FEATURED ARTICLE

GENOME REPORT

G3 Genes | Genomes | Genetics

The Genome of C57BL/6J "Eve", the Mother of the Laboratory Mouse Genome Reference Strain

Vishal Kumar Sarsani,* Narayanan Raghupathy,* Ian T. Fiddes,† Joel Armstrong,† Francoise Thibaud-Nissen,‡ Oraya Zinder,* Mohan Bolisetty,* Kerstin Howe,§ Doug Hinerfeld, Xiaoan Ruan,†† Lucy Rowe,* Mary Barter,* Guruprasad Ananda,†† Benedict Paten,† George M. Weinstock,†† Gary A. Churchill,* Michael V. Wiles,* Valerie A. Schneider,‡ Anuj Srivastava,††,1 and Laura G. Reinholdt*,1**

††The Jackson Laboratory for Genomic Medicine, Farmington, CT, *The Jackson Laboratory for Mammalian Genetics, Bar Harbor ME, †UC Santa Cruz Genomics Institute, Santa Cruz, California, Department of Biomolecular Engineering, University of California, Santa Cruz, California, ‡National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, §Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA, UK, **NanoString Technologies, Inc, and ††The Jackson Laboratory for Genomic Medicine, Farmington, CT

ORCID IDs: 0000-0002-8455-3993 (A.S.); 0000-0003-4054-4048 (L.G.R.)

ABSTRACT Isogenic laboratory mouse strains enhance reproducibility because individual animals are genetically identical. For the most widely used isogenic strain, C57BL/6, there exists a wealth of genetic, phenotypic, and genomic data, including a high-quality reference genome (GRCm38.p6). Now 20 years after the first release of the mouse reference genome, C57BL/6J mice are at least 26 inbreeding generations removed from GRCm38 and the strain is now maintained with periodic reintroduction of cryorecovered mice derived from a single breeder pair, aptly named Adam and Eve. To provide an update to the mouse reference genome that more accurately represents the genome of today's C57BL/6J mice, we took advantage of long read, short read, and optical mapping technologies to generate a *de novo* assembly of the C57BL/6J Eve genome (B6Eve). Using these data, we have addressed recurring variants observed in previous mouse genomic studies. We have also identified structural variations, closed gaps in the mouse reference assembly, and revealed previously unannotated coding sequences. This B6Eve assembly explains discrepant observations that have been associated with GRCm38-based analyses, and will inform a reference genome that is more representative of the C57BL/6J mice that are in use today.

KEYWORDS
reproducibility
reference genomes
de novo genome assembly
long read sequencing
laboratory mouse
C57BL/6J
Mus musculus domesticus

The inbred mouse strain C57BL/6 (B6) is the most commonly cited and well-characterized laboratory strain used in biomedical research. For that reason, this strain was selected by the Mouse Genome Sequencing Consortium (MGSC) to represent the laboratory mouse reference genome (Marshall 2002; Mouse Genome Sequencing Consortium *et al.* 2002) in 1999 and it is the background strain on which the Knockout Mouse Project (Dickinson *et al.* 2016) is creating and phenotyping null alleles for all protein-coding genes. The original whole genome shotgun (WGS) draft assembly of the C57BL/6 genome (MGSCv3) was later updated to a finished, clone-based assembly (Church *et al.* 2009). The finished assembly is comprised predominantly of Sanger sequencing data from two bacterial artificial clone (BAC) libraries, RPCI-23 and RPCI-24, derived from the DNA from pooled tissues of 3 females (kidney and brain) and one male (spleen and brain) mouse, respectively, representing inbreeding generation F204-F207 from production colonies at The Jackson Laboratory, hence the sub-strain designation C57BL/6J (Church *et al.* 2009). Since 2010, the Genome Reference Consortium GRC has actively maintained the mouse reference genome and produced updated assemblies, beginning with GRCm38 (GCA_000001635.2) in 2012 and its six subsequent patch releases.

doi: https://doi.org/10.1534/g3.119.400071
Manuscript received February 8, 2019; accepted for publication April 16, 2019; published Early Online April 17, 2019.

Supplemental material available at FigShare: https://doi.org/10.25387/g3.7977044.
[1]Co-corresponding authors: E-mail: laura.reinholdt@jax.org, The Jackson Laboratory for Mammalian Genetics, 600 Main St, Bar Harbor, Maine, 04609; E-mail: anuj.srivastava@jax.org, The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032

Despite being one of the best-assembled and curated mammalian reference genomes, GRCm38 still contains 523 gaps within chromosome sequences, and there are nearly 300 unresolved issues that have been reported to the GRC (https://genomereference.org). In addition to gaps, these issues include reports of localized sequence mis-assembly, missing genic and non-genic sequences, sequencing errors and suspect variation. These types of assembly issues inflate false positive rates in reference-based variant calling. For example, we reported an analysis of systemic exome variants called across a wide variety of mouse strains (including C57BL/6J) and showed that a significant fraction of these overlap with regions with annotated reference assembly issues and/or gaps (Fairfield *et al.* 2015). While a small fraction of these are expected due to private variation in the reference genome, the vast majority are likely technical artifacts resulting from unreported issues in the reference genome assembly, or regions where paralogous gene copies are not fully represented.

In an effort to close gaps and resolve other issues in the current mouse reference genome, to minimize variant calls associated with GRCm38-private variation (*i.e.*, to bring the mouse reference genome sequence closer to the C57BL/6J mice that are currently in use), to provide a *de novo* assembly representing a single individual, and to identify additional data to support unannotated genes, we used high coverage, long-read sequencing, optical mapping and short-read data to generate a *de novo* genome assembly from C57BL/6J Eve.

MATERIAL & METHODS

Sample preparation and sequencing - PacBio

Genomic DNA samples were extracted using both kidney and brain samples from the C57BL/6J Eve female (MouseID 03-03685 (The Jackson Laboratory), Strain ID 00664, born 8/27/2003, generation F223) by phenol-chloroform extraction of a nuclei-enriched pellet. DNA samples were resuspended in TE buffer to a final concentration of 300-400 ng/ul, 260/280 1.8-1.9(Taylor and Rowe 1984). The PacBio data were generated from three libraries prepared using the Pacific Biosciences SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, Menlo Park, CA, USA) using the "20-kb Template Preparation Using BluePippin Size-Selection System (15-kb Size cutoff)" protocol obtained from PacBio SampleNet. The BluePippin (Sage Science, Inc, Beverly, MA, USA) was set to collect from 7-50 kb. After sequence length QC, the resulting sized libraries were repaired using the "Procedure & Checklist- 10 kb Template Preparation and Sequencing" protocol. All libraries were sequenced using 294 SMRT cells on Pacific Biosciences RS II platform (P6C4 chemistry). One of these libraries was generated and sequenced (10X) by Pacific Biosciences using the same protocols and chemistries. The other two libraries and the remaining coverage were generated and sequenced at The Jackson Laboratory.

For IsoSeq cDNA sequencing, RNA was extracted from archived whole brain samples from C57BL/6J Eve. 1 microgram of input RNA was used to generate cDNA (Clontech SMARTER cDNA synthesis kit), cDNA was size selected (3-6 kb) by BluePippin, and 400 ng of SMRTBell library was prepared as above. The library was sequenced on the Pacific Biosciences RSII platform (P6v2 chemistry), 532,941 reads were generated with a mean insert length of 3,004 bp. Quiver (Chin *et al.* 2013) was used to predict consensus isoforms and for polishing. There were 31,076 high-quality isoforms, and 11,661 low-quality isoforms with average consensus read length of 3,042 bp.

Sample preparation and sequencing – Illumina short read

Genomic DNA was fragmented and Illumina whole genome libraries were constructed using the methods described in Hodges *et al.* (2009). Steps 1-28 were followed to produce a whole genome library that was then sequenced rather than used in the enrichment portion of the protocol. The library was quantified by QPCR and sequenced on six lanes of an Illumina HiSeq GAIIX (Illumina, San Diego, CA, USA) using a 100 base paired end sequencing protocol.

Sample preparation and sequencing – Bionano optical mapping

DNA Isolation: High-molecular weight DNA was extracted from mouse spleen. 70 mg of mouse frozen spleen tissue was place on a Petri dish over ice and chopped with a razor blade into approximately 2 mm chunks. The tissue was then transferred into a 15 mL conical. 1 mL of fixing solution (2% (v/v) formaldehyde, 10 mM Tris, 10 mM EDTA, 100 mM NaCl, pH 7.5) and left on ice for 30 min. Fixing solution was pipetted out and discarded. Tissue was washed 3 times by adding 2 MB Buffer (10 mM Tris, 10 mM EDTA, 100 mM NaCl, pH 9.4), swirling tube, and pipetting off MB Buffer. 2 mL of MB Buffer was added after the third wash. The tissue was blended using a fixed rotor-stator homogenizer (TissueRuptor, Qiagen #9001271) on high speed for 10 sec. The homogenate was transferred to a 2 mL microfuge tube and spun down at 2000 rcf for 5 min. at 4°. Supernatant was removed, pellet was resuspended in 1.5 mL of MB Buffer, and spin was repeated. Supernatant was removed and final pellet was resuspended in MB Buffer.

Resuspended cells were embedded into low-melting point agarose gel plugs, using the CHEF Mammalian Genomic DNA Plug Kit (BioRad #170-3591), following the manufacturer's instructions. Plugs were made with 10 mg, 15, mg, and 20 mg equivalents of the original starting material (mass equivalents calculated based on volume of final resuspended pellet). The plugs were incubated with Lysis Buffer (Bionano #20270) and Puregene Proteinase K (Qiagen #1588920) overnight at 50°, then again the following morning for 2 hr. (using new buffer and Proteinase K). The plug was washed, melted, and solubilized with GELase (Epicentre #G09200). The purified DNA was subjected to 4 hr. of drop dialysis (Millipore, #VCWP04700) and quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen/Molecular Probes #P11496). The plug made with 10 mg equivalents of starting material had a concentration of 286 ng/μL and was clear and viscous, so it was selected for further processing.

DNA Fluorescent Labeling: DNA was labeled according to commercial protocols using the NLRS kit (Bionano Genomics, #80001). Briefly, 300 ng of purified genomic DNA was nicked with 7 U nicking endonuclease Nt.BspQI (New England BioLabs (NEB), #R0644) at 37° for two hrs. in NEBuffer3. The nicked DNA was labeled with a fluorescent-dUTP nucleotide analog using Taq DNA Polymerase (NEB, #M0267) for one hr. at 72°. After labeling, the nicks were ligated with Taq DNA Ligase (NEB, #M0208) in the presence of dNTPs. The backbone of fluorescently labeled DNA was counterstained using the DNA Stain from the NLRS DNA Labeling Kit.

Data Collection: The labeled DNA was loaded onto Irys chips (Bionano, #20247) and inserted into the Irys instrument. The instrument automated the electrophoresis of the DNA into nanochannels, thereby linearizing them with uniform stretch throughout the molecule. The stationary molecules were then imaged, and the automated process of electrophoresis followed by imaging was repeated for multiple cycles until the desired amount of data were collected. The stained DNA molecule backbones and locations of fluorescent labels along each molecule were automatically detected using the in-house software

package, IrysView. A total of 244 Gbp (~80X coverage depth) of data were generated, using 3 Irys chips.

Processing raw data - PacBio

Quality trimming of sequenced reads: Raw data from 294 SMRT cells were imported into the SMRT portal (https://bit.ly/2S63Tav) and subreads were extracted from the raw h5 files within PacBio SMRT portal. Subreads with polymerase read were further filtered according to the following criterion: quality < 75, read length < 50, and polymerase read length < 50. Finally, after filtering, 32,210,376 subreads (mean subread length of 5,753) were extracted from 20,081,751 raw reads, providing theoretical coverage of 66X for *de novo* assembly.

Error correction: Error correction of reads was accomplished with the MinHash Alignment Process (MHAP)(Berlin *et al.* 2015) within PBcR. Continuous benchmarking of correction parameters (k-mer size, hash size, min-mer size, error rate) was done to obtain the best possible set of corrected subreads. Our analysis indicated that usage of more sensitive parameters (MhapSensitivity = highOvlErrorRate = 0.05) significantly increased run time but overall improved the quality of corrected reads.

Sequence assembly - PacBio

Corrected reads from MHAP were assembled using the Celera assembler (CA 8.3) (default parameters)(Myers *et al.* 2000), which requires ~60-70 gigabases of corrected sequence and consists of overlapper, unitigger, scaffolder and consensus steps to reconstruct genomes from corrected long reads.

Hybrid scaffolding

Single molecule high-resolution maps of the B6Eve genome were obtained using the Bionano Irys System (Das *et al.* 2010). Label positions captured in images and molecule map lengths were stored in CMAP format files (consensus map). The hybrid scaffold tool from Bionano genomics was used to further extend the scaffold size by combining the PacBio *de novo* assembly and genome map data of B6Eve. The hybrid scaffold pipeline created an alignment between the datasets and constructed super scaffolds excluding the conflicting alignments[33].

Polishing and assembly evaluation

The hybrid scaffolded assembly was polished using Quiver (Chin *et al.* 2013) to improve consensus accuracies in the range of Q60 and to reduce the high indel errors that are expected in the PacBio sequencing data (Zuo *et al.* 2012) . pbalign (https://github.com/PacificBiosciences/pbalign) was used to create a bam file from all h5 files of SMRT cells and Quiver trained on P6-C4 chemistries were used to obtain the consensus corrected assembly. This assembly was further improved by using Pilon (Walker *et al.* 2014) with default parameters, that corrects bases, fixes mis-assemblies and fills gaps provided a draft assembly and paired-end Illumina sequencing data. Nearly 32X of Illumina data from B6Eve was used as an input to Pilon to fix the bases in B6Eve assembly. The GATK variant calling pipeline (following best practices) was used to call variant using Illumina data on the B6Eve assembly to judge the improvement in overall quality at each step of polishing. Finally, the QUAST tool (Gurevich *et al.* 2013) was used (–split-scaffolds) to compare the final assembly with GRCm38. The split-scaffolds option breaks the assembly and performs reconstruction of "contigs" which were used to build the scaffolds to compare the effectiveness of scaffolding.

RefSeq transcript alignments

Murine "known" RefSeq transcripts (those with NM and NR prefixes) were queried from NCBI Entrez on September 11, 2017 and aligned to the Pilon-corrected B6Eve assembly and GRCm38 full assembly (GCF_000001635.20). From these analyses, the counts of transcripts with low quality alignments, split alignments or no alignment to the GRCm38 primary assembly unit and B6Eve were determined, as were the counts of transcripts dropped for co-location, (Taylor and Rowe 1984). From the same set of RefSeq transcripts, we additionally identified alignments to GRCm38 and to the Pilon-corrected B6Eve containing frameshifting and non-frameshifting indels in CDS (Schneider *et al.* 2017). The frameshift analysis of the pre-polished/corrected assembly used a set of known RefSeq transcripts queried on February 28, 2016.

LiftOver construction

LiftOver was performed between the Eve assembly and GRCm38 reference genome using the same species lift over construction procedure[9] outlined by University of California Santa Cruz Genome Bioinformatics Group. Same species lift over construction contained two steps a) BLAT alignments and b) chaining and netting to obtain the lift over file. Genome loci of the Eve assembly were further converted into GRCm38 coordinates using the LiftOver[8] tool from UCSC utilities.

Repeat content assessment

The repeat elements in the GRCm38.p6 (excluding ChrY and alternate loci scaffolds) and B6Eve assembly were determined by RepeatMasker (Smit *et al.* 2013–2015) trained on the mouse model by excluding RNA elements (-norna). A chi-square test was performed to identify the repeat classes that are enriched in one genome over the other.

Repeat analysis of unaligned B6Eve sequences

There were unaligned B6Eve sequences from three different alignment methods a) Cactus based alignment using UCSC Comparative Annotation Toolkit which was used for the B6Eve annotation, b) NCBI BLAST-based alignment of B6Eve to GRCm38, and c) QUAST (minimap2 aligner). A consensus set (common among all three set), was constructed using BEDTools (Matera *et al.* 2008). RepeatMasker was used to identify the repeat content in the common unaligned region. A chi-square test was used to test the differences in repeat content for each of the repeat classes of common unaligned sequence against GRCm38.

Resolving recurring variants

Recurring variants present in ≥ 75% of strains, detected in previous whole exome sequencing efforts [10] were extracted. Fixed (homozygous) variants from the mouse Collaborative Cross genome[11] project, as well as fixed variants from 24 C57BL/6J pedigrees descendent from Eve[12] were also obtained (G. Ananda and G. Churchill, unpublished data). We used LiftOver to remap the genomic coordinates, and a recurring variant was said to be resolved if the ALT allele of a recurring variant in exome data matches the REF allele in the B6Eve assembly. The analysis was restricted to only homozygous variant calls.

B6Eve annotation

B6Eve was annotated using the Comparative Annotation Toolkit (Fiddes *et al.* 2018) (https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit commit c7852b4). As input, CAT was given a progressive Cactus alignment generated with rat rn6 (GCA_000001895.4) and human GRCh38 as outgroups as well as the GENCODE VM11 annotation on mouse GRCm38. CAT was provided with extrinsic transcript information from RNA-seq as well

as IsoSeq. For mouse GRCm38, the same RNA-seq used in the CAT publication was used. RNA-seq data were generated from whole brain of a female C57BL/6J Eve descendent and aligned to the B6Eve assembly. To guide AugustusPB in detecting novel isoforms, a total of 26,188 IsoSeq full length cDNA reads were aligned to the B6Eve assembly.

Novel isoform detection

To detect novel isoforms, homGeneMapping (Konig *et al.* 2016) was used to map GENCODE VM11 annotation coordinates onto B6Eve, and these splice junction coordinates compared to AugustusPB and AugustusCGP transcript predictions filtered for IsoSeq support. Transcripts with annotation support were filtered out. The remaining candidate novel isoforms then were checked to see if they overlapped a comparatively annotated locus and if they contained either a fully novel exon or a splice site shift based on bedtools (Quinlan *et al.* 2010) intersections.

SV detection & gap filling

PacBio read alignment: Raw PacBio reads were aligned to GRCm38 using the long-read aligner NGMLR version 0.2.6. CoNvex Gap-cost alignments for Long Reads (NGMLR) (Beal *et al.* 2012a) is a long-read aligner designed to align PacBio reads with the focus on identifying structural variations. Stringent alignment requirements were used for identifying SVs: -i 0.85 argument to disregard alignments with identity with less than 85% and -R 0.5 option to ignore alignments containing less than 50% of the read length.

Structural variant calling: Sniffles (Beal *et al.* 2012a) (default parameters) was used to call SVs from the alignments produced by NGMLR. GATK was used to process Illumina WGS data from B6Eve. Best practices were used to generate a BAM file and SVS were called using Delly v0.7.7. SURVIVOR-1.0.3 (Buac *et al.* 2008) package was used to perform the integration of PacBio and Illumina calls.

Gap filling: We extracted the coordinates of the gaps from the GRCm38 chromosomes and further extended this to include the 50 Kb flanking both sides of the gap. We aligned B6Eve scaffolds to these padded regions using minimap2. We filtered the candidate alignments according to the following criterion: a) Must be the reciprocal best hit, b) Total alignment length >= 80KB, and c) Align to one unique location to the reference (extracted this information from assembly-assembly alignments). The retained alignments were further visualized in Integrated Genomic Viewer (IGV) to inspect insertion/deletion patterns around the gap region. To confirm and extract the gap spanning a B6Eve scaffold, we performed the reciprocal alignments, aligning the padded gap regions to B6Eve scaffolds, using minimap2. We filtered out candidate alignments not satisfying criteria mentioned above and visualized the retained alignments in IGV to inspect whether we observe the opposite of previously found insertion/deletion pattern. The sequence and locus of confirmed gap spanning B6Eve scaffolds were extracted and subjected to GRC internal curation.

Data availability

C57BL/6J mice and breeding generation information are available through the Jackson Laboratory. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession LXEJ00000000. The version described in this paper is version LXEJ02000000. The raw PacBio, Illumina and Bionano data used were deposited at NCBI BioProject under accession PRJNA318985. The B6Eve assembly along with annotation and an assembly hub are available at ftp link (ftp://ftp.jax.org/b6eve). Visualization of the assembly can be found at https://genome.ucsc.edu → MyData → Track Hubs → My Hubs with the following URL: ftp://ftp.jax.org/b6eve/assemblyhub/hub.txt. All supplementary figure, table, and file names and descriptions are listed in FileS1. A list of the B6Eve scaffold names mapped to GenBank accessions is available in FileS2. Supplemental material available at FigShare: https://doi.org/10.25387/g3.7977044.

RESULTS AND DISCUSSION

The Jackson Laboratory manages the rate of genetic drift through periodic replenishment of foundation breeding colonies from pedigreed, cryopreserved embryo stock (Wiles and Taft 2010) that are three generations removed from a single brother-sister breeder pair, "Adam" and "Eve" (Figure 1). This process introduces a controlled bottleneck that minimizes the accumulation of genetic change. These two individual mice capture an evolutionary snapshot in time at inbreeding generation F223 (Figure 1). Therefore, any C57BL/6J individual obtained from the production colonies at The Jackson Laboratory today is limited to a maximum of 24 inbreeding generations removed from the mice whose DNA was used to generate the C57BL/6J reference assembly, GRCm38 (Figure 1). Under the highly selective breeding paradigms employed for inbred laboratory strains, this genetic distance is sufficient for rapid fixation of 98.7% of variants, such that today's C57BL/6J mice are by definition a sub-strain of the animals from which GRCm38 is derived (Green 1981). Therefore, we chose DNA isolated from one of these individuals as the material for our *de novo* assembly with the goal of providing a genome sequence from a single individual that is not more than eight generations removed from C57BL/6J mice sourced from The Jackson Laboratory today, and that might also be used to improve the current C57BL/6J reference assembly (GRCm38). We chose C57BL/6J Eve (B6Eve) to get balanced representation of the X chromosome and the autosomes. Future efforts are focused on a *de novo* assembly of Adam, where *de novo* assembly the Y chromosome will require more specialize approaches.

Sequence assembly and evaluation

To generate data for our *de novo* assembly of the B6Eve genome, we used a range of technologies, including Pacific BioSciences (PacBio) long read technology at 66X whole genome coverage (Table S1), Illumina short read at 32X whole genome coverage, and Bionano Genomics (BNG) optical maps. The overall assembly procedure involved 1) correction of PacBio reads, 2) creation of contigs from PacBio, 3) extension of contigs to scaffolds using optical maps, 4) polishing of the assembly, and 5) further correction of the assembly using Illumina data (Figure 2).

To assess base-pair level improvements in assembly quality afforded by each of these steps, we mapped the Illumina reads of B6Eve and called variants at each step using the GATK HaplotypeCaller (Depristo *et al.* 2011) (Table S2). The scaffolded assembly yielded 1,664,599 variants, the majority of which were insertions (~70%), followed by SNPs and small deletions. This pattern is reminiscent of the error profile generated from PacBio technology (Zuo *et al.* 2012). To improve base pair accuracy, we used Quiver software (Chin *et al.* 2013) to polish the assembly and reduced the total number of variants to 505,782. We found that compared to the unpolished assembly, Quiver reduced the number of insertions to just 22%. Finally, we used B6Eve Illumina data itself to correct the polished assembly with Pilon (Walker *et al.* 2014), a tool that improves the quality of the draft assemblies using read alignment analysis. After Pilon correction, the number of variants was narrowed down to 310,205; of which 227,523 were considered high quality (PASS) by GATK HaplotypeCaller (File S3).

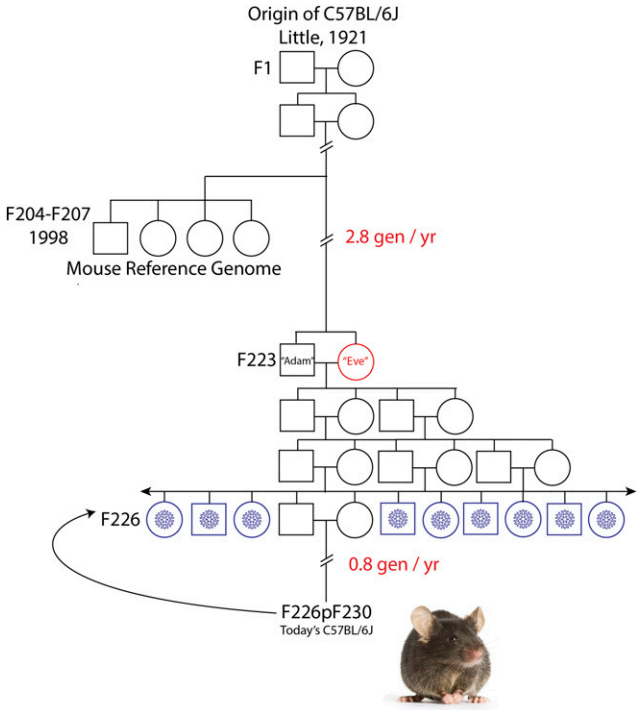

Figure 1 Origin of the inbred strain C57BL/6J. Inbred laboratory mouse strains are maintained by brother x sister mating. Filial (F) generations from which mice contributing to the reference assembly clone libraries and from which the B6Eve mouse were derived are shown. Cryopreserved embryo stock is represented by blue snowflakes at F226, 3 generations from Adam and Eve at F223. Generations subsequent to the cryopreservation event are F226p###, *e.g.*, F226p230, which means embryos cryopreserved at F226 were recovered and there were an additional 4 generations of subsequent inbreeding.

Since these sequencing data sets were generated from the same individual female, we surmised that most of these high-quality variant calls were due to technical artifacts arising from errors in the B6Eve genome assembly, alignment errors, or errors in the Illumina data. Although we couldn't exclude the possibility of true variant calls resulting from somatic variation since different tissues were used to source DNA. To explore these variants in more detail we plotted the distributions of alternate allele frequency, genotype called, and total depth relationship for these (File S4). We found that the vast majority of these variant calls were low frequency calls, which are more commonly due to technical artifacts or somatic mosaicism. To place these variants onto the mouse reference genome, we lifted them over to GRCm38 where roughly 95% (210,571 [SNP:152,486 and Indels:58,080]) could be placed onto a chromosome (File S5). We found that 62% percent of these variants mapped repeat elements making it likely that the variants are technical artifacts due to alignment issues. Many of these variants also mapped to a prominent cluster on Chr12 ~17.5 MB containing 10,464 variants within a 1.5 Mb region including exonic, intronic, and intergenic sequence (File S6). The non-random distribution of these variant calls also supports the conclusion that they were technical artifacts due to misalignment of reads in a poorly assembled region in B6Eve. A closer look at the GRCm38 assembly revealed the presence of segmental duplications in this region, which may have caused a collapse in the B6 Eve assembly.

We also called variants using Eve Illumina data aligned to GRCm38 and found significantly fewer variants (69,089 [SNP: 50,143 and Indel: 18,946]), about 10% of which were shared with the variants resulted from the mapping of Eve Illumina data to the Eve assembly (File S7). Except for a few minor clusters, including one on Chr1, these shared variants were uniformly distributed. The shared Chr1 cluster had 560 variants which were present in both variants call sets (Eve Illumina-Eve PacBio Assembly and Eve Illumina to GRCm38). Of these variants, there were two prominent clusters on Chr1: 85062206-85279614, Chr1: 88212297-88310615 with 100 and 110 variants, respectively. The first region includes the *Csprs* gene and second includes a known GRCm38 assembly issue which was also unresolved in the B6Eve assembly. Similar to the Chr12 region above, these regions on Chr1 are flanked by segmental duplications in GRCm38.

We used an automatic assembly quality evaluation tool, QUAST (Gurevich *et al.* 2013), to assess the overall quality of the Pilon corrected assembly. We found that during assembly-assembly alignment, 96.8% of the B6Eve assembly aligned to 97% of the GRCm38 (excluding ChrY and alternate sequences) reference genome chromosomal sequences and 54% of unplaced & unlocalized sequences (Figure S1). Only 166 and 2,456 B6Eve components comprising of 3.7 Mb and 7.2 Mb of sequences remained wholly or partially unaligned to the reference genome, respectively. We also found that the K-mer based completeness of B6Eve was very high at 97.4%, suggesting high coverage and per-base quality. The detailed QUAST report for complete and broken-down (breaks the assembly by continuous fragments of N's of length ≥ 10) versions of the assembly is found in File S8. Taken together, our resulting PacBio-only *de novo* B6Eve genome assembly was 2.53 Gb consisting of 14,551 contigs (longest contig = 4,574,471 and 2.3% of total contigs exceeding 1Mb) with an N50 size of 401,294 bp. Our complete PacBio-Bionano hybrid assembly yielded an N50 of 1,290,032 bp with a total assembly size of 2.79 Gb, which was a 3X improvement (in N50) over the PacBio-only assembly (Table 1).

Gene Content Analysis

We also evaluated the gene content of the B6Eve assembly as another measure of assembly quality. Akin to previous analyses (Taylor and Rowe 1984; Schneider *et al.* 2017) we aligned 36,009 RefSeq transcripts to the GRCm38 primary assembly (also excluding unplaced sequences and alternate loci scaffolds from other mouse strains that are a part of the full assembly) and to the "Piloned" B6Eve assembly versions (Table 2 and Table S3). We observed that B6Eve provides a comparable representation for total gene content, as compared to GRCm38, with only 19 non-chromosome and ChrY associated sequences having no alignment. Consistent with the more fragmented nature of the B6Eve assembly, however, a greater number of aligned RefSeq transcripts exhibited partial alignments or alignments split over multiple scaffolds than in GRCm38. We examined the co-placement of transcripts representing different genes as a proxy for measuring the collapse of segmental duplications. Although B6Eve showed a greater number of co-placed transcripts than GRCm38, these numbers were consistent with those seen in other high-quality long-read derived WGS assemblies, demonstrating the utility of this mouse assembly. To gauge the impact of the Illumina-read correction step on the quality of protein representation in the B6Eve assembly, we looked at the incidence of frameshifting indels in aligned RefSeq transcripts prior to and after this step(Florea *et al.* 2011) (Table S3). Although the Pilon corrected assembly still exhibited more frameshifts than GRCm38, we found that this step resulted in a substantial improvement in functional representation (protein coding sequence).

Reference Assembly Gap Filling

The GRCm38 chromosome assemblies contain 440 gaps (excluding centromere, short arm, and telomere gaps). We assessed whether sequences in the B6Eve assembly could resolve these gaps. Based on

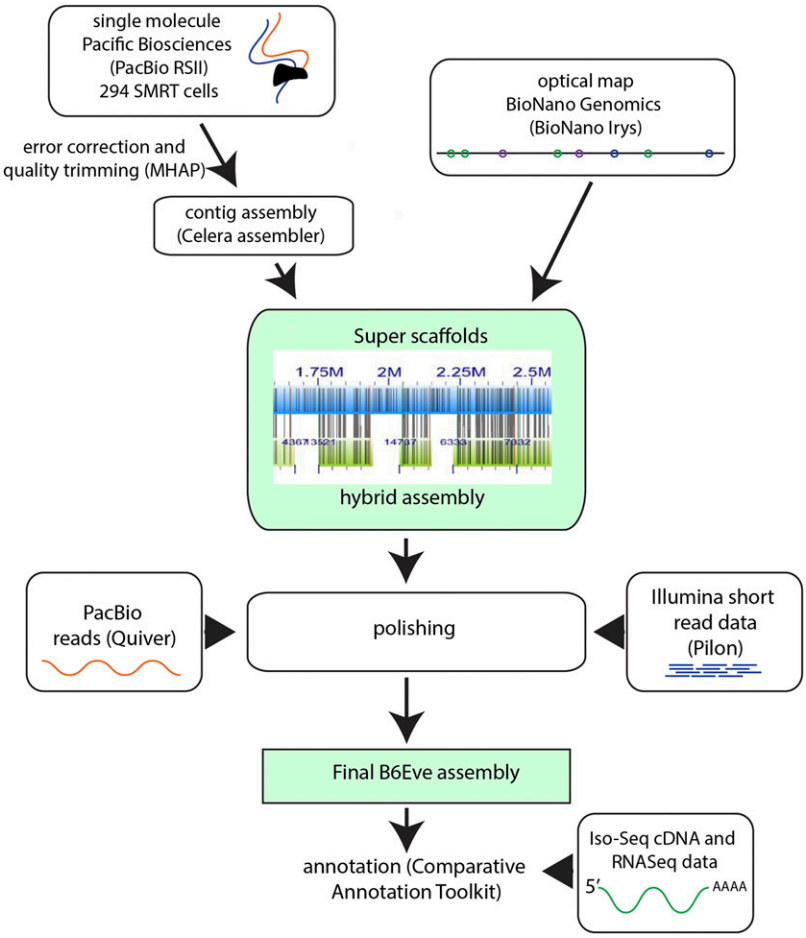

Figure 2 Schematic overview of the *de novo* assembly procedure for B6Eve. Details are described in Methods.

our gap-filling methodology (see Methods), the B6Eve assembly spanned 23 gaps in the GRCm38 chromosomes (Table S4 and Figure S2). In several instances, we observed discrepancies between the gap length reported in GRCm38 and the amount of sequence provided by the B6Eve assembly. For example, the B6Eve assembly spanned a 1,760 bp intra-scaffold gap located at Chr2:172,624,657-172,626,416 bp in GRCm38, with 1,620 bp (a 140 bp relative deletion) (Figure S2a). In other cases, B6Eve spanning sequences were longer than assembly gaps. For example, a B6Eve assembly scaffold spanned the 100 bp intra-scaffold gap at Chr1:183,334,907-183,335,006 bp in GRCm38, with 595 bp sequence (Figure S2b). These discrepancies were not unexpected, however, as the methods used to estimate reference assembly gap sizes do not always offer base-pair level resolution, and also because the GRC assigns default gap lengths when no sizing estimates are available (https://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/TPF_Specification_v1.8_20131106.docx). Consistent with prior reports that remaining reference assembly gaps are in complex genomic regions (Church *et al.* 2011), we observed that our gap spanning sequences had a repeat content of 45.9% (*vs.* 42.5% of GRCm38 total sequence) repeats, with simple repeats accounting for 18.1% (vs 2.6% in GRCm38 total sequence).

Variant analysis

Previously, we reported whole-exome sequences from of a collection of nearly 200 unique strains of spontaneous mutant mice maintained at The Jackson Laboratory (Fairfield *et al.* 2015). In our analysis of these exomes, we found that there were 855 coding variants (SNPs) common across 75% or more of the samples, which we attributed to errors with the reference genome itself due to their significant inclusion within the component-mapped boundaries of GRC incident features. We investigated the subset of these exome variants (n = 126) that were homozygous across all strains (100% allele frequency). We found that 10 (7.9%) matched the B6Eve assembly allele rather than the GRCm38 allele, supporting the assertion that high frequency alleles are putative indicators of reference assembly error (Figure 3, Table S5).

To extend this analysis to the whole genome, we performed a similar analysis using variant calls from sixty-nine multi-parent, recombinant

Table 1 Number of sequences, N50 size and assembly length for Bionano optical map, PacBio *de novo* assembly and scaffolded assemblies

	Bionano Genomics optical map	PacBio *de novo* assembly	PacBio only Hybrid	Bionano optical only Hybrid	Final Assembly (LXEJ02000000)	Improvement relative to PacBio assembly
Number of sequences	3,016	14,551	3,732	1,652	12,690	
N50 (in MB)	1.18	0.40	0.58	1.97	1.29	3.2
Assembly size (in MB)	2,482.74	2,535.01	1,820.29	2,470.31	2,789.93	1.3

Table 2 RefSeq Transcripts Alignment Table From NCBI

	GRCm38	B6Eve
Assembly accession	GCF_000001635.20*	na
Number of sequences retrieved from Entrez	36,009	36,009
Number of "alignable" sequences (B6Eve count excludes sequences from ChrY)	36,009	35,948
Number of "alignable" sequences not aligning	7	16
Number of sequences with multiple best alignments (split transcripts)	27	1,621
Number of sequences with CDS coverage < 95%	57	1,644
Number of NMs dropped at consolidation	8	284
Number of NRs dropped at consolidation	1	44
Placements with frameshifting indels (FS)†	52	Pre-correction: 8,566 Post-correction: 335
Placements with non-frameshifting indels (NFS)†	57	Pre-correction: 55 Post-correction: 32

*Transcripts were aligned to the GRCm38 full assembly (GCF_000001635.20), which includes alternate loci scaffolds from a variety of mouse strains. Counts shown in Table 2 reflect only transcript alignments to the GRCm38 primary assembly unit (GCF_000000055.19), which is comprised only of C57BL/6J sequences, unless noted.

†Frameshift counts are shown for alignments to the GRCm38 full assembly, including alternate loci scaffolds. Pre-correction: assembly prior to Quiver polishing and Pilon correction.

inbred strains (Collaborative Cross (CC) strains, a panel derived from eight founder laboratory strains) (Srivastava *et al.* 2017). We found 14,757 variants (SNPs) shared across all strains, using C57BL6/J as a reference genome. Out of 14,757 variants, 2,407 are homozygous across all strains. Consistent with the results of the exome variant analysis, 307 of these variants (12.8%) (Figure 3, Table S5) matched the B6Eve assembly allele rather than GRCm38.

Finally, we analyzed variant calls from whole genome short read sequencing of 24 recent descendants of B6Eve, representing multiple inbred lineages. We found 3,203 homozygous variants (SNPs) common across these samples; of these, 2,194 (68.5%) met minimum alignment criteria for remapping to B6Eve. Of these 2,194, we found 393 cases (12.6% of the total variation and 17.9% of net variation) (Figure 3, Table S5) where the reported alternate alleles matched the B6Eve assembly.

Taken together, our analyses identified 503 single nucleotide positions in GRCm38 (excluding variants common in three datasets) that are not representative of today's C57BL/6J mice. Two of these are non-synonymous SNPs in *Akap9 and Sfi1*. *Akap9* (A kinase anchoring protein 9) is a protein that is responsible for cytoskeletal organization and is required for formation and maintenance of the blood-testis barrier, and male fertility (Schimenti *et al.* 2013; Venkatesh *et al.* 2016). Using allele specific PCR, we confirmed the presence of the alternate *Akap9* allele in B6Eve and in randomly selected descendants of Eve, and not in an ancestor of B6Eve. Therefore, *Akap9* represents a variant that arose in and is now stably maintained in C57BL/6J mice in the years since the sequencing GRCm38. *Sfi1* is predicted to be a spindle assembly associated protein on the basis of homology with a yeast cytoskeletal protein of known function, however no phenotypic alleles have been reported in mice (Salisbury 2004). When we attempted to validate this variant, however, we discovered that the alternate allele is indeed represented in the mouse reference genome, though it aligned to an unplaced scaffold (JH584304.1: 14,038-14,339). Flanking sequence variation between this scaffold and Chr11 allowed us to design allele specific primers, with which we confirmed that both alleles are present in DNA samples from B6Eve and from randomly selected descendants of B6Eve, supporting the idea that the unplaced scaffold indeed represents C57BL/6J sequence. Previously published SV data for C57BL/6J showed that the mouse genome potentially harbors 20-30 copies of this gene (Quinlan *et al.* 2010). Therefore, the recurrent "variation" observed in this gene is likely not allelic, but due to mis-mapping of reads from paralogous gene copies to the *Sfi* locus that is currently represented on GRCm38 Chromosome 11. Paralogous gene variation may be a previously underappreciated source of variation, since we observed a relative enrichment of variants within certain genes (*e.g.*, *Tulp4*, Table S5). Previous studies have shown that GRCm38 is missing paralogous copies of many genes (Church *et al.* 2009; Church *et al.* 2011), some of which may be represented on unplaced scaffolds as we found for *Sfi1*.

Structural Variation

We aligned raw B6Eve PacBio reads to GRCm38 using NGMLR (Beal *et al.* 2012a) and called structural variants (SVs) with Sniffles (Beal *et al.* 2012a). We also aligned Illumina WGS data from B6Eve and called SVs with Delly (Beal *et al.* 2012b) (Table 3). The median size of detected duplication, deletion and inversion events from Delly were 901, 2,610 and 12,362 bp, respectively. Similarly, from PacBio data, the median size of duplication, deletion, inversion and insertion events were 432, 77, 1352, and 92 bp, respectively.

We found 12 deletion, 43 duplication and 4 inversion calls that were common in both Illumina and PacBio data (Table S6). Of these common SVs, 8 deletions, 30 duplications, and 4 inversions overlapped genes (Table S6), though mostly within noncoding intronic regions. We used DGVa to further investigate the SVs overlapping genes. Each of these were associated with multiple (21-124) DGVa entries representing germ line SV across genetically diverse inbred strains from multiple strain surveys of SV (Cutler *et al.* 2007; Graubert *et al.* 2007; Keane *et al.* 2011). Some of these regions contain genes that have been previously been shown to be subject to positive selection of copy number variants in inbred laboratory mouse strains. Our data showed that even within a strain, we detected SV in these regions, which suggests that these regions are by their very nature susceptible to rearrangements, *i.e.*, through suppressed recombination. Alternatively, recurrent SV calls could reflect either private SVs in the reference assembly, or mis-assembly of these regions.

Repeat analysis

Repetitive sequences present challenges to assembly, as highly identical repeat sequences from different genomic regions are often incorrectly assembled together. This is a particular concern for data generated from short-read technologies, which are too short to span longer repeats. One advantage of long read sequencing reads are their ability to span a greater range of genomic repeats into unique sequence, enabling resolution of repetitive regions that cannot be resolved in unlinked short read

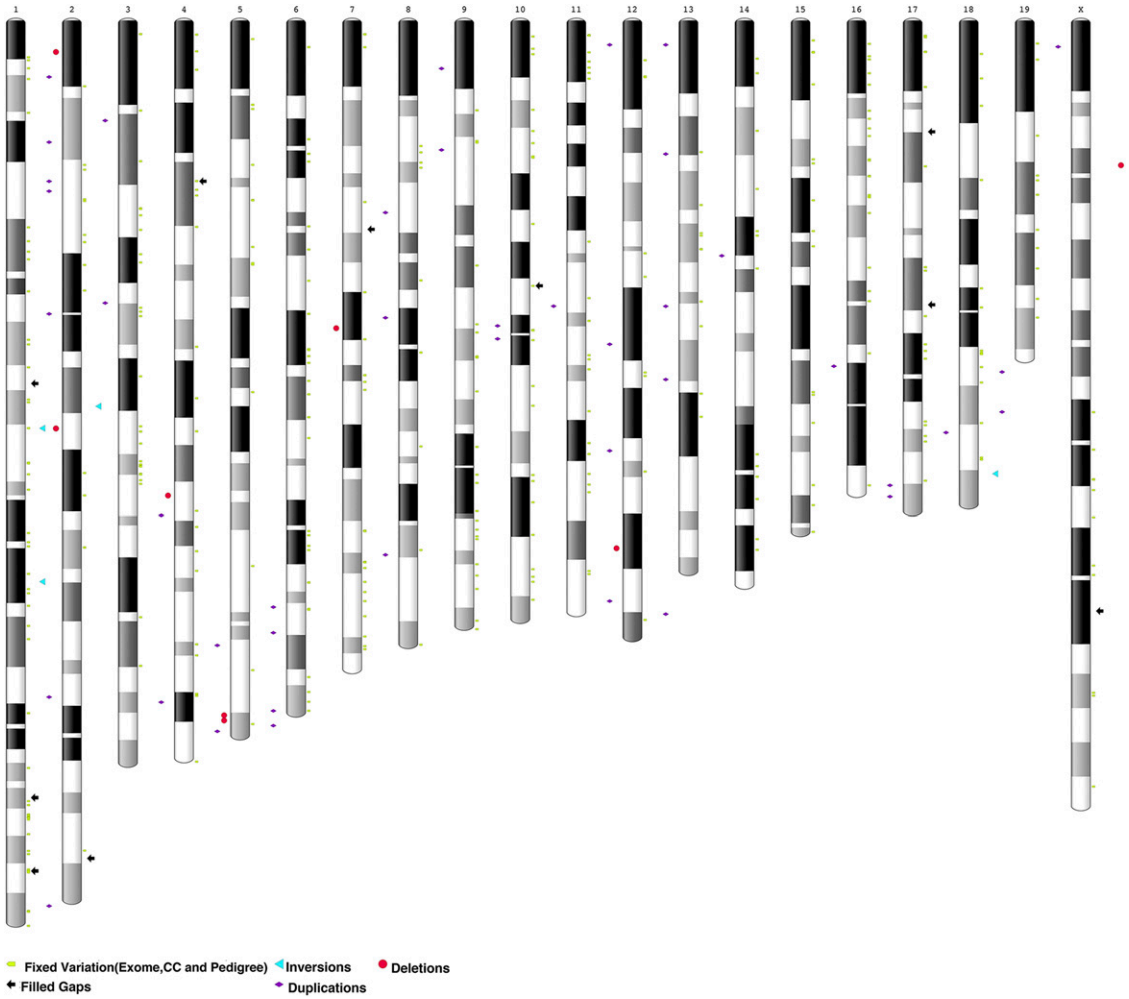

Figure 3 Ideogram of GRCm38 assembly annotated to highlight resolved gaps (*vs.* current reference), structural variants, and fixed variation using B6Eve data.

assemblies. We used RepeatMasker (Smit *et al.* 2013–2015) to compare repetitive sequence representation between B6Eve, GRCm38, and two Illumina WGS assemblies (GCA_000185125.1, GCA_000185105.2) (Locke *et al.* 2015; Milholland *et al.* 2017). This analysis revealed that the B6Eve assembly consists of 42.0% repetitive sequence (1,065,403,997 out of 2,537,631,632 bp, excluding N's). This fraction is very similar to GRCm38 (excluding the Y chromosome and alternate loci sequences), 42.5% of which is repetitive (1,088,395,156 out of 2,559,396,830 bp, excluding N's and X's). Consistent with the challenges of assembling repetitive sequence with shorter reads, the Illumina based assembly GCA_000185125.1 had 32.5% (833,318,654 out of 2,257,461,872 bp excluding N/X-runs) and GCA_000185105.2 had 33.9% (849,891,926 out of 2,279,058,378 bp excluding N/X-runs) annotated as repetitive sequence (Table S7).

We also used RepeatMasker analysis to assess repeat content in scaffolds from our B6Eve assembly that failed to align to GRCm38, as we surmised these scaffolds might contain repetitive sequences that could not be resolved with genomic clones. To do this we focused on sequences which are unaligned by all of three of the following methods a) Cactus based alignment using UCSC Comparative Annotation Toolkit b) NCBI assembly-assembly alignments and c) QUAST evaluation. The common unaligned sequences (total 6.12 MB) (File S9) between CACTUS, NCBI, and QUAST had significant enrichment for repeats relative to aligned sequences. The repeat content accounted to 77.6% (4,754,330 out of 6,128,602 bp) with the microsatellite repeat class showed significantly enrichment when compared with GRCm38 (59.9% *vs.* 0.1%, chi-square test: X^2 = 80,332,000, p-value < 2.2e-1 (Table 4). While more work is needed to determine the underlying cause of failed alignment, the enrichment of microsatellite repeats in

Table 3 Counts of various structural variation classes detected in the comparison of B6Eve Sequences to GRCm38 using PacBio and Illumina data

Technology	Duplication	Deletion	Inversion	Insertion	Trans
PacBio	229	418	36	3,394	71
Illumina	289	221	111	—	—
Common	44	12	4	—	—

Table 4 Comparison of various repeat class in common unaligned sequences with GRCm38

Repeat Class	Number of bp in common unaligned sequences (6,128,602 bp)	GRCm38 (number of bp in complete genome) excluding "alt loci" (2,559,396,830 bp excl N/X-runs)
Satellites	3,671,543 (59.91%)	3,302,550 (0.13%)
LINE1	300,944 (4.91%)	488,443,086 (18.86%)
ERVL-MaLRs	17,196 (0.28%)	113,630,025 (4.31%)
B2-B4	32,901 (0.54%)	111,079,403 (4.22%)
ERVL	16,550 (0.27%)	29,593,691 (1.12%)
hAT-Charlie	413 (0.01%)	16,625,654 (0.63%)
ERV_classI	20,863 (0.34%)	24,057,863 (0.91%)
Unclassified:	1,080 (0.02%)	8,303,004 (0.32%)
Alu/B1	101,229 (1.65%)	62,434,516 (2.37%)
TcMar-Tigger	502 (0.01%)	4,546,701 (0.17%)
ERV_classII	252,653 (4.12%)	121,444,463 (4.61%)
LTR	307,262 (5.01%)	289,477,645 (10.99%)
Simple repeats	311,973 (5.09%)	69,151,432 (2.63%)
Low complexity	25,984 (0.42%)	9,687,916 (0.37%)

these scaffolds is compelling. Microsatellite repeats are prone to slippage during DNA replication, and as a result their copy number is highly polymorphic in eukaryotic genomes; a phenomenon known as microsatellite instability (MSI). In inbred laboratory mouse strains and in the human population, mutations that change copy number occur at rates that are up to 10,000 times higher than single nucleotide mutation rates ($1\text{-}3 \times 10^{-4}$ (Beal *et al.* 2012a; Beal *et al.* 2012b; Zuo *et al.* 2012) per repeat per generation for microsatellite sequences *vs.* $2\text{-}4 \times 10^{-9}$ (Milholland *et al.* 2017) per nucleotide per generation for SNV in C57BL/6J). Similarly, in the human population, CNV are estimated to occur at rates that are 100-10,000 times higher than the point mutation rate (Hu *et al.* 2017). Taken together, CNV are a major source of intrastrain variation and divergence from isogenicity (Watkins-Chow and Pavan 2008; Locke *et al.* 2015). Therefore, failed alignment of these microsatellite containing scaffolds could be due to repeat polymorphisms that have arisen over the intervening years in C57BL/6J. Alternatively, failed alignment could be due to assembly issues in either genome.

Gene prediction

Long read sequencing of cDNAs (IsoSeq) provides full-length transcript sequences and highly accurate representations of splice junctions and isoforms. To determine if long read sequencing data of B6Eve cDNAs could support more accurate gene prediction for the mouse reference genome, we generated IsoSeq data from RNA extracted from archived B6Eve brain. We used the Comparative Annotation Toolkit (CAT)(Fiddes *et al.* 2018) to identify 107,192 transcripts (82,187 protein-coding) representing 41,669 gene loci (20,182 protein-coding). 2,426 transcript predictions had splice junctions that were novel relative to GENCODE VM11, and we found additional support for these junctions in RNA-Seq data generated from the brain of a female C57BL/6J descendent of Eve. Analysis of the transcript predictions produced by AugustusPB and AugustusCGP revealed 206 exons with splice site shifts relative to GRCm38, nine putatively novel exons and ten putatively novel loci (Table S8). Three of the novel exons detected in the IsoSeq data reveal deletions in GRCm38: (1) 640 bp in *Mia3* (Figure 4), (2) *Traf5*, and *Slc26a6* (Figure S3 and Figure S4). In support of these data, there are GRC incident reports describing deletions at each of these loci in GRCm38. Contiguity analysis in the B6Eve assembly showed that 616 genes mapped across two or more scaffolds and four genes had projections split on the same scaffold. A total of 258 protein-coding genes exhibited signs of gene family collapse, with 156 pairs of genes being resolved to the same locus.

Conclusions

The value of isogenic mouse strain backgrounds in biomedical research was recognized by geneticists in the early 20th century leading to the creation and description of the over 450 unique inbred mouse strains to date (Silver 1995; Beck *et al.* 2000). Twenty generations of sibling intercrosses are required for the generation of a new inbred strain; a breeding method that creates genomes in which more than 98% of loci are homozygous. Therefore, individuals within a generation, within the same vivarium are essentially, genetically identical. The remarkable genetic architecture of inbred laboratory mouse strains is shaped by the frequent bottlenecks required for the on-going maintenance of these strains. This accelerates genetic drift and is a major source of the often unexpected, genetic variation that can be observed across generations and/or between vivaria. For example, a reference-based alignment of the inbred laboratory strain C57BL/6J yields approximately 900 raw variant calls (SNPs/Indels), despite it being the same inbred strain as the mouse reference genome(Fairfield *et al.* 2015). While a subset of these variant calls are the expected result of genetic drift, we previously found that a significant percentage of these variants are located in regions of the reference genome where there are reported assembly issues(Fairfield *et al.* 2015) and regions that contain missing paralogs, which are a known source of false positive variation due to mis-mapping of reads (Church *et al.* 2011; Hu *et al.* 2017). A major goal of this *de novo* assembly was to generate long read sequencing data that could potentially be used resolve these regions and to provide an updated representation of the variation that is present in the most recent inbreeding generations of C57BL/6J.

The generation of the B6Eve assembly also provides important insights into the relative merits and limitations of different sequencing and assembly approaches. As we demonstrate, long-read WGS assemblies can resolve regions in which there is no coverage in clone-based assemblies. As some genomic regions are recalcitrant to cloning in vectors, alternate technologies like long-read WGS are critical to completing a gapless assembly. However, even long-read WGS assemblies of mammalian genomes are prone to collapse of highly repetitive or segmentally duplicated regions where these lengths are greater than the average read length. The impact of these collapses manifest as false positive variant calls in regions where paralogous variants were miscalled. As sequencing technologies improve and read lengths get even longer, we may reach a point at which the need for genomic clones in assembly is obviated. However, to generate very high-quality reference assemblies, it will also be important to further reduce the error rates associated with long reads, even beyond the corrections achieved with

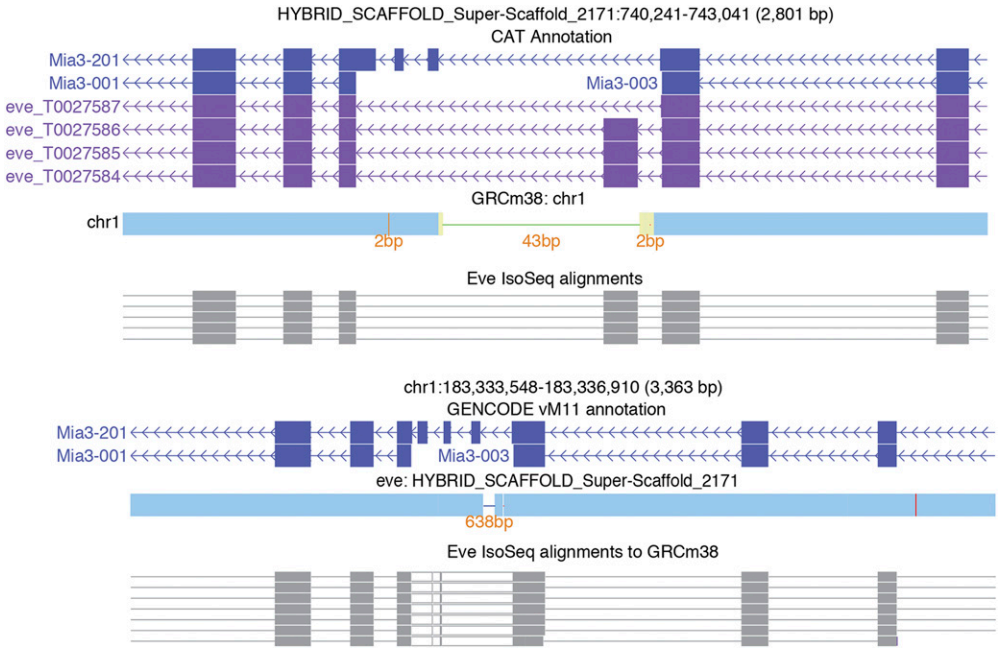

Figure 4 The *Mia3* locus from the perspective of both the B6Eve assembly (top) and the GRCm38 mouse reference (bottom). CAT annotation of B6Eve identified three isoforms with an IsoSeq supported exon not found in the reference. The cactus alignments (blue bars) show that there are 43 bp of reference sequence that does not align to B6Eve, and that there are 638 bp of B6Eve not seen in the reference. These 638 bp contain the extra exon. This result is confirmed in the B6Eve IsoSeq GRCm38 alignment, which shows an insertion (white blocks between gray exon alignments).

short reads. Currently, generation of the highest quality assemblies requires a mix of techniques.

Using variant data from our B6Eve assembly, as well as data from several other large sequencing efforts (Srivastava *et al.* 2017), we provide a "truth" call for over 500 high quality, recurring variants (SNP/Indels) that can be used to update the mouse reference genome (GRCm38.p6). We also found evidence for over 40 structural variants (inversions, deletions, and duplications) involving protein-coding genes in our B6Eve assembly compared to the reference genome. The majority of these SV calls were found in DGVa across a variety of strains, suggesting that they are likely recurrent SV calls that, similar to recurrent variants, are due to mis-assembly of paralogous sequences or reference specific SVs. Further, our data fill 23 gaps of varying length in the mouse reference genome which will be used to inform the upcoming release of GRCm39.

Our IsoSeq data provided improved/more accurate gene models with previously unrecognized splice junctions for over 2,000 genes. This is likely an underrepresentation since our analysis is limited to only those genes expressed in brain. We also found evidence for novel exons, as well as evidence for novel loci (expressed regions that lack gene annotation). This demonstrates that even in a well-curated reference genome assembly, gene annotation remains subject to change as new technologies provide improved representation of transcribed sequences and access to more highly specialized cell types.

Overall, our *de novo* assembly of Eve is not as polished as GRCm38, a clone-based assembly that benefits from more than 20 years of on-going curation and annotation, but it does provide key enhancements and a full picture of the types and sources of technical error in re-sequencing. Whole genome sequencing data are now available for hundreds of standardized laboratory inbred mouse strains (Keane *et al.* 2011; Srivastava *et al.* 2017; Lilue *et al.* 2018). These data reveal the remarkable architecture of inbred genomes, and provide a stark reminder that isogenic mouse strains are subject to genetic drift; a feature that directly conflicts with the idea of a 'reagent-grade' laboratory mouse. Careful breeding practices, cryoarchiving, and routine sequencing are key steps toward maximizing reproducibility of studies that rely on these living reagents. Ultimately, *de novo* assembly captures the full spectrum of genetic variation resident in inbred strains, some of which harbor significantly more variation than distantly related human populations. Recently, genome graphs have been used to represent "population reference genomes" as a means to improve read mapping and to minimize false positive variant calls (Rosen *et al.* 2017; Garrison *et al.* 2018). As applied to mouse genomes, this approach would ideally provide a framework for future representation of the laboratory mouse reference genome as a graph of many inbred strains upon which emergent variation can be more accurately discovered and used to guide experimental research involving laboratory mouse strains.

ACKNOWLEDGMENTS

LGR, VKS, AS, NR, and OZ were supported in part by NIH R24 OD02135 awarded to LGR and The Jackson Laboratory. The work of VAS and F. T-N was supported by the intramural research program of the National Library of Medicine, National Institutes of Health. We are grateful to the services provided by The Jackson Laboratory Genome Technologies and Computational Sciences Core, which are supported by a grant to The Jackson Laboratory Cancer Center, NCI P30 CA034196.

LITERATURE CITED

Beal, M. A., T. C. Glenn, S. L. Lance, and C. M. Somers, 2012a Characterization of unstable microsatellites in mice: no evidence for germline mutation induction following gamma-radiation exposure. Environ. Mol. Mutagen. 53: 599–607. https://doi.org/10.1002/em.21726

Beal, M. A., T. C. Glenn, and C. M. Somers, 2012b Whole genome sequencing for quantifying germline mutation frequency in humans and model species: cautious optimism. Mutat. Res. 750: 96–106. https://doi.org/10.1016/j.mrrev.2011.11.002

Beck, J. A., S. Lloyd, M. Hafezparast, M. Lennon-Pierce, J. T. Eppig *et al.*, 2000 Genealogies of mouse inbred strains. Nat. Genet. 24: 23–25. https://doi.org/10.1038/71641

Berlin, K., S. Koren, C. S. Chin, J. P. Drake, J. M. Landolin *et al.*, 2015 Assembling large genomes with single-molecule sequencing and

locality-sensitive hashing. Nat. Biotechnol. 33: 623–630. Erratum: 33: 1109. https://doi.org/10.1038/nbt.3238

Buac, K., D. E. Watkins-Chow, S. K. Loftus, D. M. Larson, A. Incao *et al.*, 2008 A Sox10 expression screen identifies an amino acid essential for Erbb3 function. PLoS Genet. 4: e1000177. https://doi.org/10.1371/journal.pgen.1000177

Chin, C. S., D. H. Alexander, P. Marks, A. A. Klammer, J. Drake *et al.*, 2013 Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10: 563–569. https://doi.org/10.1038/nmeth.2474

Church, D. M., L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein *et al.*, 2009 Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 7: e1000112. https://doi.org/10.1371/journal.pbio.1000112

Church, D. M., V. A. Schneider, T. Graves, K. Auger, F. Cunningham *et al.*, 2011 Modernizing reference genome assemblies. PLoS Biol. 9: e1001091. https://doi.org/10.1371/journal.pbio.1001091

Cutler, G., L. A. Marshall, N. Chin, H. Baribault, and P. D. Kassner, 2007 Significant gene content variation characterizes the genomes of inbred mouse strains. Genome Res. 17: 1743–1754. https://doi.org/10.1101/gr.6754607

Das, S. K., M. D. Austin, M. C. Akana, P. Deshpande, H. Cao *et al.*, 2010 Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. Nucleic Acids Res. 38: e177. https://doi.org/10.1093/nar/gkq673

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498. https://doi.org/10.1038/ng.806

Dickinson, M. E., A. M. Flenniken, X. Ji, L. Teboul, M. D. Wong *et al.*, 2016 High-throughput discovery of novel developmental phenotypes. Nature 537: 508–514. https://doi.org/10.1038/nature19356

Fairfield, H., A. Srivastava, G. Ananda, R. Liu, M. Kircher *et al.*, 2015 Exome sequencing reveals pathogenic mutations in 91 strains of mice with Mendelian disorders. Genome Res. 25: 948–957. https://doi.org/10.1101/gr.186882.114

Fiddes, I. T., J. Armstrong, M. Diekhans, S. Nachtweide, Z. N. Kronenberg *et al.*, 2018 Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. Genome Res. 28: 1029–1038. https://doi.org/10.1101/gr.233460.117

Florea, L., A. Souvorov, T. S. Kalbfleisch, and S. L. Salzberg, 2011 Genome assembly has a major impact on gene content: a comparison of annotation in two Bos taurus assemblies. PLoS One 6: e21400. https://doi.org/10.1371/journal.pone.0021400

Garrison, E., J. Siren, A. M. Novak, G. Hickey, J. M. Eizenga *et al.*, 2018 Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. 36: 875–879. https://doi.org/10.1038/nbt.4227

Graubert, T. A., P. Cahan, D. Edwin, R. R. Selzer, T. A. Richmond *et al.*, 2007 A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet. 3: e3. https://doi.org/10.1371/journal.pgen.0030003

Green, E. L., 1981 Genetics and Probability in Animal Breeding Experiments. https://doi.org/10.1007/978-1-349-04904-2

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler, 2013 QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hodges, E., M. Rooks, Z. Xuan, A. Bhattacharjee, D. Benjamin Gordon *et al.*, 2009 Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat. Protoc. 4: 960–974. https://doi.org/10.1038/nprot.2009.68

Hu, X. S., F. C. Yeh, Y. Hu, L. T. Deng, R. A. Ennos *et al.*, 2017 High mutation rates explain low population genetic divergence at copy-number-variable loci in Homo sapiens. Sci. Rep. 7: 43178. https://doi.org/10.1038/srep43178

Keane, T. M., L. Goodstadt, P. Danecek, M. A. White, K. Wong *et al.*, 2011 Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477: 289–294. https://doi.org/10.1038/nature10413

Konig, S., L. W. Romoth, L. Gerischer, and M. Stanke, 2016 Simultaneous gene finding in multiple genomes. Bioinformatics 32: 3388–3395.

Lilue, J., A. G. Doran, I. T. Fiddes, M. Abrudan, J. Armstrong *et al.*, 2018 Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nat. Genet. 50: 1574–1583. https://doi.org/10.1038/s41588-018-0223-8

Locke, M. E., M. Milojevic, S. T. Eitutis, N. Patel, A. E. Wishart *et al.*, 2015 Genomic copy number variation in Mus musculus. BMC Genomics 16: 497. https://doi.org/10.1186/s12864-015-1713-z

Marshall, E., 2002 Genome sequencing. Public group completes draft of the mouse. Science 296: 1005. https://doi.org/10.1126/science.296.5570.1005b

Matera, I., D. E. Watkins-Chow, S. K. Loftus, L. Hou, A. Incao *et al.*, 2008 A sensitized mutagenesis screen identifies Gli3 as a modifier of Sox10 neurocristopathy. Hum. Mol. Genet. 17: 2118–2131. https://doi.org/10.1093/hmg/ddn110

Milholland, B., X. Dong, L. Zhang, X. Hao, Y. Suh *et al.*, 2017 Differences between germline and somatic mutation rates in humans and mice. Nat. Commun. 8: 15183. https://doi.org/10.1038/ncomms15183

Mouse Genome Sequencing Consortium, Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562. https://doi.org/10.1038/nature01262

Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo *et al.*, 2000 A whole-genome assembly of Drosophila. Science 287: 2196–2204. https://doi.org/10.1126/science.287.5461.2196

Quinlan, A. R., R. A. Clark, S. Sokolova, M. L. Leibowitz, Y. Zhang *et al.*, 2010 Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. 20: 623–635. https://doi.org/10.1101/gr.102970.109

Rosen, Y., J. Eizenga, and B. Paten, 2017 Modelling haplotypes with respect to reference cohort variation graphs. Bioinformatics 33: i118–i123. https://doi.org/10.1093/bioinformatics/btx236

Salisbury, J. L., 2004 Centrosomes: Sfi1p and centrin unravel a structural riddle. Curr. Biol. 14: R27–R29. https://doi.org/10.1016/j.cub.2003.12.019

Schimenti, K. J., S. K. Feuer, L. B. Griffin, N. R. Graham, C. A. Bovet *et al.*, 2013 AKAP9 is essential for spermatogenesis and sertoli cell maturation in mice. Genetics 194: 447–457. https://doi.org/10.1534/genetics.113.150789

Schneider, V. A., T. Graves-Lindsay, K. Howe, N. Bouk, H. C. Chen *et al.*, 2017 Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27: 849–864. https://doi.org/10.1101/gr.213611.116

Silver, L. M., 1995 *Mouse Genetics*. Oxford University Press, New York.

Smit, A., R. Hubley and P. Green, 2013–2015 RepeatMasker, pp.

Srivastava, A., A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon *et al.*, 2017 Genomes of the Mouse Collaborative Cross. Genetics 206: 537–556. https://doi.org/10.1534/genetics.116.198838

Taylor, B. A., and L. Rowe, 1984 Genes for serum amyloid A proteins map to Chromosome 7 in the mouse. Mol. Gen. Genet. 195: 491–499. https://doi.org/10.1007/BF00341452

Venkatesh, D., D. Mruk, J. M. Herter, X. Cullere, K. Chojnacka *et al.*, 2016 AKAP9, a Regulator of Microtubule Dynamics, Contributes to Blood-Testis Barrier Function. Am. J. Pathol. 186: 270–284. https://doi.org/10.1016/j.ajpath.2015.10.007

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963. https://doi.org/10.1371/journal.pone.0112963

Watkins-Chow, D. E., and W. J. Pavan, 2008 Genomic copy number and expression variation within the C57BL/6J inbred mouse strain. Genome Res. 18: 60–66. https://doi.org/10.1101/gr.6927808

Wiles, M. V., and R. A. Taft, 2010 The sophisticated mouse: protecting a precious reagent. Methods Mol. Biol. 602: 23–36. https://doi.org/10.1007/978-1-60761-058-8_2

Zuo, B., X. Du, J. Zhao, H. Yang, C. Wang *et al.*, 2012 Analysis of microsatellite polymorphism in inbred knockout mice. PLoS One 7: e34555. https://doi.org/10.1371/journal.pone.0034555

Communicating editor: F. Pardo-Manuel de Villena