_variant	33	14
	splice_donor_variant	44	24
	stop_gained	29	10
	frameshift_variant	102	55
	stop_lost	171	92
	start_lost	4	1
	transcript_amplification	191	130
	feature_elongation	5823	2538
	feature_truncation	5555	2388
MODERATE	inframe_insertion	39	24
	inframe_deletion	183	88
	protein_altering_variant	5	2
LOW	splice_donor_5th_base_variant	44	24
	splice_region_variant	19	9
	splice_donor_region_variant	7	3
	splice_polypyrimidine_tract_variant	275	98
	start_retained_variant	4	1
	stop_retained_variant	19	12
MODIFIER	coding_sequence_variant	2389	1398
	mature_miRNA_variant	1	0
	5_prime_UTR_variant	1171	636
	3_prime_UTR_variant	4100	1808
	non_coding_transcript_exon_variant	7140	3161
	intron_variant	244001	97134
	non_coding_transcript_variant	41051	17103
	coding_transcript_variant	108	68
	upstream_gene_variant	11946	4955
	downstream_gene_variant	14283	5984
	regulatory_region_ablation	7130	3537
	regulatory_region_amplification	358	253
	regulatory_region_variant	66411	29290
	intergenic_variant	290358	117312

Table S3. The number of insertion and deletion SVs identified by VEP analysis as having a high impact (protein coding) in all 39 strains or in the 35 classical inbred strains are shown.

	39 strains		35 strains	
	# SV	# Unique Genes	# SV	# Unique Genes
Small INS	2082	1899	801	765
Large INS	822	740	518	487
Small DEL	1688	1535	705	654
Large DEL	590	536	293	270

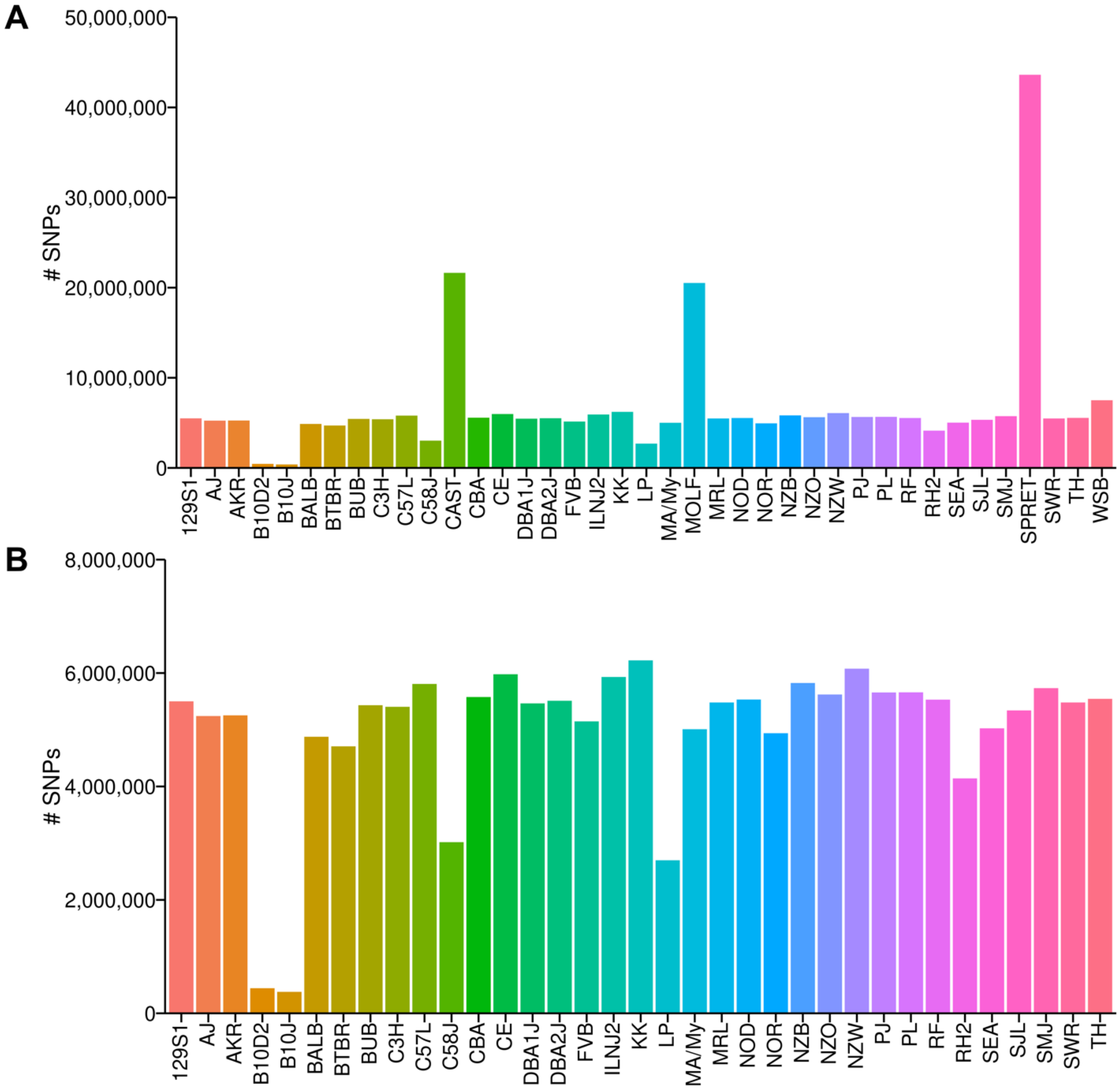


Figure S1. The number of SNP sites in (**A**) all 39 inbred strains or (**B**) in the 35 classical inbred strains are shown. SNP sites were identified by comparisons with the C57BL/6 reference sequence.

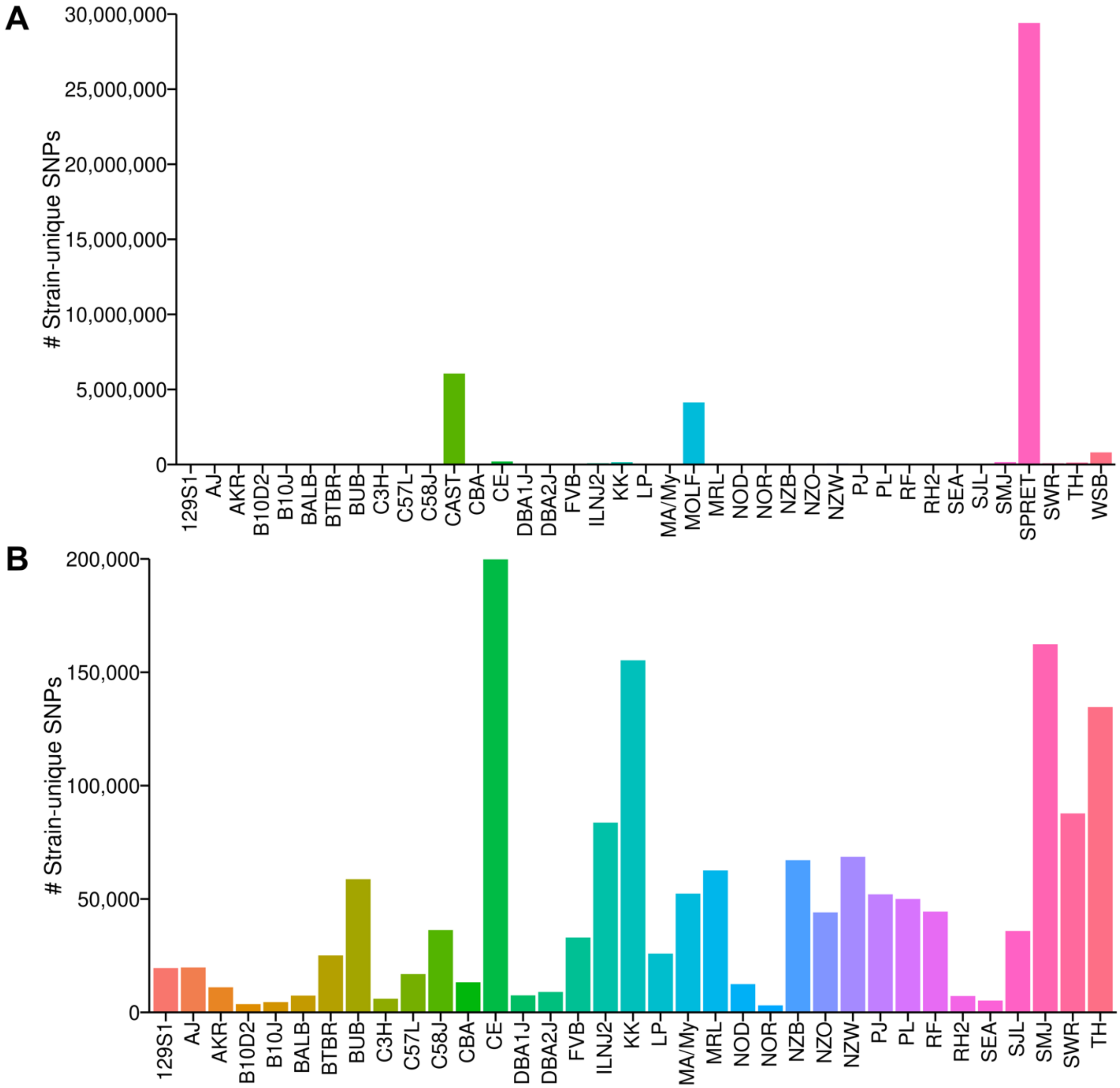


Figure S2. The number of sites with strain-unique SNP alleles in (**A**) all 39 inbred strains or (**B**) in the 35 classical inbred strains are shown. Most of the strain-unique SNPs are present in the four wild derived strains (CAST, SPRET, MOLF, WSB).

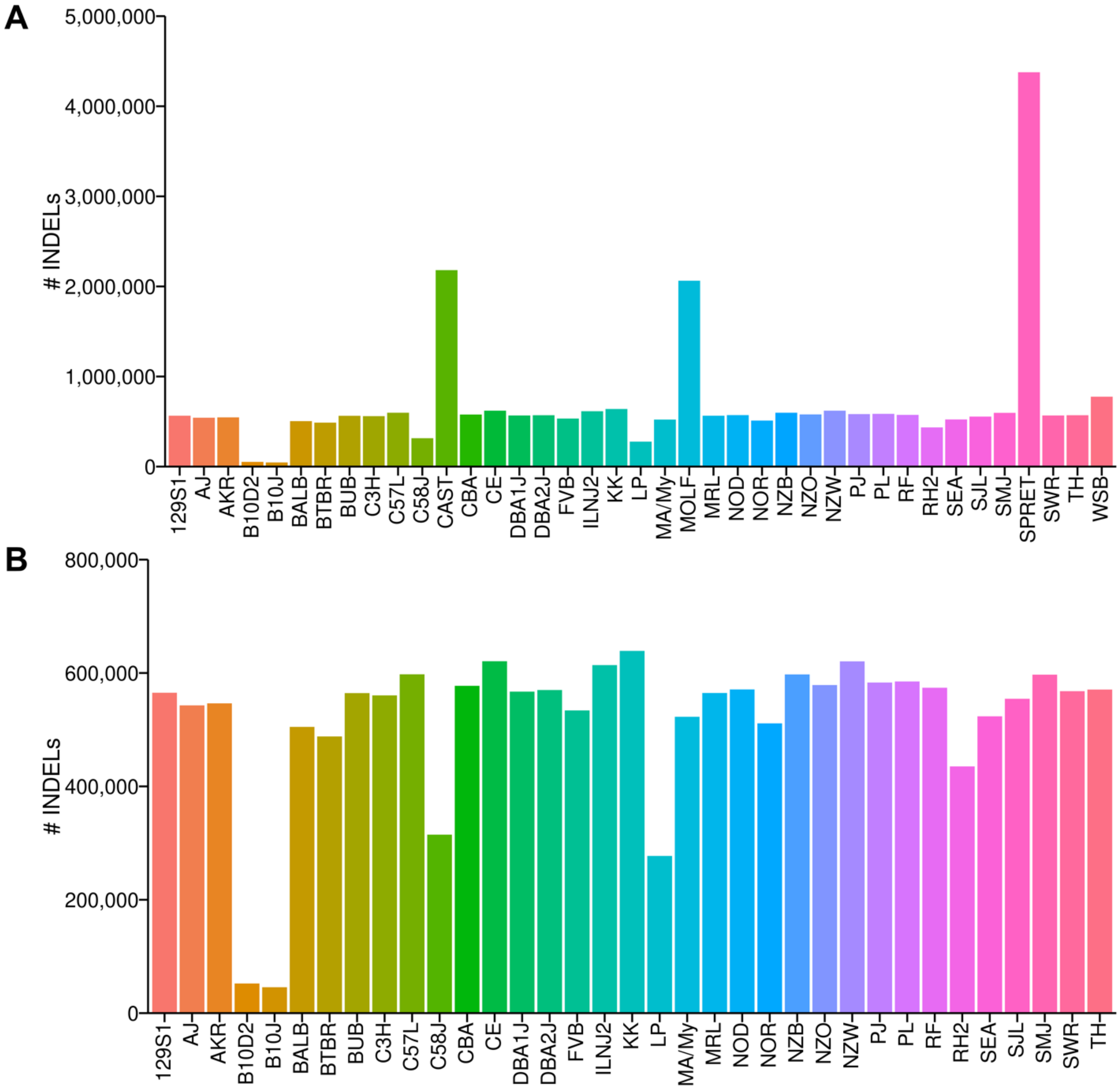


Figure S3. The number of INDEL sites identified in (**A**) all 39 inbred strains or (**B**) in the 35 classical inbred strains are shown. The INDEL sites were identified by comparisons with the C57BL/6 reference sequence.

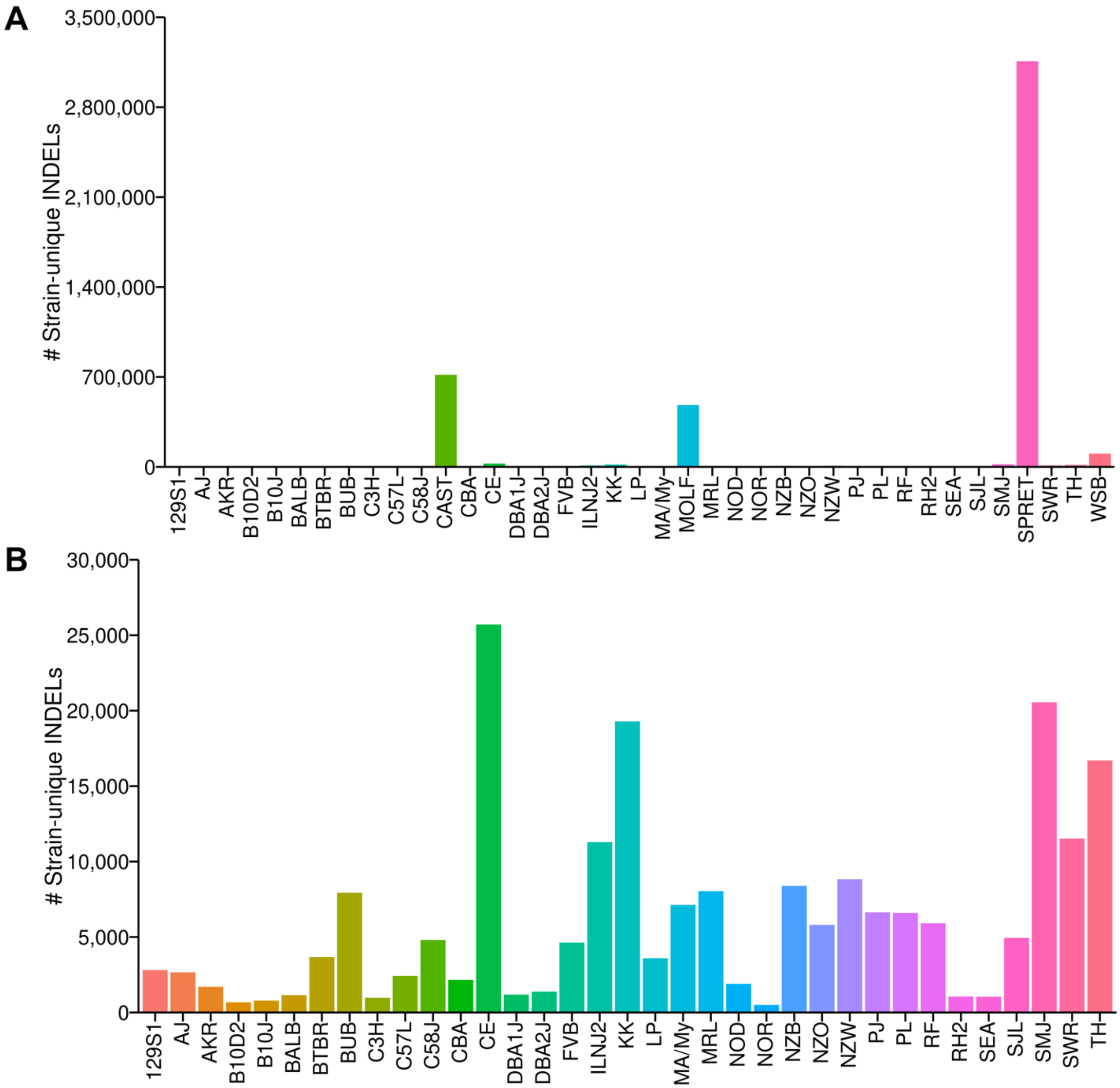
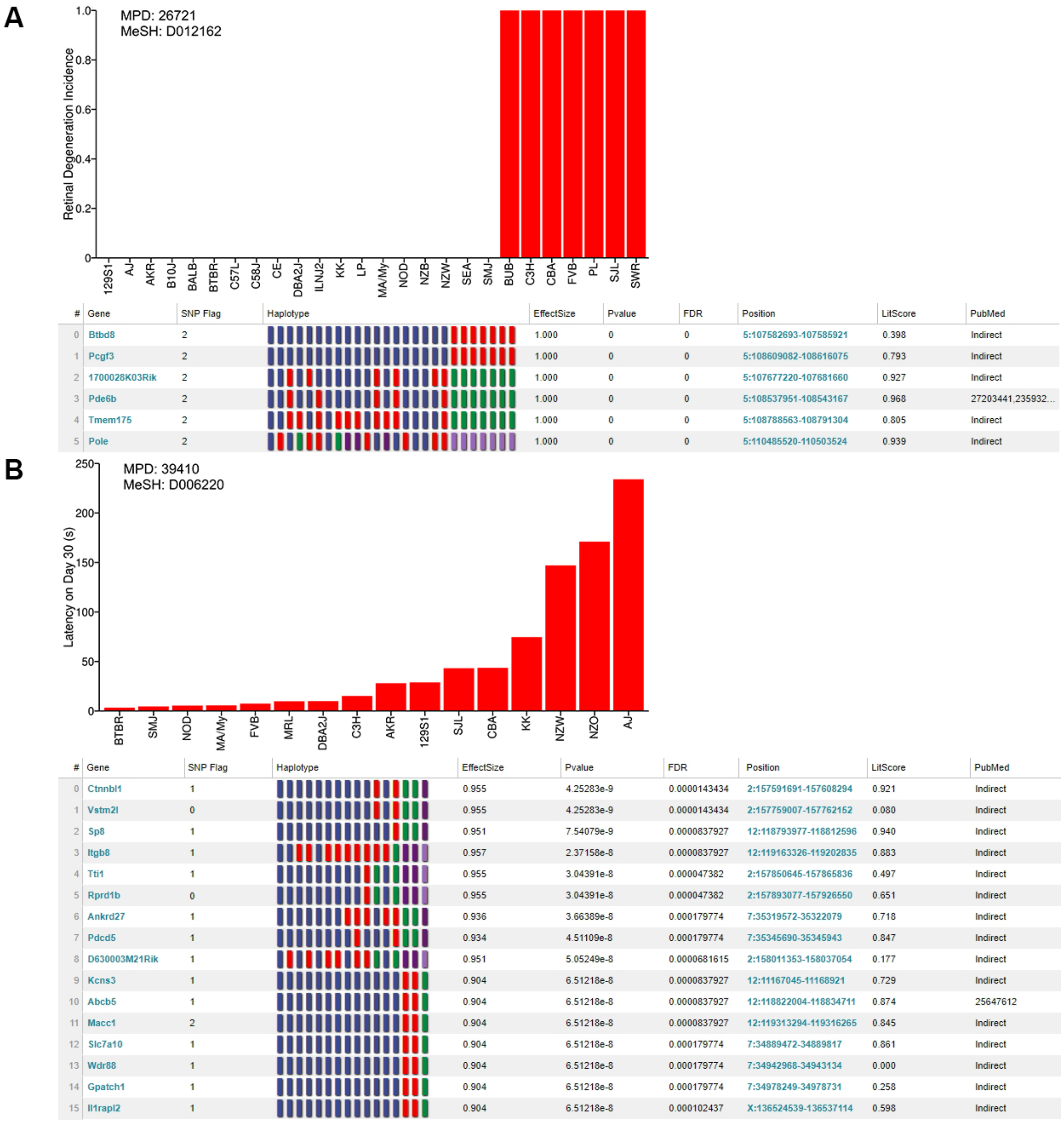


Figure S4. The number of sites with strain-unique INDEL alleles identified in (**A**) all 39 inbred strains or (**B**) in the 35 classical inbred strains. Most strain-unique INDELs are present in the four wild derived strains (CAST, SPRET, MOLF, WSB).

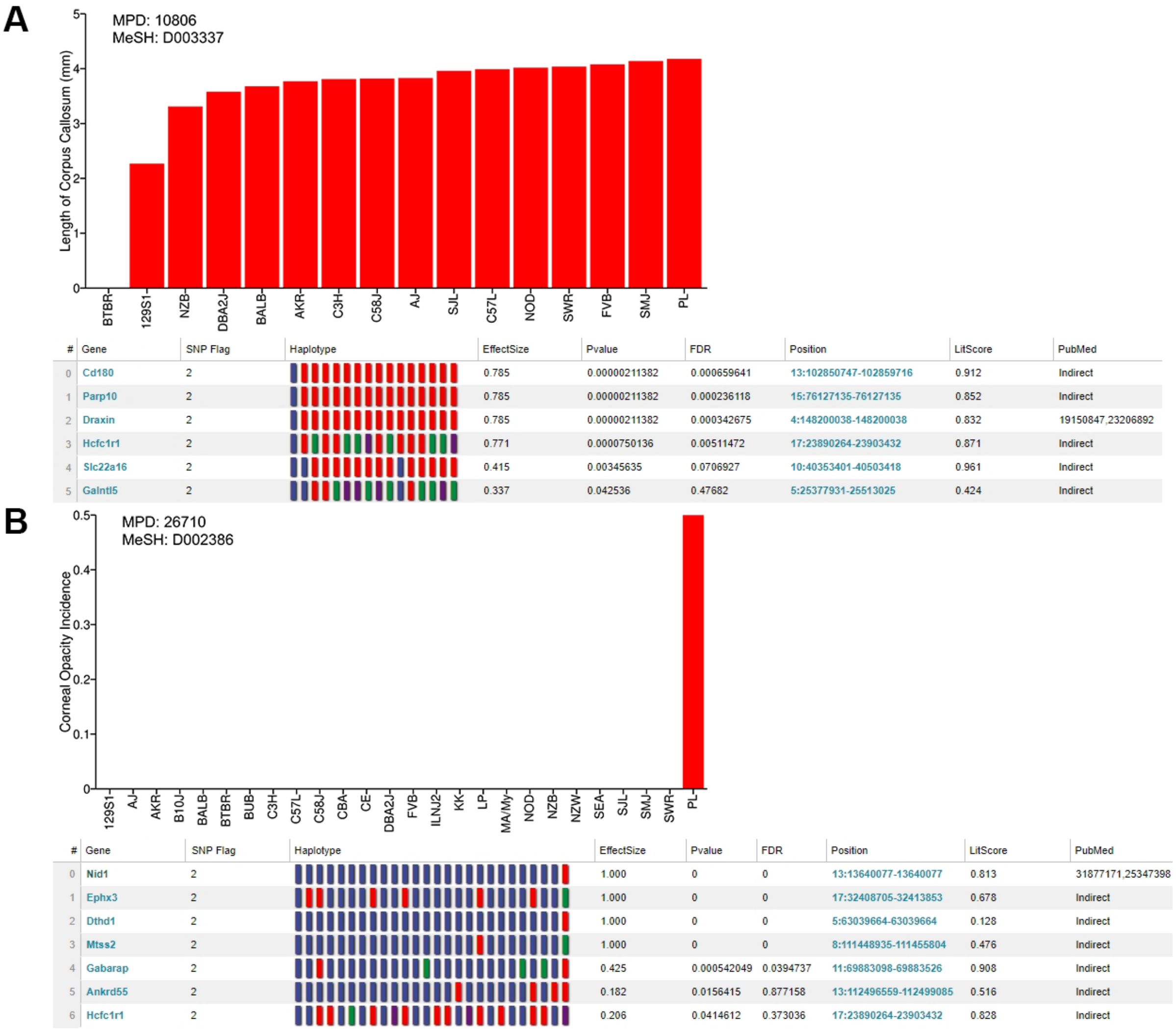


#	Gene	SNP Flag	Haplotype	EffectSize	Pvalue	FDR	Position	LitScore	PubMed
0	Btbd8	2	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥🟥🟥	1.000	0	0	5:107582693-107585921	0.398	Indirect
1	Pcgf3	2	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥🟥🟥	1.000	0	0	5:108609082-108616075	0.793	Indirect
2	1700028K03Rik	2	🟦🟦🟥🟦🟥🟦🟦🟦🟦🟦🟦🟥🟦🟥🟦🟦🟦🟥🟥🟩🟩🟩🟩🟩🟩🟩	1.000	0	0	5:107677220-107681660	0.927	Indirect
3	Pde6b	2	🟦🟦🟥🟦🟦🟥🟦🟦🟦🟦🟦🟥🟦🟥🟦🟦🟦🟦🟥🟩🟩🟩🟩🟩🟩🟩	1.000	0	0	5:108537951-108543167	0.968	27203441,235932...
4	Tmem175	2	🟦🟦🟥🟥🟦🟥🟦🟥🟥🟥🟦🟥🟥🟥🟦🟦🟦🟦🟥🟩🟩🟩🟩🟩🟩🟩	1.000	0	0	5:108788563-108791304	0.805	Indirect
5	Pole	2	🟦🟥🟦🟩🟥🟥🟦🟩🟪🟪🟥🟦🟪🟦🟥🟦🟦🟥🟥🟣🟣🟣🟣🟣🟣🟣	1.000	0	0	5:110485520-110503524	0.939	Indirect

#	Gene	SNP Flag	Haplotype	EffectSize	Pvalue	FDR	Position	LitScore	PubMed
0	Ctnnbl1	1	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟦🟥🟩🟩🟪	0.955	4.25283e-9	0.0000143434	2:157591691-157608294	0.921	Indirect
1	Vstm2l	0	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟦🟥🟩🟩🟪	0.955	4.25283e-9	0.0000143434	2:157759007-157762152	0.080	Indirect
2	Sp8	1	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟩🟩🟪	0.951	7.54079e-9	0.0000837927	12:118793977-118812596	0.940	Indirect
3	Itgb8	1	🟦🟦🟥🟥🟦🟥🟥🟥🟥🟥🟥🟥🟩🟪🟪🟣	0.957	2.37158e-8	0.0000837927	12:119163326-119202835	0.883	Indirect
4	Tti1	1	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟩🟦🟩🟪🟪🟣	0.955	3.04391e-8	0.000047382	2:157850645-157865836	0.497	Indirect
5	Rprd1b	0	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟩🟦🟩🟪🟪🟣	0.955	3.04391e-8	0.000047382	2:157893077-157926550	0.651	Indirect
6	Ankrd27	1	🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟦🟥🟥🟩🟩🟪	0.936	3.66389e-8	0.000179774	7:35319572-35322079	0.718	Indirect
7	Pdcd5	1	🟦🟦🟦🟦🟦🟦🟦🟦🟥🟦🟦🟦🟥🟩🟩🟪	0.934	4.51109e-8	0.000179774	7:35345690-35345943	0.847	Indirect
8	D630003M21Rik	1	🟦🟥🟦🟥🟦🟥🟥🟦🟥🟥🟩🟦🟩🟪🟪🟣	0.951	5.05249e-8	0.0000681615	2:158011353-158037054	0.177	Indirect
9	Kcns3	1	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟩	0.904	6.51218e-8	0.0000837927	12:11167045-11168921	0.729	Indirect
10	Abcb5	1	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟩	0.904	6.51218e-8	0.0000837927	12:118822004-118834711	0.874	25647612
11	Macc1	2	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟩	0.904	6.51218e-8	0.0000837927	12:119313294-119316265	0.845	Indirect
12	Slc7a10	1	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟩	0.904	6.51218e-8	0.000179774	7:34889472-34889817	0.861	Indirect
13	Wdr88	1	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟩	0.904	6.51218e-8	0.000179774	7:34942968-34943134	0.000	Indirect
14	Gpatch1	1	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟩	0.904	6.51218e-8	0.000179774	7:34978249-34978731	0.258	Indirect
15	Il1rapl2	1	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟩	0.904	6.51218e-8	0.000102437	X:136524539-136537114	0.598	Indirect

Figure S5. Using the SNP database developed here, the AI pipeline correctly identifies the causative genetic factors for retinal degeneration (**A**) and resistance to the extrapyramidal effects of haloperidol (**B**). The Mouse Phenome Database (MPD) has datasets examining the incidence of retinal degeneration in 26 inbred strains and the latency to move after 30 days of haloperidol treatment in 16 inbred strains, which are shown in the top graph of each panel. The AI pipeline outputs the gene symbols for that the genes it identifies as having the strongest genetic association with the phenotypic response pattern upon analysis of the published

literature. The output includes the calculated p-value for the genetic association, the genetic effect size, and the chromosome with the starting and ending position of each haplotype. The color of each square within the haplotype diagram indicates the haplotype of that strain. The literature score shows the strength of the association of that gene with the measured trait, which was performed using the indicated MeSH term. Of the genes associated with retinal degeneration, *Pde6b* has a perfect genetic association score, the highest literature association score, and it is the only gene with a direct association with retinal degeneration as indicated by the indicated PubMed identification number. Of the genes associated with susceptibility to haloperidol toxicity, *Abcb5* has a strong genetic association (genetic effect size 0.9) and is the only gene with a PubMed paper that is directly linked with the trait. *Pde6b* [26] and *Abcb5* [27] were previously shown to contain the causative murine genetic factors for retinal degeneration and susceptibility to haloperidol toxicity, respectively.



#	Gene	SNP Flag	Haplotype	EffectSize	Pvalue	FDR	Position	LitScore	PubMed
0	Cd180	2	🟦🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥	0.785	0.00000211382	0.000659641	13:102850747-102859716	0.912	Indirect
1	Parp10	2	🟦🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥	0.785	0.00000211382	0.000236118	15:76127135-76127135	0.852	Indirect
2	Draxin	2	🟦🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥	0.785	0.00000211382	0.000342675	4:148200038-148200038	0.832	19150847,23206892
3	Hcfc1r1	2	🟦🟥🟩🟥🟥🟩🟩🟪🟥🟩🟥🟥🟥🟩🟩🟪	0.771	0.0000750136	0.00511472	17:23890264-23903432	0.871	Indirect
4	Slc22a16	2	🟦🟦🟥🟥🟥🟥🟥🟥🟥🟥🟦🟥🟥🟥🟥🟥	0.415	0.00345635	0.0706927	10:40353401-40503418	0.961	Indirect
5	Galntl5	2	🟦🟦🟥🟥🟩🟪🟪🟩🟪🟩🟦🟥🟩🟩🟪🟩	0.337	0.042536	0.47682	5:25377931-25513025	0.424	Indirect

#	Gene	SNP Flag	Haplotype	EffectSize	Pvalue	FDR	Position	LitScore	PubMed
0	Nid1	2	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥	1.000	0	0	13:13640077-13640077	0.813	31877171,25347398
1	Ephx3	2	🟦🟥🟥🟦🟦🟦🟦🟥🟦🟦🟥🟦🟦🟦🟦🟦🟦🟥🟦🟦🟦🟦🟥🟦🟦🟩	1.000	0	0	17:32408705-32413853	0.678	Indirect
2	Dthd1	2	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥	1.000	0	0	5:63039664-63039664	0.128	Indirect
3	Mtss2	2	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟦🟦🟦🟦🟦🟦🟦🟩	1.000	0	0	8:111448935-111455804	0.476	Indirect
4	Gabarap	2	🟦🟦🟥🟦🟦🟦🟦🟦🟦🟦🟦🟦🟩🟦🟦🟦🟦🟦🟦🟦🟦🟩🟦🟩🟦🟥	0.425	0.000542049	0.0394737	11:69883098-69883526	0.908	Indirect
5	Ankrd55	2	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟦🟦🟦🟦🟦🟦🟥🟦🟥🟥	0.182	0.0156415	0.877158	13:112496559-112499085	0.516	Indirect
6	Hcfc1r1	2	🟦🟦🟥🟥🟦🟩🟦🟥🟦🟪🟥🟦🟦🟥🟥🟦🟪🟥🟦🟥🟦🟦🟥🟥🟦🟪	0.206	0.0414612	0.373036	17:23890264-23903432	0.828	Indirect

Figure S6. Using the INDEL database developed here, the AI pipeline correctly identifies the causative genetic factors for agenesis of the corpus callosum (**A**) and cataract formation (**B**). The Mouse Phenome Database (MPD) has datasets measuring the length of the corpus callosum in 16 inbred strains and the incidence of corneal opacity appearing in 26 inbred strains, which are shown in the top graph of each panel. The AI pipeline outputs the gene symbols for the genes it identifies as having the strongest genetic association with the phenotypic response pattern based upon analysis of the published literature. The output includes the calculated p-value for the genetic association, the genetic effect size, and the chromosome with the starting and ending position of each haplotype block. The color of each square within the haplotype diagram indicates the haplotype of that strain. The literature score shows the strength of the association of that gene with the measured trait, which was performed using the indicated MeSH term. Of the genes associated with agenesis of the corpus callosum, *Draxin* has a high impact INDEL, and it is the only gene with a direct association with the corpus

callosum as indicated by the indicated PubMed identification number. Of the genes associated with corneal opacity, *Nid1* has a strong genetic association (genetic effect size 1.0) and is the only gene with a PubMed paper that is directly linked with the trait. *Draxin* [28] and *Nid1* [29] were previously shown to contain INDELs that are the murine causative genetic factors for agenesis of the corpus collosum and cataract, respectively.

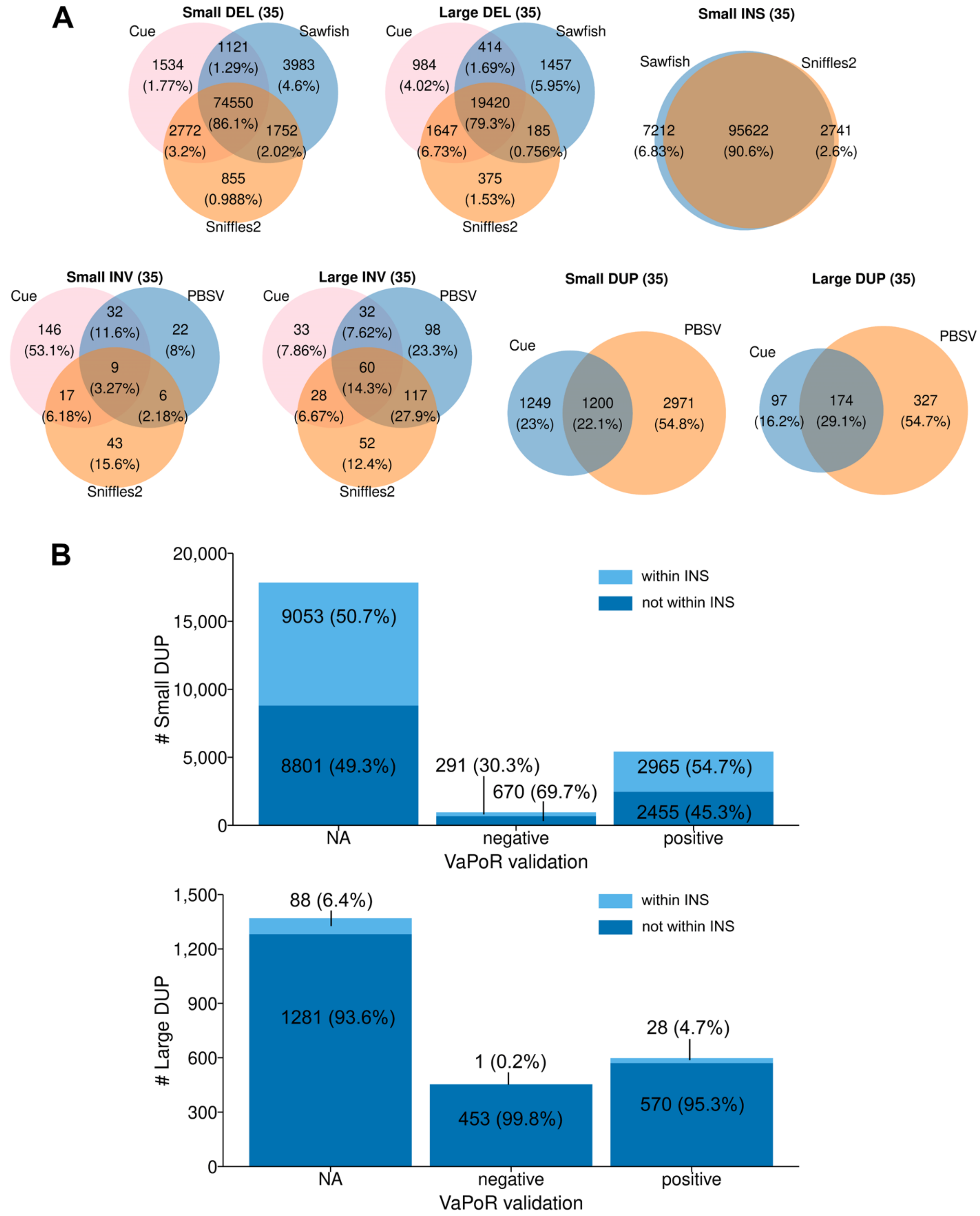


Figure S7. (**A**) SV identification programs have an acceptable level of agreement for identification of small (50 – 1000 bp) and large (1 kb – 500 kb) DELs, and for small INSs. The results obtained for the 35 classical inbred strains are shown in separate Venn diagrams. The number of each type of SV identified by each individual or combination of programs are shown within the circles (and percentages are shown within parenthesis). However, there was substantial divergence in the small and large INVs identified by Cue, PBSV and Sniffles2. Only the INVs that were validated by visual inspection (using the IGV) are counted. (**B**) Small and large DUPs identified by Cue and PBSV for the 39 inbred strains, and the VaPoR validation results are shown. Most DUPs could not be assessed (NA) by VaPoR. The assessable DUPs that passed (positive) or failed (negative) VaPoR validation are also shown. Also, 55% of the small DUPs were located within an INS, while most large DUPs were not within an INS.

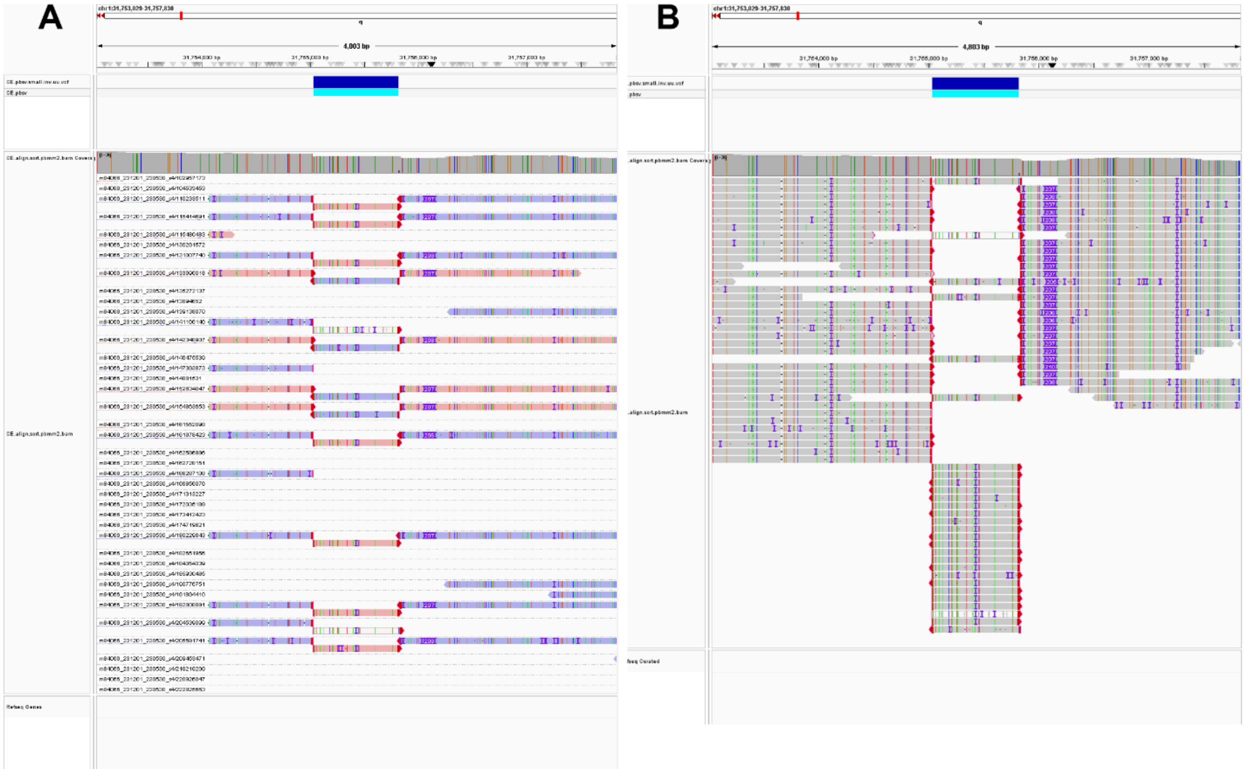


Figure S8. IGV images of an inversion (chr1:31755029-31755830) that were generated using the (**A**) altered or (**B**) conventional IGV preference settings. The inversion is more clearly seen when the altered preference parameters are used.

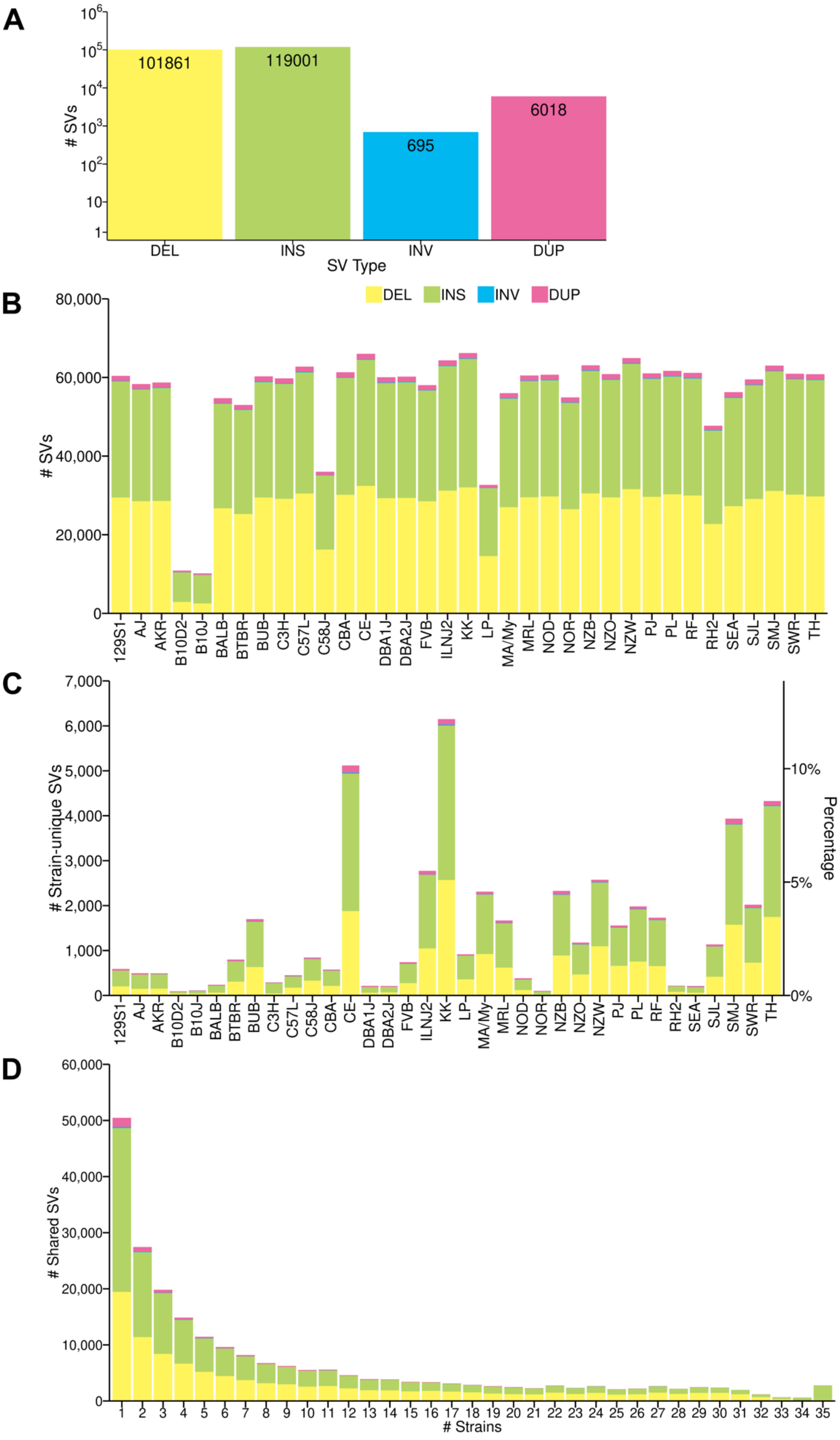

A
# SVs
10^6
10^5
10^4
10^3
10^2
10
1
101861
119001
695
6018
DEL
INS
INV
DUP
SV Type
B
DEL
INS
INV
DUP
# SVs
80,000
60,000
40,000
20,000
0
129S1
AJ
AKR
B10D2
B10J
BALB
BTBR
BUB
C3H
C57L
C58J
CBA
CE
DBA1J
DBA2J
FVB
ILNJ2
KK
LP
MA/My
MRL
NOD
NOR
NZB
NZO
NZW
PJ
PL
RF
RH2
SEA
SJL
SMJ
SWR
TH
C
# Strain-unique SVs
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
Percentage
10%
5%
0%
129S1
AJ
AKR
B10D2
B10J
BALB
BTBR
BUB
C3H
C57L
C58J
CBA
CE
DBA1J
DBA2J
FVB
ILNJ2
KK
LP
MA/My
MRL
NOD
NOR
NZB
NZO
NZW
PJ
PL
RF
RH2
SEA
SJL
SMJ
SWR
TH
D
# Shared SVs
60,000
50,000
40,000
30,000
20,000
10,000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
# Strains

Figure S9. (**A**) The number of different types of SVs [deletion (DEL), insertion (INS), inversion (INV) or duplication (DUP)] detected in the genome of the 35 classical inbred strains. The Y-axis is log-scale. (**B**) The total number of SVs detected in each strain in 35 classical inbred strains. The type of SV is indicated by the color within the bar (as shown in A). (**C**) The number (left y-axis) and percentage (right y-axis) of strain-unique SVs detected in the 35 classical inbred strains. (**D**) The number of strains with a shared SV allele of the indicated type are shown for the 35 classical inbred strains. Most of the minor SV alleles are shared by 1-3 strains.

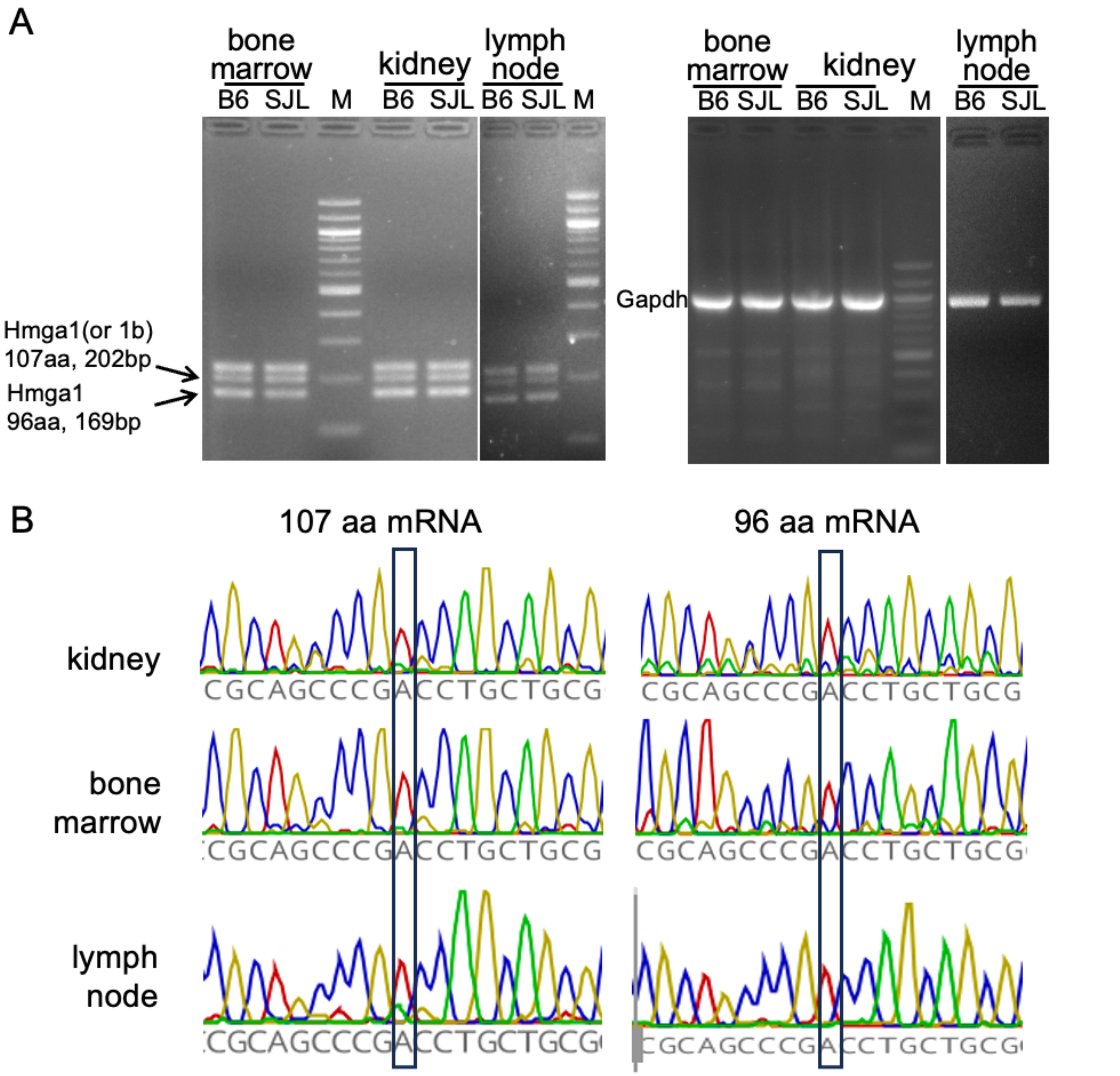


Figure S10. *Hmga1* encodes the mRNAs expressed in C57BL/6 and SJL bone marrow, kidney and lymph nodes. (**A**) RT-PCR was performed on hind limb bone marrow (femur, tibia), kidney and inguinal lymph node tissue obtained from C57BL/6 (B6) and SJL mice. The amplicons encoding the Hmga1 (96 amino acids) and Hmga1/Hmga1b 107 amino acid proteins are indicated by arrows (as in Fig. 4A). The patterns of amplicon expression in kidney, bone marrow and lymph nodes (left image) are identical to that in liver and spleen obtained from B6 and SJL mice (see Fig 4B). RT-PCR of *GAPDH* amplicons are shown as a control (right image). (**B**) RNA was prepared from kidney, bone marrow and inguinal lymph nodes obtained from C57BL/6 mice, and RT-PCR amplicons from *Hmga1* and *Hmga1b* mRNAs were generated and sequenced. The 3' UTR of *Hmga1* mRNA of C57BL/6 mice has an A, while *Hmga1b* mRNA has a G at the boxed position (as shown in Fig. 4). The sequencing results indicate that the mRNAs in C57BL/6 kidney, bone marrow and lymph nodes are encoded by *Hmga1*. This differs from the mRNAs in thymus, which are encoded by *Hmga1b (*Fig. 4*)*.