HEALTH AND MEDICINE

The genome of *Peromyscus leucopus*, natural host for Lyme disease and other emerging infections

Anthony D. Long[1]*, James Baldwin-Brown[1,2], Yuan Tao[1], Vanessa J. Cook[3], Gabriela Balderrama-Gutierrez[4], Russell Corbett-Detig[5], Ali Mortazavi[4], Alan G. Barbour[3]*

The rodent *Peromyscus leucopus* is the natural reservoir of several tick-borne infections, including Lyme disease. To expand the knowledge base for this key species in life cycles of several pathogens, we assembled and scaffolded the *P. leucopus* genome. The resulting assembly was 2.45 Gb in total length, with 24 chromosome-length scaffolds harboring 97% of predicted genes. RNA sequencing following infection of *P. leucopus* with *Borreliella burgdorferi*, a Lyme disease agent, shows that, unlike blood, the skin is actively responding to the infection after several weeks. *P. leucopus* has a high level of segregating nucleotide variation, suggesting that natural resistance alleles to Crispr gene targeting constructs are likely segregating in wild populations. The reference genome will allow for experiments aimed at elucidating the mechanisms by which this widely distributed rodent serves as natural reservoir for several infectious diseases of public health importance, potentially enabling intervention strategies.

INTRODUCTION

Peromyscus leucopus is the major reservoir for several infectious diseases in North America

The white-footed mouse *Peromyscus leucopus* is a widely distributed, abundant rodent in eastern and central United States and adjoining regions of Canada and Mexico. The species is a major reservoir or carrier for several tick-borne diseases, including the bacterial infections Lyme disease, anaplasmosis, and *Borrelia miyamotoi* relapsing fever; the malaria-like protozoan disease babesiosis; and a fatal or disabling viral encephalitis (*1*). The genus *Peromyscus* also includes the major hantavirus reservoir *Peromyscus maniculatus*, the North American deer mouse. Peromyscines cluster with hamsters, voles, and wood rats rather than murids such as *Mus musculus* and *Rattus norvegicus* (*2*). While inbred *M. musculus* has been the animal model of choice for experimental studies of Lyme disease and other infections, the house mouse is not a natural reservoir for these infections. Moreover, it differs from *P. leucopus* in manifestations of and responses to infection (*1*, *3*), but the genetic traits of *P. leucopus* that distinguish this species from *M. musculus* in this respect and make it a competent reservoir for a variety of pathogens are not known.

Lyme disease and associated zoonoses continue to increase in incidence and to spread to previously unaffected areas in North America (*4*). While antibiotics are available for treatment, there are neither human vaccines nor broadly implementable tick control measures to prevent infections. Given the key position of *P. leucopus* in the life cycles of both the tick vector and several pathogens (Fig. 1), transmission-blocking field vaccines (*5*) and gene-drive technologies to render wild *Peromyscus* resistant to infection (*6*) are increasingly seen as plausible ways to prevent human disease.

[1]Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA, USA. [2]Department of Biology, University of Utah, Salt Lake City, UT, USA. [3]Departments of Microbiology and Molecular Genetics and Medicine, University of California, Irvine, Irvine, CA, USA. [4]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA, USA. [5]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA.
*Corresponding author. Email: tdlong@uci.edu (A.D.L.); abarbour@uci.edu (A.G.B.)

RESULTS

A hybrid PacBio/Illumina assembly coupled with Hi-C scaffolding yields an assembly of chromosome length scaffolds

While attention on *P. leucopus* is justified, the dearth of genetic information and the lack of a reference genome have limited progress toward these and other goals. Accordingly, we generated Illumina and PacBio genomic DNA sequencing datasets, assembled the genome using both the hybrid and a PacBio-only approach, and merged the assemblies (*7*). This resulting assembly is 2.45 Gb in total length, with a contig N50 of 4.38 Mb. Our contigs cover between 84 and 89% of the genome, as we estimate the genome size of *P. leucopus* to be 2.7 to 2.9 Gb based on kmer counts. Cumulative contiguity plots of different assembly strategies (fig. S1) show that the Illumina/PacBio hybrid assembly is superior to the PacBio-only "s e l f"-assemblies, with the quick-merged "hybrid-hybrid" assemblies giving the highest contiguity.

We used two independently constructed Hi-C libraries with >1000× of total read span coverage for mates 10 to 200 kb apart (fig. S2) to create 24 chromosome length scaffolds representing the *P. leucopus* genome. We evaluated different scaffolders, switches, and ways of integrating over libraries and concluded that the 3d-dna scaffolder (*8*), increased total sequence coverage, and combination of Hi-C libraries gave the most robust scaffolds (fig. S3), where we use robust to mean less sensitive to read downsampling. Figure 2A and table S1 summarize the scaffolded assembly based on the expected $n = 24$ chromosomes (*9*). The 24 chromosome-sized scaffolds contain 99.5% of human coding sequences that we could align to our assembly (using Exonerate) and 3795 of 3798 (92.5%) of complete and single-copy BUSCOs (an annotated collection of genes present in all mammals), suggesting that the chromosome-sized scaffold in our assembly represents much of the euchromatic genome (fig. S4; see Materials and Methods). We compared our assembly to a genetic linkage map of *P. maniculatus* (*9*), hereafter referred to as "Kenney-Hunt." Of the 196 markers in the genetic map, 134 markers were unambiguously mapped to our chromosome length scaffolds, with 128 of the markers mapping in a consistent manner (fig. S5). On the basis of the mapping of markers and house mouse proteins aligned to scaffolds, we assigned chromosome names to our scaffolds.

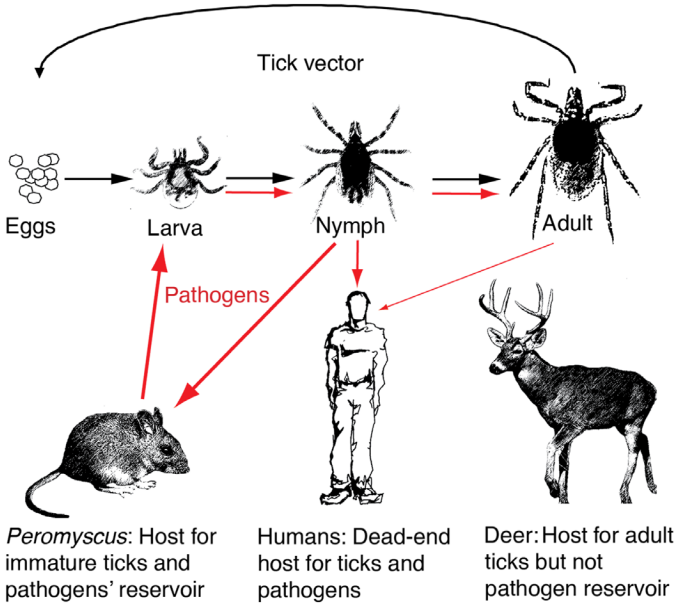

Fig. 1. Life cycle of the Lyme disease agent *B. burgdorferi*, as well as other pathogens, and its tick vector *Ixodes scapularis* in North America.

Gene-based linkage markers from a *P. maniculatus* map and synteny with rat and mouse allow assignment of scaffolds, as well as 97% of predicted genes, to named chromosomes

Regardless of the scaffolding approach used, chromosome 8 was consistently split into two scaffolds (8a and 8b) and chromosomes 16 and 21 were consistently fused. The fusion event involves genomic regions that show Hi-C contacts between the two chromosomes (fig. S6) and involve regions syntenic with rat chromosomes 20 and 9 (fig. S7). This splitting of chromosome 8 is supported by a recently updated *P. maniculatus* linkage map (*10*), referred to as "Brown" (table S1). Similarly, the joining of two chromosomes is also supported by Brown, although our assembly joins chromosomes 16 and 21, whereas Brown joins chromosomes 16 and 20. Blasting of proteins associated with gene marker names supports the fusion of chromosomes 16 and 21. Our assembly is largely consistent with current chromosome names based on *P. maniculatus* linkage maps.

An analysis of synteny between *P. leucopus* and *M. musculus* based on the positions of 62,094 orthologous gene pairs demonstrates conserved synteny for species that respectively represent the cricetine and murid families of the order Rodentia (Fig. 2B). The synteny analysis further suggests that errors in Hi-C scaffolding of the genome are likely local events and only rarely involve the transposition of contigs between chromosomes. We carried out a similar syntenic analysis with rat (fig. S8) and observe that if we depict *P. leucopus* chromosomes in the order of 3d-dna scaffold number (fig. S9), then there are several examples of telomeric regions transposed between adjacent scaffolds that are likely artifacts of the Hi-C scaffolding approach.

We identified genes using different strategies. We generated 16 RNA sequencing (RNA-seq) datasets from 11 different adult tissues, two embryo stages, two 1-day-old pups, and from both males and females (table S4). Assembling the RNA-seq datasets using Trinity (*11*) and carrying out gene predictions using Augustus (*12*) identify 609,192 transcripts. We further directly aligned the RNA-seq data to the genome using HiSAT2 (*13*) and used Strawberry (*14*) to identify 1,252,424 transcripts. We lastly used Exonerate (*15*) to align both the house mouse and the human proteome to the *P. leucopus* genome (52,756 and 112,693 alignments). The transcript predictions qualitatively agreed between methods, with subtle differences in annotations represented as University of California, Santa Cruz genome browser tracks (http://goo.gl/LwHDr5)). We took the concatenated set of transcripts from the four methods and supplied them to Augustus as a set of hints, which resulted in 19,896 gene predictions (table S3). Of these genes, 19,297 (or 97.0%) are located on the 24 chromosome-sized scaffolds.

P. leucopus harbors fewer repetitive elements but as large a repertoire for antigen recognition as the house mouse

Application of RepeatMasker flagged one-third of the genome as repetitive, with LINE1 repeats being the most abundant, occupying 10% of the genome (table S2). While the overall genome of *P. leucopus* at 2.45 Gb is 13% smaller than house mouse at 2.82 Gb (GRCm38.p6), the nonrepeat portion of the genomes are nearly identical in size (Fig. 3C). While the house mouse has more repeats overall, several LINE1 repeat subfamilies have higher copy numbers in *P. leucopus*. As expected for a rodent, *P. leucopus* is highly enriched for the *mys* family of Long Terminal Repeat (LTR) retrotransposons (Fig. 3B) (*16*, *17*).

Analysis of the prevalence of protein domains in *P. leucopus* using Pfam identified one or more protein domains in 19,056 (90%) of the genes predicted via the alignment of house mouse proteins to the genome. A comparison of Pfam domains in *P. leucopus* and house mouse (Fig. 3A) showed that while most families display nearly identical prevalence, there are 10 families that are depleted and 7 that are enriched in *P. leucopus* (chi-square, $P < 0.05$). For example, the KRAB domain, which is a repressive domain that is found in transcription factors regulating endogenous retroviruses in Muridae (*18*), is depleted in *Peromyscus*. Other depleted domains include C1_4 and C1_1, which bind phorbol ester to protein kinase C and are associated with some zinc-finger proteins (*19*).

On the other hand, some zinc finger domains are highly enriched in *Peromyscus*, suggesting expansion by duplication. This type of expansion has been noted in other zinc-finger proteins, like Prdm9 (*20*). Another expanded domain, BolA, is associated with stress responses and is believed to regulate many genes (*21*). The expanded domains zf-met, zf-Di19, and zfLYAR are involved in RNA binding, stress response to dehydration, and cell growth regulation, respectively. The conservation in numbers of domains for immune effectors, like cytokineas (e.g., TNFα), immunoglobulin domains (Ig4 and Ig5), and complement (C2 and C4), indicate that the differences between *P. leucopus* and *M. musculus* in response to infection are more likely attributable to the differences between the species transcriptional factors. The α and β domains for class II major histocompatibility complex proteins are slightly enriched in *P. leucopus*, suggesting a capacity for antigen recognition as great, if not greater, than that of *M. musculus*.

Experimental Lyme disease in *P. leucopus* shows that during persistent infection, differentially expressed genes are far more numerous in the skin than in the blood

Inasmuch as few reagents, such as antibodies to cytokines or surface markers of white cells, exist for *Peromyscus* species, the reference genome allows for more comprehensive analyses of host responses to infection or different environmental conditions than hitherto possible. As a first application, we infected six *P. leucopus* animals

A

	Draft contigs	Final assembly	Chromosome length scaffolds	Small scaffolds	Tiny scaffolds
Total base pairs	2,474,055,010	2,475,164,510	2,405,334,684	66,909,757	2,920,069
% Total	100%	100%	97.18%	2.70%	0.12%
Number of contigs/scaffolds	4216	1856	24	1170	662
Min length	1,000	1,000	46,490,564	10,022	1,000
Mean length	586,825	1,333,602	100,222,279	57,188	4,411
N50 length	3,921,871	114,273,790	NA	85,932	8,455
Max length	22,229,712	193,658,164	193,658,164	861,000	10,000

B

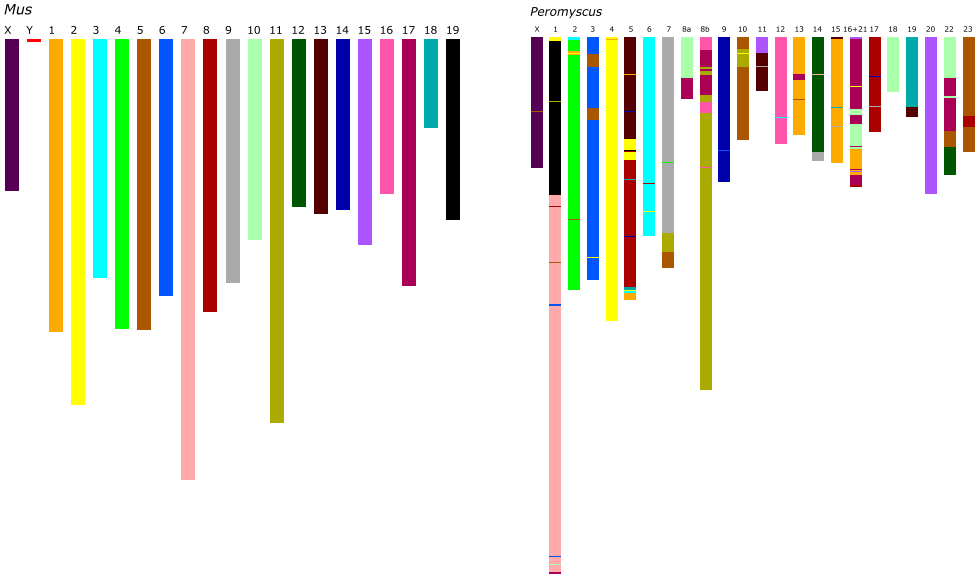

Fig. 2. *P. leucopus* assembly statistics and synteny with house mouse. (**A**) Hybrid assembly and Hi-C scaffolding summary statistics. (**B**) Syntenic blocks between *P. leucopus* and house mouse. NA, not applicable.

with the Lyme disease agent *Borreliella burgdorferi*. We profiled the responses 5 weeks after infection by RNA-seq and compared these with four animals that had been mock-infected. Previous work showed that, by 5 weeks, spirochetes are cleared from the blood while persisting in the skin (*3*). We confirmed that the six animals inoculated with bacteria 5 weeks before had *B. burgdorferi* in the skin of the ear, with genome copy numbers by quantitative polymerase chain reaction (qPCR) ranging from 500 to 4500/μg of total DNA, while *B. burgdorferi* was undetectable in mock-inoculated control animals.

In the mRNA-enriched blood RNA-seq libraries, 79 (9%) of 8384 above-background transcripts (transcript counts per million reads of >0.1) were differentially expressed (false discovery rate value of <0.05) between infected and uninfected animals (table S6; Fig. 4A). None of the differentially expressed genes (DEGs) in the blood were cytokines, chemokines, acute phase reactants, immunoglobulins, major histocompatibility antigens, Toll-like receptors, or effector enzymes of phagocytes. In contrast, 675 (18%) of 22,362 above-background genes showed differential expression in the skin libraries (table S7; Fig. 4A).

Notable among the skin up-regulated DEGs were genes for several varieties of keratin and keratin-associated protein, as well as the Sonic hedgehog protein, which has a role in hair follicle formation (*22*). In the same set of samples, transcripts for three types of tubulin (Fig. 4B), as well as cullin, desmoplakin, laminin, matrilin, and tensin, which are associated with the cytoskeleton or the extracellular matrix, showed lower expression in infected animals. Some immunoglobulin light chain genes and heat-shock proteins (Hsp70, Hsp90, and DnaJ/Hsp40) were also up-regulated in infected skin, but other genes commonly associated with an innate immunity or inflammation were not. Rather than risking bystander damage from inflammation, *Peromyscus* may contain proliferation and spread of the bacteria by alterations in the skin itself, including the number of hair follicles and extracellular matrix composition. At a standoff between the host and the pathogen, *B. burgdorferi* maintains sufficient density in the skin for efficient transmission to the next round of infesting ticks. Despite high infection prevalences in the wild, white-footed mice do not suffer high rates of morbidity, allowing *P. leucopus* to be one of the most abundant mammals on the continent.

P. leucopus has high levels of nucleotide variation and low levels of linkage disequilibrium in a wild population

To identify single-nucleotide polymorphisms (SNPs) and Insertion/Deletion (INDEL) polymorphisms in this species, we aligned the Illumina reads from the reference animal back to the reference genome. Since the short-read data are obtained from the same outbred diploid individual used for the genome assembly, we expect much of the genome

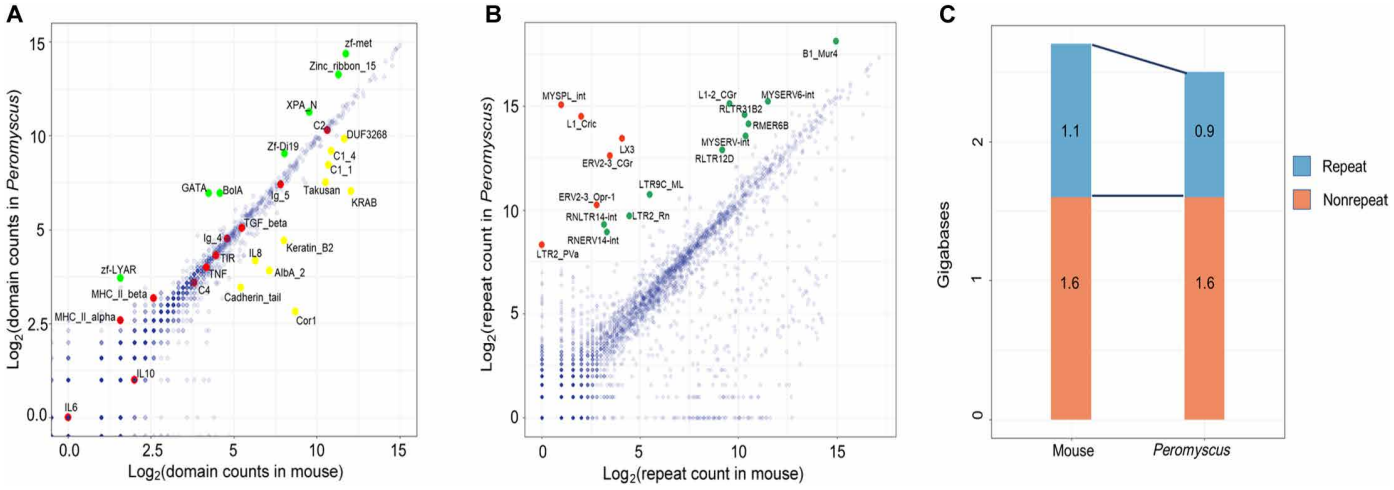

Fig. 3. An analysis of protein domains and repeat in *P. leucopus*. (**A**) Protein domains were scored using Pfam v.31 in both species to identify domains enriched in one species over the other (yellow, enriched in mouse; red, immune-related domains; green, enriched in *P. leucopus*). (**B**) RepeatMasker repeat family counts in *Peromyscus* and mouse. Repeats enriched more than 8-fold in *Peromyscus* are highlighted in green, and repeats more than 100-fold change are highlighted in red. (**C**) The difference in genome size between the *P. leucopus* and mouse is associated with a lower repeat content in *Peromyscus*.

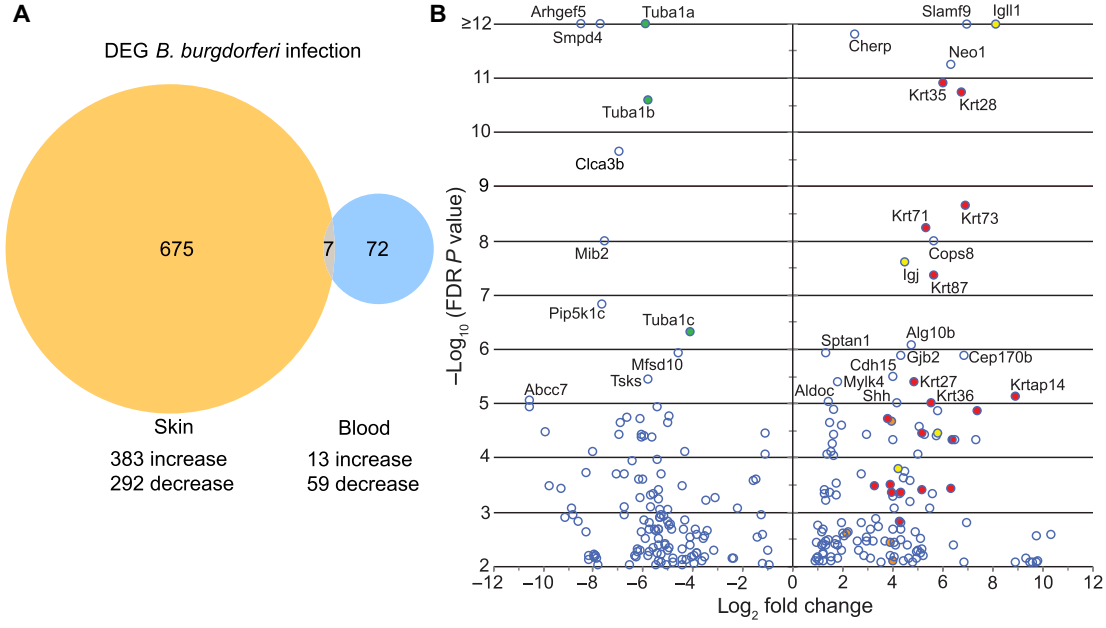

Fig. 4. DEGs in the skin and the blood of *P. leucopus* with or without *B. burgdorferi* infection after 5 weeks. (**A**) Venn diagram of the total number of DEGs [false discovery rate (FDR) $P < 0.05$] in the skin and the blood. The numbers that were increased or decreased in expression in infected animals as compared to uninfected animals are shown (table S6). (**B**) Volcano plot showing DEGs in the skin of the ear between *B. burdgorferi*–infected and uninfected *P. leucopus* at 4 weeks. The plot shows genes with both an absolute fold change of ≥2 ($\log_2$ of the fold change value on the *x* axis) and a false discovery rate *P* value of $<10^{-2}$ ($-\log_{10}$ of the *P* value on the *y* axis). If more than one DEG transcript was associated with a given gene, then the one with the lowest false discovery rate is displayed. Selected gene families are highlighted by color: tubulin (green), keratin (red), immunoglobulin light chains (yellow), and heat-shock proteins (orange). Genes with false discovery rate *P* values of $\leq 10^{-5}$ are labeled according to *M. musculus* nomenclature.

to be heterozygous and estimate a genome-wide SNP heterozygosity of 0.33% based on 9.8×10^6 total SNPs. Similarly, the rate of heterozygosity at INDELs, based on 2.0×10^6 total INDELs, was found to be 0.07%, fivefold less frequent than SNPs. The vast majority of observed SNPs and INDELs detected in the reference individual are true genetic polymorphism versus sequencing errors, since they have frequencies near 50% in the high coverage data (fig. S11). The observed level of heterozygosity in *P. leucopus* is comparable to that of wild *M. musculus castaneus* (~1.1%) and *R. norvegicus* (~0.2%) (*23*), and our estimate of heterozygosity in a *P. leucopus* individual obtained from long-standing closed colony undoubtedly underestimates heterozygosity in the wild. Fixed differences from the reference assembly lastly allow us to place lower bounds on SNP and INDEL error rates in our assembly at 4.8×10^{-6} (13,894 total SNPs).

Additional SNPs and INDELs were identified by carrying out low-pass Illumina sequencing of DNA from blood of 26 captured-and-released animals from a Connecticut population described by Tsao *et al.* (*5*). Average coverage at called SNPs ranged from 0.8× to 8.3× across individuals with a median coverage of 1.9×. At these low coverages, diploid genotype calls can be missing or inaccurate, but after filtering out alleles seen in fewer than three individuals, we still observed ~42 million total SNPs (http://goo.gl/LwHDr5). On the basis of the alignment of human genes to *P. leucopus*, 3817 SNPs are predicted to have a high impact on a gene function, suggesting that future experiments using exome capture are likely to uncover several thousand putative functional variants. Figure 5 depicts the density of SNPs in the sample of wild mice (bottom), along with smoothed heterozygosity in three wild-caught individuals (middle), and three closed colony individuals (top; blue is the reference individual). It is apparent that the reference individual has blocks of homozygosity, consistent with regions being identical by descent (IBD) in this individual from a long-standing colony; a similar pattern is apparent for the individuals genotyped from inbred strains. In contrast, wild-caught mice show many fewer and much shorter regions of IBD, consistent with very low levels of inbreeding in the wild. On the basis of the 42 million observed SNPs and a genome size of ~2.5 Gb, the proportion of bases segregating is 1.7%. This suggests that roughly 25% [Poisson density with $\lambda = 20$ base pair (bp) × 1.7%] of arbitrarily chosen 20-bp Crispr-guide RNAs are likely segregating a common SNP at their target cut site, an undesirable feature for a target candidate (*24*). A more thorough characterization of natural nucleotide variation in this species across its range will aid in the design of efficient gene drive constructs for possible application to this disease reservoir with avoidance of annotated SNPs.

Linkage disequilibrium (LD) falls off very rapidly with distance in our *P. leucopus* wild population (table S8). At 100 to 150 bp, average R^2 values are half of that observed for SNPs within 50 bp of one another, and by 25 kb, R^2 values are close to background. High levels of nucleotide variation in concert with low levels of LD imply that genome-wide association studies would be underpowered in wild *P. leucopus* animals, unless populations could be identified with more extensive LD. Hybrid zones, where *P. leucopus* geographically structured subspecies come in contact (*25*), heterogeneous stock animals, and island populations may be attractive in this regard.

DISCUSSION

A chromosome length collection of scaffolds, gene annotations, and gene lift-overs from human and mouse and a large set of intermediate frequency SNPs will accelerate studies of *P. leucopus* and thereby transmission intervention efforts directed at this species

To empower candidate gene-based studies on *P. leucopus*, we provide gene predictions, variation, and comparative alignments via a University of Santa Cruz genome hub interface (http://goo.gl/LwHDr5)). We highlight the value of an annotated genome and browser interface in fig. S12, which shows a subset of the available browser tracks for interleukin-6 (*IL6*), a proinflammatory and immunoregulatory cytokine important in the responses to viral, bacterial, fungal, and parasitic pathogens, as well as for autoimmune disease and trauma (*26*). A plausible hypothesis capable of explaining *P. leucopus*' tolerance of a pathogen's presence is a restrained inflammatory response. The *IL6* gene is located on chromosome 5 in house mouse and chromosome 3 in *P. leucopus* (table S1). In contrast to humans, the house mouse lacks a single final exon ~20 kb downstream of the remainder of the gene. *P. leucopus* RNA-seq similarly fails to identify this final exon, suggesting that the exon is absent in rodents. This being said, the final exon is conserved in an alignment between *P. leucopus* and humans, suggesting that additional experiments are needed to confirm the structure of this gene. In addition, *P. leucopus* harbors two intermediate-frequency nonsynonymous SNPs in *IL6*, which are predicted by SnpEff based on alignment of human proteins to *P. leucopus* to have a moderate phenotypic effect. Representing the genome on a public, interactive platform like the Santa Cruz Genome Browser for the research community facilitates experimental design and provides for more testing of ideas than has been possible.

For decades, *P. leucopus* and other *Peromyscus* species have been the subject of research in several fields, including aging (*27*), behavior (*28*), development (*29*), epigenetics (*30*), reproduction (*31*), and population genetics (*32*). A public health justification for the study of *P. leucopus* is this species' central position in the life cycles of pathogens and the ticks that transmit them to humans. Of additional relevance is this resilient species' remarkable capacity as an infection reservoir, with little or no obvious increase in morbidity (*1*). Understanding the biology of *P. leucopus* becomes more urgent as Lyme

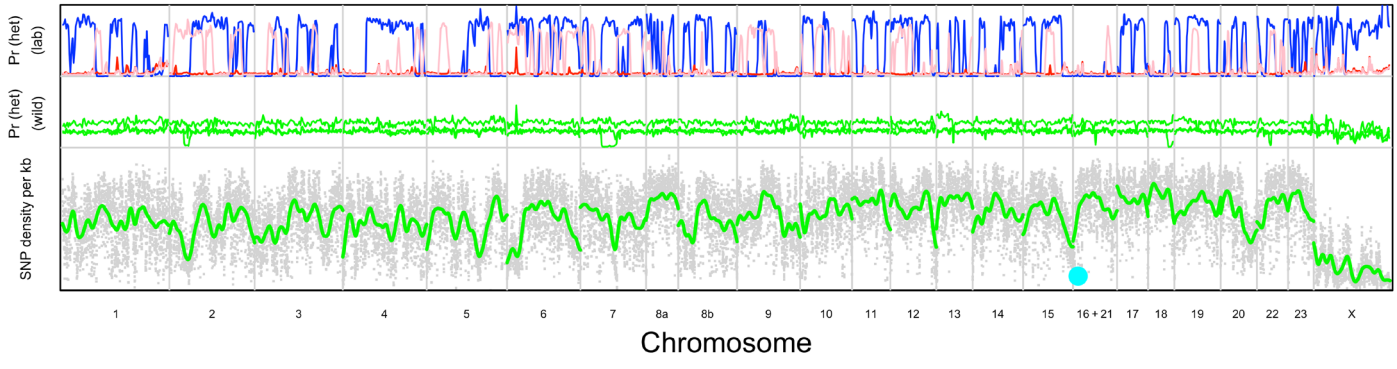

Fig. 5. Nucleotide variation in wild-caught, South Carolina colony and inbred strains of *P. leucopus*. The bottom panel depicts the physical density of SNPs in the low-pass sequenced wild individuals smoothed using a 10-Mb normal kernel. The middle panel depicts the probability of heterozygosity per SNP locus smoothed using a 2-Mb normal kernel for three wild caught animals. The top panel depicts the smoothed probability of heterozygosity per SNP locus for three colony mice. Red, inbred strain GS16A1 from the Peromyscus Genetic Stock Center; pink, outbred stock housed at the Rocky Mountain Laboratories; and blue, the reference individual described in Materials and Methods. In both RMNL and the reference individual, runs of homozygosity are regions that are identical by descent, consistent with those animals coming from closed colonies. The cyan circle is the approximate location of human leukocyte antigen located on the fusion chromosome.

disease increases in frequency and distribution and as control efforts targeting this reservoir, including transmission-blocking vaccines and gene editing, come to the fore.

To this end, we generated a high-quality genome assembly, obtained chromosome length scaffolds using Hi-C, and annotated the genome using RNA-seq data from several tissues and developmental stages and cross-species comparisons. The genome is similar in organization to the murids *M. musculus* and *R. norvegicus*. Heterozygosity levels in a wild population of *P. leucopus* in an area of high risk for Lyme disease in the northeastern United States are similar to those of wild house mouse and brown rat populations.

An annotated chromosome-scaled assembly and a catalog of millions of SNPs substantially expand the ability of *P. leucopus* to serve as a model organism, whose unique attributes for such a role were already appreciated (*32*). This genome enables evolutionary analysis of different populations, facilitates RNA-seq and other "omics" analyses of the system under a variety of experimental and natural conditions, allows for identification of genes that mediate and modulate response to pathogens, and guides the choice of target regions suitable for Crispr-based mutagenic chain reaction constructs for field-based interventions for human disease prevention. We also note that genome-wide comparisons of *P. leucopus* to humans with Lyme disease or other infections are now possible. These cross-species studies may provide insights into why some humans become disabled with *B. burgdorferi* infection while *P. leucopus* does not.

MATERIALS AND METHODS

DNA and library construction

The *P. leucopus* were specific pathogen–free, colony-bred animals of LL Stock from the Peromyscus Genetic Stock Center (PGSC) of the University of South Carolina. The rodents were euthanized by carbon dioxide overdose and intracardiac exsanguination. DNA for library construction was obtained from the liver and kidney, homogenized in liquid nitrogen, and extracted using a Qiagen Blood and Cell Culture kit with modifications suggested in (*7*). For Illumina libraries, DNA was Covaris-sheared to 500 bp and prepped using a Bioo Scientific NEXTFLEX Rapid DNA-seq kit with four cycles of PCR amplification. After size selection, we collected four HiSeq 2500 PE100 lanes for a total of 158 Gb of raw data (or roughly 54×). For PacBio libraries, we followed the study of Chakraborty *et al.* (*7*) by shearing the DNA using a 1.5-inch 24-gauge blunt-tipped needle, and the library was prepared using PacBio's SMRTbell template protocol and Blue Pippin size selection using a 15- to 50-kb cutoff. We generated several dozen libraries in this manner and collected 124 total SMRTcells of data for a total of 104 Gb with an N50 read length of 15.7 kb (~36× coverage). Two Hi-C libraries were also constructed with six cycles of PCR amplification, and PE100 was sequenced to 20× and 8× of raw coverage. On the basis of aligning back to the final assembly, this represented 341,768× and 140,112× of span coverage, respectively.

Genome assembly

We generated a De Bruijn graph-type Illumina-only assembly using the Platanus, resulting in 14.8 M contigs with an N40 of 1.2 kb. We then used DBG2OLC to generate a hybrid assembly using the Platanus contigs and raw PacBio data, resulting in an assembly consisting of 3107 contigs with an N50 of 2.6 Mb. In parallel, we produced a PacBio-only assembly using the Falcon assembler, resulting in 36,000 contigs with an N50 of 305 kb. We lastly merged the assemblies using quickmerge and polished using Quiver and Pilon, resulting in a final assembly consisting of 1762 contigs with an N50 of 4.5 Mb. We combined the two Hi-C libraries and used 3d-dna to scaffold the genome with varying numbers of suggested chromosomes (as we observed fusions and splits relative to a published linkage map for *P. maniculatus*). We settled on a Hi-C scaffolding producing 24 chromosome length scaffolds harboring a near complete set of alignable human proteins and *BUSCO* genes. We assigned chromosome names to scaffolds based on aligning sequences from a published linkage map and synteny with mouse and rat.

Informatics

Repetitive elements were annotated using RepeatMasker. Comparative syntenic analysis was carried out using a modified SynChro. Gene annotation was carried out three different ways: (i) Trinity of RNA-seq data, followed by Augustus; (ii) Hisat2 of raw RNA-seq data, followed by Strawberry; and (iii) alignment of a complete set of proteins from human, mouse, and rat to the assembly using Exonerate. Different methods gave different genes/isoforms, which are all provided as Santa Cruz Genome Browser tracks. We lastly combined all gene predictions as a set of "hints" provided to Augustus to estimate the total number of genes. We further aligned *P. leucopus*, human, rat, and mouse to one another using ProgressiveCactus and represent the resulting "snake tracks" in the same browser. We predicted the *P. leucopus* proteome using Transdecoder on the longest isoform per mouse to *Peromyscus* protein alignment and predicted protein domains using Pfam and HMMER.

Experimental infection with *B. burgdorferi*

Adult *P. leucopus* were obtained from the PGSC and maintained in the AAALAC (Association for Assessment and Accreditation of Laboratory Animal Care)–accredited University of California (UC) Irvine vivarium for 6 weeks before the experiment. Each animal was infected with an intraperitoneal injection of 50 μl of citrated blood freshly obtained from CB17 strain severe combined immunodeficiency mice (Charles River Laboratory). After 5 weeks, the animals were euthanized with carbon dioxide and terminal exsanguination, with blood and tissue samples collected. The protocol was approved by the Institutional Animal Care and Use Committee at UC Irvine and the Animal Care and Use Review Office of the United States Army Medical Research and Materiel Command.

RNA isolation and sequencing

After RNA extraction using a Qiagen RNeasy Mini kit, concentration and quality were assessed using Qubit and the Agilent 2100 Bioanalyzer and Eukaryotic RNA Nano assay (only samples passing quality control were kept). Library construction was performed using the TruSeq RNA Library Prep Kit v2. Libraries were validated using qPCR, sized using the Agilent Bioanalyzer, and sequenced as multiplex reactions as PE100s using a HiSeq4000.

RNA-seq analysis

RNA-seq data were deduplicated using PRINSEQ and trimmed using Trimmomatic. Differential expression between tissues was examined by aligning reads to the genome using TopHat and featureCounts relative to Trinity/Augustus transcripts. Statistical significance was assessed using fitNbinomGLMs in DESeq. Analysis of the experimental infection experiment was carried out using the CLC Genomics

Workbench using read pairs mapped with a length fraction of 0.7 and a similarity fraction of 0.9 to annotate Trinity/Augustus transcripts. The Expectation-Maximization (EM) estimation algorithm of the suite was used to iteratively estimate the abundance of transcripts and assign reads to transcripts according to these estimates. The differential expression between experimental conditions was assessed with an assumption of a negative binomial distribution for expression levels and a separate generalized linear model for each mRNA, including a dispersion parameter, as is implemented in the edgeR package.

SNPs and LD

We made Illumina Nextera libraries for 26 DNA samples from a natural population of *P. leucopus* collected from Lake Gaillard, CT and collected ~2× per sample worth of PE100 reads per sample. Reads were aligned to the genome using bwa-mem and merged using bamtools, and SNPs were identified using GATK. The final VCF file was filtered using VCFtools to select only SNPs, with the minor allele observed three or more times. We annotated the resulting set of SNPs using SnpEff and gene annotations based on the alignment of human proteins to the *P. leucopus* reference genome using Exonerate. We estimated LD for all chromosome 10 biallelic SNPs with a minor allele count greater than 8 and with eight or fewer nonmissing genotypes and calculated R^2 on 0/1/2 genotypes for SNPs less than 1 Mb apart.

Additional resources

A more detailed set of materials and methods are given in the Supplementary Materials. Table S9 provides accessions and links to raw and processed data. Much of the data can also be obtained and interacted with at http://goo.gl/LwHDr5. This link is dynamic and will update as the genome is improved and more genomes become available.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/7/eaaw6441/DC1

Supplementary Materials and Methods

Fig. S1. Cumulative contig coverage (Mb) as a function of contig length for different PacBio/Illumina assembly strategies.

Fig. S2. Read span coverage as a function of read span size bins for Hi-C data summed over two Hi-C libraries.

Fig. S3. Mummer dot plot for chromosome 3 assemblies.

Fig. S4. Hi-C scaffolding results in chromosome-sized scaffolds that are considerably longer than the raw contigs.

Fig. S5. A comparison of a *P. maniculatus* linkage map (orange) and the *P. leucopus* genome (blue).

Fig. S6. The Hi-C contact map for the region near the putative chromosome 16 and 21 fusion.

Fig. S7. Rat syntenic groups associated with the chromosome 16 and 21 fusion.

Fig. S8. *P. leucopus* synteny with rat.

Fig. S9. *P. leucopus* synteny with rat with *P. leucopus* chromosomes reordered to reflect the scaffold numbers produced by 3d-dna.

Fig. S10. DEGs between tissues.

Fig. S11. SNPs and INDELs identified by aligning Illumina reads from the reference individual back to the assembly.

Fig. S12. Selected tracks from the Santa Cruz Browser genomehub interface for IL-6.

Table S1. Chromosome-sized scaffold lengths and names.

Table S2. Summary of the RepeatMasker analysis of the *P. leucopus* genome.

Table S3. Summary of gene predictions.

Table S4. Samples used in the tissue RNA-seq experiment.

Table S5. Gene Ontology categories showing change between tissues.

Table S6. DEGs in the blood following infection with *B. burgdorferi*.

Table S7. DEGs in the skin following infection with *B. burgdorferi*.

Table S8. R^2 as a function of distance for chromosome 10 SNPs.

Table S9. Resources.

References (*33*–*73*)

REFERENCES AND NOTES

1. A. G. Barbour, Infection resistance and tolerance in *Peromyscus* spp., natural reservoirs of microbes that are virulent for humans. *Semin. Cell Dev. Biol.* **61**, 115–122 (2017).
2. E. R. Hall, *Mammals of North America* (John Wiley and Sons, 1979), vol. 2.
3. S. W. Barthold, D. Cadavid, M. T. Phillip, Animal models of borreliosis, in *Borrelia: Molecular Biology, Host Interaction, and Pathogenesis*, J. D. Radolf, D. S. Samuels, Eds. (Caister Academic Press, 2010), pp. 359–412.
4. C. I. Paules, H. D. Marston, M. E. Bloom, A. S. Fauci, Tickborne diseases — Confronting a growing threat. *N. Engl. J. Med.* **379**, 701–703 (2018).
5. J. I. Tsao, J. T. Wootton, J. Bunikis, M. G. Luna, D. Fish, A. G. Barbour, An ecological approach to preventing human infection: Vaccinating wild mouse reservoirs intervenes in the Lyme disease cycle. *Proc. Natl. Acad. Sci. U.S.A* **101**, 18159–18164 (2004).
6. D. A. Najjar, A. M. Normandin, E. A. Strait, K. M. Esvelt, Driving towards ecotechnologies. *Pathog. Glob. Health* **111**, 448–458 (2018).
7. M. Chakraborty, J. G. Baldwin-Brown, A. D. Long, J. J. Emerson, Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* **44**, e147 (2016).
8. O. Dudchenko, S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger, N. C. Durand, M. S. Shamim, I. Machol, E. S. Lander, A. P. Aiden, E. L. Aiden, De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
9. J. Kenney-Hunt, A. Lewandowski, T. C. Glenn, J. L. Glenn, O. V. Tsyusko, R. J. O'Neill, J. Brown, C. M. Ramsdell, Q. Nguyen, T. Phan, K. R. Shorter, M. J. Dewey, G. Szalai, P. B. Vrana, M. R. Felder, A genetic map of *Peromyscus* with chromosomal assignment of linkage groups (a *Peromyscus* genetic map). *Mamm. Genome* **25**, 160–179 (2014).
10. J. Brown, J. Crivello, R. J. O'Neill, An updated genetic map of *Peromyscus* with chromosomal assignment of linkage groups. *Mamm. Genome* **29**, 344–352 (2018).
11. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
12. M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** (Suppl 2), ii215–ii25 (2003).
13. D. Kim, B. Langmead, S. L. Salzberg, HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
14. R. Liu, J. Dickerson, Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-Seq. *PLOS Comput. Biol.* **13**, e1005851 (2017).
15. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
16. H. A. Wichman, S. S. Potter, D. S. Pine, Mys, a family of mammalian transposable elements isolated by phylogenetic screening. *Nature* **317**, 77–81 (1985).
17. M. A. Cantrell, M. M. Ederer, I. K. Erickson, V. J. Swier, R. J. Baker, H. A. Wichman, MysTR: An endogenous retrovirus family in mammals that is undergoing recent amplifications to unprecedented copy numbers. *J. Virol.* **79**, 14698–14707 (2005).
18. G. Wolf, P. Yang, A. C. Füchtbauer, E.-M. Füchtbauer, A. M. Silva, C. Park, W. Wu, A. L. Nielsen, F. S. Pedersen, T. S. Macfarlan, The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes Dev.* **29**, 538–554 (2015).
19. Y. Ono, T. Fujii, K. Igarashi, T. Kuno, C. Tanaka, U. Kikkawa, Y. Nishizuka, Phorbol ester binding to protein kinase C requires a cysteine-rich zinc-finger-like sequence. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 4868–4871 (1989).
20. P. L. Oliver, L. Goodstadt, J. J. Bayes, Z. Birtle, K. C. Roach, N. Phadnis, S. A. Beatson, G. Lunter, H. S. Malik, C. P. Ponting, Accelerated evolution of the *Prdm9* speciation gene across diverse metazoan taxa. *PLOS Genet.* **5**, e1000753 (2009).
21. C. Dressaire, R. N. Moreira, S. Barahona, A. P. Alves de Matos, C. M. Arraiano, BolA is a transcriptional switch that turns off motility and turns on biofilm development. *MBio* **6**, e02352-14 (2015).
22. C. Chiang, R. Z. Swan, M. Grachtchouk, M. Bolinger, Y. Litingtung, E. K. Robertson, M. K. Cooper, W. Gaffield, H. Westphal, P. A. Beachy, A. A. Dlugosz, Essential role for *Sonic hedgehog* during hair follicle morphogenesis. *Dev. Biol.* **205**, 1–9 (1999).
23. R. W. Ness, Y.-H. Zhang, L. Cong, Y. Wang, J.-X. Zhang, P. D. Keightley, Nuclear gene variation in wild brown rats. *G3* **2**, 1661–1664 (2012).
24. R. L. Unckless, A. G. Clark, P. W. Messer, Evolution of resistance against CRISPR/Cas9 gene drive. *Genetics* **205**, 827–841 (2017).
25. K. Nelson, R. J. Baker, R. L. Honeycutt, Mitochondrial DNA and protein differentiation between hybridizing cytotypes of the white-footed mouse, *Peromyscus leucopus*. *Evolution* **41**, 864–872 (1987).
26. F. Schaper, S. Rose-John, Interleukin-6: Biology, signaling and strategies of blockade. *Cytokine Growth Factor Rev.* **26**, 475–487 (2015).
27. G. A. Sacher, R. W. Hart, Longevity, aging and comparative cellular and molecular biology of the house mouse, Mus musculus, and the white-footed mouse, *Peromyscus leucopus*. *Birth Defects Orig. Artic. Ser.* **14**, 71–96 (1978).

28. K. R. Shorter, A. Owen, V. Anderson, A. C. Hall-South, S. Hayford, P. Cakora, J. P. Crossland, V. R. M. Georgi, A. Perkins, S. J. Kelly, M. R. Felder, P. B. Vrana, Natural genetic variation underlying differences in *Peromyscus* repetitive and social/aggressive behaviors. *Behav. Genet.* **44**, 126–135 (2014).
29. P. B. Vrana, K. R. Shorter, G. Szalai, M. R. Felder, J. P. Crossland, M. Veres, J. E. Allen, C. D. Wiley, A. R. Duselis, M. J. Dewey, W. D. Dawson, *Peromyscus* (deer mice) as developmental models. *Wiley Interdiscip. Rev. Dev. Biol.* **3**, 211–230 (2014).
30. K. R. Shorter, J. P. Crossland, D. Webb, G. Szalai, M. R. Felder, P. B. Vrana, *Peromyscus* as a mammalian epigenetic model. *Genet. Res. Int.* **2012**, 179159 (2012).
31. M. Veres, A. R. Duselis, A. Graft, W. Pryor, J. Crossland, P. B. Vrana, G. Szalai, The biology and methodology of assisted reproduction in deer mice (*Peromyscus maniculatus*). *Theriogenology* **77**, 311–319 (2012).
32. N. L. Bedford, H. E. Hoekstra, Peromyscus mice as a model for studying natural variation. *eLife* **4**, e06813 (2015).
33. C. P. Joyner, L. C. Myrick, J. P. Crossland, W. D. Dawson, Deer mice as laboratory animals. *ILAR J.* **39**, 322–330 (1998).
34. J. P. Crossland, M. J. Dewey, S. C. Barlow, P. B. Vrana, M. R. Felder, G. J. Szalai, Caring for *Peromyscus* spp. in research environments. *Lab. Anim.* **43**, 162–166 (2014).
35. N. H. Lazar, K. A. Nevonen, B. O'Connell, C. McCann, R. J. O'Neill, R. E. Green, T. J. Meyer, M. Okhovat, L. Carbone, Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res.* **28**, 983–997 (2018).
36. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
37. G. W. Vurture, F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski, M. C. Schatz, GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
38. R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, T. Itoh, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
39. C. Ye, C. M. Hill, S. Wu, J. Ruan, Z. S. Ma, DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6**, 31900 (2016).
40. S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie, E. D. Jarvis, A. M. Phillippy, Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
41. C.-S. Chin, P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz, G. R. Cramer, M. Delledonne, C. Luo, J. R. Ecker, D. Cantu, D. R. Rank, M. C. Schatz, Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
42. C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, J. Korlach, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
43. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014).
44. J. Ghurye, M. Pop, S. Koren, D. Bickhart, C.-S. Chin, Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 527 (2017).
45. G. Marçais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, A. Zimin, MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018).
46. N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, E. L. Aiden, Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
47. R. M. Waterhouse, M. Seppey, F. A. Simão, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, E. M. Zdobnov, BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2017).
48. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
49. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
50. W. J. Kent, BLAT—The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
51. G. Drillon, A. Carbone, G. Fischer, SynChro: A fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLOS ONE* **9**, e92621 (2014).
52. R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
53. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
54. B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, D. Haussler, Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).
55. B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman, A. Regev, *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
56. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
57. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
58. E. Eden, R. Navon, I. Steinfeld, D. Lipson, Z. Yakhini, *GOrilla*: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
59. E. Baum, F. Hue, A. G. Barbour, Experimental infections of the reservoir species *Peromyscus leucopus* with diverse strains of *Borrelia burgdorferi*, a Lyme disease agent. *MBio* **3**, e00434-12 (2012).
60. B. Travinsky, J. Bunikis, A. G. Barbour, Geographic differences in genetic locus linkages for *Borrelia burgdorferi*. *Emerg. Infect. Dis.* **16**, 1147–1150 (2010).
61. V. Cook, A. G. Barbour, Broad diversity of host responses of the white-footed mouse *Peromyscus leucopus* to *Borrelia* infection and antigens. *Ticks Tick Borne Dis.* **6**, 549–558 (2015).
62. Qiagen, Manual for CLC Genomics Workbench (2018).
63. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
64. A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, A. Mortazavi, A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
65. D. J. McCarthy, Y. Chen, G. K. Smyth, Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
66. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
67. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
68. M. Teng, M. I. Love, C. A. Davis, S. Djebali, A. Dobin, B. R. Graveley, S. Li, C. E. Mason, S. Olson, D. Pervouchine, C. A. Sloan, X. Wei, L. Zhan, R. A. Irizarry, A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 74 (2016).
69. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
70. D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Strömberg, G. T. Marth, BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011).
71. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
72. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
73. P. Cingolani, A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

Acknowledgments: We thank J. Crossland, M. Felder, V. Kaza, H. Kiaris, and P. Vrana of the PGSC of the University of South Carolina for discussions and advice, S. Davis of the University of South Carolina for providing specimens, and P. Sitlani for technical assistance. **Funding:** Research reported in this paper was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (NIH) under grant nos. R21 AI126037 (to A.D.L. and A.G.B.) and U54 AI065359 (to A.G.B.) and by direct support of sequencing costs by the Bay Area Lyme Foundation (Portola Valley, CA) and institutional funds of UC Irvine and the University of South Carolina. This work was also supported, in part, by the Office of the Assistant Secretary of Defense for Health Affairs through the Tick Borne Disease Research Program under award no. W81XWH-17-1-0481 (to A.G.B.). Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. R.C.-D. is supported by the Alfred P. Sloan Foundation. This work was also made possible, in part, through access to the Genomics High Throughput Facility Shared Resource of the Cancer Center Support Grant (P30CA-062203) and NIH shared instrumentation grant nos. 1S10RR025496-01, 1S10OD010794-01, and 1S10OD021718-01.
Author contributions: A.D.L. and A.G.B. conceptualized, supervised, and wrote the draft of the paper. J.B.-B. made libraries, carried out the PacBio/Illumina assembly, analyzed the tissue collection, and did gene annotation. Y.T. scaffolded the genome, validated the scaffolding,

and executed the synteny analysis. R.C.-D. made the Hi-C libraries. G.B.-G. and A.M. conceptualized and executed the Pfam and repeat analysis. V.J.C. carried out animal experiments and collected and extracted samples for the DNA sequencing and RNA-seq studies. A.D.L. carried out the population genetics analysis, and A.G.B. analyzed the differential expression experiment. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data and the code required to reproduce the results of this work are listed in table S9 as links to appropriate repositories. Additional data related to this paper may be requested from the authors.

Submitted 11 January 2019
Accepted 18 June 2019
Published 24 July 2019
10.1126/sciadv.aaw6441

Citation: A. D. Long, J. Baldwin-Brown, Y. Tao, V. J. Cook, G. Balderrama-Gutierrez, R. Corbett-Detig, A. Mortazavi, A. G. Barbour, The genome of *Peromyscus leucopus*, natural host for Lyme disease and other emerging infections. *Sci. Adv.* **5**, eaaw6441 (2019).