JCI insight

Overview of inactivating mutations in the protein-coding genome of the mouse reference strain C57BL/6J

Steven Timmermans, Claude Libert

JCI Insight. 2018;3(13):e121758. https://doi.org/10.1172/jci.insight.121758.

Technical Advance **Genetics**

Mice are extremely important as the premier model organism in human biomedical and mammalian genetic research. The genomes of several tens of mouse inbred strains have been sequenced. They have been compared to the genome of C57BL/6J, considered by convention as the reference genome. Based on a comparison of this reference genome with 36 other sequenced mouse strains, we generated an overview of all protein-coding genes that are deviant in this reference genome, compared with consensus protein-coding mouse gene sequences. We provide PROVEAN scores, reflecting the likelihood that these C57BL/6J proteins have lost function. We thus identified numerous abnormal proteins, and biological pathways, specifically present in C57BL/6J, suggesting the important caveats of this reference mouse strain, and linking candidate genes to some of the best-known phenotypes of this strain.

JCI insight

TECHNICAL ADVANCE

Overview of inactivating mutations in the protein-coding genome of the mouse reference strain C57BL/6J

Steven Timmermans[1,2] and Claude Libert[1,2]

[1]VIB Center for Inflammation Research, Ghent, Belgium. [2]Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium

Mice are extremely important as the premier model organism in human biomedical and mammalian genetic research. The genomes of several tens of mouse inbred strains have been sequenced. They have been compared to the genome of C57BL/6J, considered by convention as the reference genome. Based on a comparison of this reference genome with 36 other sequenced mouse strains, we generated an overview of all protein-coding genes that are deviant in this reference genome, compared with consensus protein-coding mouse gene sequences. We provide PROVEAN scores, reflecting the likelihood that these C57BL/6J proteins have lost function. We thus identified numerous abnormal proteins, and biological pathways, specifically present in C57BL/6J, suggesting the important caveats of this reference mouse strain, and linking candidate genes to some of the best-known phenotypes of this strain.

Introduction

In December 2017, the 15th anniversary of the publication of the mouse reference genome sequence was celebrated (1). The popular and well-known inbred strain C57BL/6J was selected to deliver the sequence of this reference genome. This event marked the start of a new era in mouse genetic research. The availability of a reference genome for the mouse model organism allowed for the identification of genetic polymorphisms between different strains and enabled phylogenetic, functional, and GWAS studies.

It is important to realize that the genome of the C57BL/6J strain is not perfect and defects that are present form an important confounding factor to any study applying (these) mice. The most comprehensive effort to date to identify all genetic polymorphisms present in inbred mouse strains is the Mouse Genomes Project (MGP) by the Wellcome Trust Sanger institute (2). In this project the genomes of 36 frequently used, prioritized inbred mouse strains were sequenced and interpreted in terms of SNPs, indels, and structural variants, found present in each strain, relative to the C57BL/6J reference strain (3). We have used this resource to provide an overview of all protein-coding variants in each inbred strain relative to C57BL/6J with an emphasis on those predicted to be defective. All data are available in a searchable database (mousepost.be). It is clear that even the closest mouse inbred strain related to the reference strain, namely C57BL/6NJ, already exhibits a number of functionally inactive genes (4).

Because of its status as a reference genome, generating a complete overview of genes that lead to defective proteins (either completely or partially) in the C57BL/6J strain is more difficult, especially for genes that show a loss of function in C57BL/6J only. Some isolated C57BL/6J-specific QTL loci and spontaneous mutants have been described, for example Nlrp12^{R1034K}, which was obtained by pairwise comparison of the C57BL/6J and C57BL/6N strains (5). We have used the MGP data to create an overview of all, potentially defective, protein-coding genes in C57BL/6J and provide insight into the biological significance of our findings.

Results

Identification and classification of C57BL/6J mutated transcripts. To detect abnormalities in the protein-coding sequences of the C57BL/6J reference strain, we compared these sequences with those of the other 36 sequenced mouse strains. To do that, we first generated consensus sequences of each protein-coding transcript of these 36 strains, and then followed a work flow (see Methods and Figure 1) for comparison of the C57BL/6J genome sequence with the newly generated consensus protein sequences and the generation

Conflict of interest: The authors have declared that no conflict of interest exists.

Submitted: April 19, 2018
Accepted: June 6, 2018
Published: July 12, 2018

Reference information:
JCI Insight. 2018;3(13): e121758.
https://doi.org/10.1172/jci.insight.121758.

of the lists of (highly) specific C57BL/6J variations. Data were obtained from the MGP and Ensembl. A consensus amino acid sequence was constructed for each protein-coding transcript and the C57Bl/6J sequences were compared to them. Mutations in C57BL/6J were then scored and classified in 3 groups. Non–stop mutation transcripts were further analyzed with Protein Variant Effect Analyzer (PROVEAN) software. All data are available as an extension on the mousepost.be website.

Between sequences of protein-coding transcripts, we discriminated 3 classes of polymorphisms: stop gain (SG) caused by a new stop codon; stop loss (SL), where the normal stop codon has been lost, and lastly; mutated (MUT), where transcripts were grouped that have nonsynonymous SNPs and small (in-frame) insertions and/or deletions.

We identified a total of 38 SG transcripts in C57BL/6J (Supplemental Table 1; supplemental material available online with this article; https://doi.org/10.1172/jci.insight.121758DS1), 2 of which were specific for this strain. The Sftpb protein is 36% shorter than in all the other mouse strains. A full KO of this gene is lethal (6). The other gene is *Zswim9* (48% shorter protein), but only one of the transcripts is affected.

Sixty-five transcripts were classified as SL in C57BL/6J (Supplemental Table 2), 6 of which are specific to the strain. The *Cilp* and *Kndc1* genes code for proteins that are respectively 66 and 737 amino acids longer in C57BL/6J than in all other mouse strains. The *Nadk2* gene has 3 transcripts, all of which encode a significantly longer protein in C57BL/6J mice. The final gene with a C57BL/6J-only SL is *Bean1*. The affected transcript encodes a protein of 42 amino acids in 35 out of 36 strains. However, in C57BL/6J the transcript codes for a protein with a length of 326 amino acids.

The third class, MUT, has a total of 5,075 transcripts assigned to it. All transcripts and indel/SNP positions were analyzed with PROVEAN (7), which predicts the chance that the sequence variation compromises protein function, based on sequence conservation, and 4,744 transcripts received valid PROVEAN scores and were retained in the results set. Of those, 1,477 had at least one event with a PROVEAN score of –2.5 or lower (Supplemental Table 3). A score of –2.5 or lower means that the mutation is predicted to have a negative impact on protein function. For comparison, the $Tlr4^{P712H}$ mutation that renders C3H/HeJ mice completely resistant to lipopolysaccharides has a PROVEAN score of –7.833 or the Tyr^{C103S} mutation leading to albinism in BALB/cJ mice has a score of –9.738 (3). Nine different genes show mutations with maximal scores (consensus sequence and C57BL/6J-unique) and low PROVEAN scores. Some have important functions, for example *Jmjd1c* (male fertility) and *Slc15a2* (renal function) (see Table 1).

The large majority of identified mutations are not specific to C57BL/6J, but are also found in a few other strains, most notably C57BL/6NJ, but other C57 strains also often share the C57BL/6J sequence (see Supplemental Table 3). For example, we can rapidly retrieve C57BL/6J mutants such as the $Cyb5r4^{Y356C}$ allele with a PROVEAN score of –6.641 (also present in 5 other strains; mousepost.be). C57BL/6J mice are known to have high fasted glucose levels and poor glucose tolerance (Mouse Phenome Database, https://phenome.jax.org/). Interestingly, this *Cyb5r4* gene, when depleted in mice, indeed leads to pancreatic abnormalities and diabetic phenotypes (8).

Functional analysis of biological significance. One of the applications of our work is to facilitate the identification of candidate genes causing or contributing to a certain phenotype or peculiarity of the C57BL/6J mice. We combined our data with phenotype information obtained from the Mouse Phenome Database, the International Mouse Phenotyping Consortium (IMPC, www.mousephenotype.org), literature and, for genes with little information in mice but having a better known human homolog, the Online Mendelian Inheritance in Man (OMIM) database. We were able to identify several candidate genes that could help explain some of the phenotypes observed in C57BL/6J mice.

First, we performed a functional analysis on the 3 classes of genes using a gene ontology (GO) overrepresentation test. We found no significant over- or underrepresented terms for the SG and SL groups. The group of MUT transcripts did give a statistically significant result, i.e., an overrepresentation for the term "tRNA metabolic process" ($P = 3.16 \times 10^{-2}$). The genes annotated with tRNA metabolic process are *Vars2*, *Qars*, *Farsa*, *Eprs*, and *Trmt13*. A full KO of these genes is usually lethal, so any loss of function in C57BL/6J will be a priori limited.

C57BL/6J bone density. It has been described that C57BL/6J mice have low bone density, most specifically the lowest cortical bone density of all mouse strains. This is mentioned on the strain summary page on the JAX website and also described in Beamer et al. (9). We can potentially link 2 genes to this phenotype: *Farsa* and *Acan*. The former codes for the α chain of the phenylalanyl-tRNA synthetase and data from the IMPC show that loss of function of this gene leads to bone abnormalities (i.e., less mineral density of the bones). *Farsa* contains 1 mutation in C57BL/6J that is predicted to be deleterious: D291G

Figure 1. Excerpt from the multiple sequence alignment of the gene *Acan*. Shown are sequences of human, rat, C57BL/6J mouse, chicken, cow, and dog. The C57BL/6J mutated position is marked by a black rectangular box.

(PROVEAN score: –4.339). This mutation is not restricted to only C57BL/6J, since all C57 strains in the database (C57BL/10J, C57BL/6NJ, C57BR/cdJ, and C57L/J) have the same substitution as well as BUB/BnJ, NZB/B1NJ, and NZO/HlLtJ. Furthermore, C57BL/6J carries one extra mutation compared with the consensus sequence, namely a His-to-Tyr substitution at position 310. This substitution is predicted to be neutral to the function of the protein (PROVEAN score: 4.304). Finally, the LEWES/EiJ and ZALENDE/EiJ strains carry 2 more mutations that are shared and are specific to them: Q141L (predicted neutral, PROVEAN score: –1.782) and Q147L (predicted deleterious, PROVEAN score: –3.443). The *Acan* gene codes for the aggrecan protein, which is an essential component of proteoglycan, and thus important for the extracellular matrix of cartilage and bone formation. The sequence that is observed in the C57BL/6J mouse has a histidine instead of a proline at position 1,938. This mutation is unique to C57BL/6J (Table 1, Supplemental Table 3, mousepost.be) and all the other inbred strains sequenced by the MGP have a proline at this position. This substitution is predicted to have a negative impact on protein function with a strong PROVEAN score of –8.415. This mutation is located in a conserved C-type lectin domain of the protein (from position 1,918 to 2,044) and the proline at this position is well conserved in other species (see Figure 2). In humans, a point mutation in this domain causes a clear stature phenotype. For mice there are many different phenotypes, depending on the type of KO, but an abnormal bone phenotype has been described (10).

C57BL/6J mice alcohol preference. It has been shown that C57BL/6J mice have high preference for alcohol, meaning they consume high amounts of it voluntarily in a 2-bottle-choice test experiment (11). Alcohol preference, or avoidance, is a complex trait with both genetic and environmental components. Several candidate alcohol preference genes have been identified in QTL studies (12), but impact/penetrance of the observed effect is in some cases strongly dependant on the genetic background of the mice used. Interestingly, none of the candidate genes were found to be mutated in the C57BL/6J strain, including those where lower expression (for example, KO) leads to increased ethanol consumption in the mice (13). However, a recent paper describes the *Adal* gene as inversely correlated with high alcohol preference (14), meaning that higher expression results in lowered preference for alcohol. Unlike the other alcohol preference genes, this gene is mutated in the C57BL/6J strain (compared with the consensus). The gene has multiple transcripts, and all lead to proteins with 1 amino acid substitution in C57BL/6J, namely an asparagine replacing a threonine with PROVEAN scores of –3.386 to –3.141 (Supplemental Table 3). If we consider the main transcript of the gene (ENSMUST00000066155), the mutation is found at position 218. We propose that this gene may help explain the high voluntary alcohol consumption that is seen in C57BL/6J mice. The mutation in *Adal* is not C57Bl/6J specific and is also found in 4 other closely related C57 strains (C57BL/10J, C57BL/6NJ, C57BR/cdJ, and C57L/J) and PWK/PhJ. Unfortunately, the amount of data about voluntary ethanol consumption in inbred mice is limited and there are no data available for PWK/PhJ and C57BL/6NJ (but the latter is assumed to be the same as 6J). The C57BL/10J, C57BR/cdJ, and C57L/J strains have been described in the past to also have a high preference for alcohol (15).

Tlr11 and Toxoplasma gondii sensitivity. In mice, the *Tlr11* gene is linked to resistance against several infectious agents, particularly infection with *T. gondii* (16, 17). The C57BL/6J strain is found to be very sensitive to infection with this pathogen, while several other strains, such as CBA/J (18) and BALB/cJ, were reported to be resistant. Our data show that the *Tlr11* gene contains a missense mutation in C57BL/6J leading to the following substitution (main transcript): F386L. This mutation has a PROVEAN score of –5.950 and is also found in BTBR, C57BL/10J, C57BL/6NJ, C57BR/cdJ, C57L/J, and C58/J, but not in any of the other strains. Note that the MOLF/EiJ strain also contains another mutation also predicted to have negative effects in this gene. While data for *T. gondii* resistance is very limited, we were unable to find evi-

Table 1. Examples of C57BL/6J transcripts from this study

TRANSCRIPT	GENE	A = CONSENSUS SCORE (=STRAINS/36)	B = C57BL/6J-LIKE STRAINS	TOTAL SCORE = A × (1 – B/36)	CONSENSUS LENGTH PROTEIN	C57BL/6J LENGTH PROTEIN	DIFFERENCE LENGTHS	LENGTHS RATIO
ENSMUST00000119139	*6330408A02Rik*	1	0	1	849	447	404	0.527
ENSMUST00000182014	*Sftpb*	1	0	1	377	241	136	0.639
ENSMUST00000048762	*Cilp*	1	0	1	1,184	1,250	66	1.056
ENSMUST00000053445	*Kndc1*	1	0	1	1,005	1,742	737	1.733
ENSMUST00000067760	*Nadk2*	1	0	1	353	452	99	1.280
ENSMUST00000100789	*Nadk2*	1	0	1	259	401	142	1.548
ENSMUST00000100790	*Nadk2*	1	0	1	331	430	99	1.299
ENSMUST00000171018	*Bean1*	0.972	0	0.972	42	326	284	7.762
TRANSCRIPT	**GENE**	**A = CONSENSUS SCORE (= STRAINS/36)**	**B = C57BL/6J-LIKE STRAINS**	**TOTAL SCORE = A × (1 – B/36)**	**AA MUTATION**	**SUPPORTING SEQUENCES**	**PROVEAN SCORE**	
ENSMUST00000051179	*Fam181b*	1	0	1	D373 C374insL	131	–9.288	
ENSMUST00000173689	*Jmjd1c*	1	0	1	P1715L	563	–8.947	
ENSMUST00000032835	*Acan*	1	0	1	P1938H	116	–8.415	
ENSMUST00000076226	*Herc2*	1	0	1	D1235G	447	–6.505	
ENSMUST00000031985	*Mkrn1*	1	0	1	Y346N	513	–6.413	
ENSMUST00000164579	*Slc15a2*	1	0	1	T210M	439	–5.190	
ENSMUST00000125537	*Zc3h7a*	1	0	1	S24P	328	–4.948	
ENSMUST00000094464	*Casz1*	1	0	1	P1087L	530	–3.963	
ENSMUST00000058865	*Pdzk1*	1	0	1	D162N	295	–2.930	

Examples of transcripts with stop-gain and stop-loss sequence variations in C57BL/6J specifically (upper part) and examples of transcripts with missense or in-frame short indel variations in C57BL/6J specifically (lower part). Other examples, as well as examples of transcripts less specific for C57BL/6J are shown in Supplemental Tables 1, 2, and 3.

dence of a strain reported to be resistant and carrying these Tlr11 mutations. The presence of this mutation may help explain the sensitivity of C57BL/6J to *T. gondii* infection, but indeed needs to be explored further.

Atypical chemokine receptor 1. Ackr1 encodes a high-affinity chemokine receptor that is uncoupled from downstream signal cascades and causes sequestration and possible degradation of its ligands. A loss of function leads to more inflammatory infiltrates in lung and liver (19). In humans, the gene is linked to resistance to some *Plasmodium* species, as some of them require this receptor to enter red blood cells (20). The Mouse Phenome Database shows that the number of infected erythrocytes is the lowest of all strains, in C57BL/6J mice after infection with *P. berghei* and *P. chabaudi* (21, 22). The gene contains 3 nonsynonymous SNPs leading to 3 amino acid changes (see Table 2). Unfortunately, no information can be found about the effect of any of those mutations on protein function.

The closely related strains C57BL/6J and C57BL/6N. Despite C57BL/6J being the reference mouse strain, the closely related C57BL/6NJ genome is of utmost importance, for example because embryonic stem cells of this strain have been used to generate thousands of KO mouse lines in the IMPC. When focusing on all the mutated hits that we identified in C57BL/6J, we found that 844 of these were equally observed in the genome of C57BL/6NJ, while 14 were different between both strains (Table 3). Equally so, 26 SG variations of C57BL/6J were retained in C57BL/6NJ but 3 were not, and in the SL variations 76 were retained and 6 were not. All those that are different in C57BL/6NJ compared with C57BL/6J had the consensus sequence, as listed in Table 3.

Discussion

The mouse inbred strain C57BL/6J is one of the strains that was established in the early decades of the 20th century. The history of the strain is interesting, and has been described in detail previously (23). In view of reproducibility of scientific data, in space and time, it is essential that scientists focus on just a few inbred stains. The C57BL/6J mouse strain has become a very popular strain, for several practical reasons and its central leading position in today's biomedical research is well established. In 2002, the publication of the first draft of the C57BL/6J genome reflected its status as reference strain (1). Since

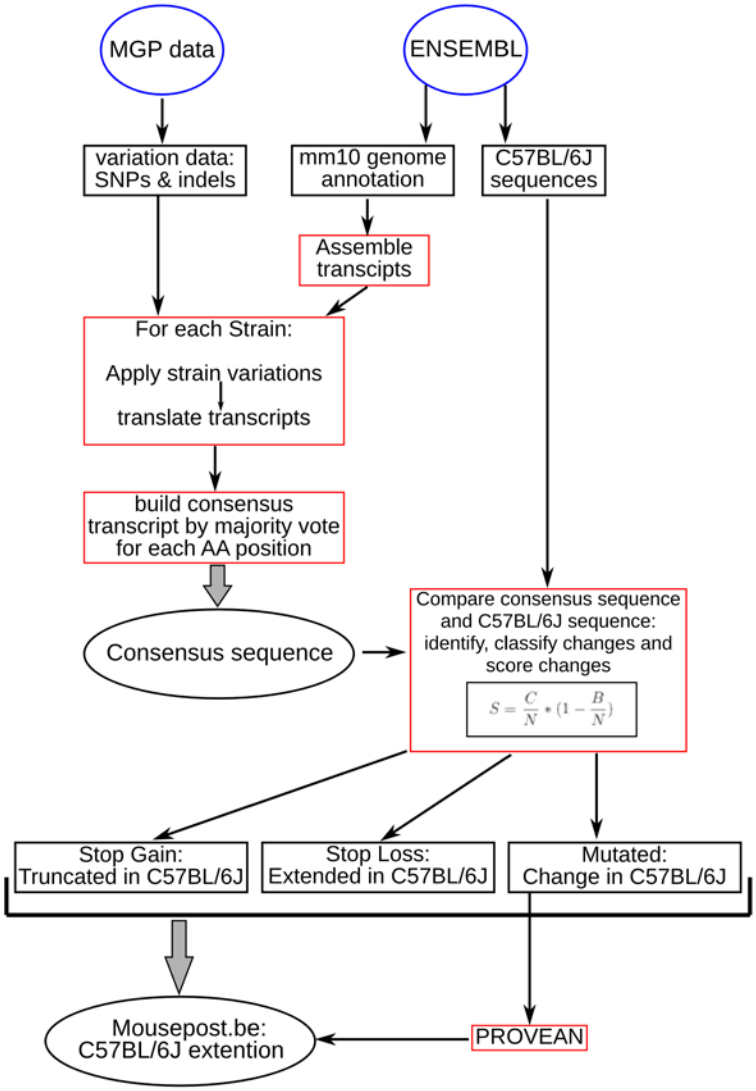

Figure 2. Schematic overview of the work flow and the comparison of the C57BL/6J genome sequence with the newly generated consensus protein sequences and the generation of the lists of (highly) specific C57BL/6J variations. Data were obtained from the mouse genomes project (MGP) and Ensembl. A consensus amino acid (AA) sequence was constructed for each protein-coding transcript and the C57BL/6J sequences were compared to them. Mutations in C57BL/6J were then scored and classified in 3 groups. Non–stop mutation transcripts were further analyzed with the Protein Variant Effect Analyzer (PROVEAN) software. All data are available as an extension on the mousepost.be website.

the publication of this genome sequence, the MGP project of the Wellcome Trust Sanger Institute has generated the genome sequences of 36 high-priority mouse inbred strains, some of which are very well appreciated and intensely studied, e.g., BALB/cJ, DBA/2J, etc. To provide a user-friendly access to the inherent richness of these genome sequences, in the past, we have focused on the protein-coding sequences of these 36 strains, and compared them all, pairwise, with the C57BL/6J reference genome (3). This has led to the searchable web page mousepost.be.

The rationale of our current report is that there is no reason to consider the C57BL/6J genome, despite being a reference genome, as a perfect genome. Therefore, we have attempted to provide an overview the abnormal protein-coding transcripts in this strain. We have chosen to compare these sequences with the other sequenced mouse strains to find the deviant ones in C57BL/6J. For that purpose, we have first generated consensus sequences of each transcript, based on a majority principle, and then compared them with the C57BL/6J ones. We have chosen not to include protein-coding transcripts of species other than mice to build these consensus sequences, e.g., rat or human.

Our work shows that the C57BL/6J mouse reference genome encodes several dozen proteins that are predicted to be defective in this strain only, and many hundreds that are defective in C57BL/6J and closely related strains. We have investigated several of these mutations and explain how they may help explain some of the best-known phenotypes of the C57BL/6J mouse strain, for example in bone density and alcohol preference. In cases where the mutation is not unique to the C57BL/6J strain it cannot always fully explain the observed phenotypes, but our results provide an interesting starting point for further research.

Table 2. Overview of the mutations present[A] in the *Ackr1* gene in C57BL/6J

Mutation	PROVEAN Score	Other Strains
D56G	−5.915	C57BL/10J, C57BL/6NJ, C57BR/cdJ, C57L/J, MOLF/EiJ, SPRET/EiJ
T194A	−3.481	C57BL/10J, C57BL/6NJ, C57BR/cdJ, C57L/J, MOLF/EiJ, SPRET/EiJ
L221P	−6.309	C57BL/10J, C57BL/6NJ, C57BR/cdJ, C57L/J, MOLF/EiJ

[A]Mutations present in the Ackr1 gene with PROVEAN scores of −2.5 or lower. Other mouse strains where the same mutation is found are also listed.

Obviously, the closely related mouse strain C57BL/6NJ shares most of the sequence-specific variation that we found in C57BL/6J, but not all of them. Some of these may also help to explain the subtle differences between C57BL/6J and C57BL/6NJ in phenotypes, as often encountered by researchers (24).

In summary, this study provides a gene-by-gene overview of the deviant sequences of protein-coding genes, specifically of the reference strain C57BL/6J (https://phenome.jax.org/) and so forms a valuable starting point to explain C57BL/6J-specific phenotypes or pathways, and may also be important from a practical point of view, for example when deciding which genetic background to use for genome editing.

Methods

Data sources: genome data. The mouse reference genome sequence and structural annotation were obtained from the Ensembl FTP server, more specifically we used GRCm38 release of the mouse reference genome (mm10) obtained from ftp://ftp.ensembl.org/pub/release86/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.primary_assembly.fa.gz. For the structural annotation of the genome we used the Ensembl gene sets in Gene Transfer Format (GTF) from Ensembl release version 86 (ftp://ftp.ensembl.org/pub/release-86/gtf/mus_musculus/Mus_musculus.GRCm38.86.gtf.gz). The sequence and annotation files were indexed with samtools faidx and tabix, respectively, to facilitate the use of data. The genome annotation file was filtered so as to only retain the genes annotated as 'protein coding'. For each exon, genomic start and stop coordinates were retained and divisions in 5′UTR, coding DNA sequence (CDS), and 3′UTR were also made. Sequences for each of these regions were extracted from the mm10 genome fasta file, sequences from genes on the '–' strand were reverse complemented.

Data sources: variation data. Data about SNPs and small indels were obtained from the MGP website. We used the REL-1505 release of the MGP data. Two variant call format (VCF) files per strain, one for SNPs and one for indels, were download from the MGP FTP site for processing. All variation data were filtered using the genomic regions obtained from the mm10 gtf file so that only SNPs/indels overlapping with the coding sequence were retained. The regions were extended 2 bp up- and downstream of each exon in order to include the splice site regions.

Consensus sequences. Consensus sequences were constructed from strain-specific sequences using a simple per-position majority vote mechanism: all sequence variations were applied to the 5′UTR, CDS, and 3′UTR regions for each gene and in each stain. Nucleotide sequences were then translated to the corresponding amino acid sequences. In the case of loss of a stop codon the 3′UTR sequence was appended to the CDS and the translation step was run again with the extended sequence. For every gene a multiple sequence alignment was constructed from the strain-specific sequences with the muscle program. Sequence alignments were processed to construct a consensus sequence by means of majority vote per position. In case of a tie, 2 sequences were accepted as being the consensus.

Identification and classification. Identification and classification of the sequences mutated in the C57BL/6J gene was performed using the consensus sequences. Three classes of sequence changes were used: 2 classes of stop codon mutants, being SG (nonsense mutations) and SL, the third class were the mutated transcripts, meaning those that had amino acid substitutions and in-frame indels. The number of transcripts belonging to each class were identified. PROVEAN software was used to predict the effect of amino acid changes (substitutions, insertions, deletions) on protein function (http://provean.jcvi.org/index.php).

Scoring. We used 2 scoring metrics and their combination to rank our results. The first metric (*M*1) takes into account the support for the consensus sequence for each position. This is expressed as the number of

Table 3. C57BL/6J-specific genes (vs. C57BL/6NJ)

Type	Gene	C57BL/6NJ equals consensus
MUT	*Slc15a1*	Y
	Kif12	Y
	Mkrn1	Y
	Acan	Y
	Lama1	Y
	Zc3h7a	Y
	Fam181b	Y
	Pdzk1	Y
	Herc2	Y
	Casz	Y
	Mroh2a	Y
	Jmjd1c	Y
	Apol11b	Y
	Olfr498	Y
SG	*G330408A02Rik*	Y
	Sftpb	Y
	1600016C10Rik	Y
SL	*Clip*	Y
	Kndc1	Y
	Nadk2	Y
	Bean1	Y
	D930048N14Rik	Y
	Muc4	Y

An overview of all genes where the specific sequence of the C57BL/6J strain is not found in the C57BL/6NJ strain. Grouping is done by event type, where MUT = mutated, SG = stop gain, SL = stop loss, as explained in the main text. In all cases, the C57BL/6NJ sequence was identical to the consensus sequence.

strains supporting the consensus amino acid (*C*) at a position versus the total number of strains (*N*), C57BL/6J excluded. The second metric (*M*2) is the number of strains that agree with the C57BL/6J sequence (*B*) versus the total number of strains (*N*), which is an indication of the variability of the position. This score is lower as the number of strains supporting the C57BL/6J sequence increases. These metrics are combined into a final score (*S*) for the mutation.

$$M1 = C/N$$

$$M2 = 1 - B/N$$

$$S = C/N \times (1 - B/N)$$

GO overrepresentation test. The GO statistical tests were performed using the PantherGO webtool with the GOslim classifications using the Fisher's exact test with FDR multiple-test correction (25, 26).

Data availability. The entire data set for this analysis can be consulted as an extension of the website mousepost.be. All essential information, including the identity of the other mouse strains that have an identical SNP/indel as C57BL/6J, can be found in this database. Three PDF files containing a selection of the data are available online as part of the supplemental information as Supplemental Tables 1, 2, and 3. These tables contain hyperlinks to Ensembl and PubMed.

Author contributions

CL provided the concept of the study. ST performed all data analyses, supervised by CL. ST and CL wrote the manuscript.

Acknowledgments

Research was funded by the Agency for Innovation of Science and Technology in Flanders (IWT), the Research Council of Ghent University (GOA program), the Research Foundation Flanders (FWO Vlaanderen), and the Interuniversity Attraction Poles Program of the Belgian Science Policy (IAP-VI-18).

Address correspondence to: Claude Libert, Technologiepark 927, 9052 Ghent, Belgium. Phone: 32.9.3313700; Email: Claude.Libert@IRC.VIB-UGent.Be.

1. Mouse Genome Sequencing Consortium, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002;420(6915):520–562.
2. Keane TM, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*. 2011;477(7364):289–294.
3. Timmermans S, Van Montagu M, Libert C. Complete overview of protein-inactivating sequence variations in 36 sequenced

mouse inbred strains. *Proc Natl Acad Sci USA*. 2017;114(34):9158–9163.
4. Kumar V, et al. C57BL/6N mutation in cytoplasmic FMRP interacting protein 2 regulates cocaine response. *Science*. 2013;342(6165):1508–1512.
5. Ulland TK, et al. Nlrp12 mutation causes C57BL/6J strain-specific defect in neutrophil recruitment. *Nat Commun*. 2016;7:13180.
6. Davé V, et al. Calcineurin/Nfat signaling is required for perinatal lung maturation and function. *J Clin Invest*. 2006;116(10):2597–2609.
7. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One*. 2012;7(10):e46688.
8. Xie J, et al. Absence of a reductase, NCB5OR, causes insulin-deficient diabetes. *Proc Natl Acad Sci USA*. 2004;101(29):10750–10755.
9. Beamer WG, Donahue LR, Rosen CJ, Baylink DJ. Genetic variability in adult bone density among inbred strains of mice. *Bone*. 1996;18(5):397–403.
10. Rittenhouse E, et al. Cartilage matrix deficiency (cmd): a new autosomal recessive lethal mutation in the mouse. *J Embryol Exp Morphol*. 1978;43:71–84.
11. Yoneyama N, Crabbe JC, Ford MM, Murillo A, Finn DA. Voluntary ethanol consumption in 22 inbred mouse strains. *Alcohol*. 2008;42(3):149–160.
12. Bice PJ, et al. Identification of QTLs influencing alcohol preference in the High Alcohol Preferring (HAP) and Low Alcohol Preferring (LAP) mouse lines. *Behav Genet*. 2006;36(2):248–260.
13. Mayfield J, Arends MA, Harris RA, Blednov YA. Genes and alcohol consumption: studies with mutant mice. *Int Rev Neurobiol*. 2016;126:293–355.
14. Lesscher HMB, Bailey A, Vanderschuren LJMJ. Genetic variability in adenosine deaminase-like contributes to variation in alcohol preference in mice. *Alcohol Clin Exp Res*. 2017;41(7):1271–1279.
15. Belknap JK, Crabbe JC, Young ER. Voluntary consumption of ethanol in 15 inbred mouse strains. *Psychopharmacology (Berl)*. 1993;112(4):503–510.
16. Yarovinsky F, et al. TLR11 activation of dendritic cells by a protozoan profilin-like protein. *Science*. 2005;308(5728):1626–1629.
17. Plattner F, et al. Toxoplasma profilin is essential for host cell invasion and TLR11-dependent induction of an interleukin-12 response. *Cell Host Microbe*. 2008;3(2):77–87.
18. Lee YH, Kasper LH. Immune responses of different mouse strains after challenge with equivalent lethal doses of Toxoplasma gondii. *Parasite*. 2004;11(1):89–97.
19. Dawson TC, et al. Exaggerated response to endotoxin in mice lacking the Duffy antigen/receptor for chemokines (DARC). *Blood*. 2000;96(5):1681–1684.
20. Miller LH, Mason SJ, Clyde DF, McGinniss MH. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N Engl J Med*. 1976;295(6):302–304.
21. Bopp SE, et al. Genome wide analysis of inbred mouse lines identifies a locus containing Ppar-gamma as contributing to enhanced malaria survival. *PLoS One*. 2010;5(5):e10903.
22. Laroque A, Min-Oo G, Tam M, Radovanovic I, Stevenson MM, Gros P. Genetic control of susceptibility to infection with Plasmodium chabaudi chabaudi AS in inbred mouse strains. *Genes Immun*. 2012;13(2):155–163.
23. Kiselycznyk C, Holmes A. All (C57BL/6) mice are not created equal. *Front Neurosci*. 2011;5:10.
24. Song HK, Hwang DY. Use of C57BL/6N mice on the variety of immunological researches. *Lab Anim Res*. 2017;33(2):119–123.
25. Thomas PD, et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*. 2003;13(9):2129–2141.
26. Thomas PD, et al. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res*. 2006;34(Web Server issue):W645–W650.