Supplementary Note

Materials and Methods

BioNano optical maps

High molecular weight DNA samples were isolated from mouse kidneys with Plug Lysis protocol according to IrysPrep® Animal Tissue DNA Isolation Soft Tissue User Guide (Protocol 30077). Megabase length genomic DNA which was further cleaned by drop dialysis. Sequence specific labelling was based on 'NLRS' procedure (Protocol 30024), using bspQI as nicking enzyme. Labelling was carried out by a limited drive nick translation process in the presence of a fluorophore-labelled nucleotide. The labelled nicks were repaired to restore strand integrity. The labeled DNA was visualized by the front end user interface software on the Irys Instrument. On average, 250GB of >150kb molecules were collected for each mouse strain. The data were de-novo assembled with the Irysview Software Package v2.5.1, resulting in genome assemblies ranging between 2000 and 5000 contigs, genome sizes of 2.4-2.7GB at 35-80x coverage. A summary of BioNano data is given in Supplementary Table 24.

Cactus whole genome alignment

The assemblies of each mouse strain were aligned to each other using the rn6 rat assembly as an outgroup with the whole-genome alignment tool progressiveCactus[7]. progressiveCactus makes use of an input phylogenetic guide tree to walk up the tree creating alignments of a few genomes at a time, imputing an ancestral genome at each internal node. This ancestral genome is then used as input to alignments higher on the tree, and in this fashion progressiveCactus progressively builds a reference-free whole-genome alignment between an arbitrary number of genomes. We used progressiveCactus commit e3c60554b3140478cbb9c3783f2288a875487f60 (git tag: msca_1508) to create the alignment used for the analysis in this paper. The following newick format guide tree was used:

```
(rn6:0.013,(SPRET_EiJ:0.002,(PWK_PhJ:0.001,(CAST_EiJ:0.001,(WSB_EiJ:1e-
05,(((NZO_HlLtJ:1e-06,(C57BL_6NJ:1e-06,C57B6J:1e-06)anc1:1e-06)anc2:1e-
06,((NOD_ShiLtJ:1e-06,FVB_NJ:1e-06)anc3:1e-06,(((DBA_2J:1e-06,(CBA_J:1e-06,C3H_HeJ:1e-
06)anc4:1e-06)anc5:1e-06,AKR_J:1e-06)anc6:1e-06,(BALB_cJ:1e-06,A_J:1e-06)anc7:1e-
06)anc8:1e-06)anc9:1e-06)anc10:1e-06,(LP_J:1e-06,129S1_SvImJ:1e-06)anc11:1e-06)anc12:1e-
06)anc13:0.0001)anc14:1e-06)anc15:1e-06)anc16:0.045)anc17;
```

The alignment runtime is linear in the number of leaves, if using a binary guide tree, but quadratic if using a star phylogeny. To reduce computation time, the subtree for the lab mice strains was semi-arbitrarily binarized with a branch length of 1e^{-06}. Some prior knowledge of strains and analysis of alignment identity in previous alignment iterations guided the binarization process.

Mitochondrial genome sequences

Whole genome sequencing produced complete sequences of the mitochondrial genome (mtDNA) for all sixteen inbred strains. Since 1979 when the first complete mouse mtDNA sequence was published[9], the field has added over 70 mtDNA sequence from mouse strains (n=60), mouse populations, or mouse cell lines. These previous works have set the circular mouse mitochondrial chromosome as between 16,295 and 16,303 bases in length. Regarding the current effort, sequences for fourteen of the strains had been published and deposited in GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Strains without a previously published sequence were NZO/HiLtJ and PWK/PhJ. All contigs with alignments to mitochondrial sequence from the mouse reference were collated and - together with read alignments - loaded into gap5[10]. The sequence was manually assessed, corrected and submitted as part of the respective main assembly.

Submission

All assembly sequences were submitted to Genbank and can be accessed as detailed in (Supplementary Table 18).

Repeat analysis

The pseudo chromosomes were analysed for repeat content using Repeatmasker open-4.0.5 using Crossmatch v0.990329 using the mouse (-species mouse) repeat library (Complete Database: 20140131).

Chromosome 11 manual sequence updates

Manual sequence curation was carried out on chromosome 11. Evidence such as genome mapping data (BioNano), transcript sequence (RefSeq, CCDSs, GENCODE lncRNAs, PacBio cDNAs), BAC/fosmid end sequence, repeat masking, marker placement and self alignments were loaded into the gEVAL[11], genome editing environment. Manual curation focussed on incomplete genes on chr11 and on severely over-/undersized gaps. Corrections included breaking scaffolds/contigs into smaller components and rearranging components, within and in between chromosomes or unplaced sequence were made.

Pseudogene annotation

We annotated the mouse pseudogenes by combining the results from in-house automatic annotation pipelines (PseudoPipe[18], RCPedia[19]) with the set obtained by lifting over the manually curated pseudogenes from the reference mouse annotation GENCODE M8 to each individual strain. PseudoPipe is a comprehensive pseudogene annotation pipeline focusing on identifying three pseudogene biotypes: processed duplicated, and unitary pseudogenes. RCPedia is a retrocopy annotation pipeline focused on identifying processed pseudogenes. In order to assure a high level of confidence in the automatic pseudogene predictions, we used as input the consensus protein coding genes between the manually annotated mouse reference genome and each strain. The pseudogenes were identified as described previously[18,19]. Each automatically annotated pseudogene is characterized by transcript biotype, genomic location and exonic structure. Next, we used the transMap lift over of the mouse reference annotation from GENCODE M8 to extract the pseudogenes shared between each strain and the reference genome. As such, we identified between 5,424 and 5,742 pseudogenes in the wild strains (SPRET/EiJ, PWK/PhJ, CAST/EiJ, WSB/EiJ) and over 6,000 pseudogenes in each of the classical strains. We intersected the liftover with the automatic annotation sets requiring at least 1bp overlap. The resulting set was further processed to assure consistency in the parent pseudogenes, biotype and exonic structure. We extend the pseudogene models to assure the maximum overlap between each annotation method. The complete pseudogene set in each strain was constructed as a consensus union between all annotation methods. Thus, the resulting set is defined by 3 confidence levels: Level 1 is indicative of pseudogenes identified by both manual liftover and automatic pipelines, Level 2 is indicative of the pseudogenes annotated only through the liftover of the manual annotation of the reference genome, while Level 3 represents the set of pseudogenes curated only by the automatic annotation workflow. The summary of the pseudogene sets in each strain is available in Supplementary Table 25.

Base accuracy

All of the paired-end sequencing reads from each strain were realigned back to the respective assembled genome using bwa mem v0.7.5, and SNPs and indels identified using samtools v1.2 (mpileup -t DP,DV -C50 -pm3 -F0.2 -d10000) and bcftools v1.2. (call -vm -f GQ). Low quality variants (-sLowQual -e"%QUAL<=10" -g3 -G10) were removed and base error rates for each strain were estimated as total variants (SNPs and indels separately) divided by the size of the respective strain genome assembly.

Local structure accuracy

PCR primer pair sequences were obtained from Yalcin *et al.*[20], including only primers for structural variants that did not involve an inversion and that uniquely align to the same chromosome and inward orientation on GRCm38. Primer alignments were carried out with bwa mem v0.7.12-r1039 with the following parameters: -A 5 -U 0 -T 10 -k 12. The list of primer pairs and their sequence is given in Supplementary Data 12.

Unplaced sequences

Summary statistics (e.g. total sequence length, total number, N50) for unplaced scaffolds were calculated for each strain using "Sanger-pathogens/assembly-stats" package (https://github.com/sanger-pathogens/assembly-stats) (Supplementary Table 2). Gene annotations for unplaced scaffolds were predicted with CEGMA pipeline[21] (Supplementary Table 3).

SNP data retrieval

To identify regions of the mouse reference genome (GRCm38) enriched for hSNPs, variation data for each of the 16 strains was retrieved from the MGP variation catalogue version 5[22]. Additionally, sequencing data generated from the C57BL/6J reference strain[23] was aligned back to GRCm38, and heterozygous SNPs identified using the same parameters described by Doran *et al*[24]. All hSNPs identified are relative to GRCm38, and only hSNPs with a genotype quality score ≥ 20 were retained for further analysis. For each strain, the total number of heterozygous SNPs available for analysis is contained in Supplementary Table 9. The density of hSNPs in adjacent 200kb windows was calculated and plotted for each strain chromosome (Figure 1b; Supplementary Figure 5).

Repeat element representation

All pass hSNP dense regions were merged into a single non-overlapping set of regions, enriched for hSNPs, using bedtools[26]. Separately, Repeatmasker v4.0.5 was used to identify all repeat sequences in the GRCm38 assembly of the mouse genome. All identified repeat sequences were then organised into repeat categories defined in Supplementary Data 13. As all hSNP dense region coordinates are based on the GRCm38 reference, repeats for each of the largest classes of categories (LINEs, LTRs, and SINEs) that intersect a hSNP region were identified. Percent divergence (indicative of age) was estimated by Repeatmasker for repeats within hSNP dense regions and compared to the equivalent repeat types, by category, outside of hSNP dense regions using Welch's Two Sample t-test.

The repeat content (i.e. count of unique repeats) within the merged hSNP dense regions for each of the above repeat categories was then calculated. To examine LINE content enrichment, we simulated the content of LINEs, from GRCm38, with the same set of merged hSNP dense regions by randomly placing those regions on the same chromosome and counting the number of overlapping LINEs (using repeatmasker IDs to uniquely identify overlapping repeats). This was repeated 1,000,000 times, and an Emperical p-value for the observed overlap was obtained by examining the simulated distribution of overlapping LINE counts. The p-value was calculated as the fraction of simulations in which the repeat content was equal to or more extreme than the observed repeat content. This process was repeated for both SINEs and LTRs separately. Calculated p-values less than 0.05 were considered significant, and thus the observed repeat type considered significantly over-represented.

Targeted reassembly and sequence improvement

Targeted manual reassembly provides advantages over traditional assembly pipelines including improved sequence gap sizing and sequence contiguity (fewer sequence gaps), especially in repeat-rich and low-complexity regions. Manual reassembly and sequence improvement of three loci (IRG, Nlrp1 and Slfn) exhibiting considerable diversity relative to the reference genome was carried out for PWK/PhJ, CAST/EiJ, WSB/EiJ, and SPRET/EiJ. The Nlrp1 locus of NOD/ShiLtJ was based on a previously published BAC sequence, NT_187026[27], which was derived from the NOD/MrkTac strain[28]. Short and long insertion Illumina reads from NOD/ShiLtJ were mapped to NT_187026 (NOD/MrkTac) with BWA-MEM v0.7.12[29]. Both NOD strains were confirmed to share the same haplotype at the *Nlrp1* locus.

In addition, reassembly and sequence improvement of the Raet1/H60 locus was carried out for all Collaborative Cross (CC) founder strains (A/J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ) excluding the reference strain, C57BL6/J. Based on CISGen Mouse Phylogeny Viewer[30] (https://msub.csbio.unc.edu/), it was observed that strains A/J, 129S1SvImJ and NOD/ShiLtJ share the same haplotype at the Raet1/H60 locus. Thus only the NOD/ShiLtJ Raet1/H60 locus reassembly was attempted among these strains. Reassembly and sequence improvement of these loci was carried out for each strain separately, using strain specific sequencing data generated using whole genome Dovetail Genomics, PacBio RSII, short read Illumina and long-insert (3kb, 6kb and 10kb) paired Illumina (protocols and methods described in previous section). The targeted reassembly pipeline involved two phases. The initial phase involved extracting sequence read data aligned to these regions from the whole genome assembly for each strain. These data were then used to assemble and improve sequence. The improved sequence was then incorporated back into the strain assembly. The pipeline is described and summarised in Supplementary Figure 16.

Dovetail *de novo* contigs (DDC)
For each strain, contigs, generated using the Dovetail Genomics *de novo* assembly pipeline, encoding these loci and approximately 100kb up and downstream are extracted. These flanking contigs can be used to anchor reassembled sequence to regions of the genome, before incorporating the improved sequence back into the updated assembly. Extracted Dovetail contigs were masked for repeats and low complexity sequence using Repeatmasker v4.0.5[31].

Merged PacBio Collection (MPC)
BLASTall v.2.2.25[32] was then used to identify strain-specific whole genome PacBio RSII reads mapped to the extracted (and repeat masked) Dovetail contigs. Pacbio reads matched to an extracted contig were retained. In addition, all strain-specific PacBio reads were mapped to the mouse reference genome (GRCm38) using BWA-MEM v0.7.12[29], and reads uniquely mapped to these loci were also retained. All retained PacBio reads were then mapped back to the extracted Dovetail *de novo* contigs, and, where possible used to fill gaps between contigs. Additionally, these reads were retained for manual sequence improvement.

Illumina *de novo* contigs (IDC)
Strain specific Illumina short reads, and their corresponding read-pair mates (mapped or unmapped), mapped to the reference genome (GRCm38) at these loci were extracted. All extracted reads were mapped to the Dovetail assembly contigs using Geneious R8 allowing a maximum of 2% base pair mismatch and up to 3% indels. Extracted reads that failed to map to the Dovetail contigs, and their corresponding mate-pairs (mapped or unmapped), were assembled into short contigs using Geneious R8[33].

Long insertion Illumina collection (LIIC)
Strain-specific long insert (3kb, 6kb and 10kb) Illumina reads mapped to the reference at these loci were extracted, and mapped to the repeat-masked reference (GRCm38) genome. Read pairs where both mates mapped to a repetitive or low complexity region were excluded, and all other mapped read-pairs were retained.

Sequence improvement around these four loci
Data extracted above was then used to manually reassemble and improve the region sequence quality surrounding the described loci. The reassembled sequence was then incorporated back into the whole genome assemblies using the steps described below.
Step 1. Beginning with the long-insertion illumina contig reassembly, the Dovetail, Illumina *de novo* and PacBio contigs are manually connected and adjusted into region scaffolds. Size of gaps were estimated using the long-insertion Illumina read-pairs.
Step 2. Illumina reads and Pacbio data were mapped to the adjusted region scaffolds with Geneious R8[33]. Mismatches and conflicts were manually checked, and errors introduced by PacBio long reads were corrected using the Illumina short read data.

Step 3. Reassembled region scaffolds were digested with BspQI restriction endonuclease *in silico* using Knickers 1.5.5[34]. Confirmed BioNano optical map digestion sites were then used to connect individual small scaffolds into a single larger scaffold. Gap-lengths were estimated using the optical map data.

Long range PCR (gap filling)
To further fill gaps in the PWK/PhJ strain assembly across the Nlrp1 locus, long range PCR was performed. Unique primer pairs up to 20kb apart were used to generate PCR products, spanning assembly gaps, that were sequenced using PacBio RSII. These sequences were then incorporated into the Nlrp1 region-scaffold for PWK/PhJ generated in step 3 above. Primers and corresponding products are summarized in Supplementary Table 26.

Fully reassembled sequence availability

Fully reassembled sequences for each locus is available through the European Nucleotide Archive[35]. Accession numbers for each locus are: IRG_PWK (186Kb, LT629149), Nlrp1_PWK(325Kb, LT629150), Slfn_PWK(369Kb, LT629151), Nlrp1_CAST(395Kb, LT629147), Slfn_CAST(370Kb, LT629148), IRG_WSB(215Kb, LT629152), Nlrp_WSB(394Kb, LT629153) and Slfn_WSB(322Kb, LT629154). The reassembled IRG locus in CAST/EiJ has previously been published[36].

Olfr members in CAST/EiJ

To assess olfactory gene family differences between the GRCm38 reference genome and the wild-derived CAST/EiJ, the CDS of all olfactory genes annotated in GENCODE vM8 and NCBI's refSeq were extracted and used to search against the CAST/EiJ genome assembly predicted genes using BLASTall[32]. Significant matches ($E < 10^{-40}$) of CAST/EiJ predicted genes to the reference genome were marked as one of:

1. Match: same chromosome and approximate location as GRCm38 reference annotation, same order of adjacent genes and <1% gaps in the CDS.
2. Match with gap: same as match, but >1% gaps in the CDS. Transcriptome sequencing data of CAST/EiJ olfactory sensory neurons[37] was used to manually fill gaps for these genes.
3. Cross: same as match, but the order of adjacent genes is altered.
4. New: no significant match to any GRCm38 annotated genes, or where many CAST/EiJ predicted genes have significant hits to a single GRCm38 gene, all passing non-primary hits are considered novel expansions of the olfr family gene tree, whereas the top matching hit is considered orthologous to the GRCm38 gene and thus not novel. All predicted novel olfr family members were manually confirmed, and compared to previously published CAST/EiJ olfactory gene predictions[37]. In addition, to further identify novel olfr family members, transcriptome sequencing data of CAST/EiJ olfactory sensory neurons was mapped to the CAST/EiJ gene CDS collection. Unmapped reads were *de novo* assembled into contigs, which were then marked as 'new'.
5. Disrupted: partial CDS loss due to genome recombination or the insertion of transposable elements.
6. Fail: GRCm38 genes which have no significant matches identified in CAST/EiJ.
7. Gene loss: For genes marked as fail, CAST/EiJ illumina short reads mapped to the GRCm38 reference at the gene's location were examined for potential gene loss. Confirmed deletions of olfr family members in the CAST/EiJ genome relative to GRCm38 are marked as 'gene loss'

High-resolution multicolour fiber-FISH

Fiber-FISH was used to investigate the underlying genomic structure at the Raet1 locus across several wild-derived and classical inbred mouse strains. Splenocytes from a C57BL/6J mouse were obtained from the Research Support Facility of the Wellcome Trust Sanger Institute (Cambridge, UK). Embryonic stem cells derived from NOD/ShiLitJ, PWK/PhJ, WSB/EiJ, CAST/EiJ, NZO/HILtJ mice were provided by the The Jackson Laboratory (ME, USA).

To highlight the presence of *Raet1* alleles in these six strains (including C57BL/6J), four fosmid clones, WI1-1159J21 (*Raet1*), WI1-1794K5 (*H60b*), WI1-867J22 (5' flanking region), WI1-2748N8 (3' flanking region), were chosen from WIBR-1 mouse Fosmid library as probes in the fibre-FISH validation experiments. For *Nlrp1* locus, three fosmids were chosen: WI1-855P1 (red, 3' end of *Nlrp1* alleles), WI1-2245E14 (green, 5' end of *Nlrp1* alleles), WI1-1427I23 (blue, 5' end of locus) and WI1-1254K19 (white, 3' end of locus). No *Nlrp1b* specific fosmid probes can be found because of high repeat density in the region. DNA fibres were prepared from agarose-embedded cells by Molecular Combing (Genomic Vision). The generation of probes by whole genome amplification, and general methods for fibre-FISH are given in detail[38].

Reference genome sequence updates informed by strains

In order to generate sequence to fill gaps in the reference genome assembly (GRCm38) that rendered genes incomplete, we identified contigs in the C57BL/6NJ assembly that corresponded to the missing sequence by investigating the genome alignments in gEVAL[11]. C57BL/6J reads (PRJNA51977) were aligned to the C57BL/6NJ assembly using bwa mem with default settings. Reads mapping to the previously identified contigs were extracted using bespoke scripts (available upon request) and assembled using geneious 9.1.3[33]. The resulting sequence was compared to (a) the C57BL/6NJ assembly to test for sanity, (b) the reference assembly to confirm overlaps with the existing gap borders, (c) the respective cDNA to identify missing exons and ensure that the generated sequence completes the fragmented gene and (d) to publicly available C57BL/6J assemblies accepted by the Genome Reference Consortium as gap fillers using NCBI Blast Genomes. For regions where (d) didn't lead to suitable alignments, the newly assembled contigs will be submitted and integrated into the reference assembly as described before[39]. If suitable publicly available contigs could be identified, they were validated as described for (a), (b) and (c).

GENCODE reference annotation updates informed by strains

Manual annotation for the whole mouse genome has not yet been completed, so the Augustus predictions were prioritised to those on chromosomes 1- 12, for which manual annotation has been completed. These predictions were further refined to those for which 75% of the introns were novel compared to GENCODE. This gave a subset of 785 predictions which were subject to manual annotation. These predictions were imported so they were viewable as a track within the Otter/Zmap annotation tools on the C57BL/6J reference genome assembly. Predictions were manually assessed and where necessary manual annotation was performed according to the HAVANA guidelines as detailed by the GENCODE project[40]. The annotation decisions that were taken are summarised in Supplementary Data 8.

Novel gene structures

The gene discovered on chromosome 11, was named as "novel protein similar to EF-hand calcium binding domain 13 EFCAB13 (Homo sapiens)". The majority of the exons were annotated based on RNAseq (Figure 3a) plus limited EST support. Introns 20 and 105 were only supported from the Augustus prediction (see previous section). These exons spliced correctly and were in frame and as such could be annotated according to the HAVANA guidelines:
ftp://ftp.sanger.ac.uk/pub/project/havana/Guidelines/Guidelines_March_2016.pdf
with the transcript annotated tagged as having an inferred exon combination. Six additional partial transcripts were also annotated based on EST evidence from C57BL/6J.

Standardized Phenotyping Pipeline for Efcab3-like KO mice

Mice underwent standardized phenotyping from 4 weeks of age, including weight curves, behavioral and morphological assessment at 9 weeks of age, intraperitoneal glucose tolerance test (ipGTT) following a 4h fast at 13 weeks, body composition assessment by dual emission x-ray absorptiometry (DEXA) at 14 weeks of age using a modified version of the MGP pipeline, using a standard chow instead of a high fat diet. At 16 weeks, random-fed mice were anesthetized using 100 mg/kg Ketamine and 10 mg/kg Xylazine and blood was collected retro-orbitally. Death was confirmed by cervical dislocation and heart removal. The standard operating procedures can be found at IMPReSS (www.mousephenotype.org/impress).
Factors thought to affect the variables were standardized as far as possible. Where standardization was not possible, steps were taken to reduce potential bias. For example, the impact of different people completing the experiment was minimized ("minimized operator") as defined in the Mouse Experimental Design Ontology (MEDO)[44] as "The process by which steps are taken to minimize the potential differences in the effector by training and monitoring of operator."

http://bioportal.bioontology.org/ontologies/MEDO/?p=summary

The data captured with the MEDO ontology can be accessed at http://www.mousephenotype.org/about-impc/arrive-guidelines

In addition, pre-set reasons were established for QC failures (e.g. insufficient sample) and detailed within IMPRESS. Data can only be QC failed from the dataset if clear technical reasons can be found for a measurement being an outlier. For mouse management purposes, the cages have both genotype and allele information, therefore the *in-vivo* screens are run unblinded. However the analysis of blood samples for clinical chemistry and haematological assessment were run blind using a barcode system. In addition, as a high-throughput screen where genes are selected for study without hypothesis and mice are studied in multiple batches, there is limited room for personal bias to influence the results. The 7 male and 8 female *Efcab3* homozygotes which entered the pipeline were tested in 6 batches (1 batch of 3 males, 1 batch of 2 males, 1 batch of 1 male and 3 females, 1 batch of 2 females, 1 batch of 1 male and 1 batch of 3 females). No mutant mice were lost during the pipeline for health or welfare reasons. With each batch of mice, a cohort of 7 age and sex matched wild-type C57BL/6NTac mice were also phenotyped. Viability of the line is assessed by genotyping a minimum of 28 offspring from heterozygous intercrosses. For *Efcab3*$^{em1(IMPC)Wtsi}$, this gave 11 homozygous mice from 34 offspring (Fisher- exact test, P=0.59). Mice were allocated to the pipeline randomly by Mendelian Inheritance.

- Weekly Weights

From 4-16 weeks of age, the mice were weighed once a week to generate a growth curve. On weeks with other testing, the weights are taken during the test in order to minimize handling and disturbing of the mice. With the ipGTT test, the weight was taken prior to the 4h fast. On weeks without testing, the mice were weighed in the animal holding room.

- Neurological and Morphological Phenotyping Assessment

At 9 weeks of age mice are assessed for gross behavioural abnormalities using a modified SHIRPA screen and their whole body morphology is examined using a standardised checklist.

- Glucose Tolerance Test

At 13 week of age, mice were singly housed and fasted for 4h after which approximately 0.5mm of the tail tip was removed with a scalpel blade and a fasting blood sample was taken. Mice were then given an intraperitoneal injection with 2g/kg glucose and further blood samples were taken at 15, 30, 60 and 120 min post-glucose injection. Blood glucose concentration were measured using Accu-chek Aviva glucose meters (Roche, Basel, Switzerland).

- DEXA

At 14 weeks of age, mice were anaesthetised with isoflurane (IsoFlo, Zoetis, UK). Body composition [fat mass (g), fat percentage estimate (%), lean mass (g), bone mineral density (g/cm^2) & bone mineral content (g)] were measured using an Ultrafocus 100 (Faxitron, Tucson, AZ, USA). Nose to tail base length measurements were performed using a ruler with 1mm graduations after DEXA measurements. An internal calibration process was performed on the Ultrafocus 100 before each imaging session.

- Blood Sample & Tissue Collection

At 16 weeks of age, blood was collected without pre-fasting by retro-orbital sinus puncture under terminal anaesthesia (100mg/kg Ketamine and 10mg/kg Xylazine, i.p.) between 08:00 and 10:30 hours. 100 µl of blood was collected into EDTA-coated paediatric tubes (Kabe Labortechnik GmbH, Numbrecht, Germany) for standard haematological analysis and flow cytometric analysis of peripheral blood leukocytes. The rest (~800 µl) was collected into heparinised paediatric tubes (Kabe Labortechnik) for plasma clinical chemistry analysis and insulin assay (MRC-CORD Mouse Biochemistry Laboratory). Selected tissues were then collected for either the Sanger Biobank or further analysis performed by external collaborators specializing in a particular organ.

- Laboratory Analyses

Within 1 hour of collection, heparinised whole blood samples were centrifuged at 5000 rcf for 10 minutes at 8°C, the plasma removed and stored at 4°C until analysis. For clinical chemistry, plasma was assessed at 37$_o$C for standard parameters on an Olympus AU400 analyser using kits and controls supplied by Beckman Coulter. Samples were also tested for haemolysis (Bernardi et al. 1996). Four parameters were measured on the Olympus AU400 analyser by kits not supplied by Beckman Coulter: non-esterified free fatty acids (NEFAC - Wako Chemical Inc., Richmond, USA), glycerol (Randox Laboratories Ltd., UK) (Champy et al. 2004), fructosamine (Roche Diagnostic, UK) and thyroxine (Thermo Fisher, MA, USA).

An automated electrical impedance cell counter was used to perform white blood cell counts (WBC ($x10^3/\mu l$)), red blood cell counts (RBC ($x10^6/\mu l$)), red blood cell distribution width (RDW(%)), haematocrit (HCT (%)), mean cell volume (MCV (fl)), mean corpuscular haemoglobin (MCH (pg)), mean corpuscular haemoglobin concentration (MCHC (g/dl)), platelet counts (PLT ($x10^3/\mu l$)) and mean platelet volume (MPV (fl)) measurements, whilst haemoglobin (HGB (g/dl)) was measured by spectrophotometry (Scil Vet abc+, Viernheim, Germany). Internal quality controls were run on a daily basis and external quality controls, assessed on a two weekly interval, were done by the RIQAS scheme (Randox, UK).

Whole blood is stained with two titrated cocktails of antibodies. Panel 1 containing CD45, TCRab, TCRgd, CD161, CD4, CD8, CD25, CD44, CD62L and KLRG1. Panel 2 containing CD45, CD19, IgD, Ly6B, Ly6G, Ly6C, CD11b and I-A/I-E. Samples are fixed and red blood cells lysed prior to acquisition on a BD LSR II flow cytometer after running automated compensation using compbeads and BD FACSDiva software. Data is analysed using FlowJo after singlet doublet discrimination, a time gate is used to exclude HTS issues and leukocytes identified with a SSC and CD45 gate.

- Statistical Analysis

A high throughput pipeline with multiple analyses and types of phenotyping data means that a single power calculation is inappropriate. Instead, the pipeline has been developed through empirically by selection of a workflow which balances rate of phenotype discovery with cost effectiveness. For phenotyping analysis, the individual mouse was considered the experimental unit within the studies. Statistical tests performed to compare genotype means, with sex as a co-variate. For continuous data, an iterative top down mixed modelling strategy was performed as described in[45],using PhenStat version 2.8.0, an R package freely available from Bioconductor[46]. The package's mixed model framework was used, and all analysis was performed by setting the argument equation to "withoutWeight" (Eq. 1) and dataPointsThreshold was set to 2. Insulin data was log-transformed prior to analysis. The model optimisation implemented will adjust for unequal variances. The genotype contribution test p value was adjusted for multiple testing to control the false discovery rate to 5%.

$$Y \sim Genotype + Sex + Genotype*Sex + (1|Batch) \qquad \text{[Eq. 1]}$$

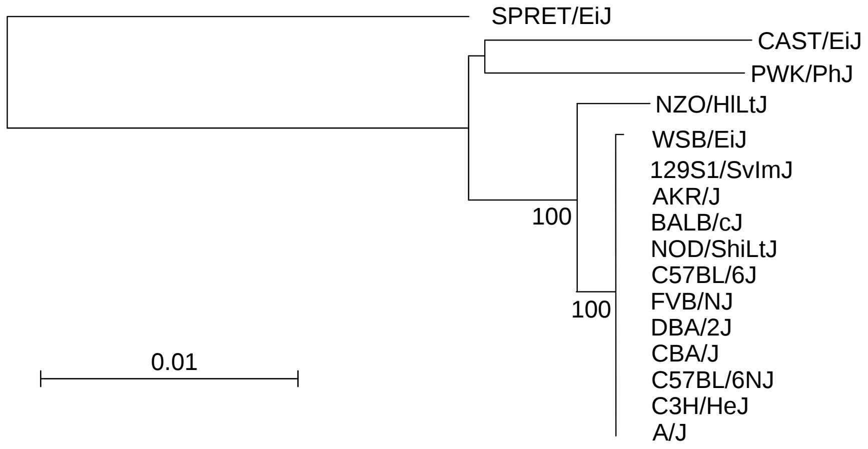
For the categorical data, a Fisher Exact Test was fitted comparing the proportions seen between wildtype and knockout mice for each sex independently using PhenStat FE Framework with the default setting. The minimum p value returned from the two tests for a variable was adjusted for multiple testing to control the false discovery rate to 5%.

References

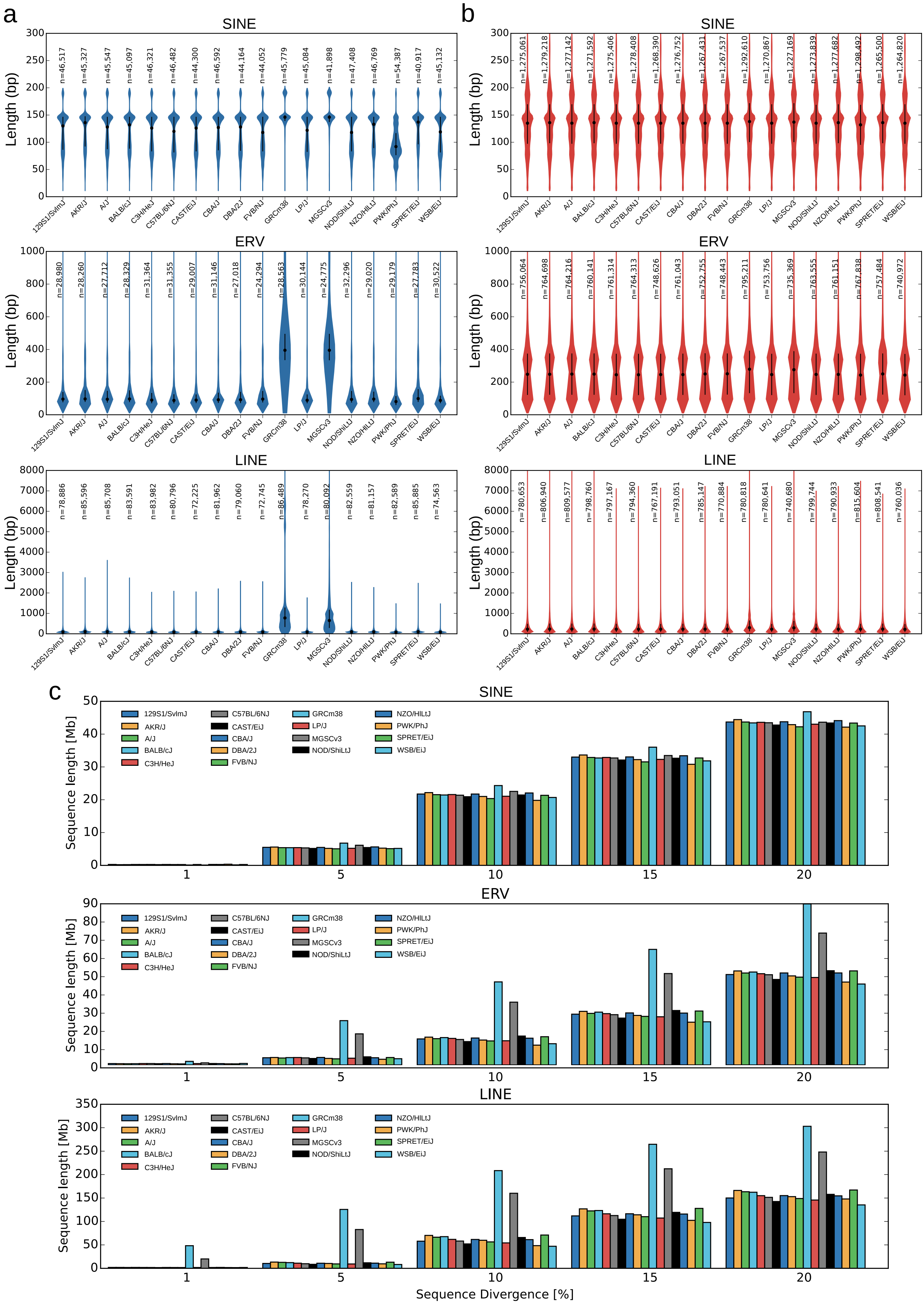
1. Park, N. *et al.* An improved approach to mate-paired library preparation for Illumina sequencing. *Methods Gener. Seq.* **1,** (2013).
2. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26,** 342–350 (2016).
3. Kirby, A. *et al.* Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. *Genetics* **185,** 1081–1095 (2010).
4. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25,** 2078–2079 (2009).
5. Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* **22,** 549–556 (2012).
6. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1,** 18 (2012).
8. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21,** 1512–1528 (2011).
9. Kolmogorov, M. *et al.* Chromosome assembly of large and complex genomes using multiple references. (2016). doi:10.1101/088435
10. Martens, P. A. & Clayton, D. A. Mechanism of mitochondrial DNA replication in mouse L-cells: localization and sequence of the light-strand origin of replication. *J. Mol. Biol.* **135,** 327–351 (1979).

11. Bonfield, J. K. & Whitwham, A. Gap5--editing the billion fragment sequence assembly. *Bioinforma. Oxf. Engl.* **26,** 1699–1703 (2010).
12. Chow, W. *et al.* gEVAL - a web-based browser for evaluating genome assemblies. *Bioinforma. Oxf. Engl.* **32,** 2508–2510 (2016).
13. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinforma. Oxf. Engl.* **24,** 637–644 (2008).
14. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7,** 62 (2006).
15. König, S., Romoth, L. W., Gerischer, L. & Stanke, M. Simultaneous gene finding in multiple genomes. *Bioinforma. Oxf. Engl.* **32,** 3388–3395 (2016).
16. Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **26,** 366–378 (2015).
17. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29,** 15–21 (2013).
18. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
19. Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinforma. Oxf. Engl.* **22,** 1437–1439 (2006).
20. Navarro, F. C. P. & Galante, P. A. F. RCPedia: a database of retrocopied genes. *Bioinforma. Oxf. Engl.* **29,** 1235–1237 (2013).
21. Yalcin, B. *et al.* The fine-scale architecture of structural variants in 17 mouse genomes. *Genome Biol.* **13,** R18 (2012).
22. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinforma. Oxf. Engl.* **23,** 1061–1067 (2007).
23. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477,** 289–294 (2011).
24. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1513–1518 (2011).
25. 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).
26. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41,** D377-386 (2013).
27. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma.* **47,** 11.12.1-34 (2014).
28. Steward, C. A. *et al.* Genome-wide end-sequenced BAC resources for the NOD/MrkTac() and NOD/ShiLtJ() mouse genomes. *Genomics* **95,** 105–110 (2010).
29. Makino, S. *et al.* Breeding of a non-obese, diabetic strain of mice. *Jikken Dobutsu* **29,** 1–13 (1980).
30. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **26,** 589–595 (2010).
31. Wang, J. R., de Villena, F. P.-M. & McMillan, L. Comparative analysis and visualization of multiple collinear genomes. *BMC Bioinformatics* **13 Suppl 3,** S13 (2012).
32. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **Chapter 4,** Unit 4.10 (2009).
33. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
34. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinforma. Oxf. Engl.* **28,** 1647–1649 (2012).
35. Shelton, J. M. *et al.* Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics* **16,** 734 (2015).
36. Toribio, A. L. *et al.* European Nucleotide Archive in 2016. *Nucleic Acids Res.* **45,** D32–D36 (2017).
37. Lilue, J., Müller, U. B., Steinfeldt, T. & Howard, J. C. Reciprocal virulence and resistance polymorphism in the relationship between Toxoplasma gondii and the house mouse. *eLife* **2,** e01298 (2013).
38. Ibarra-Soria, X. *et al.* Variation in olfactory neuron repertoires is genetically controlled and environmentally modulated. *eLife* **6,** (2017).
39. Ersfeld, K. Fiber-FISH: fluorescence in situ hybridization on stretched DNA. *Methods Mol. Biol. Clifton NJ* **270,** 395–402 (2004).
40. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27,** 849–864 (2017).
41. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
42. Skarnes, W. C. *et al.* A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474,** 337–342 (2011).
43. Boroviak, K., Doe, B., Banerjee, R., Yang, F. & Bradley, A. Chromosome engineering in zygotes with CRISPR/Cas9: Chromosome Engineering in Zygotes with CRISPR/Cas9. *genesis* **54,** 78–85 (2016).
44. Hodgkins, A. *et al.* WGE: a CRISPR database for genome engineering: Fig. 1. *Bioinformatics* **31,** 3078–3080 (2015).

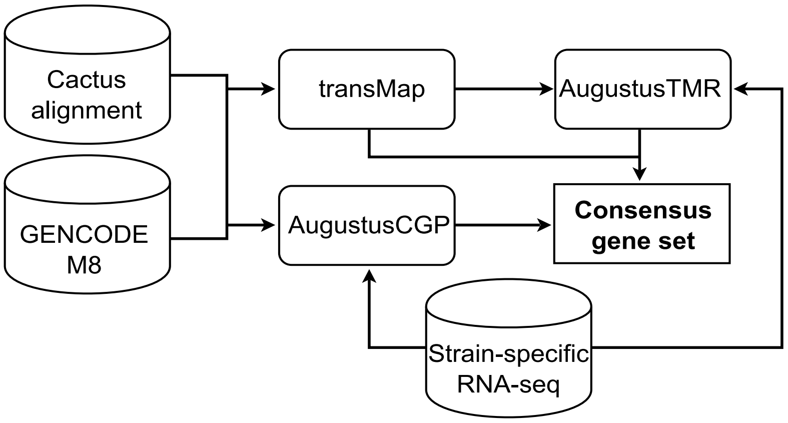
45. Tyler-Smith, C. *et al.* Where Next for Genetics and Genomics? *PLoS Biol.* **13,** e1002216 (2015).
46. Kurbatova, N., Mason, J. C., Morgan, H., Meehan, T. F. & Karp, N. A. PhenStat: A Tool Kit for Standardized Analysis of High Throughput Phenotypic Data. *PLOS ONE* **10,** e0131274 (2015).
47. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
48. Mikhaleva, A., Kannan, M., Wagner, C. & Yalcin, B. Histomorphological Phenotyping of the Adult Mouse Brain. *Curr. Protoc. Mouse Biol.* **6,** 307–332 (2016).
49. Paxinos, G. & Franklin, K. B. J. *The Mouse Brain in Stereotaxic Coordinates, 3rd ed.* (Academic Press, 2007).



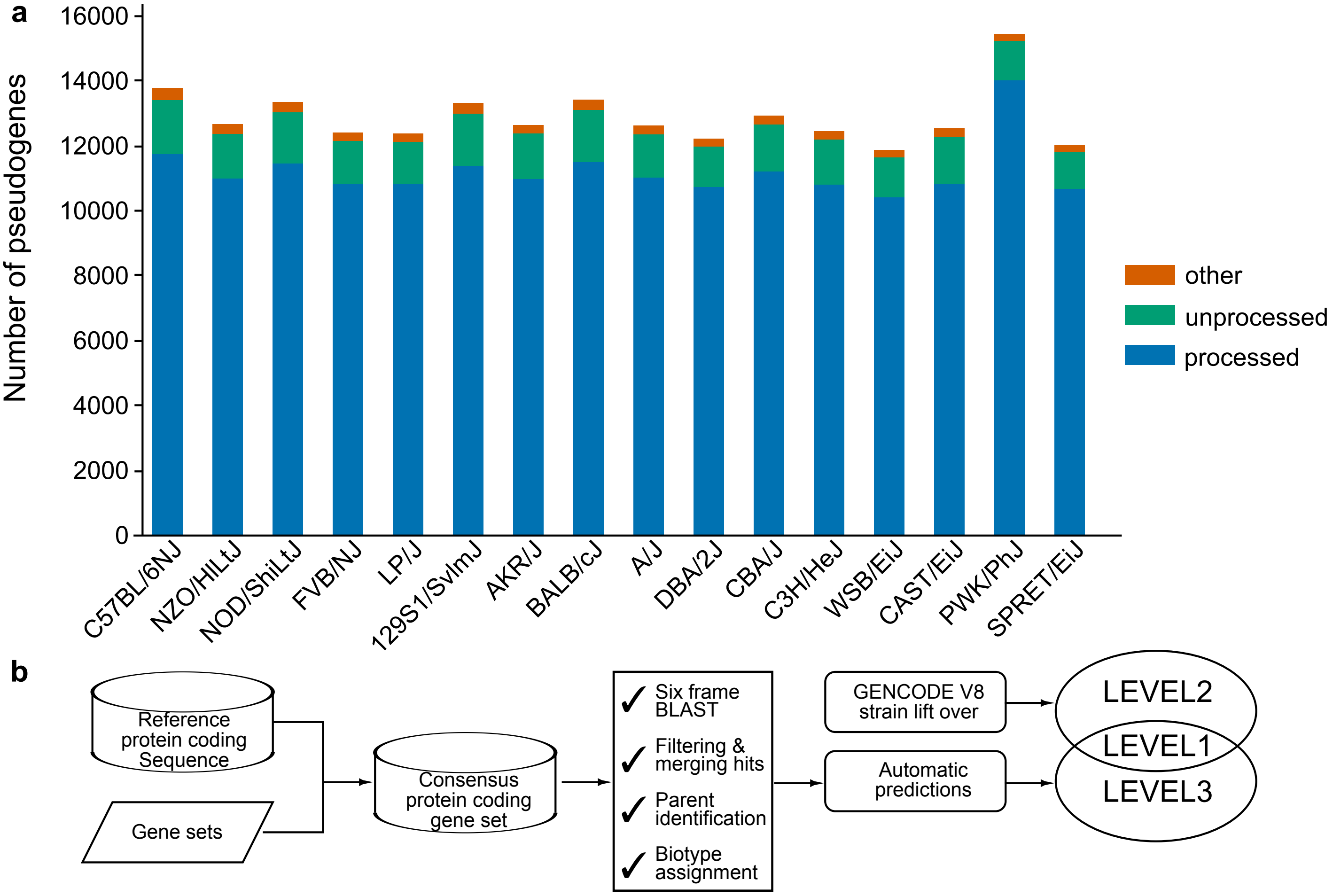
Supplementary Figure 1: Molecular Phylogenetic analysis of the mouse reference and 16 strain mitochondrial sequences. The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. Bootstraps were shown if >90.



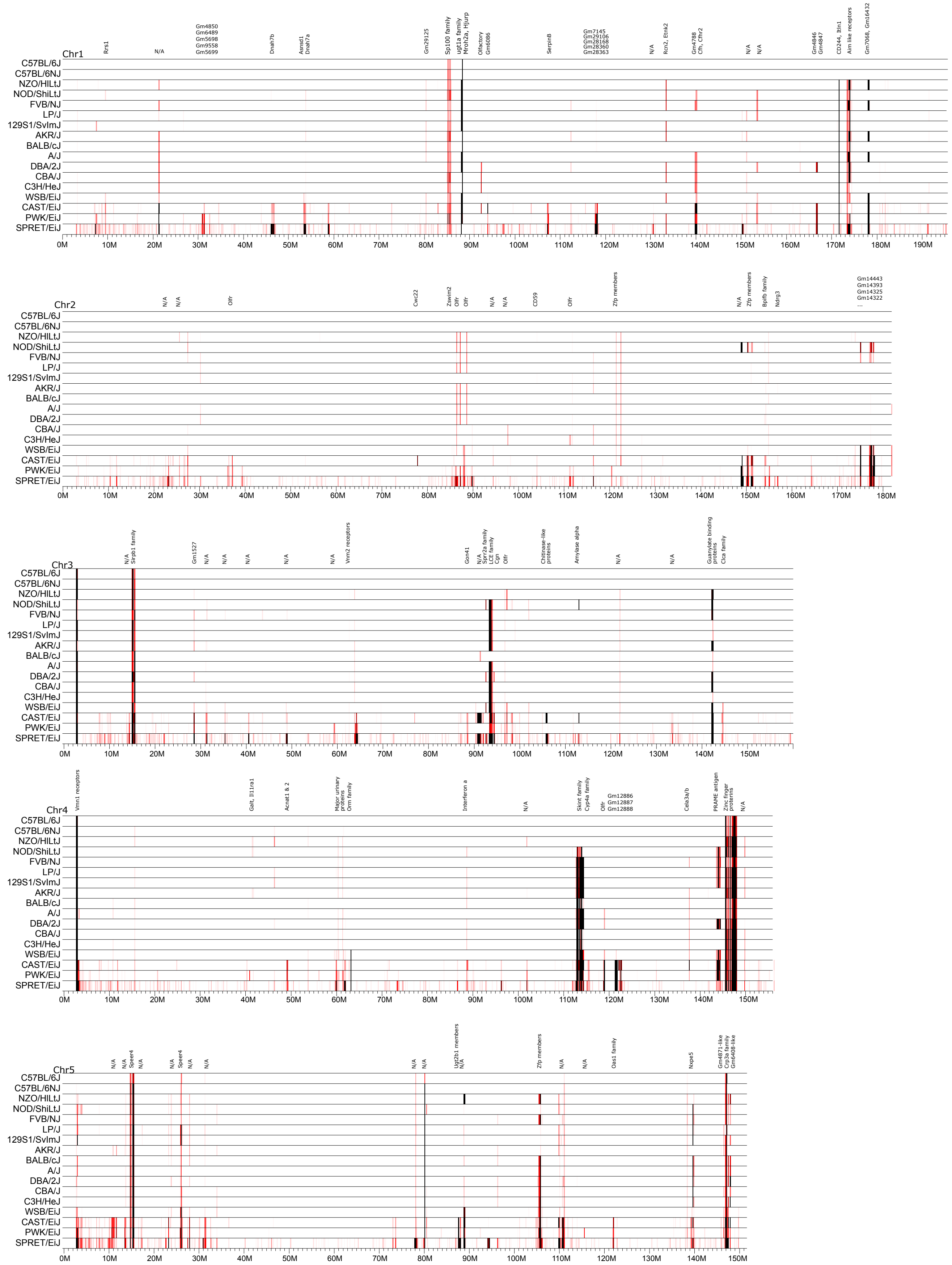
Supplementary Figure 2: Repeat elements in assembly. (a) Amount of sequence for SINE, LINE, and ERV elements in the strain assemblies, MGSCv3, and GRCm38 at less than 5% sequence divergence. Black bars show 25th and 75th percentile, dot indicate the median value. (b) Total sequence for SINE, LINE, and ERV elements inthe strain assemblies, MGSCv3, and GRCm38. Black bars show 25th and 75th percentile, dot indicate the median value. (c) Total sequence for SINE, LINE, and ERV elements at 5-20% divergence levels or less according to Repeatmasker annotations.



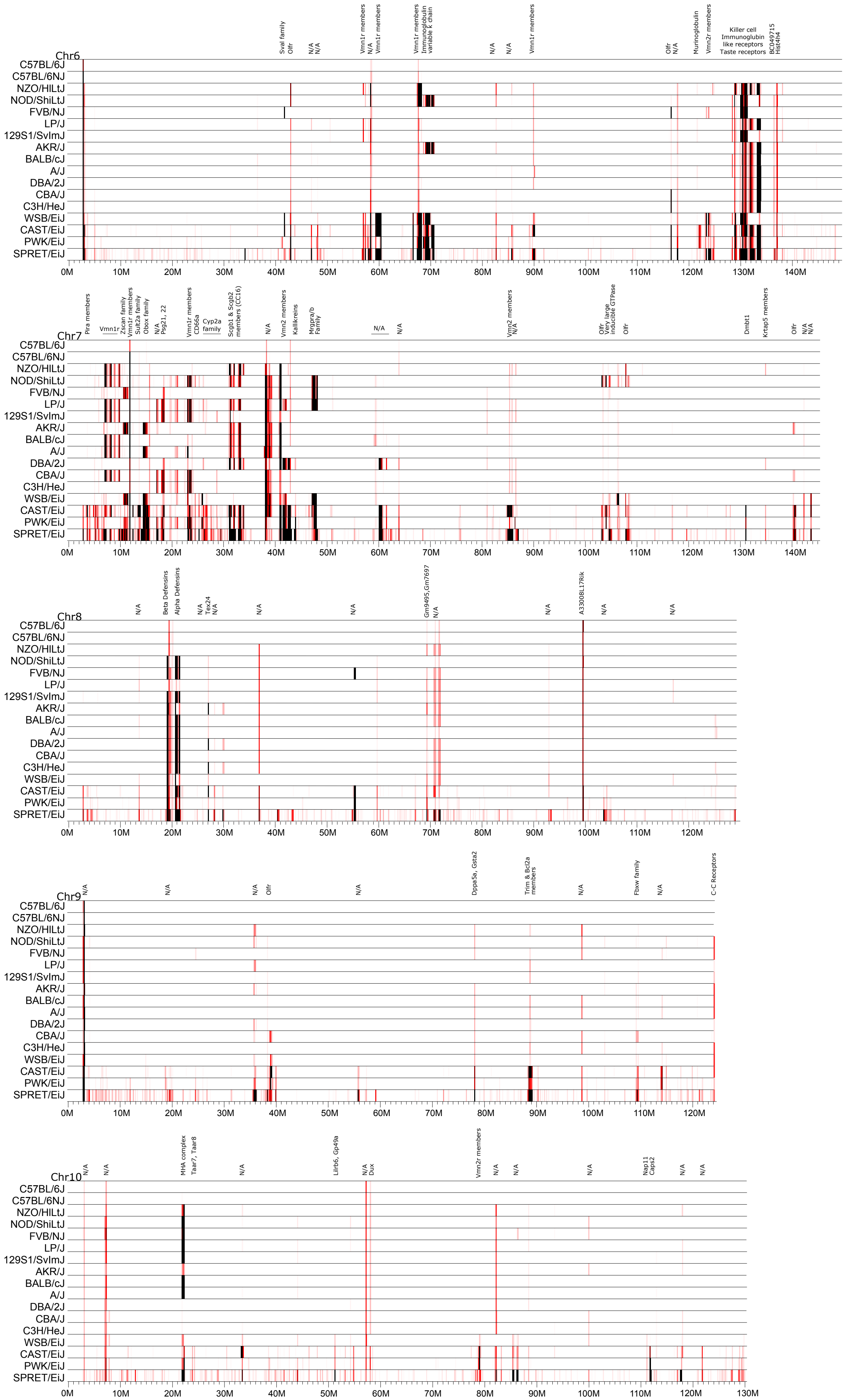
Supplementary Figure 3:Annotation pipeline. Comparative annotation flowchart with three inputs: ProgressiveCactus whole-genome alignment of all of the strains, the GENCODE M8, and a database of strain-specific RNA-seq splice junctions and expression estimates. These input to TransMap, AugustusTMR and AugustusCGP (see methods). The transcript sets output by each mode are then combined by a consensus finding algorithm into a strain specific annotation set.



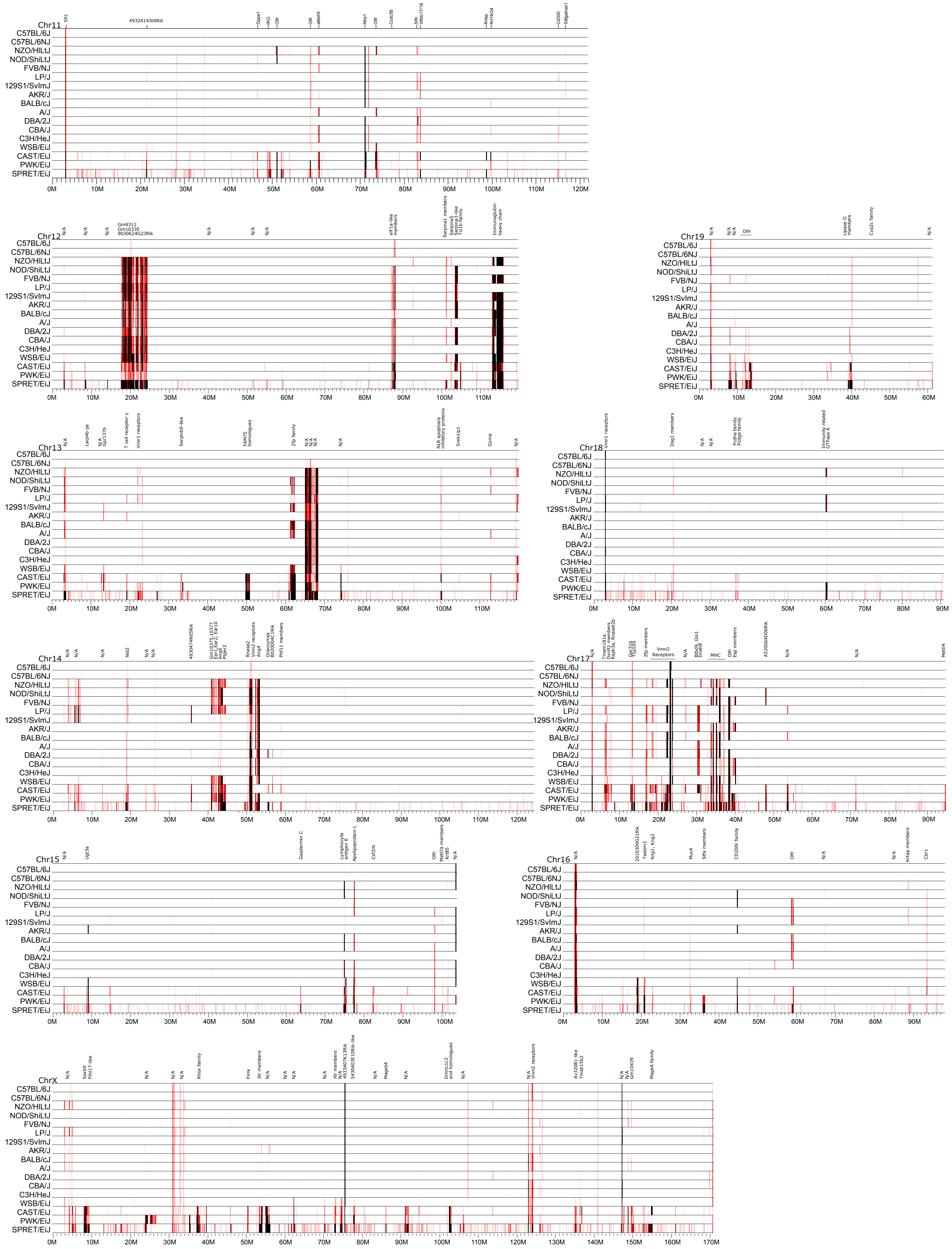
Supplementary Figure 4: (a) The number of pseudogenes predicted by our annotation pipeline for each assembly is broken down by type. Processed pseudogenes are a result of a retrotransposition event, and consequently the introns have been spliced out resulting in a monoexon-like structure. Unprocessed pseudogenes have been formed through a duplication process and follow closely the intron-exon structure of their functional homolog. A smaller number of pseudogenes are related to immune-system related genes while another small fraction of pseudogenes could not be assigned a definite biotype. These cases are grouped under the “Other” category. The number of pseudogenes predicted per genome is roughly ~13,000, with slightly smaller numbers for more distant strains (e.g. *Mus spretus, M. m. castaneus*). (b) The corresponding peptides are subjected to a six-frame BLAST against the strain genome to identify homologous loci that have not been previously annotated. Each potential match is checked for a number of disablements such as insertions, deletions, stop codons and frame-shifts. The annotated pseudogene set is intersected with the previously identified pseudogenes by lifting over the manual annotation to each strain genome. The result is a 3-tier confidence hierarchy, where level 1 refers to pseudogenes annotated by both automatic means and lift over of manual annotation, level 2 focuses on pseudogenes identified only by lifting over the manually annotated GENCODE pseudogenes, and level 3 is composed of only automatically annotated pseudogenes.



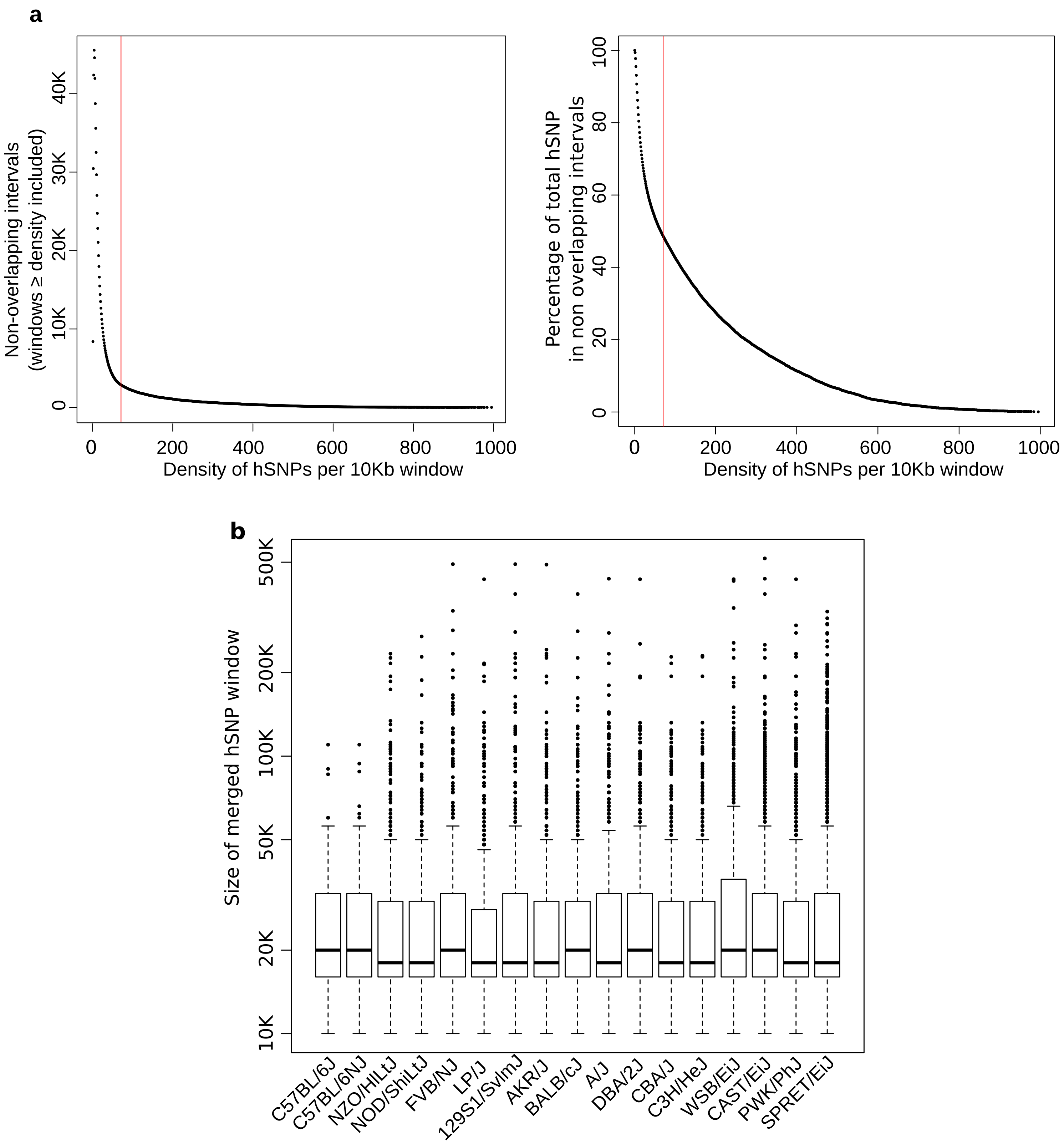
Supplementary Figure 5 (to be continued on next page)



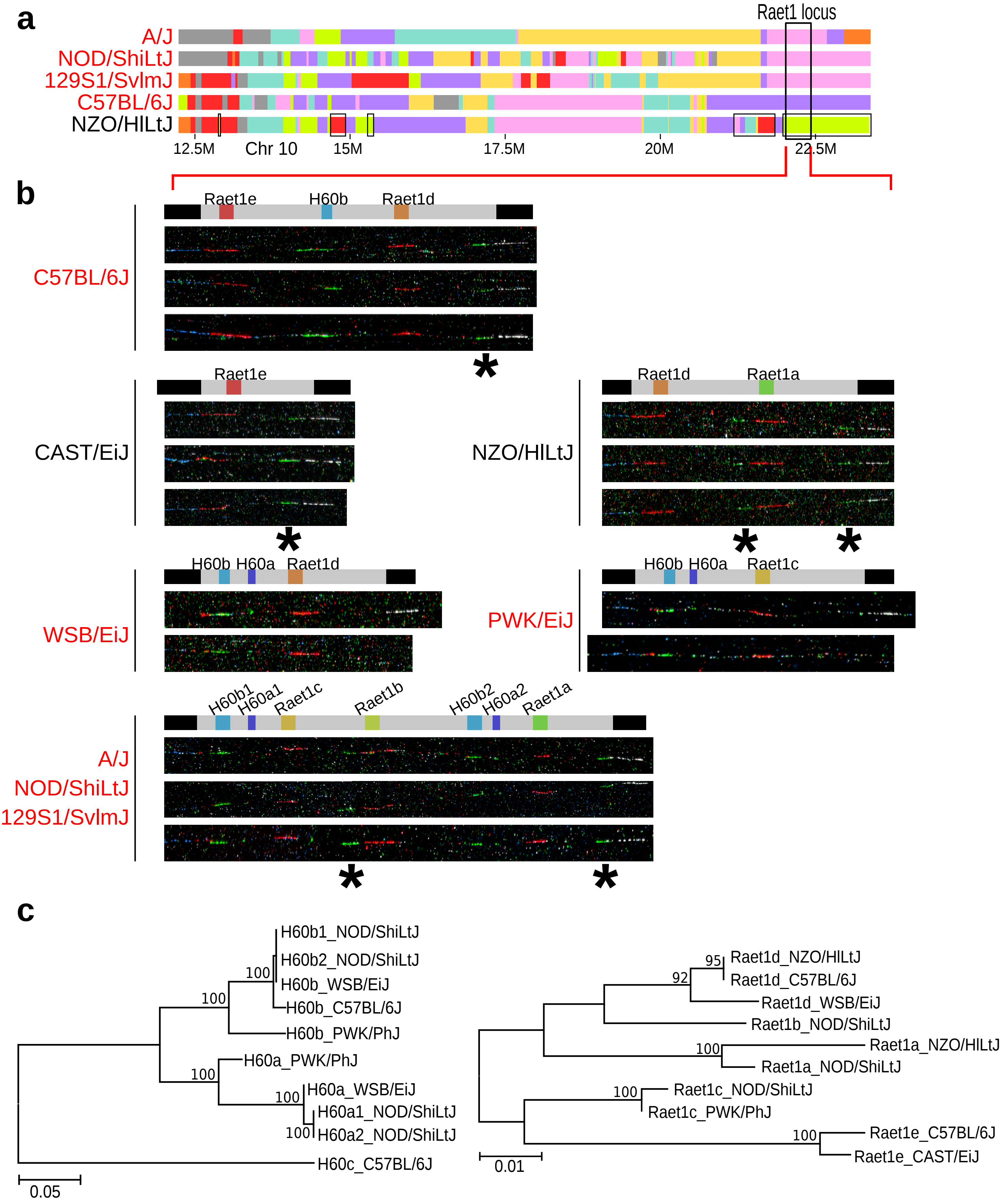
Supplementary Figure 5 (to be continued on next page)



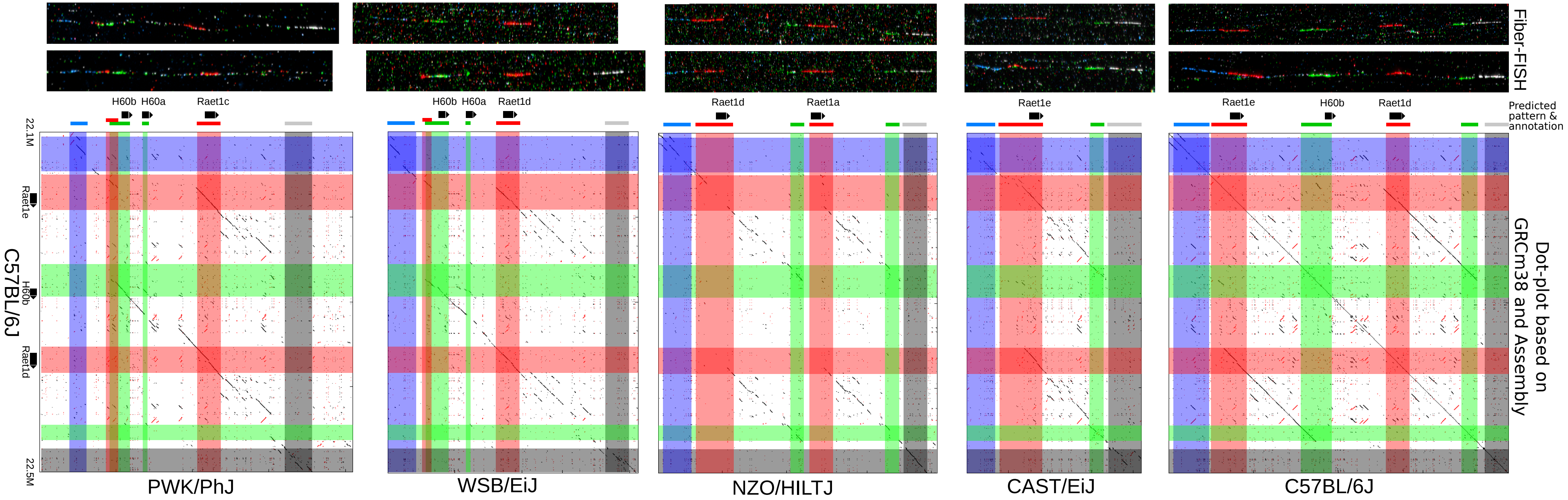
Supplementary Figure 5: Heterozygous SNP density across all chromosomes in 17 inbred mouse strains. Data generated using 200Kb adjacent windows, based on Illumina reads mapped to the GRCm38 reference genome. Heterozygous SNPs were obtained from the Mouse Genomes Project variation catalogue v5. (See Fig1b for more details).



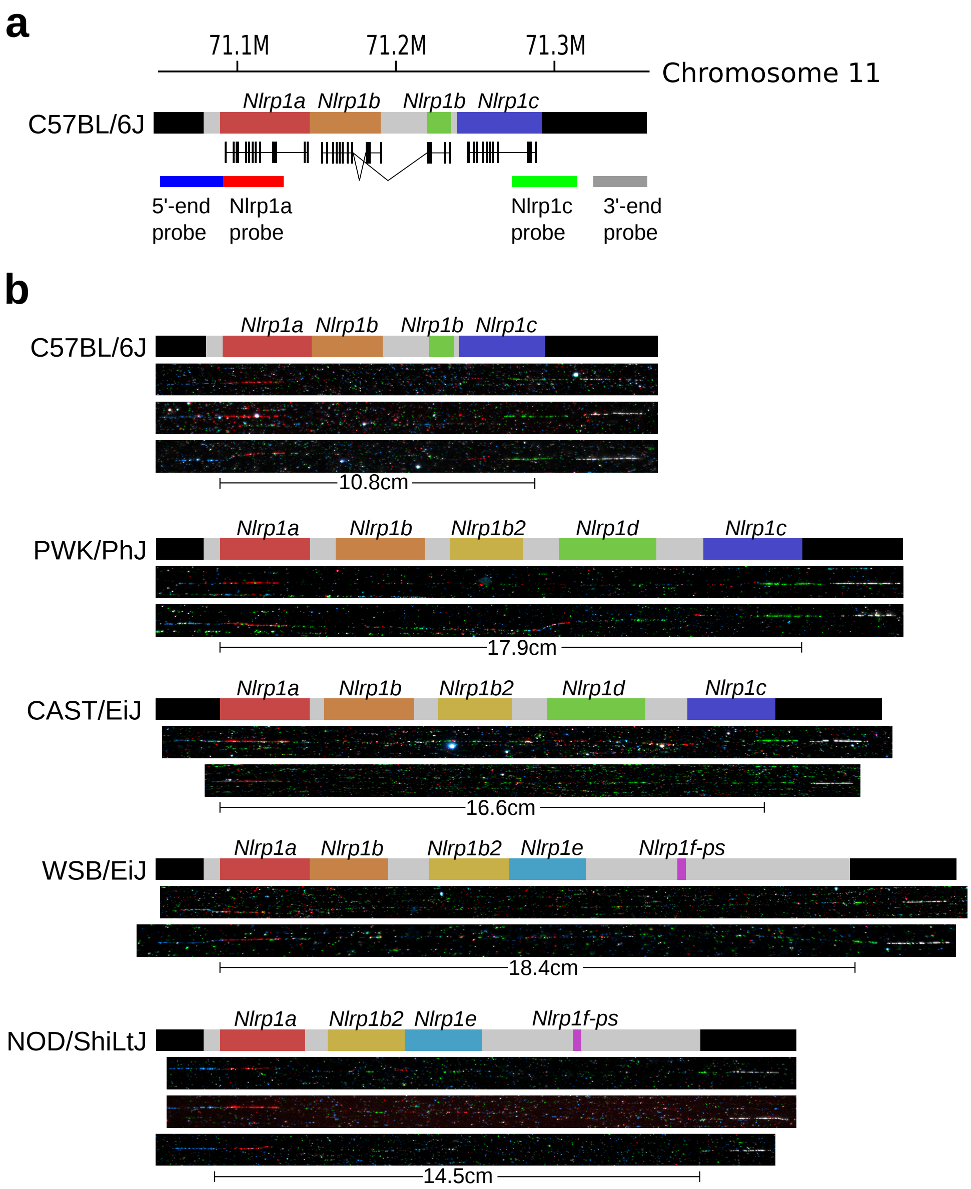
Supplementary Figure 6 (a) (left)The total number of unique non-overlapping hSNP regions at each cut-off (≥ hSNP per 10Kb windows). Regions are calculated by removing hSNP windows lower than the cut-off and collapsing the remainder into non-overlapping intervals. (right) The percentage of total hSNPs overlapping these regions. Red line indicates 5% most dense region cut off (≥71). **(b)** Boxplot containing the size distribution (y-axis) of pass hSNP dense windows for each strain (x-axis). Sample sizes of statistics are shown in supplementary table 9 (remained regions after merging / collapsing). Box in plot indicate samples from 1st quartile to 3rd quatile, with median value marked. Black dots are outliers detected by Grubbs test.



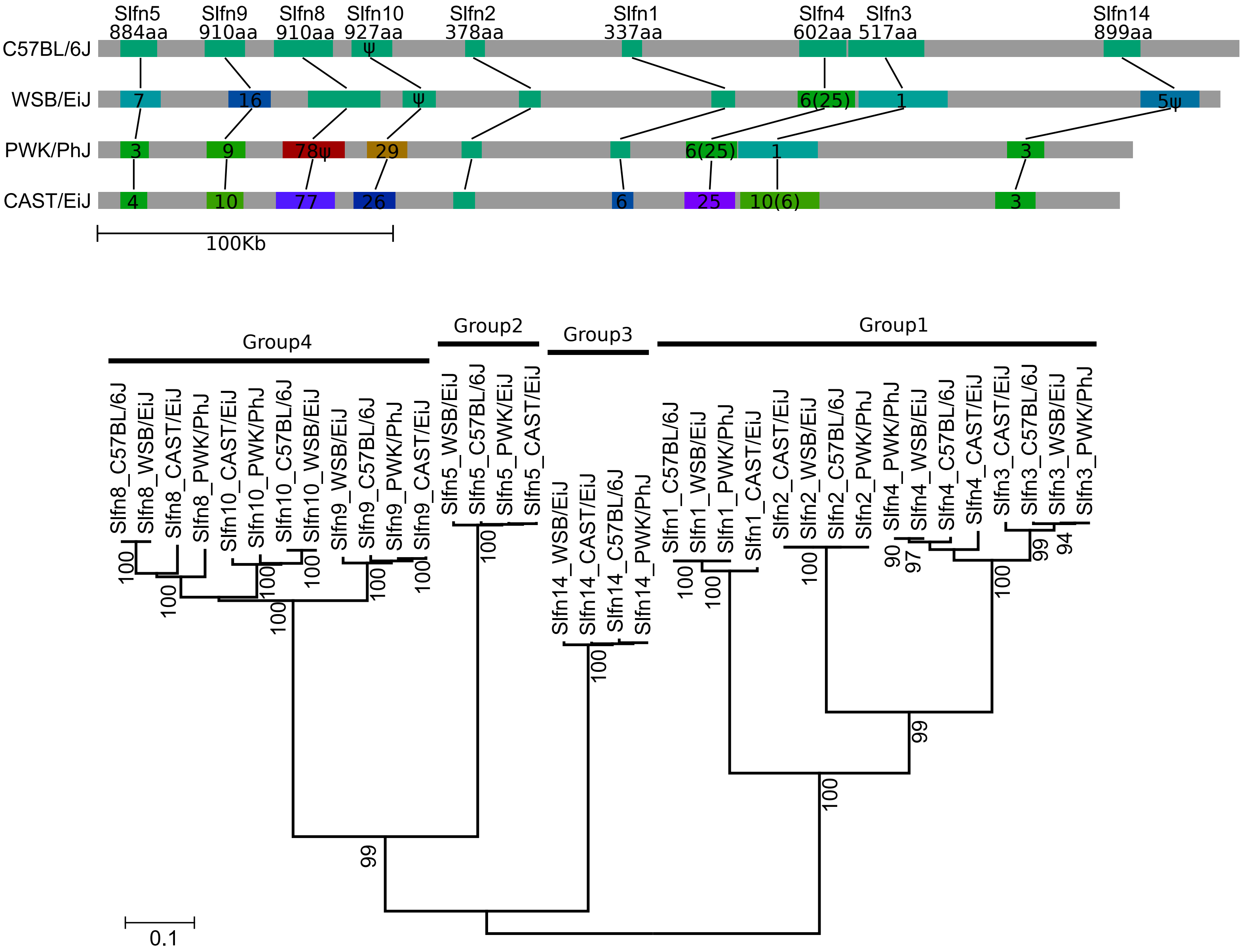
Supplementary Figure 7: De novo assembly of *Raet1/H60* locus in mouse QTL *Asprl4*. (a) Haplotype structure of the *Asprl4* QTL (90% CI) for the Colaborative Cross founder strains. Regions which is unique for *Aspergillus fumigatus* resistant strain are highlighted with boxes. Data are derived and modified from CISGen Mouse Phylogeny Viewer (https://msub.csbio.unc.edu/) Strain 129S1/SvlmJ, NOD/ShiLtJ and A/J share the same haplotype. (b) Genome structure resolved by *de novo* assembly and evidence from Fiber-FISH. Gene symbiotic colour codes are same as Figure 2b. Probe colour for Fiber-FISH photos: blue - 5'end, white - 3'end, green - *H60* alleles and adjacent region, red - *Raet1* alleles and adjacent region. Fiber-FISH hybridisation caused by a genomic duplication without H60 coding regions are marked with an asterisk ('*'). All fiber-FISH have been confirmed by at least five times identical patterns. (c) Phylogenetic tree of *H60* and *Raet1* alleles (amino acid, neighbour joining method) Note gene *H60a* is not encoded in C57BL/6 genome.



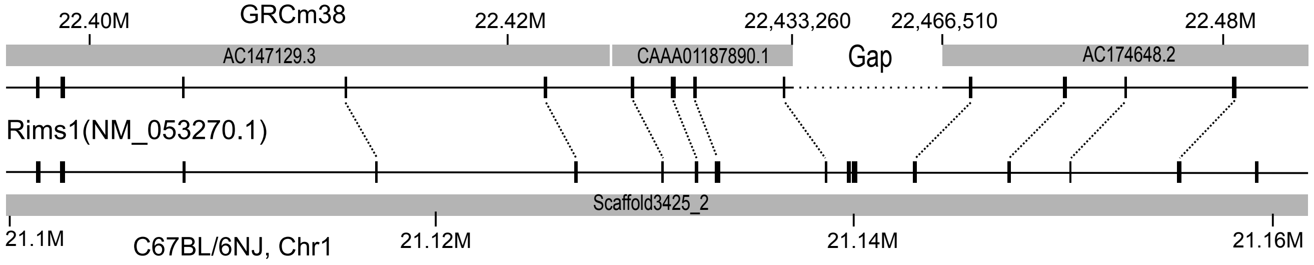
Supplementary Figure 8: Mouse strains *de novo* assemblies accurately reflects the Fiber-FISH representation of the Raet1/H60 locus. Left: Fiber-FISH observation, probe colour is same as Supplementary Figure 7. Middle: Annotated Raet1/H60 CDs and predicted pattern of Fiber-FISH based on the *de novo* sequence. Right: dot plot between GRCm38 and *de novo* assembly of other parental mouse strains with the gap size adjusted. Fiber-FISH samples shown here are identical to Supplementary Figure 7.



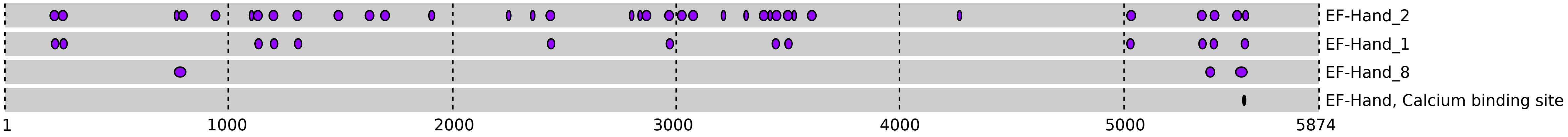
Supplmentary Figure 9, Fiber-FISH evidence for genomic structure on Nlrp1 locus. (a) Fiber-FISH probes used: blue - 5’end, white - 3’end, green - *Nlrp1c* and adjacent region, red - *Nlrp1a* alleles. No probes can be designed for *Nlrp1b* because of high concentrations of transposons in the intron. Red and green probe can bind to *Nlrp1d* and *Nlrp1e* weakly. (b) Fiber-FISH photos for five mouse strains. Gene symbiotic colour codes are same as Figure 2d. All fiber-FISH have been confirmed by at least five times identical patterns.



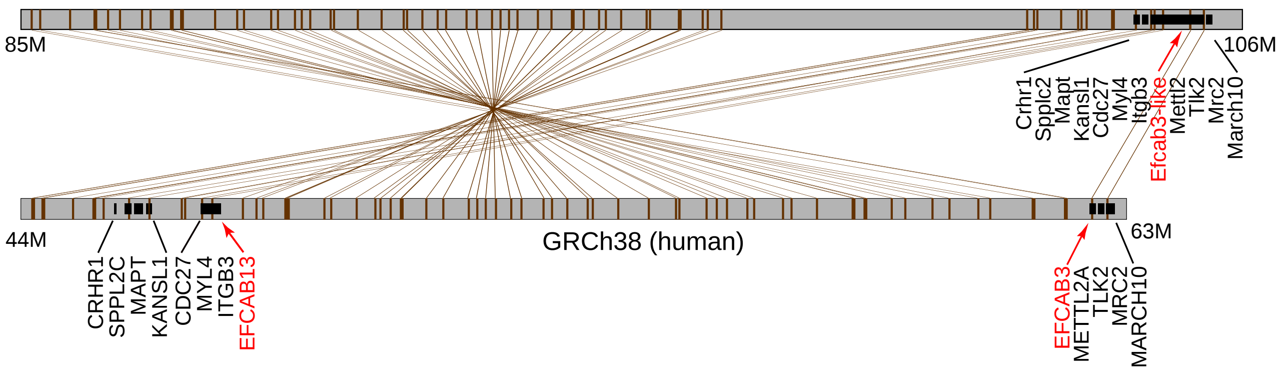
Supplementary Figure 10: Comparison of *Slfn* alleles in C57BL/6J, WSB/EiJ, PWK/PhJ and CAST/EiJ. (Top) Genome structure of *Slfn* locus in four mouse strains. Colour blocks indicate gene family members, numbers on each block indicate amino acid mismatch compared to C57BL/6J reference; number in bracket is amino acid extension caused by new start codons or loss of stop codons. Colour of each block indicate their phylogenetic relationship. "ψ" indicates pseudogenes. (bottom) Amino acid neighbour joining phylogenetic tree of *Slfn* family members. Bootstraps are shown if value >90.



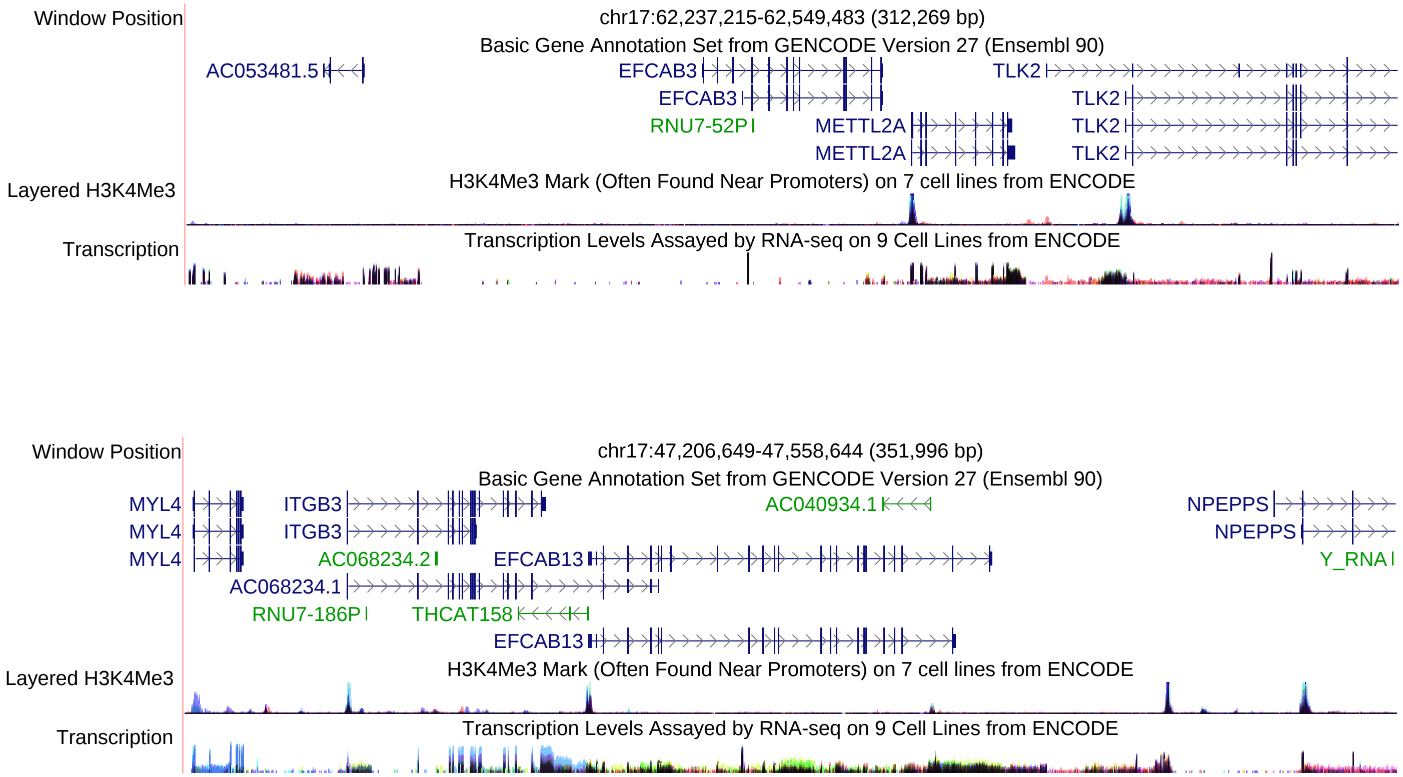
Supplementary Figure 11 Completion of Rims1 structures by the strain assemblies. Two additional exons that are located in a gap region of GRCm38 were identified in the C57BL/6NJ genome for Rims1.



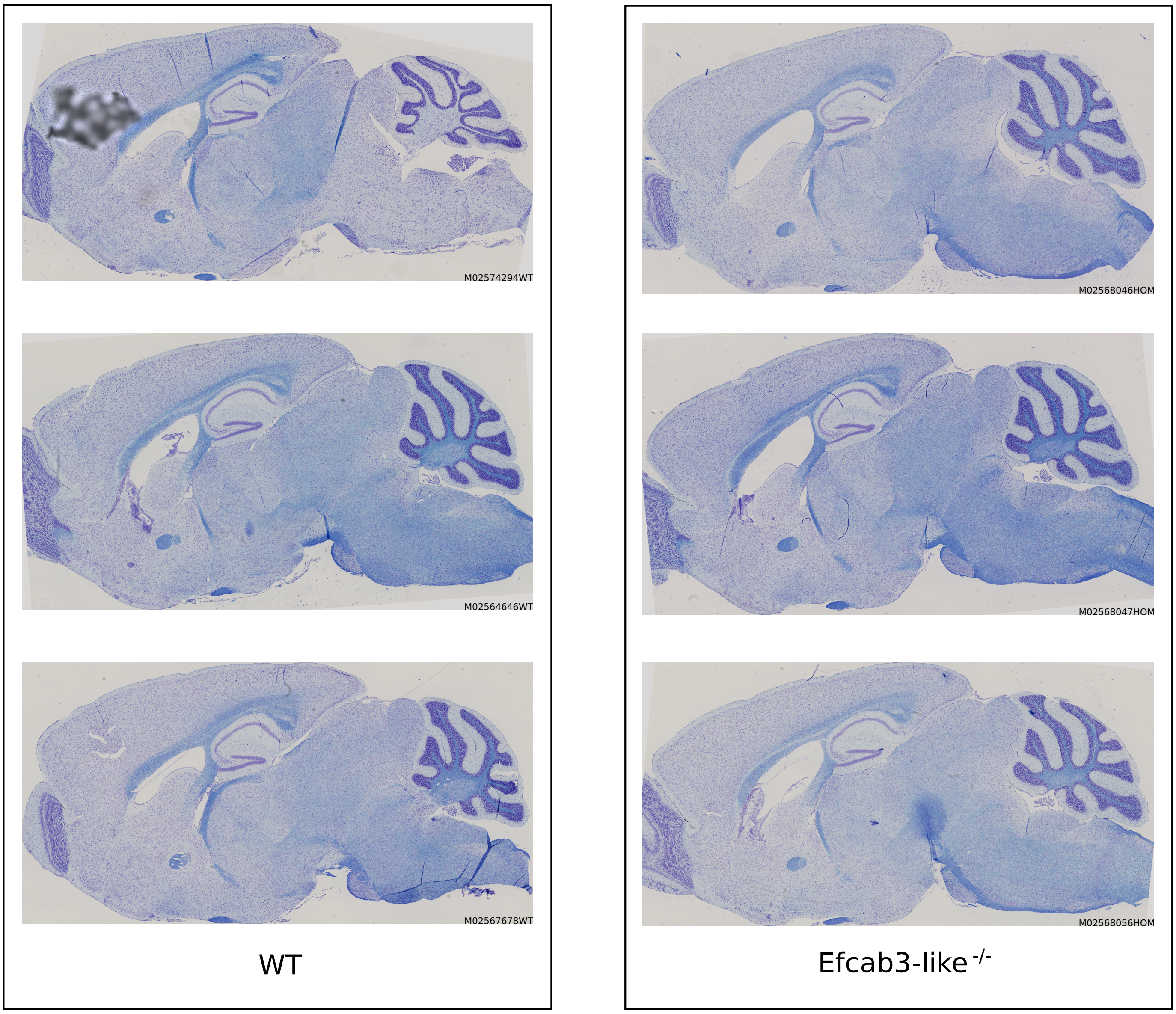
Supplementary Figure 12 Repeats and protein domains of Efcab3-like. Ef-Hand_2 (PS50222), EF-Hand_1(SM00054) EF-Hand_8(PF13883) and EF-Hand_1 Calcium binding site (PS00018) of the same Efcab3-like protein isoform were shown in four independent rows. Domain prediction was accomplished by InterPro (https://www.ebi.ac.uk/interpro/).



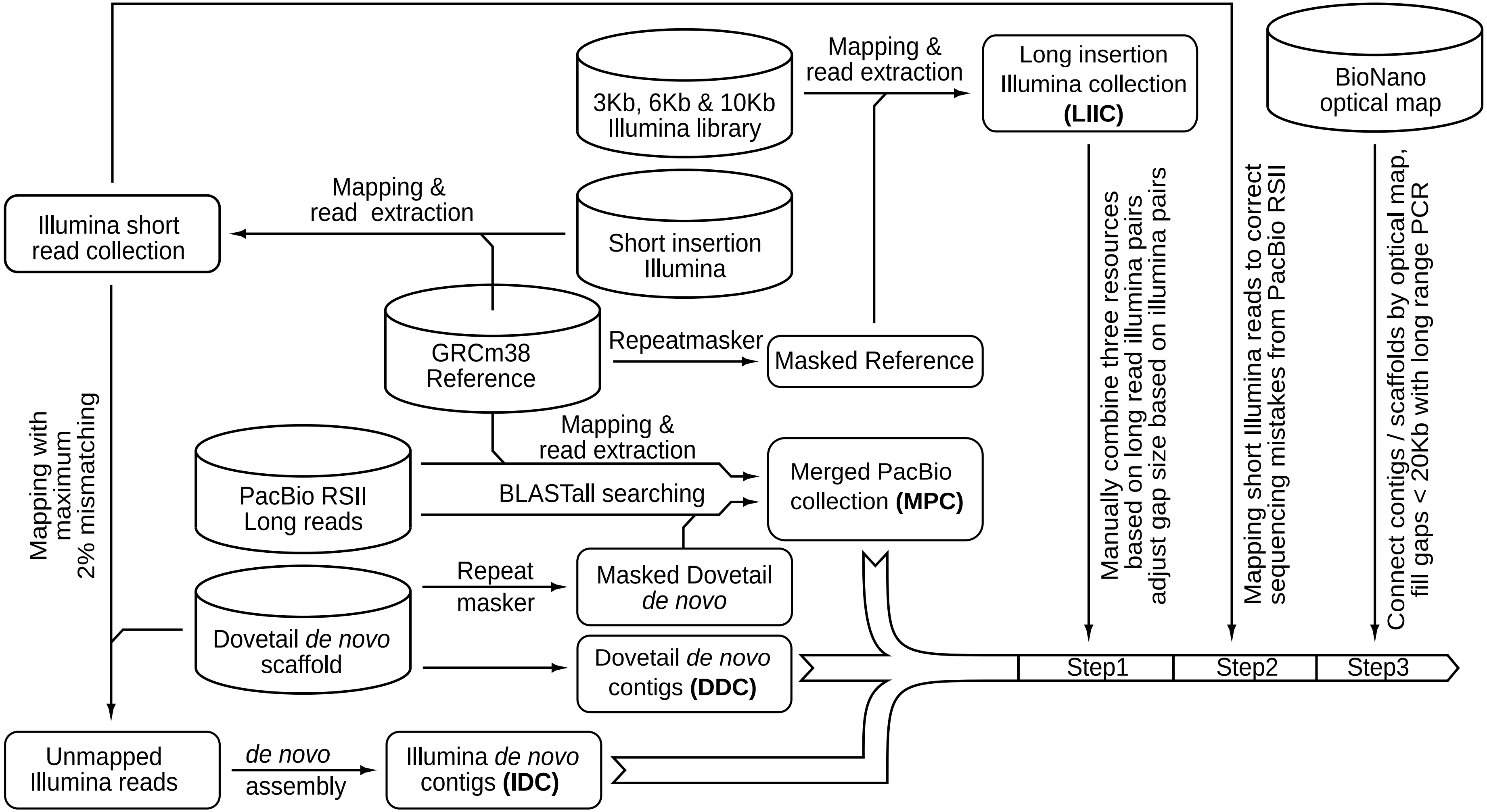
Supplementary Figure 13, a comparison of the human genes EFCAB3 and EFCAB13 and mouse *Efcab3-like*. Synteny analysis shows that EFCAB13 and EFCAB3 in human are the result of a chromosomal rearrangement which split mouse *Efcab13-like* in the *Homininae* common ancestor.


Window Position
chr17:62,237,215-62,549,483 (312,269 bp)
Basic Gene Annotation Set from GENCODE Version 27 (Ensembl 90)
AC053481.5
EFCAB3
TLK2
EFCAB3
TLK2
RNU7-52P
METTL2A
TLK2
METTL2A
TLK2
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Transcription
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
Window Position
chr17:47,206,649-47,558,644 (351,996 bp)
Basic Gene Annotation Set from GENCODE Version 27 (Ensembl 90)
MYL4
ITGB3
AC040934.1
NPEPPS
MYL4
ITGB3
NPEPPS
MYL4
AC068234.2
EFCAB13
Y_RNA
AC068234.1
RNU7-186P
THCAT158
EFCAB13
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Transcription
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE

Supplementary Figure 14 , expression of EFCAB3 and EFCAB13 in human. In human, chimpanzee and gorilla the ancestral *Efcab3-like* is broken into two pieces due to a 15Mbp inversion on chr17. The ancestral promoter can be seen on the *EFCAB13* side in the H3K4Me3 track, and expression across many tissue types in the transcription track (bottom). In contrast, the *EFCAB3* side has no promoter signal and minimal measurable expression (top).



Supplementary Figure 15. Representative sagittal brain images, double-stained with luxol fast blue and cresyl violet acetate, of matched wild type controls (WT, n = 3, left) and Efcab3-like-/- (n = 3, right), showing a larger cerebellum, enlarged lateral ventricle and increased size of the pontine nuclei.


3Kb, 6Kb & 10Kb Illumina library
Mapping & read extraction
Long insertion Illumina collection (LIIC)
BioNano optical map
Mapping & read extraction
Illumina short read collection
Short insertion Illumina
GRCm38 Reference
Repeatmasker
Masked Reference
Mapping with maximum 2% mismatching
Mapping & read extraction
PacBio RSII Long reads
BLASTall searching
Merged PacBio collection (MPC)
Repeat masker
Masked Dovetail de novo
Dovetail de novo scaffold
Dovetail de novo contigs (DDC)
Unmapped Illumina reads
de novo assembly
Illumina de novo contigs (IDC)
Manually combine three resources based on long read illumina pairs adjust gap size based on illumina pairs
Mapping short Illumina reads to correct sequencing mistakes from PacBio RSII
Connect contigs / scaffolds by optical map, fill gaps < 20Kb with long range PCR
Step1
Step2
Step3

Supplementary Figure 16: Flowchart of Targeted reassembly pipeline. Input materials are including whole genome short insertion Illumina, 3Kbp, 6Kbp and 10Kbp long insertion Illumina, PacBio RSII reads, Dovetail *de novo* and BioNano optical map (indicated by cylinders). Raw data are further processed into: Dovetail de novo contigs (DDC), Merged PacBio collection (MPC), Illumina *de novo* contigs (IDC) and Long insertion Illumina collection (LIIC). Manual adjustment and joining were performed on them, and the sequence were further polished with short and long insertion Illumina reads. Finally, BioNano optical map were used to adjust the size of long (>10Kbp) gaps.

Supplementary Table 1: Summary of the genome assemblies and strain specific gene annotation.

Strain	Contigs N50 (bp)	Scaffolds N50 (bp)	Total length (Gbp)	Unknown base (%)	Unplaced length (Mbp)	Protein coding genes	Coding transcripts	Non-coding genes	Non-coding transcripts
129S1/SvImJ	15,777	499,579	2.695	14.86	37.099	20,539	45,695	20,234	57,246
A/J	21,208	803,000	2.593	10.43	36.306	20,550	45,797	20,221	57,234
AKR/J	26,398	663,000	2.669	12.82	42.758	20,541	45,824	20,159	57,174
BALB/cJ	17,994	729,225	2.597	10.93	29.869	20,592	45,885	20,158	57,050
C3H/HeJ	15,887	780,348	2.672	13.86	2.047	20,466	45,624	19,899	56,778
C57BL/6NJ	13,800	1,402,000	2.781	17.26	24.813	20,639	45,937	20,797	57,707
CAST/EiJ	12,682	24,445,013	2.635	13.7	18.916	20,077	44,635	19,306	58,024
CBA/J	14,386	1,024,365	2.884	20.29	36.929	20,454	45,574	19,884	56,735
DBA/2J	14,576	715,000	2.578	11.22	27.352	20,530	45,731	20,026	56,869
FVB/NJ	18,499	231,286	2.557	10.56	31.154	20,479	45,548	20,027	56,883
LP/J	13,897	1,575,000	2.704	15.65	26.442	20,499	45,642	20,103	56,924
NOD/ShiLtJ	10,216	833,100	2.943	22.71	38.312	20,453	45,450	20,043	56,880
NZO/HlLtJ	13,897	455,398	2.704	15.65	34.871	20,704	46,111	20,816	57,968
PWK/PhJ	5,178	24,747,862	2.533	8.92	26.179	19,998	44,419	19,034	57,619
SPRET/EiJ	19,477	19,680,963	2.593	10.64	32.561	20,028	44,495	18,720	56,897
WSB/EiJ	12,897	783,034	2.671	15.64	17.809	20,185	44,885	19,459	56,302

Supplementary Table 2: Unplaced scaffolds

Strain	Number of scaffolds	Total length (bp)	Total size of unplaced scaffolds (bp)	Percentage of unplaced (%)	N50 scaffold size (bp)	Largest scaffold (bp)	Shortest scaffold (bp)	scaffold %N	Repeat sequences (%)
129S1/SvImJ	7133	2695574028	37099665	1.376	5944	352450	2000	28.65	71.59
A/J	4667	2593688561	36306833	1.400	14610	406133	320	30.16	66.64
AKR/J	5932	2669948688	42758867	1.601	11220	466374	442	37.91	64.25
BALB/cJ	3804	2597474342	29869912	1.150	14158	359440	1636	28.77	69.23
C3H/HeJ	4048	2672083609	29047707	1.087	11700	386019	464	40.26	70.53
C57BL/6NJ	3873	2781986799	24813488	0.892	8727	360195	309	48.15	73.15
CAST/EiJ	2956	2635074085	18916565	0.718	9514	544993	2000	32.88	77.87
CBA/J	5445	2884798194	36929761	1.280	9637	431894	319	46.87	67.46
DBA/2J	4084	2578809409	27352749	1.061	9941	359851	435	27.46	74
FVB/NJ	4992	2557464886	31154404	1.218	9414	349820	2000	18.83	72.62
LP/J	3478	2704355636	26442867	0.978	14461	430939	425	42.28	73.88
NOD/ShiLtJ	5523	2943750595	38312655	1.301	10715	393788	186	40.79	61.36
NZO/HlLtJ	7001	2664395485	34871811	1.309	5526	355679	2000	30.8	71
PWK/PhJ	3119	2533808356	26179036	1.033	11969	524517	2000	40.48	72.31
SPRET/EiJ	5383	2593026018	32561454	1.256	7465	662343	1891	35.59	69.92
WSB/EiJ	2218	2671848035	17809522	0.667	18611	395137	1049	23.81	78.88

Supplementary Table 3: Genes on unplaced scaffolds

	Total	IG_C_gene	IG_J_gene	IG_D_gene	IG_D_pseudogene	IG_V_gene	IG_V_pseudogene	TEC	TR_J_gene	TR_V_gene	TR_V_pseudogene	antisense	lincRNA	miRNA	misc_RNA	processed_pseudo	processed_transcript	protein_coding	rRNA	snRNA	snoRNA	Transcribed/processed	unprocessed	ribozyme	pseudogene	Mt_rRNA	Mt_tRNA	likely_coding	miRNA
129S1/SvImJ	335	0	0	3	1	13	13	5	1	11	2	2	8	7	1	83	2	96	1	5	10	3	68	0	0	0	0	0	0
A/J	318	0	0	14	10	0	0	2	0	15	5	5	8	9	0	63	5	114	1	2	6	2	56	1	0	0	0	0	0
AKR/J	331	0	0	15	8	0	0	1	0	5	1	2	8	4	1	88	5	130	2	0	6	1	53	0	1	0	0	0	0
BALB/cJ	338	0	0	0	0	24	15	1	0	14	5	1	6	9	1	78	2	109	0	1	8	4	60	0	0	0	0	0	0
C3H/HeJ	245	0	0	1	0	11	3	1	0	12	1	2	5	6	0	55	0	112	0	1	6	1	28	0	0	0	0	0	0
C57BL/6NJ	240	0	0	0	0	16	7	2	0	16	4	2	9	3	0	40	1	91	1	5	8	2	29	0	1	1	2	0	0
CAST/EiJ	133	2	1	0	0	6	3	1	0	0	0	0	3	6	0	42	0	40	0	2	3	1	22	0	0	0	0	1	0
CBA/J	240	0	0	1	1	10	5	1	0	8	0	1	2	8	1	60	0	103	1	3	6	0	29	0	0	0	0	0	0
DBA/2J	266	0	0	0	0	17	7	3	0	9	1	2	6	10	0	55	2	106	1	2	7	4	34	0	0	0	0	0	0
FVB/NJ	375	0	0	2	0	12	14	4	0	14	0	5	9	9	1	97	2	130	1	5	5	5	59	1	0	0	0	0	0
LP/J	204	0	0	0	0	6	6	2	2	5	2	1	10	6	1	42	3	70	0	0	7	2	39	0	0	0	0	0	0
NOD/ShiLtJ	297	0	0	0	0	10	12	0	0	14	2	6	4	8	0	74	1	98	1	7	8	2	50	0	0	0	0	0	0
NZO/HlLtJ	410	0	0	1	0	28	19	3	0	11	0	4	8	8	1	91	2	164	0	7	11	3	48	0	0	0	1	0	0
PWK/PhJ	147	0	0	0	0	5	6	1	0	14	1	0	3	2	0	31	0	46	0	3	2	1	32	0	0	0	0	0	0
SPRET/EiJ	89	0	0	0	0	3	2	0	0	0	0	0	0	0	0	15	3	39	1	3	0	0	19	0	0	0	0	0	4
WSB/EiJ	180	0	0	0	0	15	4	1	0	10	0	2	8	5	3	33	1	61	2	4	7	0	24	0	0	0	0	0	0

Supplementary Table 4: Mitochondrial DNA (mtDNA) Sequence Variation of Commonly Used Laboratory Mouse Strains.

Strain \ # of bp		C57BL/6NJ	FVB/NJ	BALB/cJ	DBA/2J	LP/J	AKR/J	NOD/ShiLtJ	A/J	CBA/J	C3H/HeJ	129S1/SvImJ	C57BL/6J†	# of Variant	SNP	Predicted AA
Locus	Position	16300	16300	16300	16300	16301	16301	16301	16300	16300	16301	16300	16299	Strains		Change
Light Strand Origin	5,182.1	:°					A*							1	_	
Cox1	5,335	T				C*								1	T5335C	I ->T
Atp8	7,778	G	T											1	G7778T	D ->Y
Cox3	8,889	G									A*			1	G8889A	A ->T
Cox3	9,348	G		A				A	A	A	A			5	G9348A	V ->I
tRNA-Gly	9,461	C											T	1	C9461T	-
tRNA-Arg	9,820.1	:									T			1	-	-
tRNA-Arg	9820-9829	9A	9A	9A	9A	10A	9A	10A	9A	9A	9A	9A	8A	3	-	-
Nd5	11,902	T		C*										1	T11902C	I ->T
Cytb	15,124	A										G		1	A15124G	I ->V
Differences vs. Reference		Reference	1	2	0	2	1	2	1	1	3	1	2			
Previous Genbank Entries	KR020497	KR020497.1	EF108338, GQ871746	EF108333 BALB/cByJ	EF108337	FJ374648	EF108332	AY533107, EF108340	EF108331	AY466499	EF108335	EF108330	DQ106412, AY172335, EF108336			
SNP Discordant with the GenBank Entries	-	-	-	T11902C	-	T5335C	:5182.1A	-	-	-	G8889A	-				

® - C57BL/6NJ was used as the reference for comparisons as this strain carries a common mtDNA haplotype.

† - C57BL/6J was not sequenced by this effort. This strain was included in comparisons due to use as a universal control strain.

* - Marks a SNP that is discordant with the mtDNA sequence that was previously published

° - A colon (:) is used to mark a missing base in the reference strain where an insertion was identified.

Supplementary Table 5: Sequence Accuracy by Samtools and Bcftools

Strain	Length	Unknown bases	SNPs	Indels	Errors per Kbp
129S1/SvImJ	2,732,659,664	411,136,651	159,923	54,379	0.092
A/J	2,630,043,295	281,603,272	185,491	67,732	0.108
AKR/J	2,712,819,408	358,705,903	183,665	80,142	0.112
BALB/cJ	2,627,356,258	292,612,086	177,423	53,543	0.099
C3H/HeJ	2,701,165,093	382,164,387	173,342	54,572	0.098
C57BL/6NJ	2,806,994,964	492,229,250	187,268	54,396	0.104
CAST/EiJ	2,653,974,351	367,107,716	192,942	58,626	0.110
CBA/J	2,921,935,818	602,869,723	166,622	56,942	0.096
DBA/2J	2,606,172,597	296,870,923	174,013	57,200	0.100
FVB/NJ	2,588,608,472	275,937,215	202,407	53,561	0.111
LP/J	2,730,912,269	434,657,747	182,429	58,264	0.105
NOD/ShiLtJ	2,982,208,028	684,342,392	177,357	48,214	0.098
NZO/HlLtJ	2,699,262,892	366,442,716	163,570	54,078	0.093
PWK/PhJ	2,595,235,564	285,549,885	165,708	52,979	0.095
SPRET/EiJ	2,625,578,457	287,409,215	215,688	60,948	0.118
WSB/EiJ	2,689,661,975	422,209,054	151,984	57,202	0.092
MGSCv3	2,580,596,378	206,479,311	520,410	272,894	0.334
GRCm38	2,730,871,774	78,088,340	38,538	14,829	0.020

Supplementary Table 6: Quality control by PCR primers

Strain	Inward	Error
A/J	569	0.048
AKR/J	570	0.047
BALB/cJ	567	0.052
C3H/HeJ	558	0.067
C57BL/6J-GRCm38	598	-
C57BL/6J-MGSCv3	538	0.100
C57BL/6NJ	593	0.008
CBA/J	569	0.048
DBA/2J	563	0.059
LP/J	570	0.047

Supplementary Table 7: cDNA PacBio concordant alignment rates (%)

	GRCm38	MGSCv3	CAST/EiJ	PWK/PhJ	SPRET/EiJ
Liver	63.07	51.70	54.37	51.69	53.01
Spleen	69.76	67.99	58.57	56.12	62.32

Supplementary Table 8: RNAseq Accession numbers

Strain	RNA-Seq					Pacbio	
	Brain	B-cell and T-cell	Lung	Spleen	Liver	Liver	Spleen
129S1/SvImJ	ERS028660		SRS502568 to SRS502584		SRS411953 to SRS411968		
A/J	ERS028666		SRS502583 to SRS502598		SRS411969 to SRS411984		
AKR/J	ERS028672						
BALB/cJ	ERS028670						
C3H/HeJ	ERS028658						
C57BL/6J				ERS819832	ERS819829	ERS716499	ERS716503
C57BL/6NJ	ERS028664						
CAST/EiJ	ERS028668, SRS875111 SRS877557 to SRS877562	ERS354502 to ERS354507 ERS354524 to ERS354528	SRS502615 to SRS502629	ERS819831	ERS819828, SRS412001 to SRS412016	ERS716498	ERS716502
CBA/J	ERS028662						
DBA/2J	ERS028659						
FVB/NJ	ERS819833						
LP/J	ERS028671						
NOD/ShiLtJ	ERS028663		SRS502630 to SRS502646		SRS412017 to SRS412032		
NZO/HlLtJ	ERS028667		SRS502645 to SRS502659		SRS412033 to SRS412048		
PWK/PhJ	ERS028661 SRS877549 to SRS877556		SRS502660 to SRS502674	ERS819830	ERS819827 SRS412049 to SRS412064	ERS716497	ERS716501
SPRET/EiJ	ERS028665				ERS819826	ERS716496	ERS716500
WSB/EiJ	ERS028669 SRS877541 to SRS877548		SRS502675 to SRS502690		SRS412065 to SRS412080		

Supplementary Table 9: hSNP density analysis

	pass het SNPs	Raw 10kb windows with 1+ hSNPs	Raw 10kb windows with 71+ hSNPs	Remained regions after merging / collapsing	hSNPs in remained regions	% hSNPs in remained regions	Total bps overlap	Total CDS bps	Overlapping genes	PantherDB defense / immunity genes
129S1/SvImJ	463053	334792	5979	530	239013	51.6168	16924530	280810	394	33
AKR/J	442255	347646	5597	522	219368	49.6022	16074522	257824	357	48
A/J	434506	324241	5737	539	224887	51.7569	16400539	261716	359	34
BALB/cJ	404719	319544	5340	506	204053	50.4184	15318506	258752	363	33
C3H/HeJ	390160	328447	4865	515	176603	45.2643	14456515	225617	323	33
C57BL/6J	69120	57727	1019	117	33747	48.8238	3090117	62769	55	1
C57BL/6NJ	116439	201313	1257	141	42321	36.3461	3756141	61538	57	1
CAST/EiJ	1128279	591370	14104	1367	552692	48.9854	40883367	667821	859	72
CBA/J	396321	310931	5136	543	185895	46.9052	15322543	247654	346	34
DBA/2J	441953	312174	6102	565	238286	53.9166	17450565	314166	411	35
FVB/NJ	446478	288587	6072	545	243127	54.4544	17228545	296115	390	40
LP/J	429677	318637	5428	599	201269	46.8419	16298599	262552	300	34
NOD/ShiLtJ	431691	347921	5574	615	199330	46.1742	16766615	294379	375	57
NZO/HlLtJ	414118	334121	5362	541	194422	46.9485	15748541	234818	341	36
PWK/PhJ	1013842	581244	11760	1264	439262	43.3265	35013264	567205	771	88
SPRET/EiJ	1895741	764285	25865	2567	926491	48.8722	75592567	1236315	1442	121
WSB/EiJ	606014	368358	8279	688	335025	55.2834	23038688	360714	503	67
Merged	4739122	6131338	123476	2862	2305651	48.6514	97120862	1575166	1828	155

Supplementary Table 10: Panther DB representation of Genes encoded in hSNP dense regions

	hSNP dense regions			GENCODE M8 CDS			PantherDB default genes		
	Gene count	Gene hits	Protein class hits	Gene count	Gene hits	Protein class hits	Gene count	Gene hits	Protein class hits
Defense/immunity (PC00090)	155	9.9%	14.0%	401	2.1%	2.7%	561	2.6%	3.4%
Nucleic acid binding (PC00171)	131	8.4%	11.8%	2135	11.3%	14.5%	2367	11.2%	14.6%
Transcription factor (PC00218)	116	7.4%	10.5%	1286	6.8%	8.8%	1457	7.0%	9.0%
Transporter (PC00227)	103	6.6%	9.3%	942	5.0%	6.4%	920	4.4%	5.8%
Receptor (PC00197)	99	6.3%	8.9%	1137	6.0%	7.7%	1086	5.2%	6.8%
Hydrolase (PC00121)	89	5.7%	8.0%	1357	7.2%	9.2%	1483	7.0%	9.2%
Signaling molecule (PC00207)	76	4.9%	6.9%	959	5.1%	6.5%	1083	5.2%	6.8%
Enzyme modulator (PC00095)	73	4.7%	6.6%	1229	6.5%	8.4%	1353	6.4%	8.4%
Transferase (PC00220)	54	3.4%	4.9%	1051	5.5%	7.2%	1164	5.6%	7.2%
Oxidoreductase (PC00176)	36	2.3%	3.2%	577	3.0%	3.9%	597	2.8%	3.8%
Cell adhesion (PC00069)	36	2.3%	3.2%	434	2.3%	3.0%	492	2.4%	3.0%
Transfer/carrier (PC00219)	30	1.9%	2.7%	361	1.9%	2.5%	364	1.8%	2.2%
Cytoskeletal protein (PC00085)	28	1.8%	2.5%	667	3.5%	4.5%	778	3.8%	4.8%
Calcium-binding (PC00060)	20	1.3%	1.8%	367	1.9%	2.5%	390	1.8%	2.4%
Ligase (PC00142)	17	1.1%	1.5%	343	1.8%	2.3%	372	1.8%	2.4%
Isomerase (PC00135)	10	0.6%	0.9%	157	0.8%	1.1%	158	0.8%	1.0%
Lyase (PC00144)	9	0.6%	0.8%	144	0.8%	1.0%	151	0.8%	1.0%
Membrane traffic (PC00150)	5	0.3%	0.5%	306	1.6%	2.1%	372	1.8%	2.4%
Cell junction protein (PC00070)	5	0.3%	0.5%	137	0.7%	0.9%	140	0.6%	0.8%
Storage protein (PC00210)	5	0.3%	0.5%	27	0.1%	0.2%	25	0.2%	0.2%
Extracellular matrix protein (PC00102)	4	0.3%	0.4%	312	1.6%	2.1%	364	1.8%	2.2%
Chaperone (PC00072)	3	0.2%	0.3%	152	0.8%	1.0%	183	0.8%	1.2%
Viral protein (PC00237)	2	0.1%	0.2%	10	0.1%	0.1%	16	0.0%	0.0%
Structural protein (PC00211)	2	0.1%	0.2%	141	0.7%	1.0%	166	0.8%	1.0%
Transmembrane receptor regulatory/adaptor (PC00226)	1	0.1%	0.1%	56	0.3%	0.4%	65	0.4%	0.4%
surfactant (PC00212)	0	0.0%	0.0%	5	0.0%	0.0%	8	0.0%	0.0%
Total genes	1567			19171			20972		
Total number protein class hits	1109			14773			16115		

Supplementary Table 11: Comparison of repeat ages found within and outside of heterozygous SNP dense pass regions

Repeat category	GRCm38 genome-wide	Mean % divergence	Sample size (n) Within hSNP dense regions	Mean % divergence	n < 1% divergence	Sample size (n) Outside hSNP dense regions	Mean % divergence	n < 1% divergence	Difference in mean ages*
LINEs	625618	17.799	27076	15.377	623	599617	17.911	14545	$2.2\text{x}10^{-16}$
LTRs	670986	20.155	29815	17.510	146	641890	20.297	2046	$2.2\text{x}10^{-16}$
SINEs	1292649	21.118	29399	20.137	14	1263279	21.141	603	$2.2\text{x}10^{-16}$

* Welch's two sample t-test that was used to compare the average ages of the repeats within and outside hSNP dense regions.

Supplementary Table 12. Repeat enrichment analysis

	All age repeats analysed			Young repeats (<1% seq divergence)		
LINEs	p-value	Total simulations	Mean content	p-value	Total simulations	Mean content
Over-representation	$1\text{x}10^{-7}$	1000000	22050.562	0.047	1000000	577.835
Under-representation	1	1000000	22050.562	0.957	1000000	577.835
SINEs						
Over-representation	0.982057	1000000	46200.103	0.971	1000000	21.843
Under-representation	0.017943	1000000	46200.103	0.049	1000000	21.843
LTRs						
Over-representation	$1\text{x}10^{-7}$	1000000	22795.038	$1\text{x}10^{-7}$	1000000	73.091
Under-representation	1	1000000	22795.038	1	1000000	73.091

For over-representation, p-values are the fractions of simulated repeat content counts were equal to or more extreme than the observed repeat content in hSNP dense regions. For under-representation, p-values are the fraction of simulated repeat content counts were equal to or less extreme than the observed repeat content in hSNP dense regions

Supplementary Table 13: Representation of novel CAST/EiJ *Olfr* alleles in *de novo* assembly

New alleles	Length	Identity	Location	Start (bp)	End (bp)	*De novo* assembly
Olfr1459_new	1088	100	chr19	9961696	9962783	Correct
Olfr1502_new	1386	100	chr19	10772704	10773434	Correct*
Olfr1480_new	1076	100	chr19	10425162	10423157	Match with gaps
Olfr1487_new	1009	-	-	-	-	No good hit
Olfr46_new	1063	99.9	chr7	137512292	137513261	Match with gaps
Olfr643_new	1079	100	chr7	98952738	98955250	Match with gaps
Olfr393_new	1049	98	chr11	73912427	73913454	Match with gaps
Olfr566_CAST	1030	100	chr7	97630898	97630781	Partial sequence
Olfr585_CAST	1135	100	chr7	97879663	97876532	Partial sequence
Olfr498_CAST	1113	100	chr7	103389301	103391680	Partial sequence
Olfr209_new	1081	100	chr16	58382240	58382512	Partial sequence
Olfr1151_new	1163	100	chr2	88767581	88776837	Partial sequence
Olfr1152_new	1031	100	chr2	88774256	88774754	Partial sequence
Olfr507_new	1096	100	chr7	103622238	103623333	Correct
Olfr635_new	1032	100	chr7	98852030	98852464	Partial sequence
Olfr911_new	1005	99.9	chr9	36511467	36512471	Correct
Olfr467_new	967	100	chr7	102812359	102812210	Correct
Olfr661_new	653	100	chr7	99656716	99657335	Correct
Olfr525_new	1054	100	chr7	137367392	137366339	Correct
Olfr384_new	1075	100	chr11	74001578	74003037	Match with gaps
Olfr394_new	998	98	chr11	74206283	74207267	Match with gaps
Olfr646_new	1027	99.71	chr7	99026945	99027973	Match with gaps
Olfr1333_new	1220	99.41	chr4	116638873	116639881	Match with gaps
Olfr285_new	1013	99.9	chr15	99936591	99937603	Match with gaps
Olfr1331_new	1071	100	chr4	116598198	116597128	Correct
Olfr747_CAST	894	100	chr14	41807383	41808366	Correct
Olfr212_new	881	97.39	chr6	118114492	118115372	Match with gaps
Olfr1402_new1	1075	-	-	-	-	No good hit
Olfr1402_new2	1103	100	chr3	97033490	97034565	Partial sequence

* Transposon element insertion disrupts the sequence, so the gene is truncated.

Twenty nine new olfactory receptors or strain specific olfr alleles were reported in mouse strain CAST/EiJ in previous study. 27 out of 29 were fully or partially assembled in *de novo* assembly (final column).

Supplementary Table 14: GENCODE reference annotation updates

GRCm38, Chr1 to Chr12, manually corrected

Annotation decision		**Biotype**	**Number**
Annotated new locus	**62**	Protein coding	19
		lncRNA	37
		Pseudogene	6
Annotated updated annotation	**272**	new coding transcript	105
		new transcript	31
		new NMD transcript	6
		other	130
Rejected	**451**	already annotated	231
		would not be annotated	185
		genomic sequence error	34
		pseudogene	1
		Total	**785**

GRCm38, Chr13 to Chr19, and ChrX

Annotation decision		**Biotype**	**Number**
Annotated new locus	**36**	Protein coding	15
		lncRNA	15
		Pseudogene	6
Annotated updated annotation	**13**	new coding transcript	0
		new transcript	3
		new NMD transcript	2
		other	8
Rejected	**20**	already annotated	2
		genomic sequence error	7
		would not be annotated	11
		Total	**69**
		Total Annotated	383
		Total Rejected	471
		Grand Total	**854**

Supplementary Table 15: Description of the 40 neuroanatomical parameters

Region ID	Parameter	Description	Unit
1	4_TB_area	Total brain area	cm^2
	4_TB_width	Width of the total brain	cm
	4_TB_height_CS1	Height of the total brain at CS1 (coronal critical section 1)	cm
	4_TB_height_CS2	Height of the total brain at CS2 (coronal critical section 2)	cm
2	4_TCTX_area	Total cortical area	cm^2
	4_M2_length	Length of the secondary motor cortex	cm
	4_M1_length	Length of the primary motor cortex	cm
3	4_Pons_height	Height of the pons	cm
4	4_TC_area	Total cerebellar area	cm^2
	4_IGL_area	Area of the internal granular layer of the cerebellum	cm^2
	4_Folia	Number of folia	digit
	4_Med_area	Area of the medial cerebellar nucleus	cm^2
5	4_LV_area	Lateral ventricle area	cm^2
6	4_cc_area	Corpus callosum area	cm^2
	4_cc_length	Total outer length of the corpus callosum	cm
	4_cc_height	Corpus callosum thickness	cm
7	4_TTh_area	Total thalamic area	cm^2
8	4_CPu_area	Caudate putamen area	cm^2
9	4_HP_area	Hippocampus area	cm^2
	4_Rad_length	Length of the radiatum layer of the hippocampus	cm
	4_Or_length	Length of the oriens layer of the hippocampus	cm
	4_TILpy_area	Area of pyramidal cells of the hippocampus	cm^2
	4_TILpy_length	Total internal length of pyramidal cell layer of the hippocampus	cm
	4_Mol_length	Length of the molecular layer of the hippocampus	cm
	4_DG_area	Dentate gyrus area	cm^2
	4_DG_length	Dentate gyrus length	cm
10	4_fi_area	Area of the fimbria of the hippocampus	cm^2
11	4_aca_area	Anterior commissure area	cm^2
12	4_sm_area	Stria medullaris area	cm^2
13	4_f_area	Fornix area	cm^2
14	4_och_area	Optic chiasm area	cm^2
15	4_VMHvl_area	Area of ventromedial nucleus of the hypothalamus	cm^2
16	4_Pn_area	Pontine nuclei area	cm^2
17	4_SN_area	Substantia nigra area	cm^2
18	4_fp_area	Area of fibre of pons	cm^2
19	4_Cg_area	Cingulate cortex area	cm^2
	4_Cg_height	Height of the cingulate cortex	cm
20	4_DS_area	Dorsal subiculum area	cm^2
21	4_InfC_area	Inferior colliculus area	cm^2
22	4_SupC_area	Superior colliculus area	cm^2

Supplementary Table 16: Mouse Efcab3-/- full raw neuroanatomical data

Sex	Male	Male	Male	Male	Male	Male
Barcode	M02574294	M02564646	M02567678	M02568046	M02568047	M02568056
Genotype	WT	WT	WT	Efcab3-/-	Efcab3-/-	Efcab3-/-
Age	16 weeks	16 weeks	16 weeks	16 weeks	16 weeks	16 weeks
4_TB_area	0.394839062	0.3893	0.3917	0.4138	0.427	0.4122
4_TB_width	0.681202975	0.6505	0.6822	0.6756	0.6641	0.6739
4_TB_height_CS1	0.557927685	0.5561	0.5498	0.5588	0.5642	0.5408
4_TB_height_CS2	0.601141271	0.6136	0.6022	0.6336	0.6647	0.6248
4_TCTX_area	0.03961	0.03494	0.0377	0.03909	0.03831	0.03848
4_M2_length	0.145630557	0.1142	0.1476	0.1405	0.1175	0.1426
4_M1_length	0.133662337	0.1339	0.1476	0.1457	0.1423	0.1293
4_Pons_height	0.234887096	0.2707	0.2523	0.2669	0.2605	0.2597
4_TC_area		0.06805	0.06007	0.08621	0.07759	0.08043
4_IGL_area	0.034195211	0.03554	0.02742	0.0429	0.04203	0.03464
4_Folia	8	8	8	8	8	8
4_Med_area		0.003572	0.004626	0.0004613	0.00185	0.002456
4_LV_area	0.009947616	0.012605403	0.008081	0.01631	0.01728	0.01707
4_cc_area	0.012429953	0.013	0.01379	0.01222	0.0143	0.01348
4_cc_length	0.554432276	0.5767	0.592	0.5935	0.6515	0.6591
4_cc_height	0.018583466	0.02268	0.0198	0.02261	0.02451	0.02145
4_TTh_area	0.035233709	0.03687	0.03318	0.04147	0.04305	0.04041
4_CPu_area	0.0002768	0.002947			0.0006813	0.003032
4_HP_area	0.018596528	0.02175	0.02002	0.01932	0.02344	0.02232
4_Rad_length	0.02904	0.03097	0.03143	0.02583	0.0384	0.03061
4_Or_length	0.012538119	0.01302	0.01243	0.01441	0.01618	0.01241
4_TILpy_area	0.00200652	0.001708	0.001784	0.001699	0.002092	0.002098
4_TILpy_length	0.208773608	0.214	0.2116	0.2299	0.2716	0.2311
4_Mol_length	0.012571168	0.01558	0.01312	0.01298	0.01713	0.01421
4_DG_area	0.002233363	0.00185	0.001854	0.001788	0.001897	0.001982
4_DG_length	0.204909557	0.2048	0.1845	0.1866	0.2202	0.2044
4_fi_area	0.005504016	0.005991	0.005328	0.005189	0.005061	0.005331
4_aca_area		0.00166		0.001458	0.001703	0.001441
4_sm_area	0.00164858	0.00115	0.004514	0.001472	0.0009963	0.001267
4_f_area	0.00130731			0.00094	0.001026	0.001155
4_och_area	0.000786433	0.0008198	0.000763	0.0007356	0.001022	0.0009324
4_VMHvl_area				0.002336	0.002719	0.00259
4_Pn_area		0.002329	0.002318	0.003424	0.003266	0.003213
4_SN_area		0.003378	0.002189	0.002665	0.002383	
4_fp_area		0.002418	0.003039	0.00348	0.002925	0.002876
4_Cg_area	0.004525	0.005272	0.007008	0.007478	0.006565	0.005457
4_Cg_height	0.07979959	0.07215	0.07297	0.08927	0.1197	0.07743
4_DS_area	0.0008697	0.001135	0.0008985	0.0008761	0.001505	0.0008949
4_InfC_area		0.01167	0.01083	0.006489	0.0126	0.01076
4_SupC_area		0.05302	0.05447	0.05629	0.05585	0.05189

Supplementary Table 17: Identifiers for the mice sequenced in this study

Strain	Jax stock no.	Generation of sequenced animal	Date of birth
C57BL/6NJ	005304	?+F8	3/1/08
FVB/NJ	001800	F95pF98	
A/J	000646	F280	6/6/08
AKR/J	000648	F256	6/3/08
BALB/cJ	000651	F226	7/30/08
C3H/HeJ	000659	F258pF262	7/31/08
CBA/J	000656	F275	7/27/08
CAST/EiJ	000928	F90pF93	6/8/08
DBA/2J	000671	F219pF224	6/10/08
LP/J	000676	F195	6/5/08
NOD/ShiLtJ	001976	F117pF121	6/18/08
NZO/HlLtJ	002105	?+F41	5/20/08
PWK/PhJ	004660	F69+3+17	6/1/08
SPRET/EiJ	001146	F78	2/25/08
WSB/EiJ	001145	?+F4	5/17/08
129S1/SvImJ	002448	F63pF65	6/29/08

Supplementary Table 18: SGA de novo assembly parameters.

SGA step	Parameters
preprocess (preprocess fastq files)	--pe-mode 1
index (index preprocessed data)	-a ropebwt –no-reverse
correct (error correction on preprocessed data)	-k 55 --learn
index (index error-corrected data)	-a ropebwt
filter (filter error-corrected data)	--low-complexity-check -x 2 -r 256
fm-merge (merge reads)	-m 65
index (index FM-merged reads)	-d 5000000
overlap (construct string graph)	-m 65
assemble (contig assembly)	-m 77 -d 0.4 -g 0.1 -r 10 -l 200
FilterFasta.pl (remove small contigs)	-n 200
bwa index (index filtered contigs)	-a bwtsw
sga-align (align reads to contigs)	N/A
samtools merge (merge refsorted and unsorted BAMs)	N/A
sga-astat.py (make astat file from refsort BAM)	-m 200
sga-bam2de.pl (insert size distance estimates)	--prefix pe -n 5 -m 200
scaffold (build scaffolds)	-u 25 -m 200
scaffold2fasta (covert to fasta file)	--write-unplaced -m 200 --use-overlap

SGA version 0.9.43
bwa version 0.5.9
samtools version 0.1.18-r572

Supplementary Table 19: SOAP scaffolding parameters.

	Paired-end			3kb mate-pair			6kb mate-pair			10kb mate-pair			40Kb fosmid ends			BAC-ends 120kb		
Strain	Average insertion	Pair-number cut off	Rank	Average insertion	Pair-number cut off	Rank	Average insertion	Pair-number cut off	Rank	Average insertion	Pair-number cut off	Rank	Average insertion	Pair-number cut off	Rank	Average insertion	Pair-number cut off	Rank
129S1/SvImJ	350	15	1	2200	6	2	5000	11	3	13500	23	4						
A/J	350	15	1	2300	7	2	5000	15	3	13000	28	4	35000	5	5			
AKR/J	350	15	1	4200	6	2	5000	11	3	13000	18	4	35000	5	5			
BALB/cJ	350	15	1	2500	5	2	5000	15	3	8000	32	4	35000	5	5			
C3H/HeJ	350	15	1	2500	5	2	5000	13	3	11000	15	4	35000	5	5			
C57BL/6NJ	350	15	1	4000	8	2	5000	13	3	14000	22	4	35000	6	5			
CAST/EiJ	350	15	1	2500	8	2	4800	13	3	13000	15	4						
CBA/J	350	15	1	2600	4	2	4800	11	3	13500	14	4	35000	5	5			
DBA/2J	350	15	1	2600	8	2	5000	16	3	6000	5	4	35000	11	5			
FVB/NJ	350	15	1	2600	7	2	5000	17	3	6000	8	4						
LP/J	350	15	1	2500	10	2	5200	12	3	15000	15	4	35000	7	5			
NOD/ShiLtJ	350	15	1	2500	13	2	5200	40	3	12500	26	4				200000	5	5
NZO/HlLtJ	350	15	1	2600	5	2	5000	13	3	12000	31	4						
PWK/PhJ	350	15	1	4000	6	2	5000	12	3	6000	5	4						
SPRET/EiJ	350	15	1	2500	6	2	5000	11	3	12000	15	4						
WSB/EiJ	350	15	1	2600	14	2	5000	14	3	14000	38	4						

Supplementary Table 20: Scaffold break parameters

	10kb minimum	40kb minimum
129S1/SvImJ	21	
A/J	25	4
AKR/J	15	4
BALB/cJ	28	4
C3H/HeJ	13	4
C57BL/6NJ	19	5
CAST/EiJ	13	
CBA/J	12	4
DBA/2J	5	10
FVB/NJ	7	
LP/J	13	6
NOD/ShiLtJ	24	
NZO/HlLtJ	28	
PWK/PhJ	5	
SPRET/EiJ	13	
WSB/EiJ	34	

Supplementary table 21: Pairs of flanking gRNAs used for the Efcab3-like CRISPR deletion including WGE IDs (https://www.sanger.ac.uk/htgt/wge/).

Location	WGE ID	Primer sequence
5’ forward	343966999	GAATTTGGAGCACATGCCTGTGG
5’ reverse	343967008	TCAAGGCTAGTCTGACTACGTGG
3’ forward	343967088	TGAGCCTATGCAGGTCACACTGG
3’ reverse	343967105	CTGATAGACGTGAGAATTTGAGG

Supplementary Table 22: Genome Assembly Submissions

BioProject	BioSample	Accession	Strain	Jax code
PRJNA310854	SAMN04489811	LVXH00000000	129S1/SvImJ	2448
PRJNA310854	SAMN04489813	LVXI00000000	A/J	646
PRJNA310854	SAMN04489815	LVXJ00000000	AKR/J	648
PRJNA310854	SAMN04489816	LVXK00000000	BALB/cJ	651
PRJNA310854	SAMN04489818	LVXL00000000	C3H/HeJ	659
PRJNA310854	SAMN04489821	LVXM00000000	C57BL/6NJ	5304
PRJNA310854	SAMN04489822	LVXN00000000	CAST/EiJ	928
PRJNA310854	SAMN04489823	LVXO00000000	CBA/J	656
PRJNA310854	SAMN04489824	LVXP00000000	DBA/2J	671
PRJNA310854	SAMN04489825	LVXQ00000000	FVB/NJ	1800
PRJNA310854	SAMN04489826	LVXR00000000	LP/J	676
PRJNA310854	SAMN04489827	LVXS00000000	NOD/ShiLtJ	1976
PRJNA310854	SAMN04489828	LVXT00000000	NZO/HiLtJ	2105
PRJNA310854	SAMN04489829	LVXU00000000	PWK/PhJ	3715
PRJNA310854	SAMN04489830	LVXV00000000	SPRET/EiJ	1145
PRJNA310854	SAMN04489831	LVXW00000000	WSB/EiJ	1146

Supplementary Table 23: Accession codes for read sequencing data used for genome assemblies, and consensus sequences for immunity loci.

Strain	Illumina reads				Fosmid ends	Dovetail	Whole genome PacBio	Reassembly		
	400bp	3kbp	6kbp	10kbp				IRG	Nlrp1	Slfn
129S1/SvImJ	ERS076385	ERS012249	ERS160570	ERS349316						
AKR/J	ERS075418 ERS212196 ERS154382	ERS012255	ERS160571	ERS361114						
A/J	ERS075416 ERS138733 ERS212195	ERS126509	ERS160572	ERS361111						
BALB/cJ	ERS076386	ERS012260	ERS160573	ERS349279						
C3H/HeJ	ERS076383	ERS012262	ERS160574	ERS349280	ERS180302 ERS180306					
C57BL/6NJ	ERS076384	ERS126512	ERS160575	ERS349310	ERS180304 ERS180303					
CAST/EiJ	ERS076381	ERS012267	ERS160576	ERS361113		ERS745821	ERS636080		LT629147	LT629148
CBA/J	ERS076379	ERS012268	ERS160577	ERS349312						
DBA/2J	ERS075663	ERS012271	ERS160578	ERS361112	ERS234606 ERS234607					
FVB/NJ	ERP000687	ERS126515	ERS160585	ERS349309						
LP/J	ERS076382	ERS012273	ERS160579	ERS349315	ERS234608 ERS234609					
NOD/ShiLtJ	ERS076389	ERS126513	ERS160580	ERS349281						
NZO/HlLtJ	ERS076387	ERS012274	ERS160581	ERS349282						
PWK/PhJ	ERS076378	ERS126511	ERS160582	ERS349311		ERS745822	ERS559170	LT629149	LT629150	LT629151
SPRET/EiJ	ERS076388 ERS138732	ERS126510	ERS160583	ERS349314		ERS745823	ERS636081			
WSB/EiJ	ERS076380	ERS126514	ERS160584	ERS349313				LT629152	LT629153	LT629154

Supplementary Table 24: BioNano genome maps summary

	Map count	Min length (Mb)	Median length (Mb)	Mean length (Mb)	N50 length (Mb)	Max length (Mb)	total length (Mb)	Mean length between labels (bp)	Median length between labels
129S1/SvImJ	3220	0.051	0.606	0.796	1.05	6.306	2564.703	7,463.49	5,692.80
A/J	2403	0.106	0.773	1.053	1.454	7.081	2531.241	7,735.42	5,905.90
AKR/J	5164	0.085	0.4	0.477	0.55	2.539	2461.811	7,550.62	5,818.25
BALB/cJ	3617	0.089	0.548	0.712	0.909	6.005	2574.398	7,496.68	5,740.20
C3H/HeJ	2559	0.083	0.779	1.008	1.345	8.842	2579.082	7,672.16	5,869
C7BL/6NJ	4428	0.054	0.467	0.573	0.694	3.437	2537.999	7,966.23	6,188.20
CAST/EiJ	2659	0.131	0.723	0.968	1.386	9.856	2574.823	7,610.07	5,813.30
CBA/J	3048	0.103	0.62	0.793	1.031	4.682	2417.411	7,774.18	6,087.80
DBA/2J	2929	0.082	0.647	0.859	1.174	5.893	2515.869	7,489.16	5,761.70
FVB/NJ	2767	0.077	0.684	0.94	1.353	7.552	2602.332	7,748.98	5,926.40
NOD/ShiLtJ	3079	0.19	0.649	0.816	1.009	4.001	2512.108	7,594.36	5,773.90
NZO/HiLtJ	2510	0.153	0.75	1.029	1.455	8.51	2583.715	7,634.20	5,858
PWK/PhJ	4520	0.082	0.452	0.559	0.675	4.22	2524.735	7,352.23	5,634.80
SPRET/EiJ	3465	0.097	0.593	0.725	0.852	6.804	2511.936	8,176.41	6,354.10
WSB/EiJ	3076	0.08	0.627	0.785	0.952	4.903	2414.256	7,969.92	6,179.40
LP/J	2999	0.121	0.638	0.911	1.303	8.126	2732.569	8,100.62	6,214.60

Supplementary Table 25: Summary of the pseudogene annotation in mouse strains

Strain	Total	Level 1	Level 2	Level 3	Processed	Duplicated	Other
129S1/SvImJ	12827	5284	1042	6501	10616	1591	620
A/J	12740	5295	997	6448	10684	1417	639
AKR/J	12914	5289	996	6629	10791	1496	627
BALB/cJ	13011	5344	939	6728	10786	1598	627
C3H/HeJ	12736	5201	917	6618	10665	1455	616
C57BL/6NJ	13205	5615	993	6597	10859	1661	685
CAST/EiJ	12404	4694	1003	6707	10216	1549	639
CBA/J	12842	5231	898	6713	10710	1494	638
DBA/2J	12409	5282	908	6219	10451	1335	623
FVB/NJ	12694	5257	977	6460	10652	1430	612
LP/J	12688	5199	1015	6474	10626	1418	644
NOD/ShiLtJ	12954	5285	937	6732	10725	1589	640
NZO/HlLtJ	12877	5592	1048	6237	10762	1465	650
PWK/PhJ	12163	4630	865	6668	10294	1325	544
SPRET/EiJ	11935	4444	980	6511	10137	1242	556
WSB/EiJ	12102	4869	873	6360	10168	1336	598

Supplementary Table 26: Long range PCR for PWK Nlrp1

External 5' primer	External 3' primer	internal 5' primer	internal 3' primer	predicted size(bp)	PCR product size(bp)	Repeat elements (%)
CTTCCAACTGGCATTAGTGAT	ACATCTGCCATCTCCTACATT	GATGGGCTTCTATCTACCTGA	CGGGAATCTAGTATTGCATGA	~2Kb	1988	58%
TGCCATGTAAGAAGATAGTTG	GGTGTATGGCTTCATTCATAG	AAGGAGCCAAGTTGATAACC	TTCAAAAGCTACATCGCTAAA	~3Kb	3567	Off target PCR
CGCCAACCCACATACCT	CTCATTGGGAGAACCTATTCA	GTGCCCAATCCTGATAGC	CAAGAATTGGATAGCCCTAGA	~10Kb	10620	88%
TACTGCCAGAAACTTGACAAC	GCCTGGAGCCAATACCT	TGCCAGAAACTTGACAACC	GAGCCAATACCTAGACGAGAC	~3Kb	3716	70%
GCCTTACAGAGATATTGCAC	GAGAAGACAGCCAACTACTTT	CCTTACAGAGATATTGCAC	ACTGTTTCCTAATAATCTGAG	~2Kb	No product	-
AGCAGGCCAATACTTATCATC	GAGTGGCCAGTGGGTTAT	TACATGCTAGTTGTCCGAAGT	GGCCAGTGGGTTATCAGT	~9Kb	12267	90%
GGGCAGGTATTGTGTTAATCT	GTCCTTCCTTATGGCATTAGA	GGGCAGGTATTGTGTTAATCT	GTCCTTCCTTATGGCATTAGA	~6Kb	No product	-
AGCCCAGTGTGTCATATCTAC	TGCCCTTACTCGGTCA	AGCCCAGTGTGTCATATCTAC	CTGCCCTTACTCGGTC	~10Kb	4863	Off target PCR
AGGACTTTGGGAGCAGTAAGA	AGAGATGCGTCCTGCTAAAC	AGGACTTTGGGAGCAGTAAGA	GAGATGCGTCCTGCTAAACA	~2Kb	2155	48%
GAGGAAGTAAACCGGACCA	TTCAGTCAGATAACGGCACAT	AAACCGGACCAGCTACTTGTA	TCTGGGCTCTGTGTCTAAGG	~5Kb	6559	Off target PCR
TACACACCGGTGCATAACTGG	TTTCAAGAACACGGGATGG	ACACCGGTGCATAACTGGC	GAACACGGGATGGTCCTAGAA	~10Kb	5731	79%
GGCCTGCTGTAGTGATT	TAACAGCCTCTCAGAGACTAT	GGCCTGCTGTAGTGATT	CAGCCTCTCAGAGACTATCAG	~2Kb	1480	89%