Published in final edited form as:
Nat Genet. 2018 November ; 50(11): 1574–1583. doi:10.1038/s41588-018-0223-8.

Sixteen diverse laboratory mouse reference genomes define strain specific haplotypes and novel functional loci

A full list of authors and affiliations appears at the end of the article.

Abstract

We report full-length draft *De novo* genome assemblies for 16 widely used inbred mouse strains and reveal extensive strain-specific haplotype variation. We identify and characterise 2,567 regions on the current mouse reference genome exhibiting the greatest sequence diversity. These regions are enriched for genes involved in pathogen defence and immunity and exhibit enrichment of transposable elements and signatures of recent retrotransposition events. Combinations of alleles and genes unique to an individual strain are commonly observed at these loci, reflecting distinct strain phenotypes. We used these genomes to improve the mouse reference genome resulting in the completion of 10 new gene structures and 62 new coding loci were added to the reference genome annotation. Notably these genomes revealed a large previously unannotated gene (*Efcab3-like*) encoding 5,874 amino acids. *Efcab3-like* mutant mice display anomalies in multiple brain regions suggesting a role in the regulation of brain development.

Keywords

mouse; genome; *de novo* assembly; allele; subspecies

Introduction

Inbred laboratory strains are broadly organised into two groups; classical and wild-derived strains1, that can be used to model the variation observed in human populations2,3. Notably, inbred laboratory strains of wild-derived origin represent a rich source of phenotypic responses and genetic diversity not present in classical strains4–6, and have been crossed with classical strains to create powerful resources such as the Collaborative Cross (CC) and Diversity Outbred Cross (DO), in which genetic traits have been mapped7.

The generation and assembly of a reference genome for C57BL/6J accelerated the discovery of the genetic landscape underlying phenotypic variation11. Using this reference, genome-

[*] Joint corresponding authors. Correspondence to TMK (tk2@ebi.ac.uk).
[+] Joint first authors

Author contributions
Study design and supervision: TMK, DJA, BP, JF. Samples and sequencing: MQ, LS, NP, LRe, AC, MDu, AF-S. Genome assembly: JLi, TMK, LG, KH, SKP, SP, JT, JW, MK, WC, FY, KW, BF, JC. Genome annotation: BP, ITF, MA, JA, MDi, DDD, DE, AF, MG, JG, JH, SN, PM, FCPN, LRo, MS, CSi, CSt, GT, RB, JLo, MSH, DT, JLi. Genome analysis: JLi, AGD, TMK, BP, ITF, MA, PD, DWL, XIS, RD, PF, CEM, RM, DTO. Knock-out generation and phenotyping: SC, CL, MT, BY.

wide variation catalogues (SNPs, short indels, and structural variation) for thirty-six laboratory mouse strains were generated12,13,,. However, reliance on mapping next-generation sequencing reads to C57BL/6J has meant that the true extent of strain specific variation is unknown. At some loci, the genetic difference between the reference and sequenced strain genomes is comparable to that between human and chimpanzees, making it hard to distinguish whether a read is mis-mapped or highly divergent. *De novo* genome assembly methods address this issue by allowing unbiased assessments of the differences between genomes.

We have completed the first draft *de novo* assemblies and strain specific gene annotation for twelve classical inbred laboratory strains (129S1/SvImJ, A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6NJ, CBA/J, DBA/2J, FVB/NJ, LP/J, NZO/HlLtJ and NOD/ShiLtJ) and four wild-derived strains representing the *M. m. castaneus* (CAST/EiJ), *Mus musculus musculus* (PWK/PhJ), *Mus musculus domesticus* (WSB/EiJ), and *Mus spretus* (SPRET/EiJ) backgrounds. This collection comprises a large and diverse array of laboratory strains, including those closely related to commonly used mouse cell lines (BALB/3T3 and L929, derived from BALB/c and C3H related strains), embryonic stem cell derived gene knockouts (historically 129-related strains)14, mouse models of human disease (such as NOD-related nude mice)15, gene knockout background strains (C57BL/6NJ)16, the founders of commonly used recombinant inbred lines such as the AKXD, BXA, BXD, CXB and CC17, and outbred mapping populations such as the DO and the heterogeneous stock (HS)18.

Results

Sequence assemblies and genome annotation

Chromosome scale assemblies were produced for 16 laboratory mouse strains using a mixture of Illumina paired-end (40-70x), mate-pairs (3, 6, 10Kbp), fosmid and BAC-end sequences (Supplementary Table 1) and Dovetail Genomics Chicago™ libraries19. Pseudo-chromosomes were produced in parallel utilising cross-species synteny alignments resulting in genome assemblies of between 2.254 (WSB/EiJ) to 2.328 Gbps (AKR/J) excluding unknown gap bases. Approximately 0.5-2% of total genome length per strain was unplaced and is composed of unknown gap bases (18-49%) and repeat sequences (61-79%) (Supplementary Table 2), and contain between 89-410 predicted genes per strain (Supplementary Table 3). Mitochondrial genome (mtDNA) assemblies for 14 strains supported previously published sequences20, although a small number of high quality novel sequence variants in AKR/J, BALB/cJ, C3H/HeJ, and LP/J conflicted with GenBank entries (Supplementary Table 4). Novel mtDNA haplotypes were identified in PWK/PhJ and NZO/HlLtJ. Notably, NZO/HlLtJ contained 55 SNPs (33 shared with the wild-derived strains) and appears distinct compared to the other classical inbred strains (Supplementary Figure 1). Previous variation catalogues have indicated high concordance (>97% shared SNPs) between NZO/HlLtJ and another inbred laboratory strain NZB/BlNJ21.

We assessed the base accuracy of the strain chromosomes relative to two versions of the C57BL/6J reference genome (MGSCv311 and GRCm382) by first realigning all of the paired-end sequencing reads from each strain back to their respective genome assemblies then using these alignments to identify SNPs and indels. The combined SNP and indel error rate was between 0.09-0.1 errors per Kbp, compared to 0.334 for MGSCv3 and 0.02 for GRCm38 (Supplementary Table 5). Next, we used a set of 612 PCR primer pairs previously used to validate structural variant calls in eight strains22. The assemblies had between 4.7-6.7% primer pairs showing incorrect alignments compared to 10% for MGSCv3 (Supplementary Table 6). Finally, alignment of PacBio long read cDNA sequences from liver and spleen of C57BL/6J, CAST/EiJ, PWK/PhJ and SPRET/EiJ showed that the GRCm38 reference genome had the highest proportion of correctly aligned cDNA reads (99% and 98%, respectively) and the strains and MGSCv3 were 1-2% lower (Supplementary Table 7). The representation of known mouse repeat families in the assemblies shows that the short repeat (<200bp) content was comparable to GRCm38 (Supplementary Figure 2a +b). The total number of long repeat types (>200bp) is consistent across all strains, however the total sequence lengths are consistently shorter than GRCm38 (Supplementary Figure 2c).

Strain specific consensus gene sets were produced using the GENCODE C57BL/6J annotation and strain specific RNA-Seq from multiple tissues23 (Supplementary Table 8, Supplementary Figure 3). The consensus gene sets contain over 20,000 protein coding genes and over 18,000 non-coding genes (Figure 1a, Supplementary Table 1). For the classical laboratory strains 90.2% of coding transcripts (88.0% in wild-derived strains) and 91.2% of noncoding transcripts (91.4% in wild-derived strains) present in the GRCm38 reference gene-set were comparatively annotated. Gene predictions from strain specific RNA-Seq (Comparative Augustus24) added an average of 1,400 new isoforms to wild-derived and 1,207 new isoforms to classical strain gene annotation sets. Gene prediction based on PacBio cDNA sequencing introduced an average of 1,865 further new isoforms to CAST/EiJ, PWK/PhJ and SPRET/EiJ. Putative novel loci are defined as spliced genes that were predicted from strain specific RNA-Seq and did not overlap any genes projected from the reference genome. On average, 37 genes were putative novel loci (Supplementary Data 1) in wild-derived strains, and 22 in classical strains. Most often these appear to result from gene duplication events. Additionally, an automated pseudogene annotation workflow, Pseudopipe25, alongside manually curated pseudogenes lifted-over from the GRCm38 reference genome, identified an average of 11,000 (3,317 conserved between all strains) pseudogenes per strain (Supplementary Figure 4) that appear to have arisen either through retrotransposition (~80%) or gene duplication events (~20%).

Regions of the mouse genome with extreme allelic variation

Inbred laboratory mouse strains are characterized by at least twenty generations of inbreeding, and are genetically homozygous at almost all loci1. Despite this, previous SNP variation catalogs have identified high quality heterozygous SNPs (hSNPs) when reads were aligned to the C57BL/6J reference genome12. The presence of higher densities of hSNPs may indicate copy number changes, or novel genes that are not present in the reference assembly, forced to partially map to a single locus in the reference12,21. Thus, their identification is a powerful tool for finding errors in genome assemblies. We identified between 116,439 (C57BL/6NJ) and 1,895,741 (SPRET/EiJ) high quality hSNPs from the MGP variation catalogue v521 (Supplementary Table 9). Focusing our analysis on the top 5% most hSNP dense regions (windows >= 71 hSNPs per 10Kbp sliding window) identified the majority of known polymorphic regions among the strains (Supplementary Figure 5) and accounted for ~49% of all hSNPs (Supplementary Table 9, Supplementary Figure 6a). After applying this cut-off to all strain-specific hSNP regions and merging overlapping or adjacent windows, between 117 (C57BL/6NJ) and 2,567 (SPRET/EiJ) hSNP regions remained per strain (Supplementary Table 9), with an average size of 18-20Kbp (Supplementary Figure 6b). Many hSNP clusters overlap immunity (e.g. MHC, NOD-like receptors and AIM-like receptors), sensory (e.g. olfactory and taste receptors), reproductive (e.g pregnancy specific glycoproteins and Sperm-associated E-Rich proteins), neuronal and behavior related genes (e.g. Itch receptors26 and γ-protocadherins27) (Figure 1b, Supplementary Figure 5). All of the wild-derived strain hSNP regions contained gene and CDS base pair counts larger than any classical inbred strain (≥503 and ≥0.36Mbp, respectively; Supplementary Table 9). The regions identified in C57BL/6J and C57BL/6NJ (117 and 141, respectively; 145 combined) intersect known GRCm38 assembly issues including gaps, unplaced scaffolds or centromeric regions (107/145, 73.8%). The remaining candidate regions include large protein families (15/145, 10.3%) and repeat elements (17/145, 11.7%) (Supplementary Data 2).

We examined protein classes present in the hSNP regions by identifying 1,109 PantherDB matches, assigned to 26 protein classes, from a combined set of all genes in hSNP dense regions (Supplementary Data 3). Defence and immunity was the largest represented protein class (155 genes, Supplementary Data 4), accounting for 13.98% of all protein class hits (Supplementary Table 10). This was a five-fold enrichment compared to an estimated genome-wide rate (Figure 1d). Notably, 89 immune-related genes were identified in classical strains, and 84 of these were shared with at least one of the wild-derived strains (Figure 1d). SPRET/EiJ contributed the largest number of strain-specific gene hits (22 genes).

Many paralogous gene families were represented among the hSNP regions (Supplementary Data 3), including genes with functional human orthologs. Several prominent examples include *Apolipoprotein L* alleles; variants of which may confer resistance to *Trypanosoma brucei*, the primary cause of human sleeping sickness28,29, IFI16 (Interferon Gamma Inducible Protein 16, a member of AIM2-like receptors), a DNA sensor required for death of lymphoid CD4 T cells abortively infected with HIV30, NAIP (NLR family apoptosis inhibitory protein) in which functional copy number variation is linked to increased cell death upon *Legionella pneumophila* infection31, and secretoglobins (Scgb members) which may be involved in tumour formation and invasion, in both human and mouse32,33. Large gene families in which little functional information is known were also identified. A cluster of approximately 50 genes, which includes hippocalcin-like 1 (*Hpcal1*) and its homologues, were identified (Chr12:18-25Mbp). *Hpcal1* belongs to the neuronal calcium sensors, expressed primarily in retinal photoreceptors, neurons, and neuroendocrine cells34. This region is enriched for hSNPs in all strains except C57BL/6J and C57BL/6NJ. Interestingly, within this region, *Cpsf3* (21.29Mbp) is located on an island of high conservation in all strains, and a homozygous C57BL/6NJ knockout produces subviable offspring35. Additional examples include another region on chromosome 12 (87-88Mbp) containing approximately 20 eukaryotic translation initiation factor 1A (*eIF1a*) homologs, and on chromosome 14 (41-45Mbp) containing approximately 100 *Dlg1*-like genes. Genes within all hSNP candidate regions have been identified and annotated (Supplementary Figure 5).

We examined retrotransposon content in hSNP dense regions on GRCm38 compared to an estimated null distribution (1 million simulations) and found a significant enrichment of both LTRs (empirical $p<1\times10^{-7}$) and LINEs (empirical $p<1\times10^{-7}$) (Supplementary Tables 11,12). Gene retrotransposition has long been implicated in the creation of gene family diversity37, novel alleles conferring positively selected adaptations38. Once transposed, TEs accumulate mutations over time as the sequence diverges39,40. For LTRs, LINEs and SINEs, the mean percent sequence divergence was significantly lower ($p<1\times10^{-22}$) within hSNP regions compared to the rest of the genome (Figure 1e). The largest difference in mean sequence divergence was between LTRs within and outside of hSNP dense regions. Examining only repeat elements with less than 1% divergence, we found these regions are significantly enriched for LTRs (empirical $p<1\times10^{-7}$) and LINEs (empirical p=0.047).

De novo assembly of complex gene families

Our data elucidated copy number variation previously unknown in mouse strain genomes, and uncovered gene expansions, contractions and novel alleles (<80% sequence identity). For example, 23 distinct clusters of olfactory receptors (ORs) were identified indicating substantial variation among inbred strains. In mouse, phenotypic differences, particularly diet and behaviour, have been linked to distinct OR repertoires41,42. To this end, we have characterised the CAST/EiJ OR repertoire using our *de novo* assembly and identified 1,249 candidate OR genes (Supplementary Data 5). Relative to the reference strain (C57BL/6J), CAST/EiJ has lost 20 ORs and gained 37 gene family members; 12 novel and 25 supported by published predictions based on mRNA derived from CAST/EiJ whole olfactory mucosa (Figure 2a, Supplementary Table 13)43.

We discovered novel gene members at several important immune loci regulating innate and adaptive responses to infection. For example, chromosome 10 (22.1-22.4Mbp) on C57BL/6J contains *Raet1* alleles and minor histocompatibility antigen members of *H60*. Raet1/H60 are important ligands for NKG2D, an activating receptor of natural killer cells44. NKG2D ligands are expressed on the surface of infected45 and metastatic cells46 and may participate in allograft autoimmune responses47. From the *de novo* assembly, six different *Raet1/H60* haplotypes were identified among the eight Collaborative Cross (CC) founder strains; three of the haplotypes identified are shared among the classical inbred CC founders (A/J, 129S1/SvImJ and NOD/ShiLtJ have the same haplotype), and three different *Raet1/H60* haplotypes were identified in each of the wild-derived inbred strains (CAST/EiJ, PWK/PhJ and WSB/EiJ) (Figure 2b, Supplementary Figures 7+8). The CAST/EiJ haplotype encodes only a single *Raet1* family member (*Raet1e*) and no *H60* alleles, while the classical NOD/ShiLtJ haplotype has four *H60* and three *Raet1* alleles. The *Aspergillus*-resistant locus 4 (*Asprl4*), one of several QTLs that mediate resistance against *Aspergillus fumigatus* infection, overlaps this locus and comprises of a 1MB (~10% of QTL) interval which, compared to other classical strains, contains a haplotype unique to NZO/HlLtJ (Supplementary Figure 7). Strain specific haplotype associations with *Asprl4* and survival have been reported for CAST/EiJ and NZO/HlLtJ, both of which exhibit resistance to *A. fumigatus* infection48 and they are also the only strains to have lost *H60* alleles at this locus.

We examined three immunity related loci on chromosome 11, *IRG* (GRCm38: 48.85-49.10Mbp), *Nlrp1* (71.05-71.30Mbp) and *Slfn* (82.9-83.3Mbp), because of their polymorphic complexity and importance for mouse survival49–51. The *Nlrp1 locus* (NOD-like receptors, pyrin domain containing) encodes inflammasome components that sense endogenous microbial products and metabolic stresses, thereby stimulating innate immune responses52. In the house mouse, *Nlrp1* alleles are involved in sensing *Bacillus anthracis* lethal toxin, leading to inflammasome activation and pyroptosis of macrophages53,54. We discovered seven distinct *Nlrp1* family members by comparing six strains (CAST/EiJ, PWK/PhJ, WSB/EiJ, SPRET/EiJ, NOD/ShiLtJ, and C57BL/6J). Each strain has a unique haplotype of *Nlrp1* members, highlighting the extensive sequence diversity at this locus across inbred mouse strains (Figure 2c). Each of the three *M. m. domesticus* strains (C67BL/6J, NOD/ShiLtJ and WSB/EiJ) carry different combinations of *Nlrp1* family members; *Nlrp1d-1f* are novel strain-specific alleles that were previously unknown. Diversity between different *Nlrp1* alleles is higher than mouse/rat diversity. For example, C57BL/6J contains *Nlrp1c* which is not present in the other two strains, while *Nlrp1b2* is present in both NOD/ShiLtJ and WSB/EiJ but not C57BL/6J. In PWK/PhJ (*M. m. musculus*), the *Nlrp1* locus is almost double in size relative to the GRCm38 reference genome, and contains novel *Nlrp1* homologues (Figure 2c), whereas in *M. spretus* (also wild-derived), this locus is much shorter than any other mouse strain. Approximately 90% of intergenic regions in the PWK/PhJ assembly of the *Nlrp1* locus is composed of TEs (Figure 2d).

The wild-derived PWK/PhJ (*M. m. musculus*) and CAST/EiJ (*M. m. castaneus*) strains share highly similar haplotypes, however PWK/PhJ macrophages are resistant to pyroptotic cell death induced by anthrax lethal toxin while CAST/EiJ macrophages are not55. It has been suggested that *Nlrp1c* may be the causal family member mediating resistance; *Nlrp1c* can be amplified from cDNA from PWK/PhJ macrophages but not CAST/EiJ55. In the *de novo* assemblies, both mouse strains share the same promoter region for *Nlrp1c*; however, when transcribed, the cDNA of *Nlrp1c*_CAST could not be amplified with previously designed primers55 due to SNPs at the primer binding site (5'…CACT-3' → 5'…TACC-3'). The primer binding site in PWK/PhJ is the same as C57BL/6J, however *Nlrp1c* is a predicted pseudogene. We found an 18 amino acid mismatch in the NBD domain between Nlrp1b_CAST and Nlrp1b_PWK. These divergent profiles suggest that *Nlrp1c* is not the sole mediator of anthrax lethal toxin resistance in the mouse, and instead that several other members may also be involved. Newly annotated members *Nlrp1b2* and *Nlrp1d*, appear functionally intact in CAST/EiJ, but were both predicted as pseudogenes in PWK/PhJ due to the presence of stop codons or frameshift mutations. In C57BL/6J, three splicing isoforms of *Nlrp1b* (SV1, SV2, and SV3) were reported55. A dot-plot between PWK/PhJ and the C57BL/6J reference illustrates the disruption of co-linearity at the PWK/PhJ *Nlrp1b2* and *Nlrp1d* alleles (Figure 2d). All of the wild-derived strains we sequenced contain full-length *Nlrp1d* and exhibit a similar disruption of co-linearity at these alleles relative to C57BL/6J (Supplementary Data 6), such that the SV1 isoform in C57BL/6J is derived from truncated ancestral paralogs of *Nlrp1b* and *Nlrp1d*, indicating that *Nlrp1d* was lost in the C57BL/6J lineage. The genome structure of the *Nlrp1* locus in PWK/PhJ, CAST/EiJ, WSB/EiJ and NOD/ShiLtJ was confirmed using Fiber-FISH (Supplementary Figure 9).

The assemblies also revealed extensive diversity at each of the other loci examined; Immunity-related GTPases (*IRGs*) and Schlafen family (*Slfn*). IRG proteins belong to a subfamily of interferon-inducible GTPases present in most vertebrates56. In mouse, IRG protein family members contribute to the adaptive immune system by conferring resistance against intracellular pathogens such as *Chlamydia trachomatis*, *Trypanosoma cruzi* and *Toxoplasma gondii*57. Our *de novo* assembly is concordant with previously published data for CAST/EiJ49, and for the first time reveal the order, orientation, and structure of three highly divergent haplotypes present in WSB/EiJ, PWK/PhJ, and SPRET/EiJ including novel annotation of rearranged promoters, inserted processed pseudogenes, and a high frequency of LINE repeats (Supplementary Data 6).

The Schlafen (Chr11:82.9M-83.3M) family of genes are reportedly involved in immune responses, cell differentiation, proliferation and growth, cancer invasion and chemotherapy resistance. In humans, SLFN11 was reported to inhibit HIV protein synthesis by a codon-usage-based mechanism58, and in non-human primates positive selection on *Slfn11* has been reported59. In mouse, embryonic death may occur between strains carrying incompatible *Slfn* haplotypes60. Assembly of *Slfn* for the three collaborative cross founder strains of wild-derived origin (CAST/EiJ, PWK/PhJ and WSB/EiJ) revealed for the first time extensive variation at this locus. Members of group 4 *Slfn* genes51, *Slfn8*, *Slfn9* and *Slfn10*, show significant sequence diversity among these strains. For example, *Sfln8* is a predicted pseudogene in PWK/PhJ but protein coding in the other strains, and the CAST/EiJ allele contains 78aa mismatches compared to the C57BL/6J reference (Supplementary Figure 10). Both CAST/EiJ and PWK/PhJ contain functional copies of *Sfln10*, which is a predicted pseudogene in C57BL/6J and WSB/EiJ. A novel start codon upstream of *Slfn4*, which causes a 25aa N-terminal extension, was identified in PWK/PhJ and WSB/EiJ. Another member present in the reference, *Slfn14*, is conserved in PWK/PhJ and CAST/EiJ but a pseudogene in WSB/EiJ (Supplementary Figure 10).

Reference genome updates informed by the strain assemblies

There are currently eleven genes in the GRCm38 reference assembly (C57BL/6J) that are incomplete due to a gap in the sequence. First, these loci were compared to the respective regions in the C57BL/6NJ assembly and used to identify contigs from public assemblies of the reference strain, previously omitted due to insufficient overlap. Second, C57BL/6J reads aligned to the regions of interest in the C57BL/6NJ assembly were extracted for targeted assembly, leading to the generation of contigs covering sequence currently missing from the reference. Both approaches resulted in the completion of ten new gene structures (e.g. Supplementary Figure 11 and Supplementary Data 7), and the near-complete inclusion of the *Sts* gene that was previously missing.

Improvements to the reference genome, coupled with pan-strain gene predictions, were used to provide updates to the existing reference genome annotation, maintained by the GENCODE consortium61. We examined the strain specific RNA-Seq (Comparative Augustus) gene predictions containing 75% novel introns compared to the existing reference annotation (Table 1) (GENCODE M8, chromosomes 1-12). Of the 785 predictions investigated, 62 led to the annotation of new loci including 19 protein coding genes and 6 pseudogenes (Supplementary Table 14, Supplementary Data 8). In most cases where a new locus was predicted on the reference genome, we identified pre-existing, but often incomplete, annotation. For example, the *Nmur1* gene was extended at its 5' end and made complete on the basis of evidence supporting a prediction which spliced to an upstream exon containing the previously missing start codon. The *Mroh3* gene, which was originally annotated as an unprocessed pseudogene, was updated to a protein coding gene due to the identification of a novel intron that permitted extension of the CDS to full-length. The previously annotated pseudogene model has been retained as a nonsense-mediated decay (NMD) transcript of the protein coding locus. At the novel bicistronic locus, *Chml_Opn3*, the original annotation was a single exon gene, *Chml*, that was extended and found to share its first exon with the *Opn3* gene.

We discovered a novel 188-exon gene on chr11 that significantly extends the existing gene *Efcab3* spanning between *Itgb3* and *Mettl2,* Figure 3a). This *Efcab3-like* gene was manually curated, validated according to HAVANA guidelines62, and identified in GENCODE releases M11 onwards as *Gm11639*. *Efcab3/Efcab13* are calcium-binding proteins and the new gene primarily consists of repeated EF-hand protein domains (Supplementary Figure 12). Analysis of synteny and genome structure revealed that the *Efcab3* locus is largely conserved across other mammals including most primates. Comparative gene prediction identified the full length version in orangutan, rhesus macaque, bushbaby and squirrel monkey. However, the locus contains a breakpoint at the common ancestor of chimpanzee, gorilla and human (*Homininae)* due to a ~15Mbp intra-chromosomal rearrangement that also deleted many of the internal EF-hand domain repeats (Figure 3b, Supplementary Figure 13). Analysis of GTEx expression data63 in human revealed that the *EFCAB13* locus is expressed across many tissue types, with the highest expression measured in testis and thyroid. In contrast, the *EFCAB3* locus only has low level measurable expression in testis. This is consistent with the promoter of the full length gene being present upstream of the *EFCAB13* version, which is supported by H3K4Me3 analysis (Supplementary Figure 14). In mice, *Efcab3* is specifically expressed during development throughout many tissues with high expression in the upper layers of the cortical plate (see URLs), and is located in the immediate vicinity of the genomic 17q21.31 syntenic region linked to brain structural changes both in mice and humans64. We used CRISPR to create *Efcab3*-like mutant mice (*Efcab3*$^{em1(IMPC)Wtsi}$; see methods) and recorded 188 primary phenotyping measures (Supplementary Data 9). We also measured 40 brain parameters across 22 distinct brain structures (Supplementary Table 15+16) (see methods). Notably, brain size anomalies were identified in *Efcab3-like* mutant mice when compared to matched wild-type controls (Figure 3c). Interestingly, the lateral ventricle was one the most severely affected brain structures exhibiting an enlargement of 65% (P=0.007). The pontine nuclei were also increased in size by 42% (P=0.001) and the cerebellum by 27% (P=0.02), these are two regions are involved in motor activity (Figure 3d, Supplementary Figure 15). The thalamus was also larger by

URLs

http://www.genepaint.org
https://github.com/ucsc-mus-strain-cactus/MouseGenomesAnnotationPipeline
https://github.com/jts/sga
http://snap.cs.berkeley.edu
https://imagej.nih.gov/ij/

19% (P=0.007). As a result, the total brain area parameter was enlarged by 7% (P=0.006). Taken together, these results suggest a mechanism of *Efcab3-like* to regulate brain development and brain size regulation from the forebrain to the hindbrain.

Discussion

The completion of the mouse reference genome, based on the classical inbred strain C57BL/6J, generated a transformative resource for human and mouse genetics. Here we generate the first chromosome-scale genome assemblies for 12 classical and 4 wild-derived inbred strains, thus revealing at unprecedented resolution the striking strain-specific allelic diversity that encompasses 0.5-2.8% (14.4-75.5 Mbp, excluding C57BL/6NJ) of the mouse genome. Accessing shared and distinct genetic information across the *Mus* lineage in parallel during assembly and gene prediction leads to the placement of novel alleles, the accurate annotation of many strain-specific gene family haplotypes, and the detection of genes lowly expressed but partially supported in all strains (Figure 3a).

Genetic diversity at gene loci, particularly those related to defence and immunity, is often the result of selection that if retained, can lead to the rise of divergent alleles in a population65. We used the presence of dense clusters of heterozygous SNPs on the C57BL/6J reference genome as a marker for extreme polymorphism and examined the *de novo* assembly to explore the underlying genomic architecture. Examining the heterozygous SNPs in C57BL/6J and C57BL/6NJ (see results), we find that the vast majority can be explained as occurring in remaining gaps or problematic regions of the reference genome. However, we are left with 6 loci (57 Kbp) enriched for hSNPs in C57BL/6J and C57BL/6NJ that do not have an obvious explanation and could be attributed to residual heterozygosity. Across all strains, hSNP regions account for between 1.5-5.5% of protein coding genes (Figure 1c) and are overrepresented with genes associated with immunity, sensory, sexual reproduction, and behavioural phenotypes (Figure 1d). Genes related to immunological processes, particularly gene families involved in mediating innate immune responses (e.g. *Raet1, Nlrp1*), exhibit great diversity among the strains reflecting strain-specific disease associations, responses and susceptibility. Interestingly, regions of strain haplotype diversity appear enriched for recent LINEs and LTR repeat elements (Figure 1e). We observed several innate immunity gene families in mice with a high density of retrotransposons, which is the likely mechanism for diversification at these loci (e.g. *Nlrp1*, Figure 2d).

The challenge of generating multiple closely related mammalian genomes and annotation required new approaches to whole-genome alignment66, comparative creation of whole-chromosome scaffolds67, and comparative approaches to simultaneous genome annotation within a clade23,24. *Mus* is the first mammalian lineage to have multiple chromosome scale genomes. Simultaneous access to many rodent species assemblies in parallel with individual level gene predictions, expression and long read data facilitated the accurate prediction of many strain specific haplotypes and gene isoforms. This approach identified previously unannotated genes, including *Efcab3-like*, one of the largest known mouse genes (5874 amino acids) which also appears conserved in mammals. Interestingly, the previously unannotated *Efcab3-like* gene is very close to the 17q21.31 syntenic region associated in humans to the Koolen-de Vries microdeletion syndrome (KdVS). Both mouse deletion models of this syntenic interval68, containing four genes (*Crhr1*, *Spplc2*, *Mapt* and *Kansl1*; Figure 3a) and an *Efcab3-like* knockout showed analogous brain phenotypes, suggesting common cis-acting regulatory mechanisms as shown previously in the context of the 16p11.2 microdeletion syndrome69. *Efacb3-like* is conserved in orangutan but reversed in gorilla, and appears to have split into two separate protein coding genes, *EFCAB3* and *EFCAB13*, in the *Homininae* lineage. Many novel genes and transcripts were identified across all of the strains, highlighting unexplored sequence variation across the *Mus* lineage. The addition of these genomes, in particular C57BL/6NJ, enabled the resolution of GRCm38 reference assembly issues, and the improvement of several reference gene annotations. The assembly and alignment of a variety of haplotypes at loci heterogenous amongst the laboratory strains allows for analysis of regions previously not placed in the reference assembly. These regions are often of variable copy number between various haplotypes[91]. In particular, the wild-derived strains represent a rich resource of novel target sites, resistance alleles, genes and isoforms not present in the reference strain, or indeed many classical strains. For the first time the underlying sequence at these loci is represented in strain-specific assemblies and gene predictions from across the inbred mouse lineage, which should facilitate increased dissection of complex traits.

Online Methods

Sequencing

All DNA was obtained from the Jackson Laboratories from female mice (Supplementary Table 17). For the paired-end libraries, 1-3 μg DNA was sheared to 100 to 1,000 bp using a Covaris E210 or LE220 (Covaris, Woburn, MA, USA) and size selected (350 to 450 bp) using magnetic beads (Ampure XP; Beckman Coulter). Sheared DNA was subjected to Illumina paired-end DNA library preparation and PCR-amplified for six cycles. Amplified libraries were sequenced using the Illumina HiSeq platform as paired-end 100 base reads according to the manufacturer's protocol. Illumina sequencing compatible Mate Pair libraries were created at 3 and 6 Kb according to the Sanger method70. The 10 Kb Illumina Nextera libraries were prepared according to the manufacturer’s instructions (Illumina Nextera Sample Preparation guide) with the addition of a size selection step on the BluePippin (Sage Science; Beverly, MA, USA).

For CAST/EiJ, PWK/PhJ, and SPRET/EiJ, a Chicago™ library was prepared as described previously19. Briefly, for each library, 500 ng of high molecular weight genomic DNA (>50Kb mean fragment size) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was then digested with MboI, the 5’ overhangs were filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, cross-links were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was sheared to ~350 bp mean fragment size, and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were then isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq to produce 2x 125 bp read pairs. The number of read pairs produced and fold physical coverage (1-50kb pairs) for each genome was: 374 Million, 34x for PWK/PhJ;

373 Million, 41x for SPRET/EiJ; and 380 Million, 77x for CAST/EiJ. Every sequencing lane was genotype checked against the mouse Hapmap SNP calls71 using the Samtools/ Bcftools v1.1 'gtcheck' command.

De novo assembly

The initial contigs were assembled from the paired-end sequencing reads using SGA v0.9.43 (see URLs)72. Parameters for assembly is listed in Supplementary Table 18.

All of the mate-pair reads were aligned to GRCm38 with bwa mem v0.7.5, and duplicate fragments were removed with GATK MarkDuplicates v3.4. The subsequent reads were used as input to SOAPdenovo273 r240 to produce genome scaffolds (parameters given in Supplementary Table 19). To detect potential scaffold mis-joins, we realigned the mate-pair library reads onto the SOAP2 scaffolds with bwa-mem v0.7.5 and walked along each scaffold (greater than 10kb in size) in 5kb step intervals, and counted the number of 10kb and 40kb (where available) spanning fragments at each interval. Scaffolds were broken in locations where there was not a minimum number of 10kb and 40kb (where available) fragments that spanned the join. Scaffold break parameters are shown in Supplementary Table 20.

For CAST/EiJ, PWK/PhJ, and SPRET/EiJ, we further scaffolded the assemblies with Dovetail Genomics long range libraries. Each input genome assembly, along with its associated Chicago library read pairs in FASTQ format, were used as input data for HiRise, a software pipeline designed specifically to scaffold genomes using Chicago library data2. Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (see URLs). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify putative mis-joins and score prospective joins. After scaffolding, shotgun sequences were used to close gaps between contigs.

Pseudo-chromosomes

The scaffolds were assembled into chromosome-scale scaffolds using Ragout v2.0. Ragout identifies large conserved regions between genomes (hierarchical synteny blocks) by combining the whole genome alignment, with a de-Bruijn graph simplification algorithm67. Assembly scaffolds are further joined into chromosomes so as to minimize the number of structural differences (such as inversions or chromosomal translocations) between references and a target genome. We used the C57BL/6J GRCm38 sequence as a single reference and found that 95% adjacent synteny block pairs from the assemblies, were also adjacent in C57BL/6J reference.

Each of the genomes were assembled into a complete set of chromosomes with less than 5% of unlocalized sequence (Supplementary Data 10). On average, 10% of synteny block adjacencies in the assembled genomes were not presented in C57BL/6J reference. Ragout classified 38% of them as valid rearrangements and the rest as mis-assemblies (which were removed).

Gene prediction and annotation

Three techniques were used to produce the gene annotation for each mouse strain. First, whole genome alignments produced by progressiveCactus66 were used as input to transMap, producing an initial set of orthologs. These initial orthologs, along with strain specific RNA-Seq (Supplementary Table 8), were input to AUGUSTUS74 one at a time to apply local strain specific refinement. A consensus finding algorithm was employed to decide between possible versions of an orthologous transcript. We also created a *de novo* set of strain-specific genes and isoforms from Comparative Augustus (AugustusCGP)24 using the strain specific RNA-Seq and the progressiveCactus alignment. A subsequent round of consensus finding was employed to incorporate these transcripts into the final consensus annotation set.

The progressiveCactus whole genome alignments were used to project annotations from GENCODE VM861 onto each of the strain specific assemblies using transMap75. These transMapped transcript alignments were evaluated by a series of binary classifiers that attempt to diagnose differences between the parent and target genome. These classifiers include evaluating if a transcript maps multiple times, the proportion of unknown bases, splice site validity, both frameshifting and non-frameshifting indels, and small alignment gaps. These comparative transcripts were given to the gene-finding tool AUGUSTUS13 as strong *hints* (external evidence) in conjunction with weaker hints derived from all available RNA-seq data for the given strain. The RNA-Seq hints were generated for each of the novel strains by aligning RNA-Seq reads to the native genome with the spliced aligner STAR16. The resulting read alignments were quality filtered by coverage (>=80%), identity (>=90%) and uniqueness, i.e. when a read mapped to multiple loci, the best alignment for that read was only kept, if the alignment score of the second best was considerably worse. For the remaining reads (approximately 70%), strain-specific *exonpart* and *intron* hints were generated. The transcripts resulting from both transMap as well as AUGUSTUS were evaluated by a consensus finding algorithm that attempts to use a combination of fidelity to the reference and a series of binary classifiers to construct a consensus gene set. See the pipeline documentation for details on this process (see $^{\text{URLs}}$).

For each transMapped transcript alignment *t*, one way to identify its structure was a pipeline component we here refer to as AugustusTMR (TM=transMap, R=RNA-Seq). The aim was to try to produce all splice forms from the reference (parent) genome that likely also exist in the target genome. In the genomic region around *t*, AUGUSTUS was set to predict a gene structure without alternative splicing using evidence from *t* itself as well as from all RNA-Seq alignments in that region. Thereby, the evidence from *t* on the location of exons, introns and start and stop codon was given a much higher weight in order to produce the original splice form, also in cases where the majority of target RNA-Seq suggests a different major splice form. However, when part of a transcript structure was unclear, e.g. an unalignable transcript part, RNA-Seq evidence could help fill in missing parts.

By design, AugustusTMR restricts gene finding to regions that align to a reference gene, and thus is not able to predict genes missing in the reference annotation or genes in unaligned regions. To find novel splice forms and genes, Augustus is run in CGP (comparative gene prediction) mode, a recent extension24 that takes a whole-genome alignment of related species or strains and simultaneously predicts coding genes in all input genomes. In *AugustusCGP* the same types of evidence can be incorporated for either a subset or all species/strains. With the genome alignment, evidence is transferred across genomes. This makes it possible to exploit the combined evidence for gene finding and to discover genes, that for example, are only weakly expressed and partially supported in the reference strain, but that have a high expression in other strains. In this application, two different types of evidence are used 1.) the RNA-Seq hints for each of the novel strains from above and 2.) annotation evidence from GENCODE VM8 for the C57BL/6J reference strain. For the latter, coding sequence (CDS) and *intron* hints were generated from the GENCODE VM8 protein-coding gene set for the reference strain.

The resulting AugustusCGP gene sets were quality filtered based on how well the exon-intron structure of a transcript was supported by the combined RNA-Seq evidence (>=80% of the introns with splice junction support and >=80% of CDS exons with a read coverage of at least 10 reads per kilobase of mRNA). One of the challenges of gene finders is to distinguish coding genes from pseudogenes and expressed non-coding genes that contain partial open reading frames. All AugustusCGP transcripts that partially aligned to a reference transcript annotated as pseudogene or non-coding gene were also discarded.

The AugustusCGP transcripts were incorporated into the consensus gene set through a subsequent round of consensus finding. Based on coordinate intersections, each transcript was assigned a putative parent gene, if possible. If multiple assignments were created, they were attempted to be resolved by finding if any gene had a Jaccard distance 0.2 greater than any other; otherwise, they were discarded. After parent assignment, they were aligned with BLAT to each coding transcript associated with the parent gene. For each AugustusCGP transcript, if it had a better match to the CDS of any of the assigned transcripts than the current consensus transcript, the latter was replaced. If the AugustusCGP transcript introduced new intron junctions supported by RNAseq, then it was incorporated as a new isoform of that gene. Finally, if the AugustusCGP transcript was not assigned to any gene, it was incorporated as a putative novel gene. This process allows for the rescue of genes lost in the first round of filtering and consensus finding, as well as the discovery of polymorphic pseudogenes in the laboratory mouse lineage.

For the strains with AugustusPB transcripts, they were combined with the AugustusCGP transcripts and placed through the same consensus finding process described above. AugustusPB transcripts that could not be confidently assigned to parent transcripts were discarded and not evaluated for novel contribution.

The consensus gene sets were subsetted into a basic gene set following the methodology used by GENCODE61. Briefly, coding transcripts were retained if they were marked as having complete end information. If no complete transcripts are present, one longest CDS is picked for the gene. For non-coding transcripts, the fewest number of transcripts to keep at least 80% of present non-coding splice junctions were retained.

Sliding window analysis

Only coordinates in which at least one strain had a hSNP call were retained. These coordinates were then used to estimate the combined density of hSNPs using a 10kb sliding window (step of 2kb) across the mouse reference genome. Windows were grouped according to number of hSNPs they contained. The windows were then ordered by density of SNP (lowest, 1 hSNP per 10Kbp window, to highest). The top 5% of hSNP dense windows was identified and a shared density cut-off per 10Kbp window calculated (equivalent to 71 hSNPs / 10Kbp window). This represented the density at which the interval content and total unique overlapping bps was observed to be clustered around distinct loci (Supplementary Figure 6a).

Strain-specific analyses

For each strain separately, the density of hSNPs in 10kb sliding windows (step of 2kb) was estimated. Only windows with greater than or equal to the shared density cut-off per window were retained. These windows were then intersected with GENCODE M8 gene annotations; the total number of unique genes and base pair positions overlapping pass windows for each strain was calculated (Figure 1c). For each strain separately, coding genes from GENCODE M815 overlapping pass heterozygote dense windows were identified. Gene-sets for each strain were then combined, and, using PantherDB76, were classified based on protein class annotations (Figure 1d-left). To establish an expected rate for each protein class, the same analysis was carried out using the entire protein coding CDS annotated gene set from GENCODE M8. Strain specific gene sets (Supplementary Data 3) and PantherDB classifications are contained in Supplementary Table 10. Genes involved in defense and immunity (the largest protein class represented by the combined gene-set) were then retrieved and, the strains that contributed genes to this protein class identified. Strain specific defense genes are listed in Supplementary Data 4. To identify defense genes from the analysis shared among classical inbred strains, and each of the wild-derived strains, each of strain-specific gene sets were merged into five categories, namely classical inbred (BALB/cJ, CBA/J, DBA/2J, C3H/HeJ, 129S1/SvImJ, A/J, C57BL/6NJ, NOD/ShiLtJ, LP/J, NZO/HlLtJ, FVB/NJ, and AKR/J), PWK/PhJ, CAST/EiJ, WSB/EiJ and SPRET/EiJ (Figure 1d-right).

Generation of *Efcab3-like* knockout mice

All mice were maintained in a specific pathogen free facility with sentinel monitoring at standard temperature (19-23°C) and humidity (55% ±10%), on a 12h dark, 12h light cycle (lights on 0730–1900) and fed a standard rodent chow (LabDiet 5021-3, 9% crude fat content, 21% kcal as fat, 0.276ppm cholesterol, LabDiet, London, UK). Both food and water were available *ad libitum*. The mice were housed for phenotyping in groups of 3-4 mice per cage in either blue line (Tecniplast Seal Safe 1285L: overall dimensions of caging: (LxWxH) 365x207x140mm, floor area 530cm^2) or green line (Tecniplast GM500: overall dimensions of caged: (LxWxH) 391x199x160mm, floor area 501cm^2) individually ventilated caging receiving 60 air changes per hour. In addition to Aspen bedding substrate, standard environmental enrichment of a nestlet and a cardboard tunnel were provided. All animals were regularly monitored for health and welfare concerns and were additionally checked prior to and after procedures. The *Efcab3-like* gene has previously been represented by two loci MGI:3651790 and MGI:1918144, corresponding to the 5' and 3' regions respectively. Both loci have been targeted using a conditional approach as part of the IKMC resource. The *Efcab3-like* gene was targeted using CRISPR/Cas9 methodology as defined in77. Briefly, the constitutive coding exon 5 (chr11:104700610-104700692, GRCm38) which is well-supported by RNAseq data in multiple tissues (ENSMUST00000212287; ENSMUSE00000376310 [ENSEMBL v90]) was deleted using the SpCas9 endonuclease to induce a frameshift mutation. Pairs of flanking gRNAs were designed using the WTSI Genome Editing (WGE) tool78 creating four gRNAs (two gRNAs 5' and two gRNAs 3' to the CE region, Supplementary Table 21). Cas9 mRNA (Trilink, San Diego, CA) together with the four gRNAs was injected into the cytoplasm of 1-cell C57BL/6NTac zygotes. Injected embryos were briefly cultured and oviductal embryo transfer performed in 0.5 days post-coital pseudopregnant female recipients (CBA/C57BL/6J). F0 mice were screened for the exon deletion by a combination of end-point PCR and loss of WT allele qPCR. Positive F0 mice were further bred with C57BL/6NTac mice. F1 mice were re-screened by PCR and breakpoints confirmed by Sanger sequencing (Supplementary data 11). A single genotype-confirmed F1 mouse was used to establish the colony ($Efcab3^{em1(IMPC)Wtsi}$) used to generate mice for phenotyping. The care and use of mice in the study was carried out in accordance with UK Home Office regulations, UK Animals (Scientific Procedures) Act of 1986 under a UK Home Office license that approved this work, which was reviewed regularly by the WTSI Animal Welfare and Ethical Review Body.

Neuroanatomical studies of Efcab3-like knockout

Neuroanatomical studies were performed blind with experimenters not knowing the genotype of the mouse, on three 16-week-old matched control male mice in C57BL/6N background and three 16-week-old homozygous knockout of Efcab3. Standard operating procedures are described in more details elsewhere[79]. Mouse brain samples were immersion-fixed in 10% buffered formalin for 48 hours, before paraffin embedding and sectioning at 5µm thickness using a sliding microtome (Leica RM 2145). One precise sagittal section was stereostatically defined as the plane Lateral +0.72 mm of the Mouse Brain Atlas. Brain sections were double-stained using luxol fast blue for myelin and cresyl violet for neurons, and scanned at cell-level resolution using the Nanozoomer whole-slide scanner (Hamamatsu Photonics, Shizuoka, Japan). Using in-house ImageJ (see [URLs]) plugins, co-variates, for example sample processing dates and usernames were collected at every step of the procedure, as well as 40 brain morphological parameters of 25 area and 14 length measurements, and the number of cerebellar folia (Supplementary Table 15). This resulted in the quantification of 22 unique brain structures, including 1) the total brain area; 2) the primary and secondary motor cortices; 3) the pons; 4) the cerebellar area, the internal granular layer of the cerebellum and the medial cerebellar nucleus; 5) the lateral ventricle; 6) the corpus callosum; 7) the thalamus; 8) the caudate putamen; 9) the hippocampus and its associated features; 10) the fimbria of the hippocampus; 11) the anterior commissure; 12) the stria medullaris; 13) the fornix; 14) the optic chiasm; 15) the hypothalamus; 16) the pontine nuclei; 17) the substantia nigra; 18) the fibers of the pons; 19) the cingulate cortex; 20) the dorsal subiculum; 21) the inferior colliculus; and 22) the superior colliculus. All samples were also systematically assessed for cellular ectopia (misplaced neurons).

Neuroanatomical data (Supplementary Table 16) were analyzed using a Student two-tailed equal variance test.

Further details of methods are given in the Supplementary Note.

Data availability Statement

The genome sequencing reads are available from the European Nucleotide Archive and the assemblies are part of NCBI BioProject PRJNA310854 (Supplementary Table 22). The genome assemblies and annotation are available via the Ensembl genome browser, and the UCSC genome browser. Sequence accessions for the three immune related loci on Chr11 are available from the European Nucleotide Archive (Supplementary Table 23).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Jingtao Lilue[1,2,+], Anthony G. Doran[1,2,+], Ian T. Fiddes[3,+], Monica Abrudan[2], Joel Armstrong[3], Ruth Bennett[1], William Chow[2], Joanna Collins[2], Stephan Collins[4,5], Anne Czechanski[6], Petr Danecek[2], Mark Diekhans[3], Dirk-Dominik Dolle[2], Matt Dunn[2], Richard Durbin[2,7], Dent Earl[3], Anne Ferguson-Smith[7], Paul Flicek[1,2], Jonathan Flint[8], Adam Frankish[1,2], Beiyuan Fu[2], Mark Gerstein[9], James Gilbert[2], Leo Goodstadt[10], Jennifer Harrow[2], Kerstin Howe[2], Ximena Ibarra-Soria[2], Mikhail Kolmogorov[11], Chris Lelliott[2], Darren W. Logan[2], Jane Loveland[1,2], Clayton E. Mathews[13], Richard Mott[14], Paul Muir[9], Stefanie Nachtweide[12], Fabio C.P. Navarro[9], Duncan T. Odom[15,19], Naomi Park[2], Sarah Pelan[2], Son K Pham[16], Mike Quail[2], Laura Reinholdt[6], Lars Romoth[12], Lesley Shirley[2], Cristina Sisu[9,20], Marcela Sjoberg-Herrera[17], Mario Stanke[12], Charles Steward[2], Mark Thomas[2], Glen Threadgold[2], David Thybert[18,2], James Torrance[2], Kim Wong[2], Jonathan Wood[2], Binnaz Yalcin[4], Fengtang Yang[2], David J. Adams[2,*], Benedict Paten[3,*], and Thomas M. Keane[1,2,21,*]

Affiliations

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

[2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK

[3]Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

[4]Institut de Génétique et de Biologie Moléculaire et Cellulaire, Centre National de la Recherche Scientifique UMR7104, Institut National de la Santé et de la Recherche Médicale U964, Université de Strasbourg, 67404 Illkirch, France

[5]Centre des Sciences du Goût et de l'Alimentation, University of Bourgogne Franche-Comté, 21000 Dijon, France

[6]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

[7]Department of Genetics, University of Cambridge, Downing Site, Cambridge CB2 3EH, UK

[8]Brain Research Institute, University of California, 695 Charles E Young Dr S, Los Angeles, CA 90095, USA

[9]Yale Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

[10]OxFORD Asset Management, OxAM House, 6 George Street, Oxford OX1 2BW

[11]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA

[12]Institute of Mathematics and Computer Science, University of Greifswald, Domstraße 11, 17489 Greifswald, Germany

[13]Department of Pathology, Immunology, and Laboratory Medicine, University of Florida, Gainesville, FL, USA

[14]Genetics Institute, University College London, Gower Street, London WC1E 6BT, UK

[15]Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge, CB2 0RE, UK

[16]BioTuring Inc., San Diego, California, CA92121

[17]Departamento de Biología Celular y Molecular, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago 8331150, Chile

[18]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK

[19]German Cancer Research Center (DKFZ), Division Signaling and Functional Genomics, 69120 Heidelberg, Germany

[20]Department of Bioscience, Brunel University London, Uxbridge UB8 3PH, UK

[21]School of Life Sciences, University of Nottingham, Nottingham, UK

Acknowledgements

This work was supported by the Medical Research Council [MR/L007428/1], BBSRC [BB/M000281/1] and the Wellcome Trust. DJA is supported by Cancer Research-UK and the Wellcome Trust. MKS is supported by a research grant from FONDECYT No.1171004 and the European Commission (EUFP7 BLUEPRINT grant HEALTH-F5-2011-282510). DTO work was supported by Cancer Research UK (20412), the Wellcome Trust (202878/A/16/Z), the European Research Council (615584). PF was supported by the Wellcome Trust (grant numbers WT108749/Z/15/Z, WT098051, WT202878/B/16/Z), the National Human Genome Research Institute (U41HG007234) and the European Molecular Biology Laboratory. The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007–2013) under grant agreement HEALTH-F4-2010-241504 (EURATRANS). CEM is supported by R01 DK074656. We thank members of the Sanger Institute Mouse Pipelines teams (Mouse Informatics, Molecular Technologies, Genome Engineering Technologies, Mouse Production Team, Mouse Phenotyping) and the Research Support Facility for the provision and management of the mice. We thank Valerie Vancollie for assistance with phenotyping data.

References

1. Beck JA, et al. Genealogies of mouse inbred strains. Nat Genet. 2000; 24:23–25. [PubMed: 10615122]
2. Church DM, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 2009; 7:e1000112. [PubMed: 19468303]
3. Svenson KL, et al. Multiple trait measurements in 43 inbred mouse strains capture the phenotypic diversity characteristic of human populations. J Appl Physiol Bethesda Md 1985. 2007; 102:2369–2378.
4. Americo JL, Moss B, Earl PL. Identification of wild-derived inbred mouse strains highly susceptible to monkeypox virus infection for use as small animal models. J Virol. 2010; 84:8172–8180. [PubMed: 20519404]
5. Ideraabdullah FY, et al. Genetic and haplotype diversity among wild-derived mouse inbred strains. Genome Res. 2004; 14:1880–1887. [PubMed: 15466288]
6. Churchill GA, et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet. 2004; 36:1133–1137. [PubMed: 15514660]
7. French JE, et al. Diversity Outbred Mice Identify Population-Based Exposure Thresholds and Genetic Factors that Influence Benzene-Induced Genotoxicity. Environ Health Perspect. 2015; 123:237–245. [PubMed: 25376053]
8. Ferris MT, et al. Modeling host genetic regulation of influenza pathogenesis in the collaborative cross. PLoS Pathog. 2013; 9:e1003196. [PubMed: 23468633]
9. Rasmussen AL, et al. Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance. Science. 2014; 346:987–991. [PubMed: 25359852]
10. Kelada SNP, et al. Integrative genetic analysis of allergic inflammation in the murine lung. Am J Respir Cell Mol Biol. 2014; 51:436–445. [PubMed: 24693920]
11. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420:520–562. [PubMed: 12466850]
12. Keane TM, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011; 477:289–294. [PubMed: 21921910]
13. Yalcin B, et al. Sequence-based characterization of structural variation in the mouse genome. Nature. 2011; 477:326–329. [PubMed: 21921916]
14. Simpson EM, et al. Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. Nat Genet. 1997; 16:19–27. [PubMed: 9140391]
15. Shultz LD, Ishikawa F, Greiner DL. Humanized mice in translational biomedical research. Nat Rev Immunol. 2007; 7:118–130. [PubMed: 17259968]
16. Skarnes WC, et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature. 2011; 474:337–342. [PubMed: 21677750]
17. Flint J, Mott R. Applying mouse complex-trait resources to behavioural genetics. Nature. 2008; 456:724–727. [PubMed: 19079048]
18. Churchill GA, Gatti DM, Munger SC, Svenson KL. The Diversity Outbred mouse population. Mamm Genome Off J Int Mamm Genome Soc. 2012; 23:713–718.
19. Putnam NH, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016; 26:342–350. [PubMed: 26848124]
20. Goios A, Pereira L, Bogue M, Macaulay V, Amorim A. mtDNA phylogeny and evolution of laboratory mouse strains. Genome Res. 2007; 17:293–298. [PubMed: 17284675]
21. Doran AG, 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. 2016; 17:167. [PubMed: 27480531]
22. Yalcin B, et al. The fine-scale architecture of structural variants in 17 mouse genomes. Genome Biol. 2012; 13:R18. [PubMed: 22439878]
23. Fiddes IT, et al. Comparative Annotation Toolkit (CAT) - simultaneous clade and personal genome annotation. 2017; doi: 10.1101/231118

24. König S, Romoth LW, Gerischer L, Stanke M. Simultaneous gene finding in multiple genomes. Bioinforma Oxf Engl. 2016; 32:3388–3395.

25. Zhang Z, et al. PseudoPipe: an automated pseudogene identification pipeline. Bioinforma Oxf Engl. 2006; 22:1437–1439.

26. Liu Q, et al. Sensory neuron-specific GPCR Mrgprs are itch receptors mediating chloroquine-induced pruritus. Cell. 2009; 139:1353–1365. [PubMed: 20004959]

27. Weiner JA, Wang X, Tapia JC, Sanes JR. Gamma protocadherins are required for synaptic development in the spinal cord. Proc Natl Acad Sci U S A. 2005; 102:8–14. [PubMed: 15574493]

28. Dummer PD, et al. APOL1 Kidney Disease Risk Variants: An Evolving Landscape. Semin Nephrol. 2015; 35:222–236. [PubMed: 26215860]

29. Capewell P, Cooper A, Clucas C, Weir W, Macleod A. A co-evolutionary arms race: trypanosomes shaping the human genome, humans shaping the trypanosome genome. Parasitology. 2015; 142(Suppl 1):S108–119. [PubMed: 25656360]

30. Monroe KM, et al. IFI16 DNA sensor is required for death of lymphoid CD4 T cells abortively infected with HIV. Science. 2014; 343:428–432. [PubMed: 24356113]

31. Boniotto M, et al. Population variation in NAIP functional copy number confers increased cell death upon Legionella pneumophila infection. Hum Immunol. 2012; 73:196–200. [PubMed: 22067212]

32. Patierno SR, et al. Uteroglobin: a potential novel tumor suppressor and molecular therapeutic for prostate cancer. Clin Prostate Cancer. 2002; 1:118–124. [PubMed: 15046703]

33. Cai Y, et al. Preclinical evaluation of human secretoglobin 3A2 in mouse models of lung development and fibrosis. Am J Physiol Lung Cell Mol Physiol. 2014; 306:L10–22. [PubMed: 24213919]

34. Braunewell KH, Gundelfinger ED. Intracellular neuronal calcium sensor proteins: a family of EF-hand calcium-binding proteins in search of a function. Cell Tissue Res. 1999; 295:1–12. [PubMed: 9931348]

35. Dickinson ME, et al. High-throughput discovery of novel developmental phenotypes. Nature. 2016; 537:508–514. [PubMed: 27626380]

36. Han JS. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mob DNA. 2010; 1:15. [PubMed: 20462415]

37. Ewing AD, et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol. 2013; 14:R22. [PubMed: 23497673]

38. Schrider DR, et al. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet. 2013; 9:e1003242. [PubMed: 23359205]

39. Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]

40. Giordano J, et al. Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS Comput Biol. 2007; 3:e137. [PubMed: 17630829]

41. Liebenauer LL, Slotnick BM. Social organization and aggression in a group of olfactory bulbectomized male mice. Physiol Behav. 1996; 60:403–409. [PubMed: 8840898]

42. Saraiva LR, et al. Combinatorial effects of odorants on mouse behavior. Proc Natl Acad Sci U S A. 2016; 113:E3300–3306. [PubMed: 27208093]

43. Ibarra-Soria X, et al. Variation in olfactory neuron repertoires is genetically controlled and environmentally modulated. eLife. 2017; 6

44. Zhang H, Hardamon C, Sagoe B, Ngolab J, Bui JD. Studies of the H60a locus in C57BL/6 and 129/Sv mouse strains identify the H60a 3’UTR as a regulator of H60a expression. Mol Immunol. 2011; 48:539–545. [PubMed: 21093919]

45. Diefenbach A, Jamieson AM, Liu SD, Shastri N, Raulet DH. Ligands for the murine NKG2D receptor: expression by tumor cells and activation of NK cells and macrophages. Nat Immunol. 2000; 1:119–126. [PubMed: 11248803]

46. O’Sullivan T, Dunn GP, Lacoursiere DY, Schreiber RD, Bui JD. Cancer immunoediting of the NK group 2D ligand H60a. J Immunol Baltim Md 1950. 2011; 187:3538–3545.

47. Ye Z, et al. Expression of H60 on mice heart graft and influence of cyclosporine. Transplant Proc. 2006; 38:2168–2171. [PubMed: 16980033]

48. Durrant C, et al. Collaborative Cross mice and their power to map host susceptibility to Aspergillus fumigatus infection. Genome Res. 2011; 21:1239–1248. [PubMed: 21493779]

49. Lilue J, Müller UB, Steinfeldt T, Howard JC. Reciprocal virulence and resistance polymorphism in the relationship between Toxoplasma gondii and the house mouse. eLife. 2013; 2:e01298. [PubMed: 24175088]

50. Levinsohn JL, et al. Anthrax lethal factor cleavage of Nlrp1 is required for activation of the inflammasome. PLoS Pathog. 2012; 8:e1002638. [PubMed: 22479187]

51. Bustos O, et al. Evolution of the Schlafen genes, a gene family associated with embryonic lethality, meiotic drive, immune processes and orthopoxvirus virulence. Gene. 2009; 447:1–11. [PubMed: 19619625]

52. Bauernfeind F, Hornung V. Of inflammasomes and pathogens--sensing of microbes by the inflammasome. EMBO Mol Med. 2013; 5:814–826. [PubMed: 23666718]

53. Boyden ED, Dietrich WF. Nalp1b controls mouse macrophage susceptibility to anthrax lethal toxin. Nat Genet. 2006; 38:240–244. [PubMed: 16429160]

54. Broz P, Dixit VM. Inflammasomes: mechanism of assembly, regulation and signalling. Nat Rev Immunol. 2016; 16:407–420. [PubMed: 27291964]

55. Sastalla I, et al. Transcriptional analysis of the three Nlrp1 paralogs in mice. BMC Genomics. 2013; 14:188. [PubMed: 23506131]

56. Hunn JP, Feng CG, Sher A, Howard JC. The immunity-related GTPases in mammals: a fast-evolving cell-autonomous resistance system against intracellular pathogens. Mamm Genome Off J Int Mamm Genome Soc. 2011; 22:43–54.

57. Taylor GA. IRG proteins: key mediators of interferon-regulated host resistance to intracellular pathogens. Cell Microbiol. 2007; 9:1099–1107. [PubMed: 17359233]

58. Li M, et al. Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. Nature. 2012; 491:125–128. [PubMed: 23000900]

59. Stremlau M, et al. The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. Nature. 2004; 427:848–853. [PubMed: 14985764]

60. Bell TA, et al. The paternal gene of the DDK syndrome maps to the Schlafen gene cluster on mouse chromosome 11. Genetics. 2006; 172:411–423. [PubMed: 16172501]

61. Mudge JM, Harrow J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. Mamm Genome Off J Int Mamm Genome Soc. 2015; 26:366–378.

62. Harrow J, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760–1774. [PubMed: 22955987]

63. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–585. [PubMed: 23715323]

64. Mouse models of 17q21.31 microdeletion and microduplication syndromes highlight the importance of Kansl1 for cognition. PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/28704368. (Accessed: 11th July 2018).

65. Lanier LL. Evolutionary struggles between NK cells and viruses. Nat Rev Immunol. 2008; 8:259–268. [PubMed: 18340344]

66. Paten B, et al. Cactus: Algorithms for genome multiple sequence alignment. Genome Res. 2011; 21:1512–1528. [PubMed: 21665927]

67. Chromosome assembly of large and complex genomes using multiple references. bioRxiv. [Accessed: 11th July 2018] Available at: https://www.biorxiv.org/content/early/2018/02/11/088435.

68. Arbogast T, et al. Mouse models of 17q21.31 microdeletion and microduplication syndromes highlight the importance of Kansl1 for cognition. PLoS Genet. 2017; 13

69. Loviglio MN, et al. Chromosomal contacts connect loci associated with autism, BMI and head circumference phenotypes. Mol Psychiatry. 2017; 22:836–849. [PubMed: 27240531]

70. Park N, et al. An improved approach to mate-paired library preparation for Illumina sequencing. Methods Gener Seq. 2013; 1

71. Kirby A, et al. Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. Genetics. 2010; 185:1081–1095. [PubMed: 20439770]

72. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012; 22:549–556. [PubMed: 22156294]

73. Luo R, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012; 1:18. [PubMed: 23587118]

74. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinforma Oxf Engl. 2011; 27:757–763.

75. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinforma Oxf Engl. 2008; 24:637–644.

76. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013; 41:D377–386. [PubMed: 23193289]

77. Boroviak K, Doe B, Banerjee R, Yang F, Bradley A. Chromosome engineering in zygotes with CRISPR/Cas9. Genes N Y N. 2000; 54:78–85.

78. Hodgkins A, et al. WGE: a CRISPR database for genome engineering. Bioinforma Oxf Engl. 2015; 31:3078–3080.

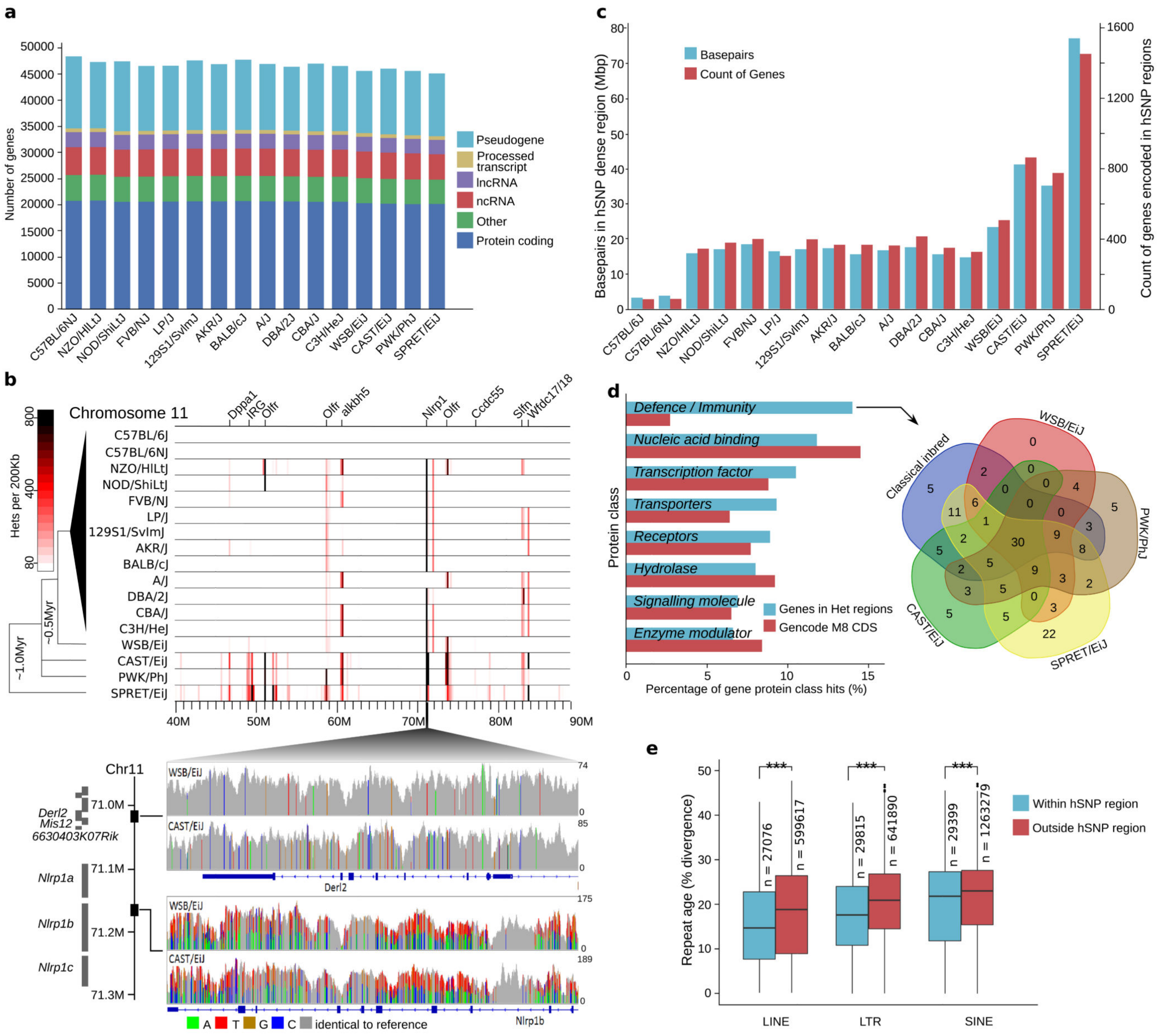

Figure 1. Genome annotation and content of strain specific haplotypes

(a) Summary of the strain specific gene sets showing the number of genes broken down by GENCODE biotype. **(b)** Heterozygous SNP density for a 50Mbp interval on chromosome 11 in 200Kbp windows for 17 inbred mouse strains based on sequencing read alignments to the C57BL/6J (GRCm38) reference genome (top). Labels indicate genes overlapping the most dense regions. SNPs visualized in CAST/EiJ and WSB/EiJ for 71.006-71.170Mbp on GRCm38 (bottom), including *Derl2*, and *Mis12* (upper panel) and *Nlrp1b* (lower panel). Grey indicates the strain base agrees with the reference, other colours indicate SNP differences, and height corresponds to sequencing depth. **(c)** Total amount of sequence and protein coding genes in regions enriched for heterozygous SNPs (relative to the GRCm38 reference genome) per strain. **(d)** Top PantherDB categories of coding genes in regions enriched for heterozygous SNPs based on protein class (left). Intersection of genes in the defence/immunity category for the wild-derived and classical inbred strains (right). **(e)** Box plot of sequence divergence (%), for LTRs, LINEs and SINEs within and outside of heterozygous dense regions. Sequence divergence is relative to a consensus sequence for the transposable element type (n=number of repeats in GRCm38, *** indicated p<0.001 using Welch's two sample t-test. Box plots show 25th and 75th percentiles, and the median value).

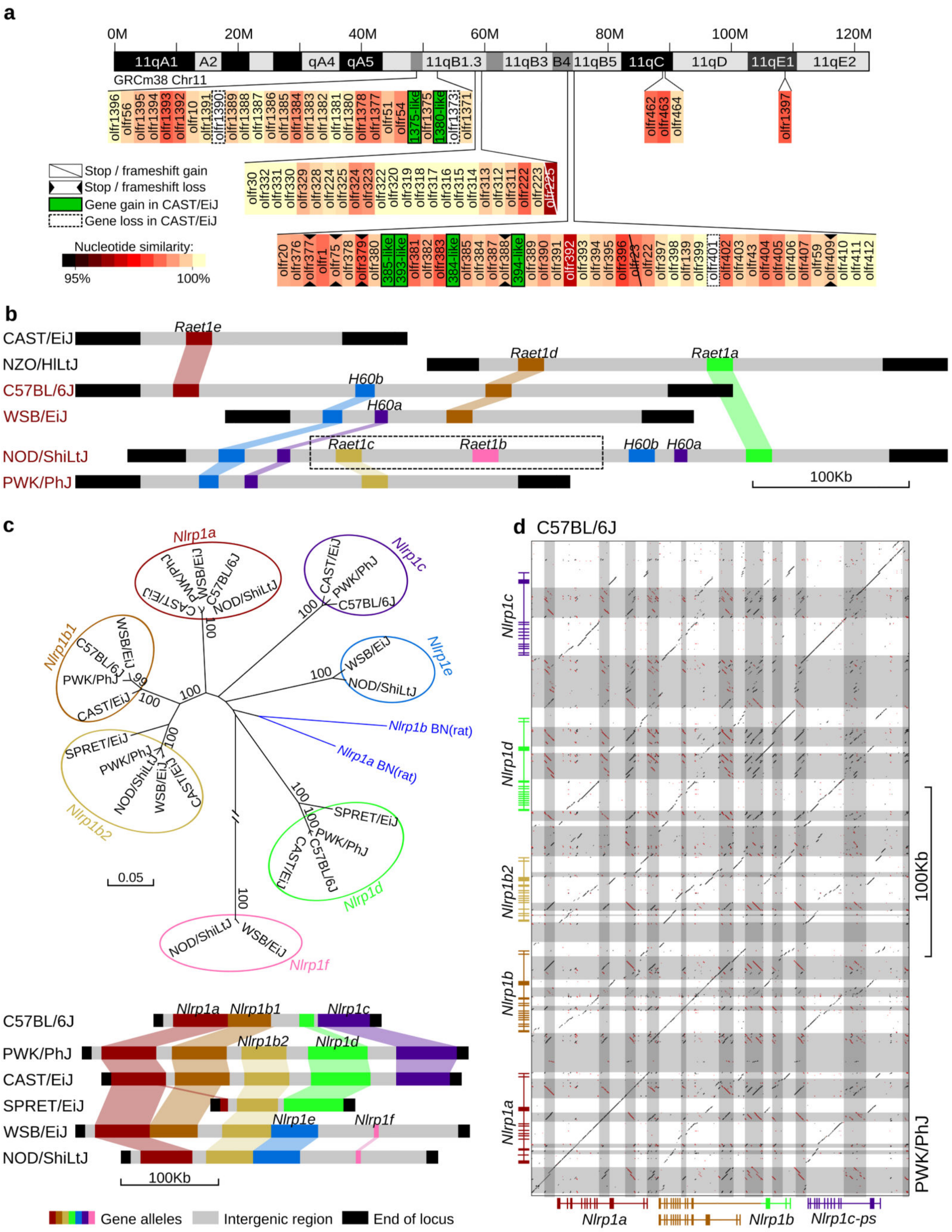
a
0M
20M
40M
60M
80M
100M
120M
11qA1
A2
qA4
qA5
11qB1.3
11qB3
B4
11qB5
11qC
11qD
11qE1
11qE2
GRCm38 Chr11
Stop / frameshift gain
Stop / frameshift loss
Gene gain in CAST/EiJ
Gene loss in CAST/EiJ
Nucleotide similarity:
95%
100%
olfr1396 olfr56 olfr1395 olfr1394 olfr1393 olfr1392 olfr10 olfr1391 olfr1390 olfr1389 olfr1388 olfr1387 olfr1386 olfr1385 olfr1384 olfr1383 olfr1382 olfr1381 olfr1380 olfr1378 olfr1377 olfr51 olfr54 1375-like olfr1375 1380-like olfr1373 olfr1371
olfr462 olfr463 olfr464
olfr1397
olfr30 olfr332 olfr331 olfr330 olfr329 olfr328 olfr224 olfr325 olfr324 olfr323 olfr322 olfr320 olfr319 olfr318 olfr317 olfr316 olfr315 olfr314 olfr313 olfr312 olfr311 olfr222 olfr223 olfr225
olfr20 olfr376 olfr377 olfr1 olfr75 olfr378 olfr379 olfr380 385-like 393-like olfr381 olfr382 olfr383 384-like olfr385 olfr384 olfr387 olfr388 394-like olfr389 olfr390 olfr391 olfr392 olfr393 olfr394 olfr395 olfr396 olfr23 olfr22 olfr397 olfr398 olfr139 olfr399 olfr401 olfr402 olfr403 olfr43 olfr404 olfr405 olfr406 olfr407 olfr59 olfr409 olfr410 olfr411 olfr412
b
CAST/EiJ
NZO/HlLtJ
C57BL/6J
WSB/EiJ
NOD/ShiLtJ
PWK/PhJ
Raet1e
Raet1d
Raet1a
H60b
H60a
Raet1c
Raet1b
H60b
H60a
100Kb
c
Nlrp1a
PWK/PhJ WSB/EiJ C57BL/6J CAST/EiJ NOD/ShiLtJ
Nlrp1c
CAST/EiJ PWK/PhJ C57BL/6J
Nlrp1b1
WSB/EiJ C57BL/6J PWK/PhJ CAST/EiJ
Nlrp1e
WSB/EiJ NOD/ShiLtJ
Nlrp1b BN(rat)
Nlrp1a BN(rat)
Nlrp1b2
SPRET/EiJ PWK/PhJ NOD/ShiLtJ WSB/EiJ CAST/EiJ
Nlrp1d
SPRET/EiJ PWK/PhJ C57BL/6J CAST/EiJ
Nlrp1f
NOD/ShiLtJ WSB/EiJ
100
99
0.05
C57BL/6J
PWK/PhJ
CAST/EiJ
SPRET/EiJ
WSB/EiJ
NOD/ShiLtJ
Nlrp1a
Nlrp1b1
Nlrp1c
Nlrp1b2
Nlrp1d
Nlrp1e
Nlrp1f
100Kb
Gene alleles
Intergenic region
End of locus
d C57BL/6J
Nlrp1c
Nlrp1d
Nlrp1b2
Nlrp1b
Nlrp1a
100Kb
PWK/PhJ
Nlrp1a
Nlrp1b
Nlrp1c-ps

Figure 2. Strain specific alleles for olfactory and immunity loci

(a) Olfactory receptor genes on chromosome 11 of CAST/EiJ. Gene gain/loss and similarity are relative to C57BL/6J. Novel members are named after their most similar homologues. **(b)** Gene order across *Raet1/H60* locus in the collaborative cross parental strains (A/J, NOD/ShiLtJ and 129S1/SvImJ share the same haplotype at this locus, represented by NOD/ShiLtJ). Strain name in black/red indicate *Aspergillus fumigatus* resistant/susceptible. Dashed box indicates unconfirmed gene order. **(c)** Novel protein-coding alleles of the *Nlrp1* gene family in the wild-derived strains and two classical inbred strains. Colours represent the phylogenetic relationships (top, amino acid neighbor joining tree of NBD domain) and the relative gene order across strains (bottom). **(d)** A regional dot plot of the *Nlrp1* locus in PWK/PhJ compared to the C57BL/6J GRCm38 reference (colour-coded same as panel (c)). Grey blocks indicate repeats and transposable elements.

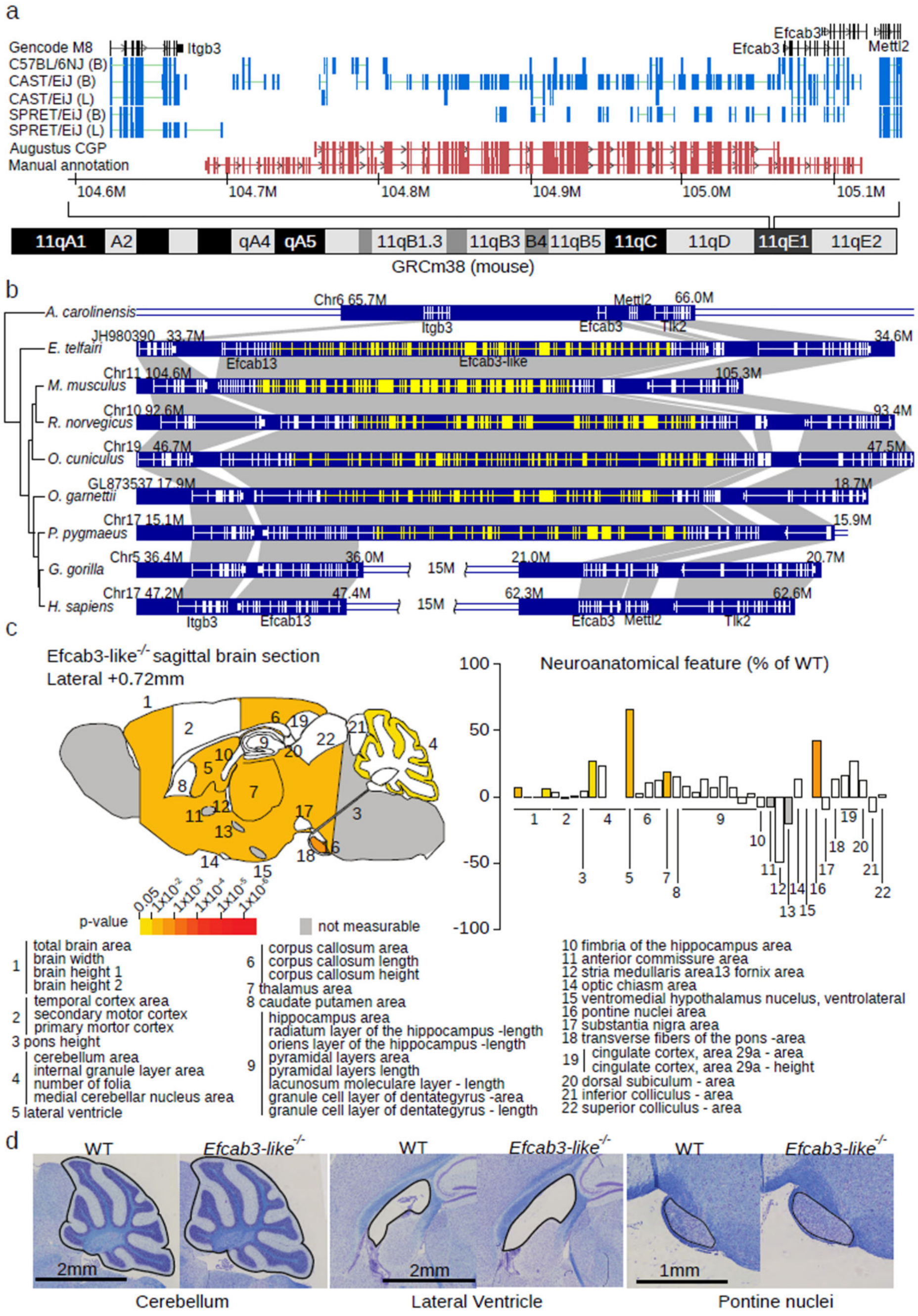

Figure 3. *Efcab3-like* locus, evolutionary history, and knockout phenotyping

(a) Comparative Augustus identified an unannotated 188 exon gene (*Efcab3-like*, red tracks). RNA-Seq splicing from two tissues (B=Brain, L=Liver, blue tracks) and five strains are displayed. Manual annotation extended this gene to 188 exons (lower red track). **(b)** Evolutionary history of *Efcab3-like* in vertebrates including genome structure and surrounding genes. The mRNA structure of each gene is shown with white lines on the blue blocks. Novel coding sequence discovered in this study is shown in yellow. Notably, *Efcab13* and *Efcab3* are fragments of the novel gene *Efcab3-like*. A recombination event happened in the common ancestor of sub-family *Homininae*, which disrupted *Efcab3-like* in gorilla and chimpanzee human. **(c)** Schematic representation of 22 brain regions plotted in sagittal plane for *Efcab3-like* mutant male mice (16 weeks of age, n=3) according to p-values (two-tailed equal variance t-test, left). Corresponding brain regions are labelled with a number that is described below the panel (Supplementary Table 15). White colouring indicates a p-value > 0.05 and grey indicates that the brain region could not be confidently tested due to missing data. Histograms showing the neuroanatomical features as percentage increase or decrease of the assessed brain regions in *Efcab3-like* mutant mice compared to matched controls (right). **(d)** Representative sagittal brain images of matched controls (left) and *Efcab3-like* mutant (right), showing a larger cerebellum, enlarged lateral ventricle and increased size of the pontine nuclei (n=3, see Supplementary Figure 15).

Table 1

Genome Reference Consortium (GRCm38) and GENCODE annotation updates informed by the strain assemblies.

Updates indicate known GRC issues solved based on C57BL/6NJ *de novo* assembly. GENCODE update is based on comparative Augustus predictions with 75% novel introns and includes annotation and predictions which occur on chromosomes 1-12.

Genome Reference Consortium (GRCm38) Update			
GRC issue solved	11	Genes completed	10
		Genes improved	1
GENCODE Update			
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