MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12301

Transcriptome resources for the white-footed mouse (*Peromyscus leucopus*): new genomic tools for investigating ecologically divergent urban and rural populations

STEPHEN E. HARRIS,* RACHEL J. O'NEILL† and JASON MUNSHI-SOUTH‡
*Program in Ecology, Evolutionary Biology, & Behavior, The Graduate Center, City University of New York (CUNY), New York, NY 10016, USA, †Molecular & Cell Biology, University of Connecticut, Storrs, CT 06269, USA, ‡Louis Calder Center-Biological Field Station, Fordham University, 53 Whippoorwill Road, Armonk, NY 10504, USA

Abstract

Genomic resources are important and attainable for examining evolutionary change in divergent natural populations of nonmodel species. We utilized two next-generation sequencing (NGS) platforms, 454 and SOLiD 5500XL, to assemble low-coverage transcriptomes of the white-footed mouse (*Peromyscus leucopus*), a widespread and abundant native rodent in eastern North America. We sequenced liver mRNA transcripts from multiple individuals collected from urban populations in New York City and rural populations in undisturbed protected areas nearby and assembled a reference transcriptome using 1 080 065 954 SOLiD 5500XL (75 bp) reads and 3 052 640 454 GS FLX + reads. The reference contained 40 908 contigs with a N50 = 1044 bp and a total content of 30.06 Megabases (Mb). Contigs were annotated from *Mus musculus* (39.96% annotated) Uniprot databases. We identified 104 655 high-quality single nucleotide polymorphisms (SNPs) and 65 single sequence repeats (SSRs) with flanking primers. We also used normalized read counts to identify putative gene expression differences in 10 genes between populations. There were 19 contigs significantly differentially expressed in urban populations compared to rural populations, with gene function annotations generally related to the translation and modification of proteins and those involved in immune responses. The individual transcriptomes generated in this study will be used to investigate evolutionary responses to urbanization. The reference transcriptome provides a valuable resource for the scientific community using North American *Peromyscus* species as emerging model systems for ecological genetics and adaptation.

Keywords: Genetic Map, *Peromyscus*, RNA-Seq, single nucleotide polymorphisms, SOLiD, transcriptome

Received 23 April 2014; revision received 26 June 2014; accepted 27 June 2014

Introduction

A major goal of ecological and evolutionary genomics (EEG) is to identify the evolutionary responses of populations to novel or divergent habitats (Renn & Siemens 2010; Pavey *et al.* 2012). Such population studies were traditionally challenging due to lack of genomic resources for nonmodel organisms. The recent advent of next-generation sequencing (NGS) has made it possible to generate population genomic resources for nearly any species using a variety of methods: low-coverage whole-genome sequencing (WGS), transcriptome sequencing (RNAseq), restriction-site-associated DNA sequencing (RADseq) or targeted sequence capture (SeqCap; McCormack *et al.* 2011; Grover *et al.* 2012; Pavey *et al.* 2012; Wolf 2013). Transcriptome sequencing is one of the most commonly used NGS approaches in nonmodel organisms because transcriptome data sets contain information on nucleotide variation and gene expression levels across tissue types, time periods or any number of ecologically relevant variables (Ekblom & Galindo 2011). RNAseq data can be used in a wide range of downstream analyses, including comparative genomics, microarray design, QTL mapping, or candidate gene identification. In this study, we report results of a RNAseq comparison among multiple populations of white-footed mice, *Peromyscus leucopus*, from ecologically divergent urban and rural habitats in the New York City (NYC) metropolitan area. We used these data to examine nucleotide variation and population structure within and among populations and habitat types, as well as to derive an annotated reference transcriptome for future examination of candidate genes involved in local adaption to urbanization.

The white-footed mouse is one of more than 50 species comprising the genus *Peromyscus*. Peromyscine

Correspondence: Jason Munshi-South, Fax: 1-914-273-6346;
E-mail: jmunshisouth@fordham.edu

rodents occur from Central America to Alaska, along extreme elevation gradients, and in multiple divergent habitats (Dewey *et al.* 2001; Bradley *et al.* 2007). *Peromyscus* spp. are one of the most well studied groups of North American mammals, including phylogenetic relationships (Bradley *et al.* 2007), karyotypes and genetic maps (Ramsdell *et al.* 2008; Kenney-Hunt *et al.* 2014), phylogeographic histories (Dragoo *et al.* 2006; Gering *et al.* 2009; Kalkvik *et al.* 2012), population genetics (Mossman & Waser 2001; Steiner *et al.* 2007; Storz *et al.* 2007; Pergams & Lacy 2007; Munshi-South & Kharchenko 2010; Rogic *et al.* 2013; Taylor & Hoffman 2014) , and decades-long population ecology studies (Wolff 1985; Vessey & Vessey 2007). This mouse is also a primary carrier of hantaviruses (Morzunov *et al.* 1998) and is implicated in spreading Lyme disease (Ostfeld 2012) and other emerging zoonotic pathogens (e.g. *Babesia*, *Anaplasma*; Keesing *et al.* 2009) in eastern North America. More recently, *Peromyscus* spp. have been developed as model systems to investigate the genetic basis of adaptation to divergent ecological conditions. For example, diversifying selection probably drove adaptive modifications in haemoglobin function between high- and low-altitude populations of *P. maniculatus* in Colorado, U.S.A. (Storz *et al.* 2007, 2009, 2010; Natarajan *et al.* 2013). In other examples, independent mutations in the *Agouti* gene probably lead to divergent coat coloration in *Peromyscus* populations in both the Nebraska Sand Hills and Florida sand dunes (Mullen & Hoekstra 2008; Linnen *et al.* 2009, 2013), while the genetic architecture of complex extended phenotypes such as burrowing behaviour have also recently been described (Weber *et al.* 2013).

Peromyscus leucopus are common residents of human-dominated environments in the eastern United States, persisting even in small, highly fragmented urban forests (Pergams & Lacy 2007; Rogic *et al.* 2013; Munshi-South & Nagy 2014). Urbanization frequently results in severe habitat fragmentation, increased exposure to diseases, toxins, and pollutants and can affect life history traits of 'urban exploiters' or 'urban adapters' by eliminating their predators and competitors (Blair 2001; Sih *et al.* 2011). New York City populations of *P. leucopus* are highly isolated from one another by urbanization and only maintain connectivity through remnant vegetated corridors (Munshi-South 2012). These populations exhibit strong genetic structure, but high levels of heterozygosity and allelic diversity within populations (Munshi-South & Kharchenko 2010), conditions potentially favourable for local adaptation. Using pooled RNA sequencing, Harris *et al.* (2013) identified nonsynonymous mutations and patterns of divergent selection in protein-coding regions of urban and rural white-footed mice. These results indicate that *P. leucopus* in NYC can serve as a useful model for investigating the population genomic implications of inhabiting novel urban ecosystems.

The removal of logistic barriers to generating genomic data sets has led to a surge in studies using a bottom-up (i.e. reverse genomics) approach to investigate the genetic basis of adaptation. Reverse genomic approaches identify potentially adaptive alleles based on signatures of selection in DNA sequences and then may further screen adaptive candidates by putative gene function (Barrett & Hoekstra 2011; Ellegren 2014). Common garden or reciprocal transplant experiments can link genotypes to phenotypes by measuring phenotypic response in individuals with known genotypes under measurable environmental conditions. These controlled experiments address the limitations of reverse genomics and are ultimately needed to understand the fitness consequences of candidate loci or alleles (Merilä & Hendry 2014). To establish a foundation for investigating adaptive changes to urbanization and provide resources for other population genomic studies, we characterized the transcriptomes of urban and rural *P. leucopus* populations using a combination of 454 and SOLiD 5500 XL sequencing. Here, we report the *P. leucopus* reference transcriptome sequences and assemblies, annotations of sequences, an extensive variant library of SNPs and SSRs, a linkage map of 4066 contigs mapped to *P. leucopus* chromosomes and initial insights into gene expression differences among populations. Despite considerable research interest in *Peromyscus* spp. among the biological community, only eight genomic or transcriptomic data sets have been generated outside of this current effort by high-throughput sequencing methods (Barrett *et al.* 2012; Peterson *et al.* 2012; Cheviron *et al.* 2012, 2013; Linnen *et al.* 2013 Harris *et al.* 2013). The SNP library and extensive transcriptome sequences developed here will facilitate future functional analyses of molecular signatures indicative of local adaptation in populations of white-footed mice as well as aid in the final annotations of draft assemblies for these emerging models (O'Neill *et al.* 1998).

Materials and methods

Specimen collection and RNA extraction

Peromyscus leucopus from urban and rural populations were trapped over a period of 1–3 nights each at six sites using four 7 × 7 m transects of 3″ × 3″ × 9″ Sherman live traps. The rural sites were chosen because they are among the largest, contiguous protected areas with relatively low human disturbance in proximity to NYC (Table 1). Adult mice were euthanized by cervical dislocation. Livers and other organs were extracted in the field and immediately placed in RNAlater (Ambion Inc., Austin, TX, USA) and stored at −80 °C until RNA

Table 1 Specimen collection locations for *Peromyscus leucopus* individuals used in this study

Code	Collection site (City, State, Year)	Latitude, Longitude	No. of individuals used for RNAseq	
			Female	Male
Rural				
WWP_BH	Wildwood State Park; Brookhaven State Park (Long Island, NY, 2012)	40° 57′ 58.4″ N, 72° 48′ 08.8″ W 40° 55′ 38.8″ N, 72° 52′ 16.8″ W	3	5
HIP	High Point State Park (Milford, NJ, 2012)	41° 18′ 22.5″ N, 74° 40′ 09.0″ W	4	4
CFP	Clarence Fahnestock State Park (Putnam Valley, NY, 2012)	41° 26′ 59.9″ N, 73° 51′ 28.8″ W	4	4
Urban				
FM	Flushing Meadows Park (Queens, NY, 2010)	40° 43′ 11.3″ N, 73° 49′ 57.4″ W	4	4
NYBG	New York Botanical Gardens (Bronx, NY, 2010)	40° 51′ 48.2″ N, 73° 52′ 32.7″ W	4	4
CP	Central Park (Manhattan, NY, 2010)	40° 47′ 47.5″ N, 73° 57′ 21.6″ W	4	4

extraction. Urban sites were located in NYC, highly fragmented and surrounded by dense urban matrix and contain genetically differentiated populations of *P. leucopus* (Munshi-South & Kharchenko 2010). RNA was extracted from a total of 48 liver samples (~15 mg) using Trizol® reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. RNA yield was measured using a Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, USA), and quality was determined using the NanoDrop® Spectrophotometer (Thermo Fisher Scientific) according to manufacturers' protocols. All animal handling procedures were approved by the Institutional Animal Care and Use Committee at Brooklyn College, CUNY (Protocol Nos. 247 and 266).

Library preparation and sequencing

Total RNA was treated with RNase-Free DNase (Promega, Wisconsin, USA) and purified using LiCL precipitation to remove genomic DNA contamination. If Qubit or Nanodrop readings indicated poor yield or low quality RNA, the total RNA was repurified and concentrated using the RNA Clean & Concentrator™-5 kit (Zymo, Irvine, CA, USA). ERCC RNA Spike-In controls (Ambion) were then added to individual total RNA for samples prepared for SOLiD sequencing. Ribosomal RNA (rRNA) was removed using the mouse Ribo-Zero™ rRNA Removal Kit (Epicentre, Madison, WI, USA), purified with the RNA Clean & Concentrator™-5 kit (Zymo), and assessed for rRNA depletion and final RNA yield using the Agilent RNA 6000 Pico Kit (Agilent, Santa Clara, CA, USA). rRNA-depleted RNA was then fragmented, reverse-transcribed into cDNA, size-selected, amplified and purified using the SOLiD™ Total RNA-Seq Kit (Thermo Fisher Scientific). The manufacturer's protocol was followed except when purifying cDNA; 80 μL of AMPure XP beads (Beckman Coulter, USA) were used. During amplification, the SOLiD RNA Barcoding Kit (Thermo Fisher Scientific) was used to add individual barcodes to each of the 48 samples. Finally, cDNA quality was assessed using the SOLiD™ Library TaqMan® Quantitation Kit (Thermo Fisher Scientific). Forty-eight barcoded cDNA samples were combined into three pools of 16 individuals each at a concentration of 500 pM, and each pool was loaded onto four lanes in one SOLiD 5500XL run conducted at the Center for Applied Genetics and Technology at UCONN. All raw sequencing files have been deposited in the GenBank Sequence Read Archive (SRA Accession no. SRP020005). Assembly files are available at the Dryad digital repository, doi: 10.5061/dryad.6hc0f.

Transcriptome assembly and annotation

Raw reads from all three pools were filtered by barcode identifier, delineated into individual *P. leucopus* libraries, further processed to remove sequencing adapters, quality trimmed (Q > 20), and subsequently filtered for length >30 bp using CUTADAPT version 1.2.1 (Martin 2011). Reads were then used to find the best overall assembly (largest N50, longest average contigs, the most contigs with unique gene annotations). One assembly strategy used reads generated from SOLiD 5500 xl sequencing. Another assembly strategy used the processed SOLiD 5500 xl reads combined with previously generated reads from Roche 454 FLX+ pyrosequencing that were previously trimmed of adaptors and poor quality nucleotides, and filtered for length >100 bp (Harris *et al.* 2013). The SOLiD sequences and the combined SOLiD/454 reads were assembled into separate contig libraries using Trinity version r2013_08_14 (Grabherr *et al.* 2011; Haas *et al.* 2013) with default settings. Both sequence library outputs from Trinity consisted of many repetitive, overlapping, and probably artefactual contigs that

required further filtering. Transdecoder version r2012-08-15 (Haas *et al.* 2013) was used to reduce assemblies to only those contigs with open reading frames (ORFs). As this method is somewhat conservative, a second filtering step was used to find contigs with single gene annotations. For this filter, the full assemblies (SOLiD and SOLiD/454 combined) and the respective Transdecoder output libraries were each searched against the UNIPROT *Mus musculus* database (Magrane and UniProt Consortium 2011) for best-hit matches using BLASTX (*E*-value cut-off = 1e-6). Contigs with unique best hits from the assemblies that were not identified using Transdecoder were added back into the Transdecoder set of contigs. The full SOLiD Trinity assembly, full SOLiD_454 Trinity assembly, filtered SOLiD_454 assembly and previously generated 454 assembly (Harris *et al.* 2013) were then compared for size and quality (Table 2). The final 'best-quality' transcriptome assembly (filtered SOLiD_454) was annotated in two ways. BLASTX was used to search the assembly against UNIPROT's *Mus musculus* database and NCBI's nonredundant (nr) protein database (*E*-value cutoff = 1e-6). The annotated *M. musculus* genes were used to verify assembled contigs and provide high-quality annotations. Using the full nr database broadens possible hits, and these additional annotations were given Gene Ontology (GO) terms using BLAST2GO version 1.0 (Conesa *et al.* 2005). The software program, KisSplice (Sacomoto *et al.* 2012) was then used with default parameters to generate a list of splicing events. Those events were then mapped back to the transcriptome assembly using megaBLAST (*E*-value cut-off = 1e-10). Information about the alternative splice variant was included in the header of parent contigs, and all such variants were tracked in downstream analyses.

SNP/SSR identification and analysis of population structure

The scripts from (De Wit *et al.* 2012) were modified (Dryad doi: 10.5061/dryad.6hc0f) and used for SNP calling and mapping each RNA-seq library to the final reference assembly (filtered SOLiD_454 assembly). Potential PCR duplicates were removed from the reads, and BOWTIE2 version 2.1.0 (Langmead & Salzberg 2012) was used with default settings for sensitive local alignment to map reads to the filtered SOLiD_454 assembly. The Genome Analysis Toolkit uses a Bayesian genotype likelihood model (GATK version 2.8, DePristo *et al.* 2011) with Variant Quality Score Recalibration to generate high-quality SNP genotypes from multi-sample alignment files. We used the recommended settings from (De Wit *et al.* 2012; Van der Auwera *et al.* 2013), for example, SNPs requiring coverage >5X, nucleotide quality >30, no strand bias (FS >35), and SNPs called from a uniquely mapped read. Additional hard filters, that is, removal of SNPs where every individual was heterozygous, overall depth >10, overall depth <350 and minor allele frequency (MAF) >0.025 were used to reduce the likelihood of variant calls from paralogues or sequencing errors.

Population structure among all six sampling sites was examined using SNMF version 0.5 (Frichot *et al.* 2014). This program uses sparse non-negative matrix factorization (sNMF) algorithms and computes least-squares estimates of ancestry coefficients. In contrast to likelihood models like STRUCTURE (Pritchard *et al.* 2000), this exploratory approach is robust to many demographic situations and does not make equilibrium population genetic assumptions, that is, Hardy–Weinberg and linkage equilibrium (Frichot *et al.* 2014). The number of putative ancestral populations tested ranged from $K = 2$ to $K = 8$, with 10 replicate runs for each value of K. A cross-entropy calculation generates masked genotypes to predict ancestry assignment error; lower values indicate better prediction of the true number of K ancestral populations (Frichot *et al.* 2014). Population structure was also investigated using smartPCA (Patterson *et al.* 2006) with default parameters to examine genetic differentiation along principal components. The significance of each principal component was calculated using the twstats program in the Eigensoft software package (Patterson *et al.* 2006). One benefit of both of these programs is the ability to include missing data. Analyses were run on a dataset including SNPs that were genotyped for at least 80% of individuals.

The filtered SOLiD_454 transcriptome assembly was searched for microsatellite repeats using MSATCOMMANDER version 1.0.8 (Faircloth 2008) with default settings, with the exception of minimum number of repeats settings as

Table 2 Assembly statistics. The SOLiD_454_Filtered assembly was used in downstream analyses

Assembly	Number of transcripts (≥100 bp)	Total bases (Mb)	Mean transcript length	N50	Number of transcripts (≥2 kb)
454_Newbler	15 004	13.42	894	1039	825
SOLiD_Full	145 072	56.57	390	395	1134
SOLiD_454_Full	143 552	62.05	432	468	2336
SOLiD_454_Filtered	40 908	30.06	734	1044	2260

follows: di = 8, tri = 8, tetra = 4, penta = 4, Hexa = 4. Primer3 (Rozen & Skaletsky 2000) was used to design forward and reverse primers from flanking sequence (Table S1).

Chromosomal assignment of transcriptome contigs

Contigs from the final transcriptome assembly were assigned to physical locations on chromosomes using scaffolds from the *P. maniculatus* genome (NCBI assembly ID: GCA_000500345.1) and a *P maniculatus*/*P. polionotus* genetic map (Kenney-Hunt *et al*. 2014).

Genes used for the genetic map were downloaded from GenBank for either *M. musculus*, *Rattus norvegicus*, or *Peromyscus* spp. Reciprocal best-hit BLAST was used with BLASTN (*E*-value cutoff 1e-6) to identify contigs from the *P. leucopus* transcriptome that correspond to a given gene marker. The full transcriptome assembly was searched against the *P. maniculatus* genome scaffolds using BLASTN (*E*-value cut-off = 1 × 10-6). A positive match was scored if it met the following criteria: an alignment length >50%, >80% identity, and the query contig matched only one location in the *P. maniculatus* scaffold database. Contigs were ordered by scaffold and by the start of the alignment in the scaffold. Genetic marker contigs with significant hits to a scaffold were labelled with their respective *P. leucopus* chromosome. All other contigs that mapped to a chromosome-defined scaffold were scored as located on the same chromosome. Full DNA sequences with chromosome placement information are provided as supplementary information (Dryad digital repository, doi: 10.5061/dryad.6hc0f).

Differential urban and rural gene expression analysis

Gene expression analysis was performed by mapping individual RNAseq data sets across six populations to the full assembled transcriptome. BOWTIE 2 conducts gapped alignment, which can be problematic for gene expression due to the increased likelihood of splitting reads and mapping to different splice variants. BOWTIE 2 does include gap penalties and imposed gap lengths, but identification of contigs containing alternatively spliced sequences and setting a minimum mean expression level was used to account for splice events. Mapped read counts for individual contigs were compared between the two urban and rural groups and among all six populations using DESEQ (Anders & Huber 2010) implemented in R version 3.0.2 (R Core Team 2013). Custom R scripts included in (De Wit *et al*. 2012) were used to format SAM files into uniquely mapped read count data for DESEQ. Read counts were normalized based on ERCC spike-in controls (Jiang *et al*. 2011) in DESEQ from known starting concentrations. The correction factors were then applied to the experimental dataset. The false discovery rate (FDR) was calculated to account for multiple testing, and a cut-off <0.05 was used to look for significantly differentially expressed genes between urban and rural populations. Genes within the top 10% FDR were also screened to investigate general patterns of gene expression between groups. Mean expression level was required to be ≥5, and differentially expressed genes were kept if they were not splice variants. The same procedure was performed on reads mapped to a second assembly where initial trimming of raw reads used a nucleotide quality cut-off of Q > 5. The contigs representing genes that were significantly differentially expressed were annotated in BLAST2GO (Conesa *et al*. 2005).

Results

Sequencing, assembly and annotation

SOLiD 5500 XL sequencing generated 1 080 065 954 reads of 75 bp from liver tissue, which were combined with 3 052 640 reads from previous 454 FLX+ pyrosequencing with an average length of 309 ± 122 bp (Harris *et al*. 2013). After trimming adapters and filtering for quality, 50.7% of sequencing reads were retained and assembled into 40 908 contigs (Table 2). The filtered SOLiD_454 assembly had the highest number of unique annotations in the UNIPROT database (Table 3) and the longest overall alignment lengths with annotated genes (Fig. 1). While 24 350 contigs were predicted to contain ORFs, 10 834 contigs were verified as genes through high-quality BLAST hits to the *M. musculus* transcriptome. Many contigs that failed the imposed BLAST filter are shorter fragments of identified genes or may represent alternative splicing (AS) events. KisSplice identified 295 AS events, and these contigs were identified and tracked through downstream analyses. However, there are probably a subset of contigs representing genes unique to *Peromyscus* that deserve further investigation. This assembly was also compared to the full nr protein database, resulting in 29 075 annotated sequences. Both the full assembly and the reduced data set with verified gene annotations from *M. musculus* are available on Dryad (doi: 10.5061/dryad.6hc0f).

Variant discovery and population structure

After read mapping and filtering, a total of 104 655 high-quality SNPs were identified across all populations using our transcriptome assembly (30.06 Mb) as a reference. There were 17 969 contigs across all data sets containing SNPs with an average of 5.8 SNPs per contig across all populations. The variant call format (VCF) file containing

Table 3 Annotation statistics from BLASTX of *Peromyscus leucopus* contigs against *Mus musculus* Uniprot (2013_11) database

BLASTX hits to Uniprot	Number of transcripts (≥100 bp)	Unique Uniprot proteins	Uniprot proteins (coverage ≥ 80%)
454_Newbler	10 553	6387	2636 (24.9%)
SOLiD_Full	41 311	9556	2510 (6.1%)
SOLiD_454	38 919	10 838	4572 (11.7%)
SOLiD_454_Filtered	38 855	10 834	4568 (11.7%)

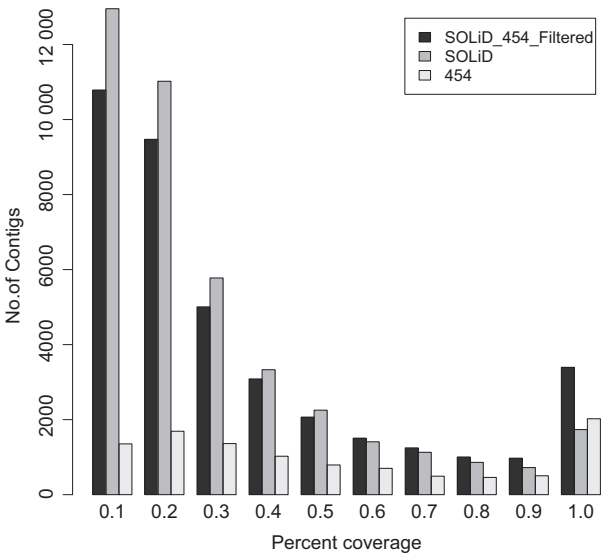

Fig. 1 Distribution of alignment coverage of full genes between subject (gene) and query (*Peromyscus leucopus* contig) in *Mus musculus* Uniprot database using BLASTX.

information on all SNPs is included in the supplementary information (Dryad digital repository, doi: 10.5061/dryad.6hc0f). A total of 609 SSRs were identified, but only 10.7% of these included appropriate flanking sequence for primer design for use in downstream population studies (Table S1). Despite the prevalence of dinucleotide repeats in rodents, tetranucleotide repeats were the most numerous in these *Peromyscus* populations, followed by di- repeats, a disparity probably caused by using protein-coding regions from transcriptome sequencing rather than genome sequence (Toth 2000).

The full SNP data set was filtered to include sites where at least 80% of individuals were genotyped at each SNP. This filtering resulted in 6449 SNP loci for the examination of population structure. Only PC1 and PC2, explaining 52.5% and 7.8% of total variance, were significant ($P \leq 0.01$) in the PCA. There was a gradient along PC1 separating individuals in rural populations from those in urban populations (Fig. 2). Individual ancestry assignment in sNMF supported these results. Assignment to two ancestral populations was highly supported (cross-entropy = 0.70) with structure occurring between urban and rural sampling sites. $K = 5$ (cross-entropy = 0.76) was also supported and showed differentiation between urban populations while individuals from rural groups showed no significant clustering (Fig. 2). Despite the small geographic distances separating urban populations (<10 km) relative to rural populations (>100 km), there was greater genetic differentiation between NYC sites than between rural sites.

Defining linkage groups

A total of 4066 contigs (9.94%) were assigned to *Peromyscus* linkage groups generated from *P. maniculatus* genome scaffolds (Pman_1.0, Assembly ID: GCA_000500345.1) and the calculated recombination frequency between markers in *Peromyscus* backcrosses (Kenney-Hunt *et al.* 2014). An average of 175.83 contigs per chromosome was relatively evenly spaced along chromosomal lengths (Fig. 3) for all linkage groups, corresponding to each of the 23 autosomal chromosomes and the X chromosome plus one additional group from chromosome 8. Of the total placed contigs, 105 carried markers with known lengths (centimorgans) based on the previously defined genetic distances within linkage groups and will thus facilitate future population genomic analyses based on patterns of linkage disequilibrium.

Interpopulation gene expression patterns

White-footed mice were assigned to either 'urban' (24 individuals) or 'rural' (23 individuals) groups based on results from the smartPCA and sNMF analyses. For pairwise population comparisons, four male and four female white-footed mice were assigned to each rural population (BH/WWP, CFP, HIP, See Table 1) and two urban populations (CP, FM, See Table 1). The third urban population, NYBG, contained seven individuals, due to poor sequencing output for one of the female mice. The fully assembled transcriptome, including nonannotated genes, was used for gene expression analysis. The parent contigs containing the 295 alternative splicing events identified above were labelled, monitored and removed if

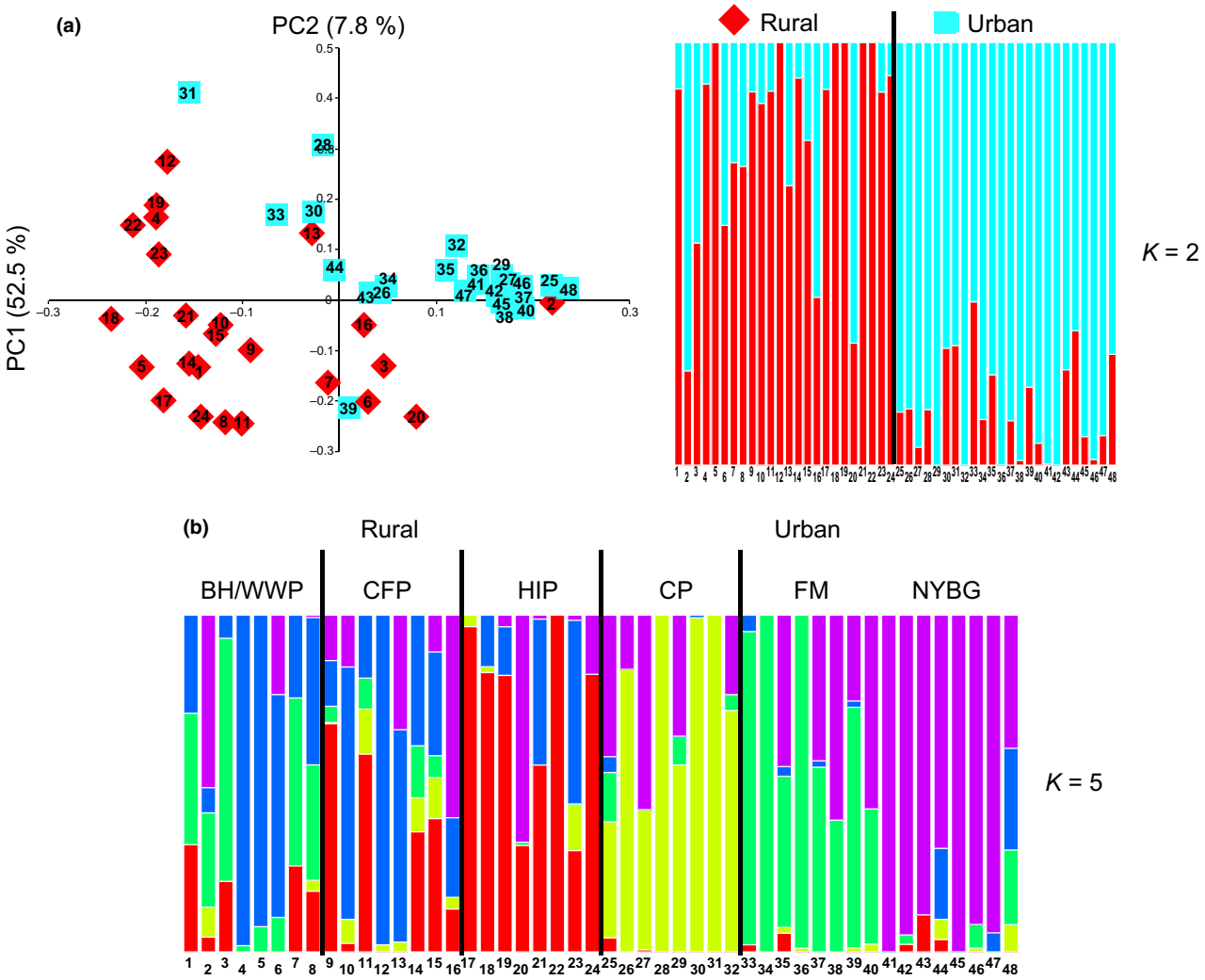

Fig. 2 Population structure analysis of 6449 SNP loci genotyped for 48 individuals from urban (n = 24) and rural (n = 24) populations using smartPCA and sNMF. (a) Individuals colour-coded as urban (blue) or rural (red) for smartPCA (left) results for principal components 1 and 2, and sNMF results for K = 2 (right). (b) Individuals sorted by sampling locality for sNMF results for K = 5. Vertical lines in sNMF plots and data points in smartPCA represent individuals. Individuals in both sNMF plots are ordered identically and numbered. smartPCA data points are numbered according to sNMF ordering.

significantly differentially expressed between urban and rural groups. After read counts were normalized using spike-in controls, genes were identified that were over- and under- expressed. This analysis was conducted separately for two sets of reads independently trimmed for Q > 20 (Q20) or Q > 5 (Q5). There was an average 4.9-fold increase (SD = 3.2) in reads used for the Q5 analysis compared with the Q20. For the Q20 data set, three genes were up-regulated in the urban group (Fig. 4). One over-expressed contig (Contig 300–1848) matched an uncharacterized protein in *Rattus norvegicus*, but without Gene Ontology information no gene function could be assigned. For the Q5 data set, there were four genes significantly (FDR > 0.05) up-regulated and two genes significantly down-regulated in urban populations. Three of the overexpressed genes had significant sequence similarity matches to *KRTAP10-4*, *art2*, and a rRNA promoter-binding protein. The functions of these genes involve immune driven resistance to senescence, rigidity of the structure of hair fibres and eukaryotic translation, respectively. One of the underexpressed genes had significant sequence similarity to *MALAT-1*, a gene involved in lung cancer metastasis. There were no significantly up-regulated genes between individual urban/urban, urban/rural or rural/rural comparisons. No down-regulated genes were identified among rural/rural comparisons, while three genes were under-expressed in individual urban/rural population

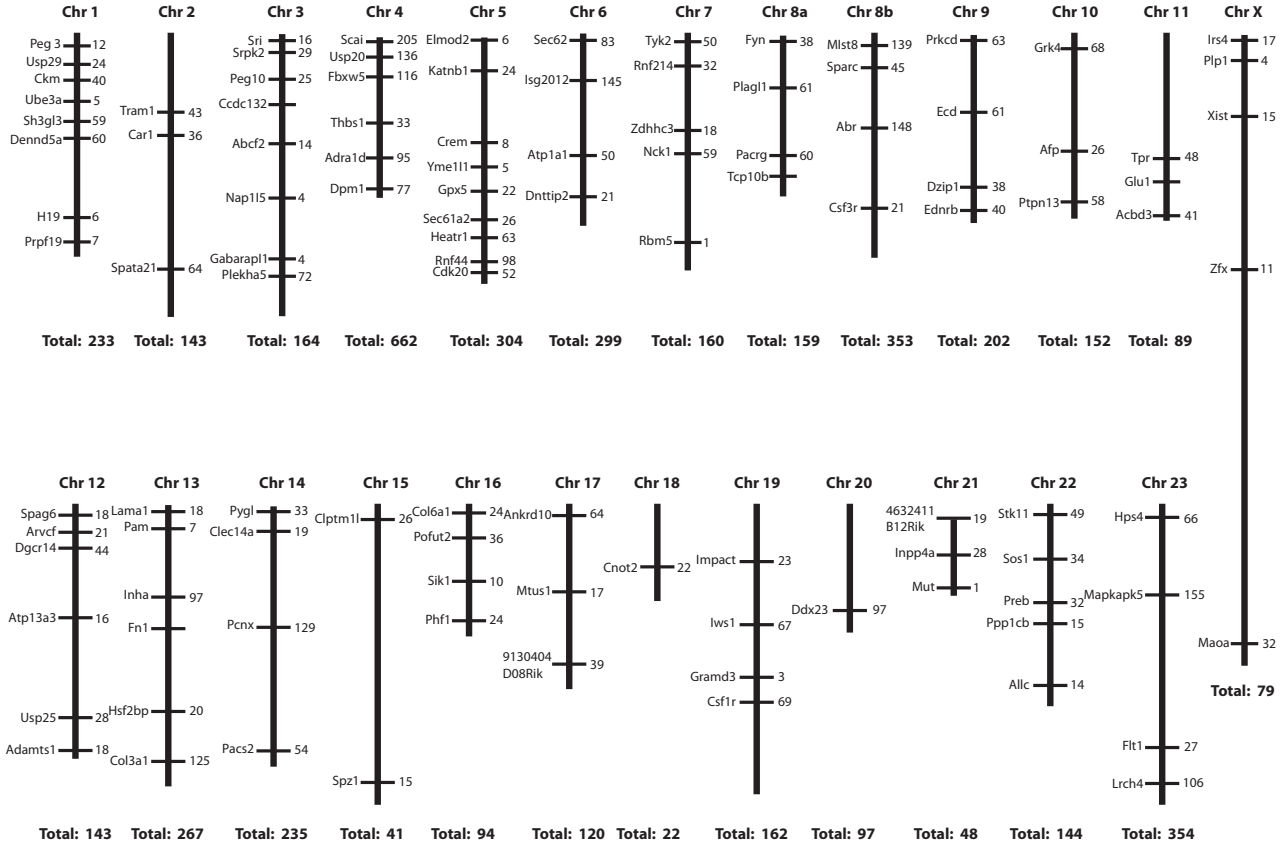

Fig. 3 Inferred locations of protein-coding genes on *Peromyscus leucopus* chromosomes. Known genetic markers from (Kenney-Hunt *et al.* 2014) are listed on the left side of each chromosome. The number on the right side of each chromosome corresponds to how many *P. leucopus* contigs from this study mapped to the *P. maniculatus* scaffold containing the genetic marker. The order of markers along the chromosome is accurate, but chromosome lengths and exact placement of genetic markers is an approximation based on Kenney-Hunt *et al.* (2014).

comparisons. One underexpressed contig represented the gene, *Zinc Transporter ZIP14*, part of the inflammatory response.

Discussion

The speed, price and efficiency of generating NGS data have ushered in a new era of genomic research. However, the 'explosion' of sequence data (Andrews & Luikart 2014) has lead to a faster accumulation of raw DNA sequence data than subsequent genomic analyses. To effectively make use of raw sequence, it is important to establish annotated libraries available to the broader scientific community. This white-footed mouse transcriptome will be useful in future studies of local adaptation, speciation, genome evolution, quantitative trait variation and investigation of the genetic basis of phenotypic traits (Vitti *et al.* 2013; Andrews & Luikart 2014; Seehausen *et al.* 2014). RNAseq is useful for targeting protein-coding regions of the genome as well as for quantifying gene expression based on normalized counts across cDNA samples from natural populations. These sequences serve as a digital measure of gene expression (Ozsolak & Milos 2010) and can be mapped to a reference genome or *de novo* transcriptome to measure differential expression or local adaptation (Lenz *et al.* 2013; Wolf 2013). The goal of *de novo* transcriptome assembly is to generate contigs containing the complete ORF for one gene, or at least sufficient coverage for accurate gene annotation. This goal can be difficult to achieve with the single-end short reads produced by SOLiD, but including longer sequence reads from 454 FLX+ pyrosequencing (Harris *et al.* 2013) significantly increased the length of resulting contigs and gene ORFs. These results are in line with previous studies on nonmodel organisms (e.g. hare, turtle, ant, oyster, tunicate) that found the best-quality transcriptome assemblies are generated when combining short and long reads (Cahais *et al.* 2012). Paired-end sequencing would further improve the transcriptome because paired sequences typically assemble better due to their known sequence distance from each other.

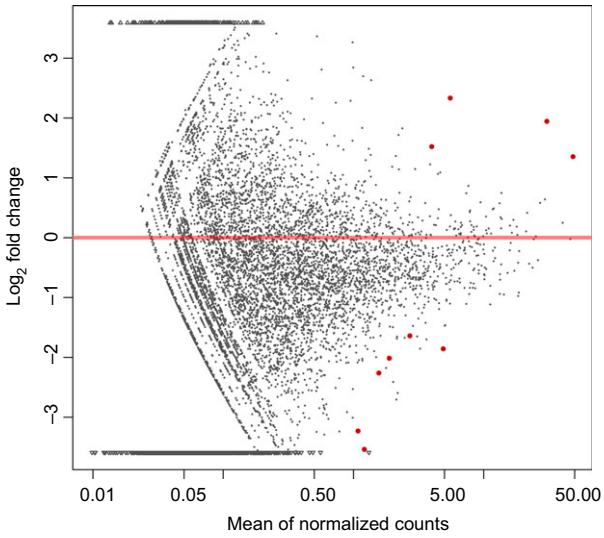

Fig. 4 Plot of normalized mean versus log2-fold change read number for the gene expression comparison between 'Rural' and 'Urban' groups using Q20 reads. Larger red circles represent genes exhibiting significant differences in expression.

Our assembled 30.06 Mb *P. leucopus* transcriptome successfully captured a large proportion of annotated protein-coding genes documented for model rodent species. The Mouse Genome Institute (MGI) lists 22 873 protein-coding genes, but we restricted our initial search to well-annotated genes with known function from the UNIPROT database (16 642 genes). Our assembly contained 98.2% of UNIPROT's manually curated genes in the *M. musculus* reference transcriptome. The size, number of contigs and N50 length are comparable with other *de novo* transcriptome assemblies using short-read sequencers (45 Mb, 57 840 contigs, N50 = 1378, carrot, Iorizzo *et al.* 2011; 54.6 Mb, 48 629 contigs, N50 = 1792, aleppo pine, Pinosio *et al.* 2014; and 103.1 Mb, 146 758 contigs, N50 = 1225, bank voles, Konczal *et al.* 2013). The sequence output for the SOLiD 5500 XL platform is relatively high with one of the lowest error rates (Glenn 2011). The resulting combination of increased depth of coverage and high confidence nucleotide calls facilitated rapid discovery of informative loci (SSRs or SNPs). The *P. leucopus* transcriptome (mean coverage = 35.7 reads per nucleotide) was mined for polymorphic bi-allelic SNP loci and simple sequence repeats. Despite stringent filtering of SNPs, a large number of variants were retained. With ~6 SNPs per contig, this transcriptome sequencing project provides a rich resource of information for population genomic analyses. The SNPs identified here are from coding regions and may not be selectively neutral, but several studies have demonstrated that synonymous polymorphisms can cautiously be used as neutral markers (Chapman *et al.* 2013; McCoy *et al.* 2013).

The assignment of genes to chromosomal linkage groups is a first in white-footed mouse genomics. Analysing patterns of linkage disequilibrium (LD) is a powerful way to identify signatures of recent selection or local adaption between geographically separated groups (Hohenlohe *et al.* 2010). Scans of LD to detect selection require known physical or genetic distances between loci (Akey 2009). Without a reference genome or genetic linkage map, LD scans are restricted to the length of contigs with only a few variable sites. The several thousand contigs within *P. leucopus* linkage groups will facilitate LD scans but must be used with caution. The draft *P. maniculatus* genome may contain rearrangements and portions that are misassembled, and there may be chromosomal inversions not accounted for here between *P. leucopus* and *P. maniculatus /polionotus* (Kenney-Hunt *et al.* 2014). In addition, there may be mapping errors between *P. leucopus* contigs and *P. maniculatus* genome scaffolds. In future analyses using these linkage groups, results should be viewed as candidate regions that should be confirmed by other tests of selection.

A subset of individuals was chosen to examine population structure within and between sampling sites. These analyses utilize allele frequency differences between populations to understand the demographic and evolutionary history of populations of interest (De Wit *et al.* 2012). Demographic histories can sometimes affect molecular data in ways that mimic signatures of selection (Nielsen *et al.* 2005), but if demographic histories can be estimated then true signatures of selection can be more confidently identified. With the sNMF analysis, individuals were assigned to two groups separated by locality (urban and rural, Fig. 2a). However, there was also support for a structure of five populations (Fig. 2b). For $K = 5$, urban individuals were assigned to separate NYC parks, while individuals from the three rural localities exhibited admixture and little to no genetic structure.

There was some admixture between urban populations, but such a result is not surprising given the relatively short time frame of urbanization in NYC. These results generally concur with the findings of Munshi-South & Kharchenko (2010) and Munshi-South (2012), but these microsatellite-based analyses examined structure only within NYC and not between NYC and surrounding rural areas. Taken together, these studies and the previous analyses indicate that urbanization in NYC has resulted in *P. leucopus* populations occupying small, highly fragmented habitat patches in the city with little to no gene flow between them. Without major modification of the NYC landscape, these populations will become increasingly differentiated from one another due to genetic drift (this study; Munshi-South & Kharchenko 2010) and local adaptation (Harris *et al.* 2013).

Surprisingly, there were only a handful of predicted genes that showed strong evidence of differential expression between urban and rural groups. This small group of genes was generally concordant with *a priori* ecological hypotheses. In the Q20 analysis, the differentially expressed genes could not be annotated with gene functions from either the *M. musculus* transcriptome or the non-redundant protein database. There were several differentially expressed genes discovered using trimmed reads with a nucleotide quality cut-off > 5. Although many bioinformatics pipelines for processing NGS data set a minimum cut-off of Q > 20, appropriate quality control of reads is important for the accuracy of specific downstream analyses (Macmanes & Eisen 2013; Zhou & Rokas 2014). Less stringent trimming criteria (Q > 5) in mapping and gene expression analysis increases the number of unique contigs and reads used to generate count data, increasing the accuracy of transcriptome wide gene expression quantification (Macmanes 2014). Using Q5, we identified three overexpressed genes. Similar to Q20 results, these included a rRNA promoter binding protein aiding in eukaryotic translation and an immunoregulatory protein, *art2*, found on the surface of T lymphocytes (Morrison *et al.* 2006). *KRTAP*, a keratin-associated protein, was identified as up-regulated in urban populations and is part of a family of keratin intermediate filaments that form in the hair cortex providing structure (Magrane and UniProt Consortium 2011). One down-regulated gene, *MALAT-1*, is a long noncoding RNA that regulates metastasis-associated genes, and when knocked-down in *M. musculus* xenografts leads to decreased tumour formation (Gutschner *et al.* 2013). Only one comparison between individual populations produced a significant result not seen in larger urban to rural tests. The gene, *ZIP14*, is responsible for the transport of zinc across membrane barriers and functions in the acute-phase response to inflammation and infection (Liuzzi *et al.* 2005). Total sequence output is of major importance for RNAseq (Wolf 2013) and while trimming reads with Q > 5 helped, more sequences per individual were probably needed to provide powerful analysis at the individual population comparison level (>10 million, Vijay *et al.* 2013).

Similar categories of genes exhibited signatures of recent positive selection in an earlier study of white-footed mouse transcriptomes from urban NYC populations (Harris *et al.* 2013). These results represent a general trend of immune function and protein modification as important to the success of urban populations of *P. leucopus*. However, these findings must be interpreted with caution. Individual white-footed mice were collected from wild populations across multiple years, at different times of the year and were not controlled for age at collection. Gene expression can be plastic across time and environmental conditions (Wolf 2013), and the lack of common garden conditions or biological replicates could have severely influenced gene expression inferences. Any differentially expressed genes may be due to age/environmental/individual variation and need to be treated as merely candidates for further investigation and hypothesis building. The purpose of this study was not to examine differences in gene expression, but a controlled field experiment employing RNASeq may yield more substantial results in the future. The site frequency spectrum, genetic differentiation and patterns of linkage disequilibrium in this dataset will be screened for signatures of directional selection by our research group in the future to further investigate genetic divergence between urban and rural populations of *P. leucopus*.

The white-footed mouse is an important emerging model species for a diverse array of ecological and evolutionary questions, and this transcriptome represents a substantial advance in the genomic resources available to *Peromyscus* researchers. While other large-scale sequencing projects for *Peromyscus* have been accomplished or are underway, only the raw data have been made publicly available. The large number of full-length annotated genes with known homology to model rodents, high-quality SNP library and preliminary genetic map presented and made publicly available here will facilitate comparative genomics studies and provide the basis for future population genomic analyses of *P. leucopus*. Highly isolated urban populations of white-footed mice may be experiencing selective pressures from the urban environment. Genomewide SNP data and the genetic map will be used with genome scans to identify outlier genes based on extreme genetic differentiation, allele frequencies indicative of selective sweeps or linkage disequilibrium in long haplotype blocks. These resources will facilitate our understanding of the genetic basis of adaptation and add to the growing body of research on the ecological and evolutionary consequences of urbanization.

Acknowledgements

We thank the New York State Department of Environmental Conservation, New Jersey Division of Fish and Wildlife, the Natural Resources Group of the NYC Department of Parks and Recreation, and the Central Park Conservancy for access to study sites. We thank Miranda Gonzalez, Justine Julien, Bo Reese, Eugene Kharonov, and Bartosz Baszynski for their assistance in the field and lab. Anna Santure, Zac Cheviron and two anonymous reviewers provided many helpful comments that improved this manuscript. This study was funded by grants from the National Science Foundation (DEB 0817259) and National Institute of General Medical Sciences/National Institutes of Health (1R15GM099055-01A1) to JMS, and a National

Science Foundation Graduate Research Fellowship to SHE. The NSF provided funding for RO; work was performed on instrumentation at the Center for Applied Genetics and Technology (UCONN).

References

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, **19**, 711–722.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

Andrews KR, Luikart G (2014) Recent novel approaches for population genomics data analysis. *Molecular Ecology*, **23**, 1661–1667.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Barrett T, Clark K, Gevorgyan R, Gorelenkov V *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*, **40**, D57–D63.

Blair RB (2001) Birds and butterflies along urban gradients in two ecoregions of the U.S. In: *Biotic Homogenization* (eds Lockwood JL & McKinney ML), pp. 33–56. Kluwer Academic Publishers, New York, NY.

Bradley RD, Durish ND, Rogers DS *et al.* (2007) Toward a molecular phylogeny for *Peromyscus*: evidence from mitochrondrial cytochrome-b sequences. *Journal of Mammalogy*, **88**, 1146–1159.

Cahais V, Gayral P, Tsagkogeorga G *et al.* (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular Ecology Resources*, **12**, 834–845.

Chapman MA, Hiscock SJ, Filatov DA (2013) Genomic divergence during speciation driven by adaptation to altitude. *Molecular Biology and Evolution*, **30**, 2553–2567.

Cheviron Z, Bachman GC, Connaty AD, McClelland GB, Storz JF (2012) Regulatory changes contribute to the adaptive enhancement of thermogenic capacity in high-altitude deer mice. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 8635–8640.

Cheviron Z, Connaty A, McClelland G, Storz J (2013) Functional genomics of adaptation to hypoxic cold-stress in high-altitude deer mice: transcriptomic plasticity and thermogenic performance. *Evolution*, **68**, 48–62.

Conesa A, Götz S, García-Gómez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, **21**, 3674–3676.

Consortium TU (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, **42**, D191–D198.

De Wit P, Pespeni MH, Ladner JT *et al.* (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, **12**, 1058–1067.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Dewey MJ, Dawson WD, Carolina S (2001) Deer mice: "the Drosophila of North American Mammalogy". *Genesis*, **29**, 105–109.

Dragoo JW, Lackey JA, Moore KE *et al.* (2006) Phylogeography of the deer mouse (*Peromyscus maniculatus*) provides a predictive framework for research on hantaviruses. *The Journal of General Virology*, **87**, 1997–2003.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

Faircloth BC (2008) msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources*, **8**, 92–94.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics*, **196**, 973–983.

Gering EJ, Opazo JC, Storz JF (2009) Molecular evolution of cytochrome b in high- and low-altitude deer mice (genus *Peromyscus*). *Heredity*, **102**, 226–235.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis1. *American Journal of Botany*, **99**, 312–319.

Gutschner T, Hämmerle M, Eissmann M *et al.* (2013) The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research*, **73**, 1180–1189.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Harris SE, Munshi-South J, Obergfell C, O'Neill R (2013) Signatures of rapid evolution in urban and rural transcriptomes of white-footed mice (*Peromyscus leucopus*) in the New York metropolitan area (N Johnson, Ed.). *PLoS One*, **8**, e74938.

Hohenlohe PA, Phillips PC, Cresko WA (2010) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International Journal of Plant Sciences*, **171**, 1059–1071.

Iorizzo M, Senalik DA, Grzebelus D *et al.* (2011) De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. *BMC Genomics*, **12**, 389.

Jiang L, Schlesinger F, Davis C (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, **21**, 1543–1551.

Kalkvik HM, Stout IJ, Doonan TJ, Parkinson CL (2012) Investigating niche and lineage diversification in widely distributed taxa: phylogeography and ecological niche modeling of the *Peromyscus maniculatus* species group. *Ecography*, **35**, 54–64.

Keesing F, Brunner J, Duerr S *et al.* (2009) Hosts as ecological traps for the vector of Lyme disease. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 3911–3919.

Kenney-Hunt J, Lewandowski A, Glenn TC *et al.* (2014) A genetic map of *Peromyscus* with chromosomal assignment of linkage groups (a *Peromyscus* genetic map). *Mammalian Genome*, **25**, 160–179.

Konczal M, Koteja P, Stuglik MT, Radwan J, Babik W (2013) Accuracy of allele frequency estimation using pooled RNA-Seq. *Molecular Ecology Resources*, **14**, 381–392.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Lenz T, Eizaguirre C, Rotter B, Kalbe M, Milinski M (2013) Exploring local immunological adaptation of two stickleback ecotypes by experimental infection and transcriptome-wide digital gene expression analysis. *Molecular Ecology*, **22**, 774–786.

Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. *Science (New York, N.Y.)*, **325**, 1095–1098.

Linnen CR, Poh Y-P, Peterson BK *et al.* (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science*, **339**, 1312–1316.

Liuzzi JP, Lichten LA, Rivera S *et al.* (2005) Interleukin-6 regulates the zinc transporter Zip14 in liver and contributes to the hypozincemia of the acute-phase response. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 6843–6848.

Macmanes MD (2014) On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, **5**, 13.

Macmanes MD, Eisen MB (2013) Improving transcriptome assembly through error correction of high-throughput sequence reads. *PeerJ*, **1**, e113.

Magrane M, UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database: The Journal of Biological Databases and Curation*. doi: 10.1093/database/bar009.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet*, **17**, 10–12.

McCormack JE, Maley JM, Hird SM *et al*. (2011) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Phylogenetics and Evolution*, **62**, 397–406.

McCoy RC, Garud NR, Kelley JL, Boggs CL, Petrov D (2013) Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population. *Molecular Ecology*, **23**, 136–150.

Merilä J, Hendry AP (2014) Climate change, adaptation, and phenotypic plasticity: the problem and the evidence. *Evolutionary Applications*, **7**, 1–14.

Morrison AR, Moss J, Stevens LA *et al*. (2006) ART2, a T cell surface mono-ADP-ribosyltransferase, generates extracellular poly(ADP-ribose). *The Journal of Biological Chemistry*, **281**, 33363–33372.

Morzunov SP, Rowe JE, Ksiazek TG *et al*. (1998) Genetic analysis of the diversity and origin of hantaviruses in *Peromyscus leucopus* mice in North America. *Journal of Virology*, **72**, 57–64.

Mossman CA, Waser PM (2001) Effects of habitat fragmentation on population genetic structure in the white-footed mouse (*Peromyscus leucopus*). *Canadian Journal of Zoology*, **79**, 285–295.

Mullen LM, Hoekstra HE (2008) Natural selection along an environmental gradient: a classic cline in mouse pigmentation. *Evolution*, **62**, 1555–1570.

Munshi-South J (2012) Urban landscape genetics: canopy cover predicts gene flow between white-footed mouse (*Peromyscus leucopus*) populations in New York City. *Molecular Ecology*, **21**, 1360–1378.

Munshi-South J, Kharchenko K (2010) Rapid, pervasive genetic differentiation of urban white-footed mouse (*Peromyscus leucopus*) populations in New York City. *Molecular Ecology*, **19**, 4242–4254.

Munshi-South J, Nagy C (2014) Urban park characteristics, genetic variation, and historical demography of white-footed mouse (*Peromyscus leucopus*) populations in New York City. *PeerJ*, **2**, e310.

Natarajan C, Inoguchi N, Weber R *et al*. (2013) Epistasis among adaptive mutations in deer mouse hemoglobin. *Science*, **340**, 1324–1327.

Nielsen R, Williamson S, Kim Y *et al*. (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

O'Neill R, Szalai G, Gibbs R *et al*. (1998) *White Paper Proposal for Sequencing the Genome of Peromyscus*. National Human Genome Research Institute, http://www. genome.gov/pages/research/sequencing/seqproposals/peromyscus.pdf.

Ostfeld R (2012) *Lyme Disease: The Ecology of a Complex System*. Oxford University Press, New York, NY.

Ozsolak F, Milos P (2010) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, **12**, 87–98.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.

Pavey SA, Bernatchez L, Aubin-Horth N, Landry CR (2012) What is needed for next-generation ecological and evolutionary genomics? *Trends in Ecology & Evolution*, **27**, 673–678.

Pergams ORW, Lacy RC (2007) Rapid morphological and genetic change in Chicago-area *Peromyscus*. *Molecular Ecology*, **17**, 450–463.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, **7**, e37135.

Pinosio S, González-Martínez SC, Bagnoli F *et al*. (2014) First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, *Pinus halepensis* Mill. *Molecular Ecology Resources*, **14**, 846–856.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsdell CM, Lewandowski AA, Glenn JLW *et al*. (2008) Comparative genome mapping of the deer mouse (*Peromyscus maniculatus*) reveals greater similarity to rat (*Rattus norvegicus*) than to the lab mouse (*Mus musculus*). *BMC Evolutionary Biology*, **8**, 65.

Renn SCP, Siemens DH (2010) Ecological genomics – changing perspectives on Darwin's basic concerns. *Molecular Ecology*, **19**, 3025–3030.

Rogic A, Tessier N, Legendre P, Lapointe F-J, Millien V (2013) Genetic structure of the white-footed mouse in the context of the emergence of Lyme disease in southern Québec. *Ecology and Evolution*, **3**, 2075–2088.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365–386.

Sacomoto GAT, Kielbassa J, Chikhi R *et al*. (2012) KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, **13**, S5.

Seehausen O, Butlin RK, Keller I *et al*. (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Sih A, Ferrari MCO, Harris DJ (2011) Evolution and behavioural responses to human-induced rapid environmental change. *Evolutionary Applications*, **4**, 367–387.

Steiner CC, Weber JN, Hoekstra HE (2007) Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS Biology*, **5**, e219.

Storz JF, Sabatino SJ, Hoffmann FG *et al*. (2007) The molecular basis of high-altitude adaptation in deer mice. *PLoS Genetics*, **3**, e45.

Storz JF, Runck AM, Sabatino SJ *et al*. (2009) Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse hemoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 14450–14455.

Storz JF, Runck AM, Moriyama H, Weber RE, Fago A (2010) Genetic differences in hemoglobin function between highland and lowland deer mice. *The Journal of Experimental Biology*, **213**, 2565–2574.

Taylor ZS, Hoffman SMG (2014) Landscape models for nuclear genetic diversity and genetic structure in white-footed mice (*Peromyscus leucopus*). *Heredity*, **112**, 588–595.

Toth G (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, **10**, 967–981.

Van der Auwera GA, Carneiro MO, Hartl C *et al*. (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, **11**, 11.10.1–11.10.33.

Vessey S, Vessey KB (2007) Linking behavior, life history and food supply with the population dynamics of white-footed mice (*Peromyscus leucopus*). *Integrative Zoology*, **2**, 123–130.

Vijay N, Poelstra J, Künstner A, Wolf J (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology*, **22**, 620–634.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*, **47**, 97–120.

Weber JN, Peterson BK, Hoekstra HE (2013) Discrete genetic modules are responsible for complex burrow evolution in *Peromyscus* mice. *Nature*, **493**, 402–405.

Wolf JBW (2013) Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular Ecology Resources*, **13**, 559–572.

Wolff JO (1985) Comparative population ecology of *Peromyscus leucopus* and *Peromyscus maniculatus*. *Canadian Journal of Zoology*, **63**, 1548–1555.

Zhou X, Rokas A (2014) Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. *Molecular Ecology*, **23**, 1679–1700.

Conceived and designed the experiments: S.E.H., J.M.S., and R.O. Performed the experiments: S.E.H., and J.M.S. Analysed the data: S.E.H., and J.M.S. Contributed reagents/materials/analysis tools: J.M.S., R.O., and S.E.H. Wrote the manuscript: S.E.H., J.M.S., and R.O.

Data Accessibility

DNA sequences: NCBI SRA: SRP020005

Transcriptome information: DRYAD entry doi: 10.5061/dryad.6hc0f

Table S1. SSR sequence and primers (doc file).

Table S2. Chromosomal order and assignment (csv file).

Contig sequences for *P. leucopus* transcriptome (fasta file).

SNP information (vcf file).

Annotation information, BLAST results, GO information (.dat file for BLAST2GO).

R scripts, bash scripts, and full analysis workflow (txt file).