Using whole-genome sequences of the LG/J and SM/J inbred mouse strains to prioritize quantitative trait genes and nucleotides

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:16121027

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Using whole-genome sequences of the LG/J and SM/J inbred mouse strains to prioritize quantitative trait genes and nucleotides

Igor Nikolskiy, Donald F Conrad, Sung Chun, Justin C Fay, James M Cheverud and Heather A Lawson

Abstract

**Background:** The laboratory mouse is the most commonly used model for studying variation in complex traits relevant to human disease. Here we present the whole-genome sequences of two inbred strains, LG/J and SM/J, which are frequently used to study variation in complex traits as diverse as aging, bone-growth, adiposity, maternal behavior, and methamphetamine sensitivity.

**Results:** We identified small nucleotide variants (SNVs) and structural variants (SVs) in the LG/J and SM/J strains relative to the reference genome and discovered novel variants in these two strains by comparing their sequences to other mouse genomes. We find that 39% of the LG/J and SM/J genomes are identical-by-descent (IBD). We characterized amino-acid changing mutations using three algorithms: LRT, PolyPhen-2 and SIFT. We also identified polymorphisms between LG/J and SM/J that fall in regulatory regions and highly informative transcription factor binding sites (TFBS). We intersected these functional predictions with quantitative trait loci (QTL) mapped in advanced intercrosses of these two strains. We find that QTL are both over-represented in non-IBD regions and highly enriched for variants predicted to have a functional impact. Variants in QTL associated with metabolic (231 QTL identified in an F16 generation) and developmental (41 QTL identified in an F34 generation) traits were interrogated and we highlight candidate quantitative trait genes (QTG) and nucleotides (QTN) in a QTL on chr13 associated with variation in basal glucose levels and in a QTL on chr6 associated with variation in tibia length.

**Conclusions:** We show how integrating genomic sequence with QTL reduces the QTL search space and helps researchers prioritize candidate genes and nucleotides for experimental follow-up. Additionally, given the LG/J and SM/J phylogenetic context among inbred strains, these data contribute important information to the genomic landscape of the laboratory mouse.

**Keywords:** Mouse models, Quantitative trait loci (QTL), Complex traits, Identity-by-descent (IBD), Predicted deleterious mutations, Quantitative trait gene (QTG), Quantitative trait nucleotide (QTN), Candidate loci, Whole genome sequence, Small nucleotide variant (SNV)

Background

The LG/J (large) and SM/J (small) strains of inbred mice were independently derived from selection experiments for large and small body size at 60 days, respectively [1]. LG/J was created from a pool of albino mice obtained from a commercial breeder over a nine-month period and selected for increased body size for over 50 generations [2,3].

SM/J was created from a pool of mice derived from four crosses of seven inbred strains: dilute brown agouti (dba), silver chocolate (sv ba), black and tan (a), pink-eyed, short-eared dilute brown agouti (ps e dba), albino (c), cinnamon spotted (bs), and agouti (a) [4]. It is unclear whether any of these strains are related to current laboratory inbred strains. The SM/J strain is kept heterozygous at the agouti locus by mating black animals (a/a) with their white-bellied agouti (A/a) siblings. Attempts to fix the SM/J strain or its Recombinant Inbred Strain offspring for the SM/J allele at agouti have resulted in strain failure [5].

© 2015 Nikolskiy et al; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Subsequent to selection, the LG/J and SM/J strains are fully inbred and have been maintained at the Jackson Laboratory since the 1950s by brother-sister mating. They are at the extremes of the adult body size distribution among the common laboratory inbred strains and have been profitably studied for genetic variation in adult body size and growth [6-18]. Early genetic studies of these two strains found differences in body size are caused by many genes of individually small effects [8]. This result has been confirmed in many quantitative trait loci (QTL) mapping studies. The strains have remained phenotypically stable over time, except that the SM/J strain is now about 6g heavier than it was in Chai’s studies [19].

LG/J and SM/J differ in many complex traits in addition to size and growth. The parental strains and their crosses differ in skeletal morphology, including the size and shape of the skull [20-22], mandible [23-30], tooth morphology [31,32], long-bone lengths [33-37], and a variety of other skeletal elements as well as bone biomechanical and structural properties [38,39]. They also differ for a variety of metabolic traits including obesity, diabetes, and serum cholesterol, triglycerides, and free fatty acids levels [40-51]. The SM/J strain responds more strongly than LG/J to a high-fat diet for these metabolic traits. The two strains also differ for maternal genetic effects on offspring growth and offspring metabolic traits [52-56] and their cross has been useful in mapping parent-of-origin genetic effects on metabolic traits [48-51,53,57].

In addition to these diverse metabolic and skeletal phenotypes, the LG/J strain shows the rare ability to regenerate tissues after injury. The LG/J strain regenerates ear pinna tissues after a 2mm hole-punch [58-61] while the SM/J strain does not. The LG/J strain also can regenerate damaged articular cartilage [62] and is protected from post-traumatic osteoarthritis [63].

Others have mapped a variety of behavioral phenotypes using a LG x SM cross, including prepulse inhibition [64] and methamphetamine sensitization and locomotor activity [65]. Additional studies have investigated blood cell parameters [66] and skeletal muscle weight and fiber types [67,68].

Here we describe the whole-genome sequences of the LG/J and SM/J inbred mouse strains, adding two more sequences to the collection of whole-genomes of laboratory mice [69,70]. We integrate these sequences with quantitative trait loci (QTL), and illustrate how top-down (phenotype to genomic sequence) and bottom-up (genomic sequence to phenotype) approaches can be used to identify candidate quantitative trait genes (QTG) and prioritize positional candidate quantitative trait nucleotides (QTN) for further mechanistic studies. Results from QTL studies using crosses of LG/J and SM/J will be more interpretable given these whole-genome sequences.

Results and discussion

LG/J and SM/J whole-genome sequences

Sequence was generated from DNA isolated from 1 LG/J female and 1 SM/J female as described below. Variants (comprised of small nucleotide variants (SNVs) including single nucleotide polymorphisms (SNPs) and small insertion-deletions (Indels) and structural variants (SVs) including deletions, insertions and inversions) were discovered based on whether they were the same as, or different from, the C57BL/6J reference sequence (mm10, NCBI build 38). Greater than 90% of reads for each strain could be uniquely mapped to the reference genome and, based on a genome size of 2.5G, our coverage is approximately 35X and 30X for LG/J and SM/J, respectively. Overviews of annotated variants identified for each strain individually, as well as those that are polymorphic between LG/J and SM/J, are described in Tables 1 and 2 and in Additional files 1 and 2. SNVs identified for the LG/J strain are available in Additional files 3 and 4 and for the SM/J strain in Additional files 5 and 6. Structural variant positions and classifications for each strain are available in Additional file 7. Polyphoric variants and their genomic context are illustrated in Figure 1. Quality was assessed by comparing our SNP calls with SNPs in dbSNP that were called from unpublished low-coverage sequence generated independently via two separate library preparations for LG/J (~18X) and SM/J (~14X) (Table 3).

We compared SNPs identified in LG/J and SM/J to those identified in other mouse strains sequenced at high coverage and identified novel variants (Table 1). Pairwise distances were calculated for LG/J and SM/J from each of these strains and a phylogenetic tree was constructed using all 20 mouse sequences (Additional files 8 and 9). The LG/J and SM/J ancestries are expected given what we know about their origins and previous studies of inbred mouse phylogenies [71]. It has been suggested that the SM/J strain may be related to the current DBA strains, DBA/1J and DBA/2J, as one of the seven strains contributing to the base population that was selected for smaller body weight was a dba strain named after its dilute brown agouti coat color [72]. Comparison of SNPs between LG/J or SM/J and the sequenced DBA/2J shows no specific relationship between SM/J and DBA/2J. However, the percent shared SNPs between inbred mouse strains can be biased by the use of C57BL/6J as the reference genome since variants may be missed due to alignment quality favoring bases that match the reference.

Functional predictions

Deleterious amino acid mutations in LG/J and SM/J were predicted using three independent methods: LRT [73], PolyPhen-2 [74] and SIFT [75] as implemented in VEP [76]. The LRT compares the probability that a
A codon has evolved under a conserved model to the probability that a codon has evolved under a neutral model. The conserved model allows a codon to have evolved under negative selection and the neutral model assumes the rates of nonsynonymous and synonymous mutations are not significantly different. PolyPhen-2 uses both sequence- and structure-based predictive features to characterize the functional importance of a mutation, making use of non-redundant protein databases. SIFT is based on the principles of protein evolution and has been applied to a variety of organisms ranging from bacteria to humans. The algorithm is based on sequence homology and uses a median conservation score to measure protein conservation. The numbers of deleterious amino acid predictions for each strain are listed in Table 4 and predicted scores are provided in Additional file 10. The intersections of these predictions for SNPs that are polymorphic between LG/J and SM/J are illustrated in Figure 2. It is worth noting that the three different methods give largely non-overlapping results, which is consistent with a similar analysis using human genomes [73]. This result implies that our ability to reliably annotate whole genome sequences is still somewhat lacking.

In addition to SIFT predictions, VEP identifies variants falling in non-coding and potentially regulatory regions that may impact a gene’s expression. An overview of potentially functional regulatory variants is listed in Table 4 and positions and scores are provided in Additional files 11 and 12.

### LG/J, SM/J identical by descent regions

We identified genomic regions that are highly conserved between LG/J and SM/J using a hidden Markov model (Figure 1 and Additional file 13). These regions are defined as stretches of sequenced DNA ≥50kb in length containing little to no polymorphism between the strains. These regions are likely to be identical by descent (IBD), and are non-randomly distributed throughout the genome, being more clustered than expected by chance (p < 2.2e-16, Wald-Wolfowitz test). We classified 39% of the LG/J and SM/J sequenced genomes as IBD.

### LG x SM QTL and whole-genome sequences

We integrated all of these data – IBD regions, polymorphisms and functional predictions – with 272 published QTL for metabolic traits (obesity, diabetes and serum-lipids) and for bone-growth traits (femur, humerus, radius and tibia length) mapped in very advanced intercrosses of LGxSM [37,48-51]. LOD scores range from 2.4 to 11.89 and are significant for their respective study. Additional file 14 describes these QTL, including their genomic coordinates, and their integration with various sequence characterizations. Figure 1 illustrates these QTL in relation to the sequence parameters we have generated. We find that 27% of all QTL interrogated in this study are covered by IBD bases, whereas 39% would be expected if they were randomly distributed. We find that both trait-specific QTL peaks (the point of the QTL with the lowest probability of a chance association with a specific trait) and QTL-specific peaks (accounting for pleiotropy, where more than one trait maps to a locus) are over-represented in non-IBD regions (χ² = 42.6, df = 1, p = 6.6x10^-11 for trait-specific peaks and χ² = 16.8, df = 1, p = 4.03x10^-65 for QTL-specific peaks). Further, we find that QTL regions are more likely to harbor variants that are predicted to have a functional impact by comparing the number of amino-acid changing variants predicted to be damaging by at least one algorithm and the number of variants predicted to fall in high-impact positions of transcription factor binding sites to empirical distributions of numbers generated from 1000 sets of randomly chosen non-QTL, non-IBD regions of similar size (p < 2.2e-16 for both amino acid changing and TFBS variants, Additional file 15).

Integrating these data provides an opportunity to examine variation between LG/J and SM/J from both bottom-up (genomic sequence to phenotype) and top-down (phenotype to genomic sequence) perspectives with the goal of identifying plausible quantitative trait nucleotides (QTN) for testing mechanistic hypotheses. To illustrate a bottom-up approach, we focus on a SNP identified in the SM/J strain (chr13:104,041,257 A → C) falling in a cMyb TFBS (JASPAR ID MA01001) with a predicted high-impact (Figure 3). This position overlaps regulatory elements including H3K36 and H3K4 histone marks and a DNase1 hypersensitive site (ENSMUSR00000276453). The SNP falls in an intronic region of the oligopeptidase

---

**Table 1 Sequence and variants overview**

<table>
<thead>
<tr>
<th>Strain</th>
<th>Total reads</th>
<th>% Mapped to reference</th>
<th>Coverage</th>
<th>SNPs</th>
<th>Strain-Specific SNPs*</th>
<th>Indels</th>
<th>Strain-Specific Indels*</th>
<th>SVs</th>
<th>Strain-Specific SVs*</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG/J</td>
<td>9.6x10⁸</td>
<td>91.72</td>
<td>35X</td>
<td>4,663,723</td>
<td>146,814</td>
<td>1,127,072</td>
<td>56,469</td>
<td>12,802</td>
<td>1,361</td>
</tr>
<tr>
<td>SM/J</td>
<td>7.7x10⁸</td>
<td>92.67</td>
<td>30X</td>
<td>4,792,568</td>
<td>227,132</td>
<td>1,127,458</td>
<td>65,935</td>
<td>15,564</td>
<td>2,255</td>
</tr>
</tbody>
</table>

*Refers to strain-specific variants with respect to 18 other available mouse genome sequences.

---

**Table 2 Genomic overview of structural variants**

<table>
<thead>
<tr>
<th>Variant</th>
<th>LG/J</th>
<th>SM/J</th>
<th>Polymorphic*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletion</td>
<td>7,697</td>
<td>9,889</td>
<td>11,206</td>
</tr>
<tr>
<td>Insertion</td>
<td>5,038</td>
<td>5,590</td>
<td>7,698</td>
</tr>
<tr>
<td>Inversion</td>
<td>94</td>
<td>159</td>
<td>184</td>
</tr>
</tbody>
</table>

*Refers to polymorphisms between LG/J and SM/J.
neurolysin, Nln (NM_029447). Nln knockout mice have been shown to be more insulin sensitive and glucose tolerant and to have increased gluconeogenesis relative to littermate controls [77]. In mammals, the liver is the main site of gluconeogenesis. A microarray analysis of hepatic gene expression shows Nln to be highly significantly differentially expressed between LG/J and SM/J, with the SM/J strain’s expression barely detectable [78]. This gene falls within the support intervals of a QTL associated with basal serum glucose levels in an F_{16} advanced intercross between LG/J and SM/J and with QTL for bone-growth traits mapped in an F_{34} advanced intercross between these two strains.

Table 3 Comparison of high-coverage SNPs to low-coverage sequence in dbSNP

<table>
<thead>
<tr>
<th></th>
<th>LG/J</th>
<th>SM/J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of matched positions and variants</td>
<td>3,693,085</td>
<td>3,798,743</td>
</tr>
<tr>
<td>Number of mismatched variants at matched positions</td>
<td>20,266</td>
<td>44,582</td>
</tr>
<tr>
<td>Number of SNPs in dbSNP not called in high coverage genomes</td>
<td>411,859</td>
<td>391,672</td>
</tr>
<tr>
<td>New SNP variants identified in high coverage sequence</td>
<td>942,372</td>
<td>941,383</td>
</tr>
</tbody>
</table>

Figure 1: Circos plot illustrating the integration of polymorphic SNVs, SVs, and regions that are predicted to be identical-by-descent (i.e. harboring little to no variation) between the LG/J and SM/J inbred strains. These data are combined with QTL for metabolic traits mapped in an F_{16} advanced intercross between LG/J and SM/J and with QTL for bone-growth traits mapped in an F_{34} advanced intercross between these two strains.
gene (QTG) and this SNP is an attractive quantitative trait nucleotide (QTN) for further mechanistic studies of Nln involvement in glucose metabolism.

To illustrate a top-down approach, we focus on a QTL mapped in an F34 generation of the LGxSM AI that is associated with variation in tibia length at chr6: 20,650,821-23,746,386 (Figure 4). There are 12 protein coding genes and 10 RNA genes falling within the QTL support interval [37]. Two of these genes, Wnt16 (NM_053116) and Ptprz1 (NM_001081306), are involved in limb development and/or bone formation [79,80]. Variants in a third gene, Cped1 (NM_001081351), have recently been associated with variation in bone mineral density in a GWAS [81]. 23% of this QTL is covered by bases falling in IBD regions, but these three genes are located in non-IBD portions of the QTL. Three nonsynonymous SNPs fall in Cped1 and are predicted to have functional consequences by at least one of the algorithms used: LRT, PolyPhen or SIFT. A fourth nonsynonymous SNP is predicted to have functional consequences by all three algorithms making it a highly attractive QTN. The SNP falls in exon 14 of Ptprz1 (ENSMUSE00000619494, position 23016230 C → A; P1676H). This exon is highly conserved. Sanger sequencing confirmed the SM/J variant and the amino acid changing polymorphism between the strains. Ptprz1 encodes the protein R-PTP-Z, which is thought to modulate osteoblast metabolism through dephosphorylating Src, which plays a key role in osteoblast activities such as adhesion and differentiation [82,83]. The amino acid change occurs in the SM/J strain, and the only other strain of sequenced mouse carrying this variant is NZO/HILtJ. NZO/HILtJ is a common laboratory strain used as a model of metabolic syndrome because of its extreme obesity and hyperglycemic phenotype [84]. It bears no special relationship with SM/J, despite sharing this particular variant and a similar metabolic phenotype on a high-fat diet.

Conclusions

Here we describe the whole-genome sequences for the LG/J and SM/J inbred mouse strains. LG/J and SM/J are frequently compared in QTL mapping studies because of their great phenotypic diversity, and because this diversity is normally distributed in intercrosses of these two strains. This makes LG x SM an ideal model system for studying the genetic architecture of normal variation in complex traits, because this most closely mimics that found in human populations, where most variation is the result of many interacting genes of individually small effects. Comparison with previous imputation methods using SNPs called from low-coverage sequence for LG/J and SM/J indicated that there were many unreported variants in these strains [51,85]. Our sequence provides higher resolution, which allows us to capture more of the genomic sequence variation found in these strains with greater certainty.

Because we are interested in identifying candidate variants for functional follow-up, our first order of business was to identify regions of the LG/J and SM/J genomes that do not vary between the strains. Identification of IBD genomic regions is a powerful way to narrow existing QTL support intervals, as the regions that do not vary between the strains are unlikely to cause variation in the mapped trait. Further, identification of IBD and non-IBD regions between two strains is a powerful tool for testing mechanistic hypotheses and for planning focused candidate gene studies [86]. We classified 39% of the LG/J and SM/J sequenced genomes as IBD, and we are able to use this classification to narrow QTL support intervals by orders of magnitude when eliminating stretches of DNA that contain little to no variation between the parental strains within QTL support intervals.
Eliminating these regions does not explicitly rule them out as harboring potentially causal variants, however focusing on the non-IBD sequence within QTL support intervals -- especially on the variants predicted to have functional consequences -- allows one to prioritize the so-called 'low-hanging fruit'. Other commonly used strain pairs, such as C57BL/6J and DBA/2J, have been sequenced but have not been subject to such close evaluation of IBD genomic regions within QTL.

We integrated the rich source of QTL results generated by very advanced intercrosses of LG x SM for metabolic traits (231 QTL associated with variation in obesity, diabetes and serum-lipids) mapped in the F\textsubscript{16} and for bone-growth traits (41 QTL associated with variation in femur, humerus, radius and tibia) mapped in the F\textsubscript{34} generations with the LG/J and SM/J genomic sequences. The median QTL interval in the F\textsubscript{16} data is \approx 3.5mB and for the F\textsubscript{34} it is \approx 2.0mB (Additional file 14). We find that \approx 29\% of F\textsubscript{16} and \approx 19\% of F\textsubscript{34} QTL intervals are covered by IBD regions. Subtracting these IBD regions from QTL reduces the genomic search space for both generations, resulting in a comparable \approx 1.2mB median number of per QTL bases for both F\textsubscript{16} and F\textsubscript{34} generations. We find that QTL from both generations have proportionally equivalent numbers of SNVs, bases covered by SVs, and bases predicted to have functional consequences (Additional file 14). Further, we find that QTL regions as a class are more likely to harbor variants predicted to have functional consequences relative to randomly selected non-QTL and non-IBD genomic regions of similar size (Additional file 15).

The traits associated with the QTL interrogated here belong to different classes of quantitative phenotype, namely metabolic traits mapped in the F\textsubscript{16} generation and developmental traits mapped in the F\textsubscript{34} generation. Having QTL for the same trait replicate in multiple generations of a mapping study increases the probability that the QTL is causal. However, given the high costs (both financial and time-wise) of breeding and maintaining large populations

Figure 3 Connecting a putative quantitative trait nucleotide (QTN) with a candidate quantitative trait gene (QTG). A: A SNP identified in the SM/J strain (chr13:104041257 A→C) falls in a highly informative position (position 2) of a predicted cMyb TFBS. B and C: This SNP falls in an intron of Nln, a protein-coding gene associated with gluconeogenesis. Nln falls in a QTL associated with variation in basal glucose levels in an F\textsubscript{16} generation of a LG x SM advanced intercross and is highly significantly differentially expressed between the LG/J and SM/J strains. C: This variant overlaps multiple regulatory elements and is a strong candidate for mechanistic studies of Nln function in glucose metabolism.
Figure 4 (See legend on next page.)
of mice for many generations, incorporating whole-genome sequence parameters as we have illustrated here can be used to extract compelling information from QTL mapped in earlier generations of intercrosses, even when the support intervals span many mega-bases of sequence. Thus whole-genome sequence data should be made part of the toolkit used to inform the design, execution and follow-up of QTL mapping studies.

We have highlighted variants falling in a QTL on chromosome 13 associated with basal glucose levels and a QTL on chromosome 6 associated with tibia length (Figures 3 and 4) to illustrate how all of these data – whole genome sequence variants, IBD regions, QTL, and functional predictions – can be integrated with each other and with other public datasets to identify and prioritize QTG and QTN. Identifying such variants within QTL in mice can facilitate translational research for correlated traits in human studies, and potentially uncover genetic underpinnings of disease phenotypes [87,88]. Our initial description and analysis of the LG/J and SM/J whole genomes offers new data that can be used to address fundamental questions about the molecular nature of quantitative phenotypic variation in an important model system.

**Methods**

**Ethics statement**

All animal care and handling procedures conformed to IACUC guidelines.

**DNA isolation and library construction**

DNA was isolated from the livers of one adult female LG/J and one adult female SM/J mouse using the Qiagen DNAeasy Blood and Tissue kit (Qiagen, West Sussex, UK). Genomic DNA was sonicated to an average size of 175 bp. The fragments were blunt ended, had addition of “A” base to 3’ end, and had Illumina’s sequencing adapters ligated to the ends. The ligated fragments underwent amplification incorporating a unique indexing sequence tag. The resulting libraries were sequenced on 6 lanes using the Illumina HiSeq-2500 as paired end reads extending 101 bases from both ends of the fragments.

**Alignment and SNV detection**

Reads were aligned to the GRC m38 (mm10) mouse reference genome using NovoAlign-2.08.02 and a BAM file was produced for each strain (http://www.novocraft.com).

The gene annotation model used was Ensembl *Mus musculus* GRC 38.72. Variants were called from each strain’s BAM file using two variant discovery tools: SAMtools pileup and FreeBayes Bayesian genetic variant detector [89,90]. Results were merged and only variants with a minimum read depth of 3 and a quality score of at least 20 for SNPs and at least 50 for indels were included in the final set of SNVs for analysis.

**Quality assessment and Sanger sequencing candidate variant**

SNPs identified for LG/J and SM/J were compared to SNPs called from independently generated, low-coverage sequence available on dbSNP [91] using custom python scripts. DNA was isolated from the livers of 4 female LG/J and 4 female SM/J animals to validate the candidate amino acid changing mutation P1676H in exon 14 of Ptprz1. The following primers were designed: Forward 5’-GCT CCA TGG CCA CTA TCT TTA CTC-3’ and reverse 5’-CAA TTC ATG CCT CAA GGT GAC TGC-3’. Sanger sequencing was performed by Genewiz (South Plainfield, NJ).

**Comparison with variants identified in other sequenced mouse strains**

VCF files for SNPs discovered in 18 mouse strains’ whole-genomes were downloaded (http://www.sanger.ac.uk/resources/mouse/genomes) and compared to LG/J and SM/J SNPs using custom python scripts.

**Structural variant prediction**

Structural variants (SV) were identified using a local implementation of the SVmerge pipeline [92]. SVmerge simplifies SV discovery and integration using multiple SV discovery algorithms, producing a single consensus SV callset. The SVmerge pipeline was run with 4 calling algorithms: Breakdancer 1.0, pindel 0.2.4q, SECluster and cnD [92-95]. After merging the results of these 4 callers into a single consensus callset, the Velvet genome assembler was used to attempt breakpoint assembly for all SVs in the callset [96].

**Deleterious mutation prediction**

**LRT**

Deleterious amino-acid changing mutations were predicted using a likelihood ratio test (LRT) as described in...
Chun and Fay (2009). The algorithm was modified to compare the genomes of LG/J and SM/J each to the reference C57BL/6 J. P-values were calculated by comparing twice the log-likelihood ratio of the two models to a \( \chi^2 \) distribution with \( df = 1 \). Deleterious amino acid changing mutations were predicted using an LRT cutoff of \( p < 0.001 \) while controlling for false positives and false negatives as previously described and modified for mouse genomic sequence [73].

**PolyPhen-2**

PolyPhen-2 was downloaded and run locally to predict deleterious mutations in the LG/J and SM/J genomes. The algorithm supports analysis of mouse proteins using prebuilt human models if the option ‘-n mouse’ is specified when preparing the local copy of UniProtKB (www.uniprot.org). Deleterious amino acid changing mutations were predicted for both strains using a recommended false discovery rate (FDR) cutoff of 20% [74].

**VEP and SIFT**

VEP was downloaded and run locally to predict deleterious mutations in the LG/J and SM/J genomes using alignments built using the TrEMBL 39.8 protein database. Deleterious amino acid changing mutations were predicted for both strains using a score of \( < 0.05 \) and a median conservation cutoff of 3.25. Regulatory annotations are based on the Ensembl regulatory build, which integrates data from ENCODE and several other large-scale projects. Transcription factor binding site (TFBS) variant scoring is provided for regulatory regions that have ChIP-seq data to support binding predictions. This is done using the MOODS software [97], which assigns significance scores by matching polymorphisms against motifs in the JASPAR database [98]. The output was filtered for TFBS sites that are classified as ‘Highly Informative’ by the software.

**IBD region identification**

To identify regions of shared ancestral background, we clustered segments of observed polymorphisms using a two-state hidden Markov model. For each state, we modeled two types of observations: 1) the number of polymorphisms in a 50kb window; and 2) the observation of a SV in a 50kb window. The count of polymorphisms is expressed as a Poisson variable, while the occurrence of a SV is a Binomial variable. Parameters for this model were estimated using the EM algorithm implemented in the depmixS4 package in the R programming language [99]. A Wald-Wolfowitz test of randomness was performed using the adehabitatLT package in R [100].

**Data integration**

The QTL interrogated in this study are from previously published studies that mapped to the mm9 (NCBI build 37) mouse reference genome. QTL peaks and support intervals were converted to mm10 (GRCh38) using the Batch Coordinate Conversion (liftOver) tool [101]. Sequence data and results generated here were integrated with other published and publically available data as indicated using custom Python and R scripts.

**Data access**

We have made these data available to the community through multiple data portals: LG/J (SAMN03075510) and SM/J (SAMN03075514) raw reads have been submitted to the NCBI Sequence Read Archive and BAM files have been submitted to the Wellcome Trust mouse genomes portal (http://www.sanger.ac.uk/resources/mouse/genomes). Other results from this study are available as Additional files 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15.

**Additional files**

- **Additional file 1:** LG/J, SM/J and LG/J x SM/J polymorphic SNPs overview.
- **Additional file 2:** LG/J, SM/J and LG/J x SM/J polymorphic indels overview.
- **Additional file 3:** SNPs in LG/J discovered by integrating output from SamTools pileup and FreeBayes variant discovery tools.
- **Additional file 4:** Indels in LG/J discovered by integrating output from SamTools pileup and FreeBayes variant discovery tools.
- **Additional file 5:** SNPs in SM/J discovered by integrating output from SamTools pileup and FreeBayes variant discovery tools.
- **Additional file 6:** Indels in SM/J discovered by integrating output from SamTools pileup and FreeBayes variant discovery tools.
- **Additional file 7:** Structural variants found in LG/J and SM/J.
- **Additional file 8:** Distances between the LG/J and SM/J strains and other sequenced strains.
- **Additional file 9:** Phylogenetic tree of 20 sequenced mouse strains.
- **Additional file 10:** Amino acid changing mutations and corresponding deleterious predicted scores using SIFT/polyphen/LRT for both LG/J and SM/J strains.
- **Additional file 11:** VEP output of variants in predicted regulatory features for both LG/J and SM/J.
- **Additional file 12:** VEP output of variants in predicted regulatory features for both LG/J and SM/J.
- **Additional file 13:** Predicted IBD regions in LG/J and SM/J sequenced genomes.
- **Additional file 14:** QTL metrics and sequence integration.
- **Additional file 15:** Distribution of the number of variants predicted to have functional consequences in QTL intervals and in 1000 sets of randomly selected non-QTL regions of similar size.

**Abbreviations**


**Competing interests**

The authors declare that they have no competing interests.
Authors’ contributions
IN, DC, SC, JF, and HL generated and analyzed data. JC and HL wrote the manuscript. All authors edited and approved the final manuscript.

Acknowledgements
This work was supported by NIH-NIDDK K01 DK095003 and P30 DK056341 to HAL and in part by NIH R01 MH10180 to DFC and by NIH-NCCR ULI RR20492 to the genome technology access center at Washington University in St Louis. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the national institutes of health.

Author details
1Department of Genetics, Washington University School of Medicine, Campus Box 8108 660 S Euclid Ave, St Louis, MO 63110, USA. *Division of Genetics, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA, USA. 2Department of Biology, Loyola University, Chicago, IL, USA.

Received: 28 October 2014 Accepted: 28 April 2015

Published online: 28 May 2015

References
42. Cheverud JM, Ehrich TH, Kenney JP, Pletscher LS, Semenkovich CF. Genetic evidence for discordance between obesity- and diabetes-related...


