Population Genomics of White-Footed Mice (Peromyscus leucopus) in New York City

Stephen Edward Harris
Graduate Center, City University of New York

Recommended Citation
Harris, Stephen Edward, "Population Genomics of White-Footed Mice (Peromyscus leucopus) in New York City" (2015). CUNY Academic Works.
https://academicworks.cuny.edu/gc_etds/965

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
Population Genomics of White-Footed Mice (*Peromyscus leucopus*) in New York City

By

Stephen Edward Harris

A dissertation submitted to the Graduate Faculty in Biology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2015
This manuscript has been read and accepted for the Graduate Faculty in Biology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Jason Munshi-South

Date

Chair of Examining Committee

Laurel Eckhardt

Date

Executive Officer

Ana C. Carnaval

Michael J. Hickerson

Stephane Boissinot

Steven J. Franks

Supervisory Committee

The City University of New York
Abstract

Population Genomics of White-Footed Mice (Peromyscus leucopus) in New York City

By

Stephen Edward Harris

Advisor: Dr. Jason Munshi-South

Urbanization significantly alters natural ecosystems. New York City (NYC) is one of the oldest and most urbanized cities in North America, but still maintains substantial populations of some native wildlife. The white-footed mouse, Peromyscus leucopus, is a common resident of NYC’s forest fragments, and isolated populations may adapt in response to novel urban ecosystems. Using pooled transcriptome-wide RNAseq data, individually barcoded transcriptome-wide RNAseq data, and genome-wide RADseq data, I found genetic differentiation between urban and rural P. leucopus populations and evidence suggestive of local adaptation. I compared genome and transcriptome-wide SNP data in P. leucopus from relatively large urban parks surrounded by dense urban infrastructure to large rural sites representative of native habitat. First, I built a publicly available genomic resource for P. leucopus, and then looked at patterns of genetic differentiation in protein-coding DNA sequences that showed divergence between urban and rural populations. I also looked at the unique demographic history of urban and rural populations of P. leucopus using coalescent simulations and the site frequency spectrum (SFS). Historical demographic inference supported a scenarios of post-glacial sea level rise that led to isolation of mainland and Long Island populations. I also found that several urban parks in NYC represent distinct P. leucopus populations, and the estimated divergence times for these parks are consistent with patterns of urbanization in NYC. I then looked at transcriptome sequences within urban P. leucopus to look for signatures of genetic differentiation and selective
sweeps due to positive selection, and then associated outliers with environmental measures of urbanization. The majority of candidate genes were involved in metabolic functions, especially dietary specialization. Candidate loci I identified suggest that populations of *P. leucopus* are using novel food resources in urban habitats or metabolizing nutrients differently. Ultimately, the data indicate that cities may represent novel ecosystems with selective pressures from urbanization leading to adaptive responses in populations of *Peromyscus leucopus*. 
Acknowledgements

I would like to 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, the Central Park Conservancy, Ellen Pehek, and Jessica Schuler for access to NYC study sites. I also thank the following persons and institutions for providing invaluable assistance in the field and lab including Paolo Cocco, Julie Sesina, Diane Jacob, Miranda Gonzalez, Justine Julien, Bo Reese, Eugene Kharonov, and Bartosz Baszynski. Funding for this research was provided in part by the National Science Foundation (DEB 0817259), National Institute of General Medical Sciences / National Institutes of Health (1R15GM099055-01A1), a National Science Foundation Graduate Research Fellowship, and a CUNY Research Competition Grant. I also give special thanks for the support, advice and encouragement of friends and fellow CUNY colleagues Terry Demos, J.T. Boehm, John Robinson, Alexander Xue, Susan Tsang, Xin Chen, Tyler Joseph, Diego Alvarado-Serrano, Miguel Pinto, Elliot Aguilar, and Ana Cristina Fazza. I give a special thank you to research collaborators at UCONN, Rachel O’Neill and Nathaniel Jue for the advice and training in Next Generation Sequencing techniques. In addition, a special thanks to my family for all their support including Cindy Harris, Robert Harris, Bryan and Jamie Harris, Megan and Brian Rosenthal, Elizabeth Dupuy, Alain Dupuy, Laetitia Dupuy, Clement Dupuy, my amazing wife, Elodie Dupuy, and our daughter, whose impending birth forced me to finish this dissertation in time. Finally, I thank my advisor Jason Munshi-South for his guidance and training on how to actually be a scientist, and the members of my Ph.D. committee, Ana Carolina Carnaval, Michael J Hickerson, Stephane Boissinot, and Steven J Franks, for their sound advice, insight, interest, and support throughout the course of this study.
Table of Contents

Abstract........................................................................................................................................iv
Acknowledgements..................................................................................................................vi
Table of Contents..................................................................................................................vii
List of Tables................................................................................................................................viii
List of Figures...........................................................................................................................ix
List of Appendices....................................................................................................................x

Chapter 1:
Signatures of rapid evolution in urban and rural transcriptomes of white-footed mice
(Peromyscus leucopus) in the New York metropolitan area.......................................................1

Chapter 2:
Transcriptome resources for the white-footed mouse (Peromyscus leucopus): new genomic
tools for investigating ecologically divergent urban and rural populations...........................45

Chapter 3:
Genome-wide SNP data reveal that urbanization impacts the demographic history of a
New York City rodent...................................................................................................................74

Chapter 4:
Scans for positive selection reveal candidate genes involved in local adaptation to
urbanization................................................................................................................................106

Appendix 1.1..............................................................................................................................149
Appendix 1.2..............................................................................................................................150
Appendix 1.3..............................................................................................................................151
Appendix 1.4..............................................................................................................................152
Appendix 1.5................................................................................................................................155
Appendix 2.1................................................................................................................................158
Appendix 3.1................................................................................................................................163
Appendix 3.2................................................................................................................................164
Appendix 3.3................................................................................................................................165
Appendix 4.1................................................................................................................................167
Appendix 4.2................................................................................................................................172
Bibliography..............................................................................................................................173
List of Tables

**Table 1.1** Results of transcriptome assembly using three different approaches ......................32

**Table 1.2** BLASTN search results of three *P. leucopus* transcriptome assemblies against reference cDNA libraries from *Mus* and *Rattus* .................................................................33

**Table 1.3** Over-represented GO terms for individual tissue types from Fisher’s Exact tests (FDR ≤ 0.5) in Blast2Go .................................................................34

**Table 1.4** Candidate loci exhibiting $p_N/p_S > 1$ .................................................................36

**Table 1.5** McDonald-Kreitman tests for candidate genes with $p_N/p_S > 1$ ..........................38

**Table 2.1** Specimen collection locations for *Peromyscus leucopus* individuals used in this study .................................................................67

**Table 2.2** Transcriptome assembly statistics .................................................................68

**Table 2.3** Annotation statistics from BLASTx of *Peromyscus leucopus* contigs against *Mus musculus* Uniprot (2013_11) database .................................................................69

**Table 3.1** Inferred demographic parameters with 95% confidence values from parametric bootstrapping for all models .................................................................99

**Table 4.1** Summary statistic averages for six *P. leucopus* populations ..................................140

**Table 4.2** Average pairwise $F_{ST}$ among six *P. leucopus* populations ..............................141

**Table 4.3** Results for selection from Bayescan and associations with environmental variables across urban and rural populations ..............................................142

**Table 4.4** Results for selection from SweeD and associations with environmental variables across urban and rural populations ..............................................144
List of Figures

Figure 1.1 Frequency of contig lengths for three transcriptome assembly methods………………..39

Figure 1.2 Transcriptome alignment to reference rodent genomes…………………………………40

Figure 1.3 Annotation of final reference transcriptome……………………………………………..41

Figure 1.4 Over-represented GO terms from pairwise tissue comparisons (FDR ≤0.05)………42

Figure 1.5 Non-synonymous (p_N) SNP substitutions plotted vs. synonymous (p_S) substitutions for 354 genes……………………………………………………………………….43

Figure 1.6 Location and number of individuals collected from five populations in the NYC metropolitan area………………………………………………………………………………44

Figure 2.1 Distribution of alignment coverage………………………………………………………70

Figure 2.2 Population structure analysis of RNAseq data…………………………………………71

Figure 2.3 Inferred locations of protein-coding genes on P. leucopus chromosomes……………72

Figure 2.4 Plot of normalized mean versus log2 fold change read number for the gene expression comparison between ‘Rural’ and ‘Urban’ groups using Q20 read……..73

Figure 3.1 Map of NYC region…………………………………………………………………….101

Figure 3.2 Three main demographic histories tested in fastsimcoal2…………………………102

Figure 3.3 Scatterplot generated from discriminant analysis of principle components………103

Figure 3.4 Population structure analysis of RADseq data………………………………………..104

Figure 3.5 Box plots of parametric bootstrap results for inferred demographic parameters…..105

Figure 4.1 Map of NYC and surrounding area showing included sample localities…………146

Figure 4.2 Bayescan and SweeD results………………………………………………………….147

Figure 4.3 Population values for three environmental variables and allele frequencies for three contigs found as outliers in a genome scan and GEA test…………………..148
List of Appendices

Appendix 1.1 Frequency distribution of depth of coverage (reads / contig)…………………………149

Appendix 1.2 Distribution of species with the most top-hit BLASTX results in Blast2Go……150

Appendix 1.3 Sequencing and assembly statistics for Newbler cDNA transcriptome assembly………………………………………………………………………………151

Appendix 1.4 Full list of over represented GO terms for all tissue pairwise comparisons……152

Appendix 1.5 Candidate loci with $p_N/p_S$ between 0.5 and 1……………………………………155

Appendix 2.1 Microsatellite repeats identified…………………………………………………………158

Appendix 3.1 sNMF results for K=2 (highest support) of ≤ 15% missing
SNP dataset (3,664 SNPs)………………………………………………………………………………163

Appendix 3.2 Membership probability plots generated from compoplot for
global, MM, and LI groups of individuals……………………………………………………………164

Appendix 3.3 TPL file and EST file (input files for fastsimcoal2)……………………………………165

Appendix 4.1 Blast2GO table with BLASTX hits and top three supported
Gene Ontology terms for outlier genes from Bayescan and SweeD……………………………………167

Appendix 4.2 KEGG analysis for biochemical pathways containing more than
one outlier contig……………………………………………………………………………………………………172
Chapter 1

Signatures of rapid evolution in urban and rural transcriptomes of white-footed mice

(Peromyscus leucopus) in the New York metropolitan area

Abstract

Urbanization is a major cause of ecological degradation around the world, and human settlement in large cities is accelerating. New York City (NYC) is one of the oldest and most urbanized cities in North America, but still maintains 20% vegetation cover and substantial populations of some native wildlife. The white-footed mouse, Peromyscus leucopus, is a common resident of NYC’s forest fragments and an emerging model system for examining the evolutionary consequences of urbanization. In this study, I developed transcriptomic resources for urban P. leucopus to examine evolutionary changes in protein-coding regions for an exemplar ‘urban adapter’. I used Roche 454 GS FLX+ high throughput sequencing to derive transcriptomes from multiple tissues from individuals across both urban and rural populations. From these data, I identified 31,015 SNPs and several candidate genes potentially experiencing positive selection in urban populations of P. leucopus. These candidate genes are involved in xenobiotic metabolism, innate immune response, demethylation activity, and other important biological phenomena in novel urban environments. This study is one of the first to report candidate genes exhibiting signatures of directional selection in divergent urban ecosystems.

Keywords: Peromyscus leucopus, transcriptome, next-generation sequencing, RNAseq, positive selection
Introduction

Urbanization dramatically alters natural habitats (Shochat et al. 2006), and its speed and intensity will increase as over two-thirds of the world’s human population is predicted to live in urban areas by 2050 (United Nations 2011). Understanding how natural populations adapt to ecologically divergent urban habitats is thus an important and immediate goal for urban ecologists and evolutionary biologists. Few ecological and evolutionary studies are conducted in urban environments (Martin et al. 2012), but recent attitude shifts and technological advancements have removed many of the obstacles to working on urban wildlife. Multiple studies have demonstrated that urban areas are biologically diverse, productive, and viable (Pickett et al. 2011), and the development of next generation sequencing (NGS) has facilitated the generation of genomic resources for uncharacterized species in natural environments (Storz & Hoekstra 2007; Rice et al. 2010; Hohenlohe et al. 2011). Understanding the genetic basis of adaptation in successful urban species will aid in future conservation efforts and provide insights into the effects of other anthropogenic factors, such as global climate change and evolutionary trajectories in human-dominated environments (White et al. 2005; Grimm et al. 2008; Pickett et al. 2011).

Cities typically experience a substantial decrease in biodiversity of many taxonomic groups as urban ‘avoiders’ disappear, accompanied by a rise in urban ‘exploiters’ that are primarily non-native human commensals such as pigeons or rats. Urban ‘adapters’ are native species that favor disturbed edge habitats such as urban forest fragments, relying on a combination of wild-growing and human-derived resources (Blair 2001; McKinney 2002, 2006). This last group is of primary interest for examining genetic signatures of recent evolutionary change in novel urban environments. Severe habitat fragmentation is one of the primary impacts
of urbanization and often leads to genetic differentiation between populations (Wandeler et al. 2003a; Shochat et al. 2006; Bjorklund et al. 2010). Introductions of new predators and competitors alter ecological interactions (Peluc et al. 2008), and new or more abundant parasites or pathogens influence immune system processes (Sih et al. 2011). Air, water, and soil pollution typically increase in local urban ecosystems, and selection may favor previously-rare genetic variants that more efficiently process these toxins (Yauk et al. 2008; Whitehead et al. 2010; Francis & Chadwick 2012). Recent studies provide some evidence of local adaptation and rapid evolution in urban patches. Using a candidate gene approach, Mueller et al. (Mueller et al. 2013a) found consistent genetic divergence between behavioral genes for circadian behavior, harm avoidance, migratory behavior and exploratory behavior in multiple urban-rural population pairs of the common blackbird, Turdus merula. Examining phenotypes, Brady (Brady 2012) found rapid adaptation to roadside breeding pond conditions in the salamander, Ambystoma maculatum, and Cheptou et al. (Cheptou et al. 2008) reported a heritable increase in production of non-dispersing seeds in the weed, Crepis sancta, over 5-12 generations in fragmented urban tree pits. The genetic architecture of the phenotypes under selection has not been described for either of these urban ‘adapters’, but outlier scans of transcriptome sequence datasets are one promising approach (Sun et al. 2012).

Peromyscus spp. are an emerging model system for examining evolution in wild populations (Storz et al. 2007b; Linnen et al. 2009; Weber et al. 2013), but large-scale genomic resources are not yet widely available. The genus contains the most widespread and abundant small mammals in North America, and Peromyscus research on population ecology, adaptation, aging, and disease has a long, productive history (Metzger 1971; O’Neill et al. 1998; Vessey & Vessey 2007; Wang et al. 2008; Ungvari et al. 2008). An increasing number of studies have
demonstrated that *Peromyscus* spp. rapidly (i.e. in several hundreds to thousands of generations) adapt to divergent environments. These examples include adaptation to hypoxia in high altitude environments (Storz *et al.* 2007b) and adaptive variation in pelage color on light-colored soil substrates (Mullen & Hoekstra 2008a; Linnen & Hoekstra 2009; Linnen *et al.* 2009). Presently, *P. leucopus* is the sole *Peromyscus* spp. in New York City (J. Munshi-South, unpublished data) and searches of the Mammal Networked Information System (MANIS) database indicate that *P. maniculatus* has not occurred in NYC for several decades. In NYC, *P. leucopus* occupies most small patches of secondary forest, shrublands, and meadows within NYC parklands (Mullen & Hoekstra 2008b; Puth & Burns 2009). The smallest patches in NYC often contain the highest population densities of white-footed mice (Ekernas & Mertes 2007), most likely due to ecological release and obstacles to dispersal (Nupp & Swihart 1996; Barko *et al.* 2003).

Consistently elevated population density in urban patches compared to surrounding rural populations is predicted to result in density-dependent selective pressures on traits related to immunology, intraspecific competition, and male-male competition for mating opportunities, among others (Lankau 2010; Lankau & Strauss 2011).

White-footed mouse populations in NYC exhibit high levels of heterozygosity and allelic diversity at neutral loci within populations, but genetic differentiation and low migration rates between populations (Munshi-South & Kharchenko 2010; Munshi-South 2012). This genetic structure contrasts with weak differentiation reported for *Peromyscus* spp. at regional scales (Yang & Kenagy 2009), or even between populations isolated on different islands for thousands of generations (Degner *et al.* 2007; Ozer *et al.* 2011). High genetic diversity within and low to nonexistent migration between most NYC populations suggests that selection can operate efficiently within these geographically isolated populations, either on standing genetic variation
or de novo mutations. In this study I take steps to develop *P. leucopus* as a genomic model for adaptive change in urban environments.

Pooling mRNA from multiple individuals is an effective approach to transcriptome sequencing that avoids the prohibitive cost of sequencing individual genomes (Futschik & Schlötterer 2010; Boitard *et al.* 2012). While pooling results in the loss of genetic information from individuals, the ability to identify SNPs in a population increases due to the inclusion of multiple individuals in the pool (Gompert & Buerkle 2011). By analyzing SNPs within thousands of transcripts, it is feasible to identify candidate genes underlying rapid divergence of populations in novel environments (Stinchcombe & Hoekstra 2008; Ungerer *et al.* 2008; Bonin 2008; Rice *et al.* 2010). Certain statistical approaches, such as the ratio between non-synonymous and synonymous (*p*<sub>N</sub>*⁄*<sub>S</sub>) substitutions, can be applied to pooled transcriptome data to identify potential signatures of selection between isolated populations (Ungerer *et al.* 2008; Sloan *et al.* 2012; Sun *et al.* 2012). If positive selection is acting on a codon, then non-synonymous mutations should be more common than under neutral expectations (Nielsen & Yang 1998; Ellegren 2008).

Here, I describe the results of de novo transcriptome sequencing, annotation, SNP discovery, and outlier scans for selection among urban and rural white-footed mouse populations. I used the 454 GS FLX+ system to sequence cDNA libraries generated from pooled mRNA samples from multiple tissues and populations. Several de novo transcriptome assembly programs were used and the contribution of specific tissue types to the transcriptome assembly was examined. I then identified coding region SNPs between urban and rural populations, and scanned this dataset for signatures of positive selection using *p*<sub>N</sub>*⁄*<sub>S</sub> between populations and McDonald-Kreitman tests between multiple species. I report several candidate genes potentially
experiencing directional selection in urban environments, and provide annotated transcriptome datasets for future evolutionary studies of an emerging model system.

Materials and Methods

Ethics statement

All animal procedures were approved by the Institutional Animal Care and Use Committee at Brooklyn College, CUNY (Protocol No. 247), and adhered to the Guidelines of the American Society of Mammalogists for the Use of Wild Mammals in Research (Sikes & Gannon 2011). Field work was conducted with the permission of the New York State Department of Environmental Conservation (License to Collect or Possess Wildlife No. 1603) and the New York City Department of Parks and Recreation.

Study Sites and population sampling

*P. leucopus* were trapped and collected from each of four urban and one rural site (*N* = 20-25 / population) for sequencing and analysis (total *N* = 112; Fig. 6). The four urban sites (Central Park, Flushing Meadows-Willow Lake, New York Botanical Gardens, and the Ridgewood Reservoir) were chosen due to their large area, isolation by dense urban matrix, high population density of mice, substantial genetic differentiation, and genetic isolation from other populations (Munshi-South & Kharchenko 2010; Munshi-South 2012). The rural site, Harriman State Park located ~68 km north of Central Park, is one of the largest contiguous protected areas nearby and the most likely representative of a non-urban population of mice in proximity to NYC. Mice were trapped over a period of 1-3 nights at each site using four 7x7 transects of 3”x3”x9” Sherman live traps. Mice were killed by cervical dislocation and immediately dissected in the
field. Livers, gonads and brains were extracted, rinsed with PBS to remove any debris from the surface of the tissue, and immediately placed in RNALater® (Ambion Inc., Austin, TX) on ice before transport and storage at -80°C. These tissue types were chosen for initial analysis due to their wide range of expressed gene transcripts (Yang et al. 2006) and potential roles in adaptation to urban conditions.

**RNA extraction and cDNA library preparation**

Total RNA was extracted and cDNA libraries were pooled for all five populations for four multiplexed plates of 454 sequencing. The first plate of sequencing was normalized to produce equalized concentrations of all transcripts present, potentially allowing enhanced gene discovery and greater overall coverage of the transcriptome (Babik et al. 2010). However, the normalization process introduces additional steps and biases in library preparation (Babik et al. 2010), and resulted in a relatively low number of total high-quality 454 reads. Normalization attempts to reduce the amount of overly abundant transcripts, but this process removes the possibility of measuring expression changes and can indiscriminately remove some rare reads, potentially reducing the number of unique contigs and called variants in downstream analyses (Camp et al. 2012). Thus, non-normalized libraries were prepared using a modified protocol for the last three 454 plates.

For plate 1, total RNA was isolated from ~60 mg of liver (eight males and eight females / population), ~60 mg of testis (eight males / population), and ~60 mg of ovaries (eight females / population) for two populations using RNaqueous® kits (Ambion, Austin, TX). An Agilent Bioanalyzer was used to determine concentrations before individual RNA extracts were pooled in equimolar amounts by population and organ type. The pools were then enriched for mature
mRNA using the MicroPoly(A)Purist™ kit (Ambion, Austin, TX). Next, mRNA pools were reverse-transcribed using the SMARTer™ cDNA synthesis kit (Clontech, Mountain View, CA), and normalized using the Trimmer-Direct cDNA normalization kit (Evrogen, Moscow, Russia). Then, normalized cDNA pools were sequenced with multiplex identifiers using standard 454 FLX Titanium protocols. This pilot plate contained cDNA pools for Harriman State Park and Flushing Meadows-Willow Lake.

For plates 2-4, total RNA using Trizol® reagent (Invitrogen, Carlsbad, CA) was extracted from ~70 mg of brain tissue (four males and four females / population), ~70 mg of testes (eight males / population), and ~15 mg of liver (four males and four females / population). After DNase treatment (Promega, Madison, WI) and pooling individual samples in equimolar amounts by population and tissue, the samples were treated with the RiboMinus™ Eukaryote kit (Invitrogen, Carlsbad, CA) to reduce ribosomal RNA. RNA pools were then reverse-transcribed using the Roche cDNA synthesis kit (Roche Diagnostics, Indianapolis, IN) and sequenced with multiplex identifiers using standard 454 FLX Titanium protocols. Plate 2 included brain cDNA pools for all five populations, plate 3 included liver and testis cDNA for Central Park, Ridgewood Reservoir, New York Botanical Gardens, and Harriman State Park, and plate 4 included liver cDNA pools from all five populations. All raw sequencing files have been deposited in the GenBank Sequence Read Archive (SRA) under accession number SRP020005.

**Transcriptome assembly**

Two methods were used to assemble the best transcriptome from all four 454 plates: Cap3 (Huang & Madan 1999) a long-read assembler that performs well in transcriptome assemblies (Cahais et al. 2012), and Roche’s proprietary software, Newbler (Version 2.5.3), that was
designed specifically for assembling 454 sequencing reads with additional features for cDNA sequence. Newbler’s cDNA options assemble reads into contigs, followed by assembly into larger ‘isotigs’ representing alternatively-spliced transcripts. Isotigs are then clustered into larger ‘isogroups’ representing full-length genes. Transcriptome assembly was attempted with the full set of reads using Cap3 and Newbler with cDNA options, but due to computational limitations the full dataset could not be assembled with either software program. I addressed this issue by first assembling sequences from all four plates with Newbler using the genome assembly settings and default parameters after trimming 454 adaptors and barcodes from the reads. Reads that were either ‘assembled’ or ‘partially assembled’ in this pilot run were filtered and used as input for cDNA assemblies in Newbler or Cap3. These reads were filtered from the raw sff files using a locally-installed instance of Galaxy (Blankenberg et al. 2010). Before the cDNA assembly, nucleotides with poor quality scores, primer sequences, and long poly(A) tails were removed using cutadapt (Version 1.2.1 2012, (Martin 2011)) and the trim-fastq.pl perl script implemented in Popoolation (Kofler et al. 2011). The filtered fastq files were then used as input for Cap3 or Newbler with the cDNA assembly option, using default parameters for both assemblies. These assemblies (1. genome assembly with Newbler, 2. cDNA assembly with Newbler, and 3. cDNA assembly with Cap3) were compared to identify the best full reference transcriptome for downstream analysis.

For analyses of individual tissues, separate cDNA assemblies were performed. Tissues were barcoded, and sequence reads originating from liver, gonads, or brains were parsed from the raw 454 sequencing reads. These datasets were small enough to be assembled separately as tissue-specific transcriptomes in Newbler using the cDNA option with default parameters.
Population-specific transcriptomes were also assembled, using the same methodology, to examine population-specific statistical signatures of selection.

**Alignment to model rodent genomes**

*Peromyscus* assemblies were initially characterized and annotated by performing two separate analyses using *Mus musculus* and *Rattus norvegicus* genomic resources. The first analysis was used to determine the number of likely genes in each assembly. BLASTN searches were performed against *Mus musculus* (NCBI Annotation Release 103) and *Rattus norvegicus* (NCBI build 5.1) cDNA reference libraries downloaded from NCBI. BLASTN matches were considered significant when sequence identity was greater than 80%, alignment length was at least 50% of the total length of either the query or subject sequence, and the e-value was less than $10^{-5}$. While significant, these hits may not be ideal for population genomic analyses due to inclusion of paralogous gene matches, matches between multi-gene families, and false positive orthologous gene matches. In order to identify individual isotigs representing a single gene with known function useful for statistical analysis, BLASTN results were further filtered by including query hits that matched only one subject ID (i.e. gene) and *vice versa*. These contigs were annotated as ‘Gene Candidates’.

The distribution of *P. leucopus* isotigs across model rodent genomes was analyzed. All *P. leucopus* isotigs were mapped to chromosomes in the *Mus* (GRCm38) and *Rattus* (RGSC 5.0) reference genomes. Default BLAT parameters were used with an exception for aligning mRNA to genomes across species (-q=rna -t=dnax, (Kent 2002)), and best BLAT hits were parsed based on percent identity and score (# match – # mismatch).
Mapping and SNP discovery

To generate a SNP library for downstream population genomic analysis, 454 reads were first mapped to the Newbler cDNA assembly using the BWA-SW (http://bio-bwa.sourceforge.net/) alignment algorithm for long read mapping (Li & Durbin 2010). I only used trimmed reads from the final assembly, removed singletons before mapping to reduce false positive SNP calls from sequencing errors or duplicate reads, and included reads with a mapping quality > 20 in SAMtools. The SAM file from BWA-SW was used in the SAMtools package (v. 0.1.17, (Li et al. 2009)) to call SNPs using the mpileup command with a maximum coverage cutoff of 200. The SNP calling pipeline implemented in SAMtools uses base alignment quality (BAQ) calculations to generate likelihoods of genotypes, can overcome low coverage by using sequence information from multiple samples to call variants, and uses Bayesian inference to make SNP calls with high confidence (Li et al. 2009; Nielsen et al. 2011; Altmann et al. 2012). In addition to the default parameters in SAMtools, I included stringent additional filters by removing any potential INDELs, only including SNPs with a phred quality (Q-value) ≥ 20, a minimum occurrence of two, and coverage ≤ 200 to exclude alignment artifacts, duplicates, and paralogous genes (Altmann et al. 2012).

Functional annotation of transcriptomes

The reference transcriptome was annotated by performing a BLASTX search to identify homologous sequences from the NCBI non-redundant protein database, and then GO terms associated with BLASTX hits were retrieved using the annotation pipeline in Blast2GO (Conesa et al. 2005; Götz et al. 2008). Tissue-specific assemblies were also annotated in Blast2GO, and Fisher’s Exact Test was used to examine whether GO terms were over-represented between pairs
of tissue types. Each pairwise tissue comparison (liver, brain, gonad) was analyzed for over-
representation, and significant results were identified with a False Discovery Rate (FDR) ≤ 0.05.

**Prediction of Open Reading Frames (ORFs) and p\textsubscript{N}/p\textsubscript{S} calculations**

Regions containing ORFs were identified using BLASTX searches of our assembled contigs
against the NCBI non-redundant protein database. Only best hits with an e-value ≤ 10\textsuperscript{-5}, and
when query transcripts hit only one subject sequence and *vice versa*, were kept. From these
results, a general feature file (GFF) was manually created indicating the start and stop
coordinates, strand information, and reading frame from the BLASTX results. Within these
protein coding regions, putative ORFs were identified when a start codon was found and the
reading frame was greater than 150 bp long. The Perl script, Syn-nonsyn-at-position.pl,
implemented in Popoolation v. 1.2.2 (Kofler *et al.* 2011) was used to define population-specific
SNPs obtained from the SAMtools analysis above as either non-synonymous or synonymous.

The ratio of non-synonymous (p\textsubscript{N}) to synonymous (p\textsubscript{S}) SNP substitutions (p\textsubscript{N}/p\textsubscript{S}) was
calculated between individual Newbler cDNA population assemblies to identify coding
sequences potentially experiencing directional selection in urban *P. leucopus* populations. For
each population pair, the fastaFromBed command in bedtools (Quinlan & Hall 2010) was used to
filter contigs and generate a fasta file of putative ORFs (identified above) for each population
assembly. The USEARCH (http://www.drive5.com/usearch/) clustering and alignment software
for genomic datasets (Edgar 2010) was used to create pairwise alignments between all
population ORFs using an e-value ≤ 0.001. Signatures of selection between aligned ORFs were
identified using KaKs_Calculator1.2 (Model GY) (Zhang *et al.* 2006) to calculate the ratio of
non-synonymous (p\textsubscript{N}) to synonymous (p\textsubscript{S}) SNPs in each population pair. Only transcripts with at
least three SNPs, an ORF length greater than 150 bp, and no in-frame stop codons were included. The mean number of SNPs per ORF was $1.4 \pm SE = 2.9$. A three SNP threshold was chosen to avoid bias as $Ka / Ks$ calculations lose statistical power as the number of substitutions per ORF decreases (Montoya-Burgos 2011). The maximum likelihood method was used that accounts for evolutionary characteristics (i.e. ratio of transition / transversion rates, nucleotide frequencies) of our transcriptome datasets. Contigs with elevated $p_N / p_S$ ratios were then annotated in Blast2GO as above.

Candidate genes were screened for evidence of recombination, and additional signatures of natural selection were examined using McDonald-Kreitman tests. I used BLASTN searches to find orthologous mRNA sequences from multiple species for each candidate gene. For recombination analysis, multiple mammals were used and always included *Cricetulus griseus*, *Rattus norvegicus*, or *Mus musculus*. Orthologous sequences were codon-aligned using MACSE (Ranwez et al. 2011) and then scanned for evidence of recombination using a GARD analysis implemented in the Data Monkey webserver (Kosakovsky Pond et al. 2006; Delport et al. 2010). For McDonald-Kreitman tests, orthologous genes between *Peromyscus leucopus* and *Rattus norvegicus* or *Cricetulus griseus* were codon aligned with MACSE (Ranwez et al. 2011). Non-overlapping datasets of polymorphisms within *P. leucopus* and fixed genetic changes between species were then generated. The McDonald-Kreitman test was performed with these data using DnaSP v.5.10.1 (Librado & Rozas 2009). Fasta files of assembled contigs / isotigs, vcf files of SNP marker data, BLAST2GO files of functional annotations, and output files from population genetics tests are available on the Dryad digital repository (doi: http://dx.doi.org/10.5061/dryad.r8ns3/3).
Results

Sequencing and comparison of assembly methods

454 Sequencing of four full plates of *P. leucopus* cDNA libraries made from liver, brain, and gonad tissue produced 3,052,640 individual reads with an average length of 309 ± 122 bp (median = 341, Interquartile Range (IQR) = 188 bp). While the initial Newbler genomic assembly and Cap3 assembly produced more contigs, the mean length and N50 for both sets of contigs were lower than the Newbler cDNA assembly (Table 1). The Cap3 assembly and the genomic assembly included a much higher proportion of shorter contigs than the cDNA assembly (Fig. 1). Coverage was calculated for all three assemblies, and all had similar median read coverage per contig (Newbler Genomic, median = 4.7 reads, IQR = 4.6; Newbler cDNA, median = 4.9 reads, IQR = 4.1; Cap3, median = 5.0 reads, IQR = 7.0, Fig. S1).

After filtering BLASTN searches against *Mus musculus* and *Rattus norvegicus* cDNA libraries, there was an average for all assemblies of 13,443 hits to known genes. The Cap3 assembly and Newbler genomic assembly produced the most hits, but the average alignment length was longest for the Newbler cDNA assembly (Table 2). Of the total number of contigs for each assembly, the Newbler cDNA assembly had the highest proportion (47%) of ‘Gene Candidates’ followed by the Cap3 assembly (42%) and the Newbler genomic assembly (41%). Assessments important for looking at $p_N/p_S$ (longest average length of contigs, largest N50 value) and for reducing false positives (largest proportion of hits to one gene with known function) supported the assertion that Newbler’s cDNA assembly produced the best quality reference transcriptome, and all further analyses used this assembly.

cDNA transcriptome assembly
The final reference *P. leucopus* Newbler cDNA assembly produced 17,371 contigs with an average length of 613 ± 507 bp. These contigs were assembled into 15,004 isotigs and 12,464 isogroups with a combined length of 13,421,361 bp. Isotigs were constructed from an average of 1.6 contigs and isogroups from an average of 1.2 isotigs. The contribution of sequence reads from individual tissues to the final reference transcriptome was not equal. Liver and brain cDNA libraries produced higher numbers of total reads and a greater proportion of assembled reads compared to ovary and testis libraries. The average read coverage of contigs for each tissue type varied, but coverage from liver sequences was highest with nearly 2X more compared to brain, testes, and ovaries (Table S1). Among all contigs assembled, 70% contained reads from plate 1 (normalized), 57% contained reads from plate 2 (non-normalized), 79% contained reads from plate 3 (non-normalized), and 89% contained reads from plate 4 (non-normalized). Comparison of normalized (Plate 1) and non-normalized (Plates 2-4) cDNA libraries indicated that non-normalization produced nearly twice as many total sequencing reads as compared to normalization, and non-normalized plates were able to sequence rare transcripts at a similar rate compared to the normalized plate (Table S1).

**Mouse and rat genome comparisons**

Assembled mRNA transcripts from *P. leucopus* successfully mapped to both *Mus* and *Rattus* reference genomes and were distributed across all chromosomes for both references (Fig. 2). There were 9,418 best BLAT hits between *P. leucopus* contigs and known *Mus* genes and 8,786 best hits with *Rattus* genes. The latest cDNA references include 35,900 genes for *Mus* (mm10) and 29,261 genes for *Rattus* (rn5), suggesting that full or partial coding sequence from approximately one-third to one-fourth of the *P. leucopus* transcriptome was sequenced. Given
that many of the 15,000 contigs I assembled from our raw sequencing data may represent *Peromyscus*-specific genes not found in model rodent databases, the real proportion of the sequenced transcriptome may be much higher.

**Functional annotation**

Among isotigs from the reference *P. leucopus* transcriptome, 11,355 (75.7%) had BLASTX hits to known genes, and 6,385 (42.6%) mapped to proteins and were annotated with known biological functions (GO terms) from protein databases. Top sources for these annotations were the model rodents *Cricetulus griseus* (3,686 BLASTX hits, 24.5%), *Mus musculus* (2,914 BLASTX hits, 19.4%), and *Rattus norvegicus* (1,671 BLASTX hits, 11.1%, Fig. S2). For cDNA assemblies of individual organs, the ovary transcriptome (1,589 isotigs) had the highest proportion (73.9%) of assembled contigs with GO annotations (Fig. 3). Liver (6,240 isotigs) and testes (5,728 isotigs) produced the largest number of total assembled contigs with similar proportions having GO term annotations (65.6% and 64.6%, respectively). The brain transcriptome (2,613 isotigs) included a lower number of assembled contigs and percent GO annotation (56.8%; Fig. 3).

One-tailed Fisher’s Exact tests (False Discovery Rate (FDR) ≤ 0.05) indicated that liver had the most GO terms that were significantly over-represented compared to the other tissue types (Fig. 4). 1,320 annotations in liver were overrepresented in both liver to brain and liver to gonad comparisons, and there were 69 overlapping annotations in brain to gonad and brain to liver comparisons (Fig. 4). Gonads had the least number of annotations (five) commonly overrepresented in both brain and liver comparisons (Fig. 4). When reduced to their most-specific terms, pairwise comparisons detected 64 over-represented GO annotations for liver
when compared to both of the other tissues, 20 for brain, and five for gonads (Table 3). Over-represented GO terms in liver were related to metabolic processes including ATP binding, GTP binding, NADH dehydrogenase, and electron carrier activity. Over-represented GO terms in brain included regulation of behavior, actin binding, ion channel activity, motor activity, and calcium ion binding. Significantly different gonad annotations were related to reproduction, cilia (for sperm locomotion), the cell cycle, transcription regulation, and epigenetic regulation of gene expression (See Table 3 and Table S2 for full list of overrepresented GO annotations in all pairwise comparisons).

**SNP calling and calculation of \( p_N/p_S \)**

After mapping the reads used in the assembly back to the Newbler cDNA reference transcriptome, 31,015 SNPs were called in 7,625 isotigs. The distribution of SNPs per isotig ranged from 1 – 78 (mean = 4 ± 5.4; median = 2). ORFs were identified in 11,704 isotigs comprising 5.6 Mb of sequence, and 2,655 putative ORFs contained 4,893 SNPs. Of these SNPs, 1,795 (36.6%) were classified as non-synonymous and 3,098 (63.3%) were classified as synonymous. Aligned ORFs were used to calculate \( p_N/p_S \) between each pair of populations. The majority of the ORFs did not exhibit statistical signatures of positive selection (overall mean ± SE \( p_N/p_S = 0.28 ± 0.56 \)). For the 2,307 pairs of homologous cDNA sequences between populations that contained predicted ORFs, did not contain in-frame stop codons, and had greater than or equal to three SNPs, \( p_N/p_S \) values for 11 (0.5%) contigs exceeded 1.0 (Table 4, Fig. 5). The proportion of genes with \( p_N/p_S > 1.0 \) is comparable to similar studies; Sun et al. (Sun et al. 2012) found that 0.4% of genes in their *Pomacea canaliculata* dataset were positively selected, Renaut et al. (Renaut et al. 2010a) reported 0.5% in *Coregonus clupeaformis*, and Wang et al
(Wang et al. 2012) reported 1.8% in Bemisia tabaci. Four contigs (0.2%) exhibited $p_N/p_S$ values $> 1.0$ in urban to urban comparisons and 7 contigs (0.3%) in urban to rural population comparisons. 42 (1.8%) contigs were found with $p_N/p_S$ between 0.5 and 1 (Table S3, Fig. 5); $p_N/p_S$ greater than 0.5 is a less conservative filter for detecting positive selection, especially when using truncated ORFs (Swanson et al. 2004; Elmer et al. 2010).

Different genes showed strong ($p_N/p_S > 1$) signatures of selection when urban populations were compared to other urban populations than when urban and rural populations were compared. Candidate genes identified from the ORF pairs (i.e. $p_N/p_S > 1$) in urban to rural comparisons were related to metabolic processes (including xenobiotic metabolism), reproduction, and demethylation (Table 4). Three genes were involved in metabolic processes: cytochrome P450 2A15 (xenobiotic metabolism, HP_contig01783, $p_N/p_S = 1.89$), camello-like 1 (HP_contig00870, $p_N/p_S = 1.74$), and aldo-keto reductase family 1, member C12 (Xenobiotic metabolism, HP_contig01919, $p_N/p_S = 1.18$). Our analysis also identified a reproductive gene, histone H1-like protein in spermatids 1 (HP_contig02656, $p_N/p_S = 1.07$) that is involved in transcriptional regulation during spermatogenesis. The gene phd finger protein 8 (HP_contig01778, $p_N/p_S = 1.12$), codes for a demethylase that removes methyl groups from histones.

Candidate genes in urban to urban population comparisons were primarily involved in immune system processes. One of these genes is involved in regulating the innate immune response, alpha-1-acid glycoprotein 1 (CP_contig00748, $p_N/p_S = 1.97$), by modulating innate immune response while circulating in the blood, respectively. The other immune system genes are involved in blood coagulation and inflammation, serine protease inhibitor a3c (CP_contig00256, $p_N/p_S = 1.76$) and fibrinogen alpha chain (CP_contig00473, $p_N/p_S = 1.23$). I
also identified \textit{solute carrier organic anion transporter family member 1A5} (CP\_contig01204, $p_N/p_S = 1.55$) that facilitates intestinal absorption of bile acids and renal uptake and excretion of uremic toxins.

For the 22 contigs with $p_N/p_S$ between 0.5 and 1 for urban to rural comparisons, genes are primarily involved in the innate immune response, metabolic processes, and methylation activity, and some of these genes are involved in the same biological pathways as genes listed above for contigs that exhibited $p_N/p_S > 1$ (Tables 4, S3). For the 20 contigs with $p_N/p_S$ between 0.5 and 1 for urban pairwise comparisons, genes are primarily involved with the innate immune response, metabolic processes (including xenobiotic), and reproductive processes.

Candidate genes were scanned for evidence of recombination using a phylogenetic framework. The Genetic Algorithm Recombination Detection (GARD) analysis identified no evidence of recombination in any potential candidate genes. Would-be breakpoints were identified in the genes \textit{Translocation protein SEC62}, \textit{Histone H1-like protein in spermatids 1}, \textit{Aldo-keto reductase family 1 member C12}, \textit{Fibrinogen alpha chain}, \textit{Solute carrier organic anion transporter family member 1A5}, and \textit{Serine protease inhibitor a3c}, but Kishino-Hasegawa testing implemented in the Data Monkey web server found the signal most likely resulted from evolutionary rate variation as opposed to recombination.

McDonald-Kreitman tests were then performed to examine potentially adaptive evolution between species in all the identified candidate genes. \textit{P. leucopus} was compared to \textit{R. norvegicus}, and \textit{C. griseus} when \textit{Rattus} sequences were not available. This approach minimized the number of multiple mutations at individual sites, but results were very similar when the orthologous candidate genes were compared to any rodent with available orthologous gene
sequence. Excess adaptive change (diversifying selection between species) was not identified in any of the candidate genes. For four genes, 39S ribosomal protein L51, PHD finger protein 8, Cytochrome P450 2A15, and Solute carrier organic anion transporter 1A, the ratio of non-synonymous to synonymous polymorphisms within *P. leucopus* was significantly higher than the ratio for divergent genetic changes between species (Table 5). While there were more non-synonymous polymorphisms than synonymous polymorphisms in the remaining seven genes, results were not significantly different from expected neutral evolution.

**Discussion**

**De novo transcriptome assembly and characterization**

Compared to other NGS technologies, 454 transcriptome sequencing provides longer read lengths ideal for *de novo* assembly (Metzker 2010) and is especially useful for organisms without extensive genomic resources like *P. leucopus* (Vera et al. 2008; Meyer et al. 2009; Renaut et al. 2010a; Santure et al. 2011; Sloan et al. 2012). I compared the relative merits of two established long-read assembly programs, CAP3 and Newbler, for assembling our transcriptomes (Mundry et al. 2012; Cahais et al. 2012). Despite the substantially fewer megabases per run generated by 454 FLX+ compared to Illumina or SOLiD sequencing (Glenn 2011b), I still ran into computational limitations during assembly when using options for cDNA sequence. Similar to Cahais et al. (Cahais et al. 2012), I had the most success after compressing the raw reads into a smaller number of partially assembled sequences using a genome assembler followed by another assembly method better suited for transcriptome data. While the CAP3 assembly produced more contigs, the Newbler v. 2.5.3 transcriptome assembly performed better based on assessments useful for downstream population genomic analyses (e.g. number of long contigs and average
contig length). Newbler performed well at assembling full-length cDNA contigs, and our results are in line with Mundry et al.’s (Mundry et al. 2012) findings that Newbler outperformed other assembly programs in simulated experiments. The N50 value reported here is comparable to de novo Newbler cDNA assemblies for other organisms: N50 = 1,735 bp in *Oncopeltus fasciatus*, Ewen-Campen et al. (Ewen-Campen et al. 2011); N50 = 1,333 bp in *Silene vulgaris*, Sloan et al. (Sloan et al. 2012); N50 = 1,588 bp in *Spalax galili*, Malik et al. (Malik et al. 2011); and N50 = 854 bp in *Arctocephalus gazella*, Hoffman & Nichols (Hoffman & Nichols 2011).

I sequenced samples using normalized and non-normalized cDNA pools and examined the influence each protocol had on gene discovery. Following sequencing of the first normalized plate, I used a new protocol from Roche that excluded normalization of libraries. Surprisingly, I found that normalization did not necessarily improve the number of uniquely assembled contigs. Theoretically, normalization reduces the sequencing of overly abundant transcripts and increases the discovery of rare sequences (Christodoulou et al. 2011; Davey et al. 2011), but normalization does not disproportionately influence gene discovery when enough sequencing coverage is achieved (Vijay et al. 2013). I found that read coverage per transcript increased for our non-normalized plates compared to the normalized pilot plate. However, Ekblom et al. (Ekblom et al. 2012) suggest that differences in technologies and sequencing effort may ultimately affect comparisons between normalized and non-normalized cDNA libraries, and any differences I identify may be due to different protocols used to extract RNA and prepare pooled libraries.

**Mapping to rodent genomes**

The mammalian laboratory models *Mus* and *Rattus* have extensively annotated genomes that provide a good substitute reference for other rodent sequencing projects. The New World
Peromyscus and Old World Mus and Rattus lineages last shared a common ancestor ~25 million years ago (Steppan et al. 2004). Deep divergence and high rates of chromosome evolution across these lineages (Mlynarski et al. 2010) may have affected the percentage of identified homologous gene transcripts. Ramsdell et al. (Ramsdell et al. 2008) found the Peromyscus genome to be more similar to Rattus than Mus due to an enhanced level of genome rearrangement in Mus compared to ancestral muroids. Our results support these findings given that most Peromyscus transcripts mapped to different chromosomes (96.1%) between Mus and Rattus. Our homologous gene matches between Peromyscus and Rattus also represented a higher proportion (30.1%) of total Rattus genes than homologous gene matches between Peromyscus and Mus (25.7%). Non-homologous hits and mapping differences between reference genomes may also be due to highly variable or alternatively spliced transcripts, contamination by genomic DNA, or inclusion of low-quality data (Ferreira de Carvalho et al. 2013), although our assembly methods included measures to limit the influence of these artifacts.

**Functional annotation and tissue comparisons**

Over 75% of our assembled contigs produced significant BLASTX hits to known genes in NCBI’s nonredundant (nr) protein database. This rate of annotation is similar to studies on other non-model species with genomic information available from closely-related model organisms, e.g. 66% in the rodent Ctenomys sociabilis (MacManes & Lacey 2012) and 79.7% in the plant Silene vulgaris (Sloan et al. 2012). These rates are much higher than some other organisms with few model relatives, such as 19.58% in a bat, Artibeus jamaicensis, (Shaw et al. 2012), 18% in a butterfly, Melitaea cinxia, (Vera et al. 2008), and 29.2% in the gastropod, Pomacea canaliculata, (Sun et al. 2012). Phylogenetic analyses support Peromyscus spp. and Cricetulus spp. as
members of a monophyletic clade that diverged separately from *Mus* and *Rattus* (Steppan et al. 2004), and *C. griseus* represented the highest proportion of BLASTX top-hits (Figure S2, Supplementary Material). Laboratory use of *C. griseus* is not as prevalent as *Mus* or *Rattus*, but Chinese hamster ovary (CHO) cell lines are commonly used in vitro to produce biopharmaceuticals (Becker et al. 2011), and a draft genome has also been sequenced (Xu et al. 2011). Research on protein pathways and interactions within CHO cell lines provides a future resource for investigating functional consequences of divergent genes between urban and rural populations of *P. leucopus*.

Transcriptome studies in model rodents provide useful context for understanding how much of each tissue-specific transcriptome I sequenced in this study. Yang et al. (Yang et al. 2006) used microarray analysis to identify 12,845 active genes in *Mus* liver, and RNA-Seq using an Illumina HiSeq 2000 on *Rattus* liver identified 7,514 known genes (Chapple 2012). Our gene discovery was between 40-60% of these previously reported liver transcriptomes. In brain tissue, 4,508 genes were identified in *Mus* by Yang et al. (Yang et al. 2006), and Chrast et al. (Chrast et al. 2000) report ~4,000 genes identified in *Mus* brain tissue. The 2,610 gene annotations from our brain cDNA libraries represent between 60-65% of the full *P. leucopus* brain transcriptome. Microarray analysis of testis RNA identified up to 13,812 known genes (Shima et al. 2004) in *Mus*, and 454 sequencing of cDNA libraries from *C. griseus* identified 13,187 annotations in ovary (Becker et al. 2011). UniGene (Pontius et al. 2003) includes 8,946 genes for *Mus* testis, 5,285 for *Mus* ovaries, 4,355 for *Rattus* testis, and 5,093 for *Rattus* ovaries. The only cDNA library established in UniGene for *Peromyscus* spp. includes 635 putative genes from testis (Glenn et al. 2008). Our assembled libraries from gonad tissue fall within these ranges, and non-annotated transcripts could represent *Peromyscus*-specific genes. To recover 100% of each
tissue transcriptome, samples would need to be prepared at various developmental stages and under various environmental conditions.

Fisher’s Exact Tests allowed us to identify annotated transcripts over-represented in one tissue compared to the others. The brain transcriptome of the social rodent, *C. sociabilis*, exhibited highly expressed genes involved with behavior and signal transduction (MacManes & Lacey 2012). Over-represented GO terms in *P. leucopus* brain tissue were related to similar major functions in the brain, including regulation of behavior, cellular signaling, actin binding, ion transport and channel activity, motor activity, and calcium ion binding. In liver, over-represented GO terms were largely dedicated to metabolic processes including ATP binding, GTP binding, NADH dehydrogenase, and electron carrier activity. There were also several GO terms related to the immune response, hematopoietic processes, and nutrient binding; these annotations are supported by microarray and RNA-seq analyses of liver in mouse and rat, respectively (Yang et al. 2006; Chapple 2012).

**SNP discovery and characterization**

Without a reference genome, aligning reads to assembled transcripts and assigning mismatches as SNPs (Barbazuk et al. 2007) is an acceptable substitute for generating sequence polymorphisms for non-model species (Collins et al. 2008; Renaut et al. 2010a; Sloan et al. 2012). Difficulties may persist in distinguishing true SNPs from false positives created by sequencing errors or misaligned reads, and alignment of reads to paralogous genes can also generate false SNPs. Identifying true SNPs depends on assembly quality, filtering criteria of nucleotide mismatches during alignment, and statistical models used to call nucleotide variants

I used conservative filtering criteria when calling SNPs to minimize false positives. SAMtools (Li et al. 2009) excels at SNP detection with low sequence coverage by comparing multiple samples simultaneously (Nielsen et al. 2011; Altmann et al. 2012). I also filtered variants based on thresholds of quality and minimum occurrence, and restricted maximum coverage to filter out false positive SNPs from paralogous genes. Excluding transcripts with the highest coverage after mapping limits problems with gene duplications (McCormack et al. 2011a). The thresholds I used for minimum SNP occurrence and nucleotide quality reduce error rates by several orders of magnitude for pooled data, ensuring the reliability of SNP libraries for downstream analyses (Kofler et al. 2011). Our SNP library represents highly confident variant calls and will serve as an important resource for future population genetic studies of urban and rural populations of *P. leucopus*. I cannot completely rule out paralogous genes or misalignments in our transcriptome assemblies, and thus future work will require sequencing of transcripts from multiple individuals to validate SNP calls in candidate genes of particular interest.

**Positive selection and the transcriptome**

I used the ratio of non-synonymous to synonymous substitution rates \( p_N/p_S \) to identify candidate genes that may have experienced positive selection in urban populations of *P. leucopus*. Using SNPs to calculate \( p_N/p_S \) ratios in ORFs from assembled transcriptomes can be a fruitful method for identifying the operation of natural selection on individual loci (Nielsen & Yang 1998; Oleksyk et al. 2010; Hohenlohe et al. 2011). This approach has recently been
used to identify genes under positive or purifying selection between cichlid fish lineages in Nicaragua (Elmer et al. 2010), between lake whitefish species pairs (Renaut et al. 2010b), and within an invasive gastropod (Sun et al. 2012). Studies traditionally identify positive selection in genes with $p_N/p_S > 1.0$. I used this cutoff value, but also identified sequence pairs with $p_N/p_S$ between 0.5 and 1.0 to avoid overlooking relevant non-synonymous substitutions in candidate genes that might be of interest for individual re-sequencing projects. Lack of full-length ORFs can decrease $p_N/p_S$ values when some non-synonymous substitutions are not sampled (Swanson et al. 2004; Elmer et al. 2010). The $p_N/p_S$ index can also be used when samples have been pooled prior to sequencing (Baldo et al. 2011), unlike summary statistics that rely on allele frequencies (Sloan et al. 2012).

I used McDonald-Kreitman tests to further elucidate patterns of evolution in candidate genes. This method can identify adaptive changes between species and primarily detects selection processes occurring thousands or even millions of years in the past. I calculated a neutrality index (NI) as $(p_N/p_S)/(d_N/d_S)$ to look at deviations from neutral expectations (McDonald & Kreitman 1991; Smith & Eyre-Walker 2002). While I detected an excess of non-synonymous polymorphisms within *P. leucopus* in genes with functions including demethylation, xenobiotic metabolism, and innate immunity, I did not find evidence of positive selection between species. While these patterns could suggest purifying selection preventing the fixation of harmful mutations (Andersen et al. 2012) or indicate balancing selection acting to maintain favorable alleles in different populations (Storz et al. 2007b), interpretation should proceed cautiously due to limitations of polymorphism data generated from pooled transcriptomes. The inability to assign individual allele frequencies when identifying polymorphisms leads to an ascertainment bias towards high within-species $p_N/p_S$ ratios.
compared to interspecies ratios, and this bias may explain the lack of NI values < 1 (positive selection). These results could be interpreted as the result of balancing selection whereby different alleles are favored in different urban populations, however, which would seem consistent with the ecology of these relatively isolated populations. Individual resequencing of mice from multiple populations will remove the ascertainment bias, uncover more polymorphisms, and allow the use of more powerful tests to study recent selective pressures in urban populations.

Many ecological changes arising from urbanization may drive local adaption to novel conditions in fragmented urban populations, and I made several predictions about the types of adaptive traits present in urban habitats from current literature. Genes involved in divergence of urban and rural populations of white-footed mice are likely associated with quantitative traits affected by crowded (i.e. high population density) and polluted urban environments (life history, longevity, reproduction, immunity, metabolism, thermoregulatory and/or toxicological traits). I identified candidate genes ($p_N/p_S > 1$) that supported these predictions between urban and rural populations of mice, but also between individual urban populations. The urban matrix is a strong enough barrier to dispersal that white-footed mouse populations in individual city parks may experience highly localized selective pressures in addition to selective pressures that are general to urban environments (Munshi-South 2012).

New predators, competitors, parasites, and pathogens can drive local adaptation of traits, especially those related to immunity, in novel urban environments (Peluc et al. 2008; Sih et al. 2011). I identified candidate genes involved in the innate immune system and activation of the complement pathway to identify pathogens. Additionally, two candidate genes were identified in comparisons of urban populations that function in blood coagulation and inflammation. The
innate immune system is a biochemical pathway that removes pathogens by identifying and killing target cells (Kosiol et al. 2008), and positive selection is found to act on pathogen recognition genes within the complement activation pathway (Sackton et al. 2007). The introduction of invasive species, population growth of ‘urban exploiters’, and increased traffic, trade, and transportation within cities can introduce large numbers of novel pathogens (Bradley & Altizer 2007), and white-footed mice in NYC may be evolving to efficiently recognize them and respond immunologically. I also identified several genes involved in metabolism that were divergent between populations, and a gene expressed during spermatogenesis that was divergent between urban and rural populations. Rapid evolution has been identified in reproductive proteins between Peromyscus spp. affecting spermatogenesis, sperm competition, and sperm-egg interactions (Turner et al. 2008), and the intensity of sperm competition and reproductive conflict may be increasing in dense P. leucopus populations in NYC.

Increasing air, water, and soil pollution are all typical impacts of urbanization (Yauk et al. 2008; Whitehead et al. 2010; Francis & Chadwick 2012). One potential marker of increased exposure to pollutants is hypermethylation of regulatory regions of the genome (Somers et al. 2002; Yauk et al. 2008; Janssens et al. 2009; Somers & Cooper 2009). Positive selection may also be acting on genes involved in xenobiotic metabolism. Heavy metals including mercury, lead, and arsenic occur at increased concentrations within NYC park soils (S. Harris, unpublished data), and McGuire et al. (McGuire et al. 2013) found lower pH and higher concentrations of heavy metals in NYC parks compared to green roofs. PCB resistance was identified in multiple populations of Fundulus heteroclitus (Whitehead et al. 2010), and Wirgin et al. (Wirgin et al. 2011) also found rapid PCB resistance in tomcod from the Hudson River through positive selection. In urban to rural comparisons I found two potential toxicological
candidate genes: one gene involved in metabolizing foreign chemical compounds (i.e. xenobiotics), and a demethylase that removes methyl groups from histone lysines.

Comparing candidate genes from all pairwise analyses with $p_N/p_S$ between 0.5 and 1 reveals several additional patterns. Proteins were identified that function in the alternative pathway, which acts continuously in an organism without antibody activation to clear foreign pathogens (Carroll 2004), and supports our conclusion that positive selection ($p_N/p_S > 1$) is acting on the innate immune system in these populations. Four *cytochrome p450* genes, 2d27-like, family 2 subfamily B, subfamily polypeptide 13, and 2a15, exhibited $p_N/p_S$ between 0.5 and 1 in urban populations and between urban and rural populations. *Cytochrome p450 2a15* was also found to have a $p_N/p_S > 1$ and McDonald-Kreitman tests found significantly more polymorphisms within *P. leucopus* than between species (Table 4, Table 5). The *cytochrome p450* family of genes plays a major role in xenobiotic metabolism, including detoxification in variable environments (Su & Ding 2004; Büntge 2010). Patterns of divergence and positive selection have been robustly identified in *cytochrome p450* genes in natural populations of both *Mus musculus* ingesting toxins through their diet and *Tetrahymena thermophila* exposed to toxic environments (Fu *et al.* 2009; Büntge 2010). *P. leucopus* in NYC populations may be experiencing different dietary demands and exposure to pollutants, leading to selective pressures on detoxifying genes like the cytochrome p450 gene family.

Alternatively, genetic differences between urban and rural populations may result from genetic drift rather than selection. Given a small effective population size, alleles may become fixed between populations through chance alone (Hohenlohe & Arnold 2008). A null distribution of genetic variation can be generated if the neutral evolutionary history of a population is known, and genetic variation can be attributed to selection if the genetic pattern
falls beyond the extremes of the null distribution (Ovaskainen et al. 2011). Additionally, the same signature of selection seen in multiple distinct populations experiencing the same selection regime is additional evidence that distinguishes genetic drift from selection. I will differentiate between drift and selection in future work by examining genetic divergence between multiple urban and rural populations at these candidate loci and by inferring the demographic history of P. leucopus in NYC under a null model of neutral evolution. I must also be cautious when inferring the function of candidate genes after identifying statistical signatures of positive selection.

While a \( p_N/p_S \) ratio > 1.0 can represent positive selection, it may also occur due to relaxation of purifying selection, and individual codons within a gene can have an excess of non-synonymous substitutions due to random biological processes (Hughes 2007). However, current statistical tests address these issues and are generally robust in identifying positive selection (Zhai et al. 2012). In the case of a single population, \( p_N/p_S > 1 \) may not represent positive selection. Kryazhimskiy & Plotkin (Kryazhimskiy & Plotkin 2008) demonstrated that the relationship between \( p_N/p_S \) and selection is radically different when samples originated from the same population; \( p_N/p_S \) actually decreases in response to positive selection. To infer selection between two samples using \( p_N/p_S \), samples must come from reproductively isolated populations with fixed substitutions (Kryazhimskiy & Plotkin 2008). All samples used to calculate \( p_N/p_S \) for this study came from reproductively isolated and genetically structured populations (Munshi-South & Kharchenko 2010). I assembled transcriptome datasets individually for each population to identify fixed substitutions between populations and avoid randomly segregating SNPs in \( p_N/p_S \) analyses. Indices such as \( p_N/p_S \) identify genes with previously unknown signatures of selection, but candidates still need to be studied in a controlled setting to identify phenotype and function (Zhai et al. 2012).
The ability of $p_N/p_S$ and McDonald-Kreitman tests to detect genes under positive selection is limited in some situations, so it is likely that I have missed many candidate genes. Additionally, such analyses do not identify adaptive variation in gene regulatory regions as opposed to transcribed cDNA (Prud’homme et al. 2007). Ratios such as $p_N/p_S$ may also vary widely when there are relatively few mutations per gene (Hughes 2007; Renaut et al. 2010a). Given strong selection within populations, however, it is plausible that multiple substitutions may rise to high frequency or become fixed within a few hundred generations (i.e. in the timeframe of divergence for urban and rural populations of white-footed mice). The candidate genes identified herein can be confirmed in future work using the reference genome of $P. maniculatus$ (sequenced and currently being assembled) and multiple tests of selection that provide more statistical power and higher resolution when identifying types and age of selection in single candidate genes (Grossman et al. 2010; Li et al. 2012). These emerging resources will allow us to validate many of our predicted polymorphisms, identify paralogous genes with greater certainty, and perform more powerful tests of selection by providing genetic distances and genomic coordinates for our sequenced contigs. Our ongoing work in this system uses these external resources with our new transcriptomic and genomic libraries from individual mice from several urban, rural, and suburban populations. These ongoing studies employ multiple outlier statistics based on the allele frequency spectrum and linkage disequilibrium to examine recent selection in both coding and non-coding regions of urban white-footed mouse genomes.
Table 1. Results of transcriptome assembly using three different approaches.

<table>
<thead>
<tr>
<th>Assembly Method</th>
<th>No. Contigs</th>
<th>Mean Contig Length (bp)</th>
<th>Median Contig Length (bp)</th>
<th>N50*</th>
<th>Length (Mb)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newbler genome(^a)</td>
<td>20,570</td>
<td>630 ± 504</td>
<td>516</td>
<td>830</td>
<td>12.95</td>
</tr>
<tr>
<td>Cap3(^b)</td>
<td>27,497</td>
<td>653 ± 380</td>
<td>566</td>
<td>732</td>
<td>17.95</td>
</tr>
<tr>
<td>Newbler cDNA(^c)</td>
<td>15,004 (Isotigs)</td>
<td>895 ± 752</td>
<td>683</td>
<td>1,039</td>
<td>13.42</td>
</tr>
</tbody>
</table>

\(^a\)Newbler v. 2.5.3 large genomic assembly of total set of raw sequencing reads

\(^b\)Cap3 assembly using ‘assembled’ or ‘partially assembled’ reads from Newbler genome assembly

\(^c\)Newbler v. 2.5.3 cDNA assembly using ‘assembled’ or ‘partially assembled’ reads from Newbler genome assembly

*N50, The value where half the assembly is represented by contigs of this size or longer

**Total assembly length in Megabases.
Table 2. BLASTN search results of three *P. leucopus* transcriptome assemblies against reference cDNA libraries from *Mus* and *Rattus*.

<table>
<thead>
<tr>
<th>Assembly Method</th>
<th>Total Significant Hits; <em>Mus</em></th>
<th>Total Significant Hits; <em>Rattus</em></th>
<th>Gene Candidates, <em>Mus</em></th>
<th>Gene Candidates, <em>Rattus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Newbler genome</td>
<td>12,932</td>
<td>12,807</td>
<td>8,568 (708 bp)</td>
<td>8,080 (714 bp)</td>
</tr>
<tr>
<td>Cap3</td>
<td>17,333</td>
<td>16,792</td>
<td>11,662 (623 bp)</td>
<td>10,938 (638 bp)</td>
</tr>
<tr>
<td>Newbler cDNA</td>
<td>10,699</td>
<td>10,094</td>
<td>7,048 (823 bp)</td>
<td>6,814 (847 bp)</td>
</tr>
</tbody>
</table>

* = Average alignment length in base pairs

Total significant hits represent sequence identity ≥ 80%, alignment length ≥ 50% of the total length of either the query or subject sequence, and e-value ≤ 10⁻⁵. Gene candidates represent significant hits where one query sequence matches one subject gene and *vice versa*. 
Table 3. Over-represented GO terms for individual tissue types from Fisher’s Exact tests (FDR ≤ 0.5) in Blast2Go.

<table>
<thead>
<tr>
<th>GO term</th>
<th>FDR</th>
<th># Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Liver</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP binding</td>
<td>5.31E-24</td>
<td>184</td>
</tr>
<tr>
<td>zinc ion binding</td>
<td>5.93E-20</td>
<td>154</td>
</tr>
<tr>
<td>transcription factor complex</td>
<td>3.91E-19</td>
<td>148</td>
</tr>
<tr>
<td>electron carrier activity</td>
<td>8.53E-18</td>
<td>251</td>
</tr>
<tr>
<td>structural constituent of ribosome</td>
<td>5.51E-15</td>
<td>117</td>
</tr>
<tr>
<td>soluble fraction</td>
<td>2.35E-12</td>
<td>97</td>
</tr>
<tr>
<td>microsome</td>
<td>1.53E-10</td>
<td>83</td>
</tr>
<tr>
<td>protein homodimerization activity</td>
<td>2.75E-10</td>
<td>81</td>
</tr>
<tr>
<td>oxygen binding</td>
<td>1.97E-09</td>
<td>93</td>
</tr>
<tr>
<td>perinuclear region of cytoplasm</td>
<td>9.92E-09</td>
<td>69</td>
</tr>
<tr>
<td>GTP binding</td>
<td>7.64E-08</td>
<td>62</td>
</tr>
<tr>
<td>GTPase activity</td>
<td>2.82E-05</td>
<td>42</td>
</tr>
<tr>
<td>ubiquitin-protein ligase activity</td>
<td>2.82E-05</td>
<td>42</td>
</tr>
<tr>
<td>NADH dehydrogenase (ubiquinone) activity</td>
<td>5.01E-05</td>
<td>40</td>
</tr>
<tr>
<td>drug binding</td>
<td>6.65E-05</td>
<td>39</td>
</tr>
<tr>
<td>sequence-specific DNA binding</td>
<td>6.65E-05</td>
<td>39</td>
</tr>
<tr>
<td>double-stranded DNA binding</td>
<td>8.90E-05</td>
<td>38</td>
</tr>
<tr>
<td>mitochondrial respiratory chain complex I</td>
<td>1.18E-04</td>
<td>37</td>
</tr>
<tr>
<td>transcription coactivator activity</td>
<td>1.18E-04</td>
<td>37</td>
</tr>
<tr>
<td>catalytic step 2 spliceosome</td>
<td>1.58E-04</td>
<td>36</td>
</tr>
<tr>
<td><strong>Brain</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>protein complex</td>
<td>1.27E-06</td>
<td>569</td>
</tr>
<tr>
<td>plasma membrane</td>
<td>4.30E-92</td>
<td>567</td>
</tr>
<tr>
<td>signal transduction</td>
<td>2.15E-39</td>
<td>525</td>
</tr>
<tr>
<td>cytosol</td>
<td>1.79E-08</td>
<td>411</td>
</tr>
</tbody>
</table>
cell differentiation    5.07E-28  372
anatomical structure morphogenesis 1.89E-30  291
cell death    1.78E-06  247
cell-cell signaling 2.79E-61  232
ion transport    3.12E-17  209
cytoplasmic membrane-bounded vesicle 1.33E-22  197
golgi apparatus 1.51E-10  168
cytoskeleton organization 9.13E-13  145
cellular homeostasis 9.82E-16  134
behavior    6.72E-28  133
calcium ion binding 7.69E-13  109
actin binding 3.54E-15  93
response to abiotic stimulus 4.97E-08  88
protein kinase activity 1.61E-03  77
ion channel activity 5.21E-17  62
motor activity 8.38E-06  48

Gonads

nucleic acid binding 1.87E-08  1101
nuclear chromosome 9.86E-06  119
reproduction 1.92E-06  680
RNA binding 6.70E-04  637
viral reproduction 1.74E-02  339

GO terms have been reduced to their most specific terms. Only common GO terms over represented for one tissue compared to the other two tissues are shown. The top 20 terms are shown, see Table S2 for full list of GO annotations
Table 4. Candidate loci exhibiting $p_N/p_S > 1$

<table>
<thead>
<tr>
<th>Sequence name</th>
<th>$p_N/p_S$</th>
<th>Gene name</th>
<th>Gene function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pairwise</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Urban:Rural</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban:Urban</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Comparisons</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP_contig01773</td>
<td>1.01</td>
<td>SEC62</td>
<td>Translocation protein Post-translational protein translocation into the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39S ribosomal</td>
<td>endoplasmic reticulum; plasma membrane protein</td>
</tr>
<tr>
<td>HP_contig02632</td>
<td>1.05</td>
<td>protein L51</td>
<td>Histone H1-like protein in spermatids Transcriptional regulation and / or</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>chromatin remodeling through DNA binding during</td>
</tr>
<tr>
<td>HP_contig02656</td>
<td>1.07</td>
<td>1</td>
<td>spermatogenesis</td>
</tr>
<tr>
<td>HP_contig01778</td>
<td>1.12</td>
<td>PHD finger protein 8</td>
<td>Aldo-keto reductase Xenobiotic metabolism; oxidation-reduction process</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>family 1, member</td>
</tr>
<tr>
<td>HP_contig01919</td>
<td>1.18</td>
<td>C12</td>
<td>Metabolic process; mitochondrial inner membrane</td>
</tr>
<tr>
<td>HP_contig00870</td>
<td>1.74</td>
<td>Camello-like 1</td>
<td>Metabolic process; testosterone 7a-hydroxylase</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>protein</td>
</tr>
<tr>
<td>CP_contig00473</td>
<td>1.23</td>
<td>Fibrinogen alpha chain</td>
<td>Glycoprotein circulating in the blood; functions in</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>blood coagulation and part of the most abundant component of blood clots</td>
</tr>
<tr>
<td>CP_contig01204</td>
<td>1.55</td>
<td>family member 1A5</td>
<td>Membrane protein; transports hormones; facilitates</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>intestinal absorption of bile acids and renal uptake of</td>
</tr>
<tr>
<td>CP_contig00256</td>
<td>1.76</td>
<td>Serine protease</td>
<td>Indoxyl sulfate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bind to proteases and inhibit proteolysis; often</td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inhibitor a3c</td>
<td>involved in blood coagulation and inflammation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alpha-1-acid</td>
<td>Transport protein in the blood stream; binds and distributes synthetic drugs throughout body;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP_contig00748</td>
<td>modulates innate immune response</td>
<td></td>
<td></td>
</tr>
<tr>
<td>glycoprotein 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5. McDonald-Kreitman tests for candidate genes with $p_{N}/p_{S} > 1$. Comparison of the amount of polymorphisms in candidate ORFs to that of the divergence in orthologous genes between *Peromyscus* and *Rattus norvegicus*. P-values were generated from Fisher’s Exact Test.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Polymorphisms</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>synonymous</td>
<td>Synonymous</td>
</tr>
<tr>
<td></td>
<td>($P_{n}$)</td>
<td>($P_{s}$)</td>
</tr>
<tr>
<td></td>
<td>Ratio</td>
<td>($P_{n}/P_{s}$)</td>
</tr>
<tr>
<td></td>
<td>Non-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>synonymous</td>
<td>Synonymous</td>
</tr>
<tr>
<td></td>
<td>($D_{n}$)</td>
<td>($D_{s}$)</td>
</tr>
<tr>
<td></td>
<td>Ratio</td>
<td>($D_{n}/D_{s}$)</td>
</tr>
<tr>
<td></td>
<td>Neutrality</td>
<td>P-value</td>
</tr>
<tr>
<td>Translocation protein</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>SEC62</td>
<td>0.64</td>
<td>3.11</td>
</tr>
<tr>
<td>39S ribosomal protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L51</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Histone H1-like</td>
<td></td>
<td></td>
</tr>
<tr>
<td>protein in spermatids 1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>PHD finger protein 8</td>
<td>0.47</td>
<td>8.53</td>
</tr>
<tr>
<td>Aldo-keto reductase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>family 1, member C12</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>PHD finger protein 8</td>
<td>0.49</td>
<td>2.67</td>
</tr>
<tr>
<td>Cytochrome P450</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2A15*</td>
<td>1.78</td>
<td>2.24</td>
</tr>
<tr>
<td>Fibrinogen alpha chain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anion transporter 1A5</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Solute carrier organic</td>
<td>0.57</td>
<td>5.29</td>
</tr>
<tr>
<td>Serine protease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inhibitor a3c*</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Alpha-1-acid</td>
<td>1.32</td>
<td>1.27</td>
</tr>
<tr>
<td>glycoprotein 1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.55</td>
<td>2.59</td>
</tr>
</tbody>
</table>

*= McDonald Kreitman test used *Cricetulus griseus*
Figure 1. Frequency of contig lengths for three transcriptome assembly methods. Inset:
Zoomed-in view of frequency of longer assembled contigs from 1,500-3,000 bp. Blue line =
Newbler cDNA, Red line = Newbler genome, Green line = Cap3.
Figure 2. Transcriptome alignment to reference rodent genomes. Number and distribution of contigs from *P. leucopus* transcriptome (Newbler cDNA assembly) that aligned to each chromosome of the. (a) *Rattus norvegicus*. Blue = total number of genes per chromosome for *Rattus*. Red = number of aligned *Peromyscus* isotigs per *Rattus* chromosome. (b) *Mus musculus*. Blue = total number of genes per chromosome for *Mus*. Red = number of aligned *Peromyscus* isotigs per *Mus* chromosome.
Figure 3. Annotation of final reference transcriptome. Number of assembled *P. leucopus* contigs from four different tissue types that had significant hits with known proteins on BLASTX, and GO term annotations from reference databases using Blast2Go; Blue = Total number of contigs, Red = BLASTX hits, Green = number of annotated contigs.
Figure 4. Over-represented GO terms from pairwise tissue comparisons (FDR ≤0.05). (a) Comparison of brain transcriptome to liver and gonad. (b) Comparison of liver to brain and gonad. (c) Comparison of gonad to liver and brain.
Figure 5. Non-synonymous ($p_N$) SNP substitutions plotted vs. synonymous ($p_S$) substitutions for 354 genes. Each circle represents one unique assembled contig. (a) Pairwise comparisons for all urban populations. (b) Pairwise comparisons for urban to rural populations. The dashed line denotes $p_N / p_S = 1$, and circles above the line ($p_N / p_S > 1$) indicate candidates for positive selection. The solid line shows the slope for $p_N / p_S = 0.5$. 


Figure 6. Location and number of individuals collected from five populations in the NYC metropolitan area. Urban populations are in shades of blue; light blue = male; dark blue = female. Rural population in orange and brown; orange = male; brown = female. Areas shaded red on the map indicate degree of urbanization (i.e. permeable surface cover such as roads and rooftops) and green areas indicate vegetation cover from the 2006 National Landcover Database. (CP = Central Park; NYBG = New York Botanical Gardens; RR = Ridgewood Reservoir; FM = Flushing Meadows-Willow Lake; HP = Harriman State Park).
Chapter 2

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

Abstract

Genomic resources are important and attainable for examining evolutionary change in divergent natural populations of non-model species. I 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. I 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 = 1,044 bp and a total content of 30.06 Megabases (Mb). Contigs were annotated from comparisons to *Mus musculus* (39.96% annotated) Uniprot databases. I identified 104,655 high quality single nucleotide polymorphisms (SNPs) and 65 single sequence repeats (SSRs) with flanking primers. I 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:** transcriptome, SOLiD, RNA-Seq, *Peromyscus*, SNPs, Genetic Map
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 non-model 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. 2011b; Grover et al. 2012; Pavey et al. 2012; Wolf 2013). Transcriptome sequencing is one of the most commonly used NGS approaches in non-model organisms because transcriptome datasets 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, I 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 metropolitan area. I 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 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.* 2011), population genetics (Mossman & Waser 2001; Steiner *et al.* 2007; Storz *et al.* 2007a; 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 likely drove adaptive modifications in hemoglobin function between high- and low-altitude populations of *P. maniculatus* in Colorado, U.S.A. (Storz *et al.* 2007a, 2009, 2010; Natarajan *et al.* 2013). In other examples, independent mutations in the *Agouti* gene likely lead to divergent coat coloration in *Peromyscus* populations in both the Nebraska Sand Hills and Florida sand dunes (Mullen & Hoekstra 2008a; Linnen *et al.* 2013), while the genetic architecture of complex extended phenotypes such as burrowing behavior have also recently been described (Weber *et al.* 2013).

*P. 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 (NYC) 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 favorable for local adaptation. Using pooled RNA sequencing, Harris et al. (2013) identified non-synonymous 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 datasets 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, I characterized the transcriptomes of urban and rural *P. leucopus* populations using a combination of 454 and SOLiD 5500 XL sequencing. Here I report the *P. leucopus* reference transcriptome sequences and assemblies, annotations of sequences, an extensive variant library of SNPs and SSRs, a linkage map of 4,066 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 datasets have been generated outside of this current effort by high-throughput sequencing methods (Peterson *et al.* 2012; Cheviron *et al.* 2012, 2013; Linnen *et al.* 2013; Harris *et al.* 2013, NCBI’s BioProject Database, 2014). 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**

*P. leucopus* from urban and rural populations were trapped over a period of 1-3 nights each at 6 sites using four 7x7 m transects of 3” x 3” x 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) and stored at -80°C until RNA 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) 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 re-purified 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 ul 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. Final 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 4 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 number 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 likely 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 for best-hit matches using BLASTx (E-value cutoff = 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 non-redundant (nr) protein database (E-value cutoff = 1e-6). The annotated *Mus 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 cutoff = 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 (VQSR) to generate high quality SNP genotypes from multi-sample alignment files. I used the recommended settings from (De Wit et al. 2012; Van der Auwera et al. 2013), e.g. SNPs requiring coverage > 5X, nucleotide quality > 30, no strand bias (FS > 35),
and SNPs called from a uniquely mapped read. Additional hard filters, i.e. 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 paralogs 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, i.e. 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 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 *Mus 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 cutoff = 1x10^-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 labeled 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 datasets 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 cutoff < 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 cutoff 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 *Mus musculus* transcriptome. Many contigs likely 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 likely 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 dataset with verified gene annotations from *Mus 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 datasets containing SNPs with an average of 5.8 SNPs per contig across all populations. The variant call format (VCF) file containing 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 likely caused by using protein-coding regions from transcriptome sequencing rather than genome sequence (Toth 2000).

The full SNP dataset was filtered to include sites where at least 80% of individuals were genotyped at each SNP. This filtering resulted in 6,449 SNP loci for 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 4,066 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 were 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.

**Inter-population 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 non-annotated genes, was used for gene expression analysis. The parent contigs containing the 295 alternative splicing events identified above were labeled, monitored, and removed if 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 to the Q20. For the Q20 dataset, three genes were up-regulated in the urban group (Fig. 4). One overexpressed 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 dataset, 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 fibers, 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 comparisons. One underexpressed contig represented the gene, Zinc Transporter ZIP14, part of the inflammatory response.

Discussion

The speed, price, and efficiency of generating next generation sequencing data have ushered in a new era of genomic research. However, the “explosion” of sequence data (Andrews & Luikart 2014) has led 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 non-model 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.

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 I restricted our initial search to well-annotated genes with known function from the Uniprot database (16,642 genes). This assembly contained 98.2% of Uniprot’s manually-curated genes in the *M. musculus* reference transcriptome. The size, number of contigs, and N50 length is comparable to 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 = 1,792, 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 2011a). 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 (McCoy et al. 2013; Chapman et al. 2013).

The assignment of genes to chromosomal linkage groups is a first in white-footed mouse genomics. Analyzing 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. 2010b). 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 timeframe 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 *Mus musculus* transcriptome or the non-redundant protein database. There were several differentially expressed genes discovered using trimmed reads with a nucleotide quality cutoff > 5. Although many bioinformatics pipelines for processing NGS data set a minimum cutoff 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, I 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 (The Uniprot Consortium 2014). One down-regulated gene, MALAT-1, is a long noncoding RNA that regulates metastasis-associated genes, and when knocked-down in Mus musculus xenografts leads to decreased tumor 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 likely needed in order 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, 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 confirm genetic divergence between urban and rural populations of *P. leucopus*.

**Conclusion**

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. Genome-wide 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.
Table 1. Specimen collection locations for *Peromyscus leucopus* individuals used in this study

<table>
<thead>
<tr>
<th>Code</th>
<th>Collection Site (City, State, Year)</th>
<th>Latitude, Longitude</th>
<th># of Individuals Used for RNAseq</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Female</td>
</tr>
<tr>
<td>Rural</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWP_BH</td>
<td>Wildwood State Park; Brookhaven State Park (Long Island, NY, 2012)</td>
<td>40° 57.973', -72° 48.147'</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40° 55.646', -72° 52.282'</td>
<td></td>
</tr>
<tr>
<td>HIP</td>
<td>High Point State Park (Milford, NJ, 2012)</td>
<td>41° 18.375', -74° 40.150'</td>
<td>4</td>
</tr>
<tr>
<td>CFP</td>
<td>Clarence Fahnestock State Park (Putnam Valley, NY, 2012)</td>
<td>41° 26.998', -73° 51.480'</td>
<td>4</td>
</tr>
<tr>
<td>Urban</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FM</td>
<td>Flushing Meadows Park (Queens, NY, 2010)</td>
<td>40.719800, -73.832620</td>
<td>4</td>
</tr>
<tr>
<td>NYBG</td>
<td>New York Botanical Gardens (Bronx, NY, 2010)</td>
<td>40.863404, -73.875743</td>
<td>4</td>
</tr>
<tr>
<td>CP</td>
<td>Central Park (Manhattan, NY, 2010)</td>
<td>40° 47'47.54&quot;, -73° 57'21.64&quot;</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 2. Assembly statistics. The SOLID_454_Filtered assembly was used in downstream analyses.

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Number of transcripts (≥100bp)</th>
<th>Total bases (Mb)</th>
<th>Mean transcript length</th>
<th>N50</th>
<th>Number of transcripts (≥2kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>454_Newbler</td>
<td>15,004</td>
<td>13.42</td>
<td>894</td>
<td>1039</td>
<td>825</td>
</tr>
<tr>
<td>SOLiD_Full</td>
<td>145,072</td>
<td>56.57</td>
<td>390</td>
<td>395</td>
<td>1134</td>
</tr>
<tr>
<td>SOLiD_454_Full</td>
<td>143,552</td>
<td>62.05</td>
<td>432</td>
<td>468</td>
<td>2336</td>
</tr>
<tr>
<td>SOLiD_454_Filtered</td>
<td>40,908</td>
<td>30.06</td>
<td>734</td>
<td>1044</td>
<td>2260</td>
</tr>
</tbody>
</table>
Table 3. Annotation statistics from BLASTx of *Peromyscus leucopus* contigs against *Mus musculus* Uniprot (2013_11) database.

<table>
<thead>
<tr>
<th>BLASTX hits to Uniprot</th>
<th>Number of transcripts (≥100bp)</th>
<th>Unique Uniprot proteins</th>
<th>Uniprot proteins (coverage ≥ 80%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>454_Newbler</td>
<td>10553</td>
<td>6387</td>
<td>2636 (24.9%)</td>
</tr>
<tr>
<td>SOLiD_Full</td>
<td>41311</td>
<td>9556</td>
<td>2510 (6.1%)</td>
</tr>
<tr>
<td>SOLiD_454</td>
<td>38919</td>
<td>10838</td>
<td>4572 (11.7%)</td>
</tr>
<tr>
<td>SOLiD_454_Filtered</td>
<td>38855</td>
<td>10834</td>
<td>4568 (11.7%)</td>
</tr>
</tbody>
</table>
**Figure 1.** Distribution of alignment coverage of full genes between subject (gene) and query *(Peromyscus leucopus contig)* in *Mus musculus* Uniprot database using BLASTx.
Figure 2. Population structure analysis of 6,449 SNP loci genotyped for 48 individuals from urban \((N = 24)\) and rural \((N = 24)\) populations using smartPCA and sNMF. A. Individuals color-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.
Figure 3. Inferred locations of protein-coding genes on *P. 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).
**Figure 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.
Chapter 3

Genome-wide SNP data reveals urbanization impacts the demographic history of a New York City rodent

ABSTRACT

The recent post-glacial history of the New York City (NYC) metropolitan area is relatively well documented, but understanding how urbanization shapes genome-wide diversity and species distributions is still largely unexplored. Urbanization is often characterized by extreme fragmentation of native habitats with roads and human infrastructure forming effective barriers to dispersal and isolated populations. I used coalescent based simulations to infer the demographic history of white-footed mice, *Peromyscus leucopus*, from the site frequency spectrum (SFS). Using a double-digest restriction site-associated DNA sequencing (ddRADseq) approach, I generated thousands of RAD loci for white-footed mice collected along an urban-to-rural gradient. I assigned individuals to evolutionary clusters using population structure analyses and subsequently used these results to inform demographic modeling. Recent divergence times, population sizes and their changes through time, and migration were inferred from our large genomic SNP dataset. Our modeling results indicate that post-glacial sea level rise led to isolation of mainland and geographic Long Island populations. These models also indicate that several urban parks in NYC represent distinct *P. leucopus* populations, and the estimated divergence times for these parks are consistent with patterns of urbanization in New York City. Taken together, our results suggest that urbanization is a driver of population divergence on par
with large-scale geological processes, and anthropogenic habitat fragmentation can influence the evolutionary history of organisms across urban landscapes.

**Keywords:** ddRAD-seq, white-footed mouse, *Peromyscus leucopus*, demography, composite likelihood, coalescent, site frequency spectrum
Introduction

Environmental changes over time lead to variable fitness between phenotypes, and thus drive directional selection in populations (Maccoll 2011). Currently, human activity significantly impacts all of Earth’s ecological processes (Corlett 2015), and urbanization is a particularly potent driver of ecosystem change around the world. In 2011, over half of the world’s population lived in urban areas, and in the next four decades the urban population will increase to over two-thirds of all humans (Heilig 2011). By understanding the impact of urbanization on the demographic history of other species, I will uncover basic population genomic responses of organisms to human-driven change and establish baselines for examining future evolutionary responses (Donihue & Lambert 2014). While the inference of demographic history from genomic data is well documented in model organisms (Vonholdt et al. 2010; Huber et al. 2014; Sousa et al. 2014), the advent of inexpensive genomic sequencing technologies (Baird et al. 2008) has made it feasible to investigate the demographic history from genomic data in non-model organisms (Nadachowska-Brzyska et al. 2013; Toonen et al. 2013; Moura et al. 2014; Trucchi et al. 2014; Lozier 2014). Using genome-wide neutral SNP data, I investigate the effect of urbanization in the New York City (NYC) metropolitan area on historical demography of the white-footed mouse, Peromyscus leucopus.

Historical demographic processes influence the observed patterns of genetic variation found in contemporary populations (McCoy et al. 2013). Quantifying historical demographic parameters like population size, population expansions or contractions, population structure, gene flow between populations, occurrences of bottlenecks, and time of population divergences, can be used to explain the distribution and evolutionary potential of contemporary populations. Genetic drift is the random change in allele frequencies from random sampling between
generations (Wright 1931), and by assuming drift under the neutral theory of evolution (Kimura 1968), I can infer these past demographic events based on the current patterns of allele frequencies. Several methods have recently been developed that use allele frequency patterns to infer demographic history from genome-wide SNP data in non-model organisms, including those that use approximate Bayesian computation (ABC) (Cornuet et al. 2008; Lopes et al. 2009), diffusion approximation (Gutenkunst et al. 2009), and maximum likelihood estimation (Adams & Hudson 2004; Lukic & Hey 2012; Excoffier et al. 2013). Often, the estimated patterns of genetic variation generated under a demographic model are compared to observed genome-wide patterns using the site frequency spectrum (SFS). The SFS is a distribution of the number of segregating sites and can be the full SFS when ancestral versus derived states are known or the folded SFS, which uses major and minor allele frequencies.

Using an ABC framework for parameter estimation, Robinson et al. (2014) used > 1,000 loci generated from low-coverage whole genome sequencing from Palearctic oak gallwasp populations to test hypotheses about gene flow and population divergence after the Pleistocene glacial maxima. Using demographic models inferred from previous studies, they looked at 40 summary statistics to capture patterns in allele frequencies and found that ABC-based inference supported previous demographic findings and that parameter inference was robust across models. However, they noted that an increase in the number and length of loci would produce more accurate inferences. The diffusion approximation of the SFS implemented in dadi (Gutenkunst et al. 2009) was used by Trucchi et al. (2014) to infer the recent demographic history of king penguin populations around Antarctica. They used RADseq to generate > 20,000 SNPs and found population expansions corresponded with warming periods estimated from ice cores.

Fastsimcoal2 uses the composite-likelihood to infer demographic parameters from the SFS and
was used by Excoffier et al. (2013) to reconstruct human demography in Africa under relatively complex scenarios. They used > 100,000 SNPs and tested scenarios of migration, expansion, and bottlenecks in up to 10 populations. Their demographic parameters converged on estimates and were in line with demographic histories estimated from previous studies. It is important to note that these studies used some combination of historical or geological records in addition to genetics to support estimates and answer questions about past demographic events from the present.

I investigated the demographic history of white-footed mice occupying contemporary forest fragments in NYC and the surrounding area. NYC and the surrounding region is particularly suited for inferring historical demography because of the city’s recent, rapid, and well-documented urbanization over the past 125 years. NYC’s geologic (Isachsen 2000), ecological (Sanderson & Brown 2007; Sanderson 2009) cultural and political history (Caro 1975; Burrows & Wallace 1998) have also been meticulously recorded and this information can inform demographic models for testing. NYC also has clearly defined urban green spaces that are delimited by anthropogenic and natural barriers, and thus occupied by independently-evolving populations of species with poor mobility through the urban matrix (Munshi-South & Kharchenko, 2010). Natural barriers include the Hudson and East Rivers that separate the mainland portion of the city (i.e. Bronx) from Manhattan and Long Islands. The establishment of Long Island did not begin until the retreat of the late Wisconsin glacier which, at its southern most extent, covered NYC and the majority of Long Island (Lewis & Stone 1991). The glacier began to retreat northward around 21,000 years before present (ybp), and by 15,000 ybp the glacier was several hundred kilometers north of NYC (Lewis 1995). For the next few thousand years, white-footed mice and other taxa would have slowly migrated north into the NY region.
and Long Island from southern refugia (Smith 1957), depending on ecological succession and suitable plant community development. During this time period, *P. leucopus* presumably maintained high levels of gene flow until sea level rise from glacial melt finally separated Long Island from mainland NY between 12,000 and 15,000 ybp (Lewis 1995). Anthropogenic barriers were not erected until after European settlement of the area around 1600 BCE (Sanderson & Brown 2007). During early phases of urbanization before the American Revolution (1609-1790) in NYC, the only green spaces within the built city were parade grounds, cemeteries, market places, farms, or private estates with highly manicured flora and fauna, unlike the forest fragments that occur in contemporary urban parks. In the mid-19th century a park movement gained support in order to alleviate crowding in the city (Christen 1988). The development of NYC parks followed a similar pattern. Heavily used plots of land, like Prospect and Central Parks, were taken over by city officials and completely transformed for aesthetic purposes (Rosenzweig & Blackmar 1992). Private estates that functioned as agricultural land, plantations, or country retreats were slowly acquired by the NYC government and redesigned as vegetated parkland (Christen 1988). The remnant fauna populations in this green space, surrounded by a dense urban infrastructure, likely recovered from bottlenecks caused by urban fragmentation as the parks developed into mature forests.

The white-footed mouse is a prevalent small mammal and likely urban adapter (i.e. a native species found at higher densities in urbanized areas (McKinney 2006)), in NYC’s forest fragments. In this study I estimate the demographic history of *P. leucopus* in NYC to test hypotheses about population expansion and divergence in response to urbanization. White-footed mice persist in small, fragmented urban forests (Pergams & Lacy 2007; Rogic *et al.* 2013; Munshi-South & Nagy 2014), and exchange migrants only through vegetated corridors between
isolated NYC parks (Munshi-South 2012). Substantial genetic structure at microsatellite loci exists between NYC parks (Munshi-South & Kharchenko 2010), and there is evidence of divergence and selection in genes underlying functional traits from urban populations of NYC white-footed mice (Harris et al. 2013).

Documenting adaptive responses of populations in human-dominated landscapes is important for understanding large-scale species responses to our changing climate (Donihue & Lambert 2014). However, an understanding of the historical demography of populations is a prerequisite for rigorously examining adaptation. Demographic history can affect genetic variation in ways that mimic natural selection and adaptive responses. For example, unidentified background selection can cause misidentification of the null demographic model because selection against deleterious mutations increases the number of low frequency variants. This problem in turn may lead to false negative calls of exponential population growth that produce a similar pattern (Bank et al. 2014). During a demographic bottleneck, the reduction in population size increases the number of intermediate frequency alleles, similar to the effect caused by balancing selection maintaining allelic variation (Biswas & Akey 2006). Population expansions and positive selection are also difficult to disentangle, as both lead to an increase in low frequency alleles (Nielsen 2005; Akey 2009). Unless the demographic history is inferred from a neutrally evolving data set and accounted for in the null model, then scans for signatures of selection from protein coding sequences may suffer from very high false positive rates (Nielsen et al. 2009).

Urbanization in NYC is a relatively recent phenomenon and may have generated patterns of genetic variation resembling natural selection (Przeworski et al. 2005). I can model and test past evolutionary histories to infer demographic parameters (Gutenkunst et al. 2009; Lukic &
Hey 2012; Excoffier et al. 2013), and through coalescent-based simulations, include these demographically-driven genetic patterns in our neutral model when identifying adaptive evolution. This study is the first to examine the impact of urbanization on demography using patterns of genome-wide variation in wild populations. For this study, I used genome-wide SNP data from a double-digest restriction-site associated DNA sequencing (ddRADseq) (Peterson et al. 2012) protocol from 23 white-footed mouse populations sampled along an urban-to-rural gradient (Munshi-South et al. 2015). I then used evolutionary clustering analyses to identify NYC parks that contain unique populations of white-footed mice. I inferred demographic parameters from the SFS using a composite-likelihood and coalescent simulation approach implemented in fastsimcoal2 to determine whether population divergence was due to urbanization in NYC. I used results from demographic modeling to answer questions about the impact of urbanization on white-footed mice. How do population sizes change in response to habitat fragmentation during urbanization? What natural and anthropogenic barriers drive population divergence in mouse populations in NYC? How do human built barriers to dispersal affect gene flow between urban white-footed mouse populations? This resulting demographic history can then be used to inform future studies of selection in NYC populations of P. leucopus.

Materials and Methods

Sampling and DNA extraction

Two previous studies (Harris et al. 2015) , Munshi-South et al. 2015) sampled individual white-footed mice between 2009 and 2013 from 23 separate localities. Sites were chosen to represent a rural to urban gradient (Fig. 1). Rural sites were defined as large tracts of relatively undisturbed natural habitat, and urban sites were fragmented habitat surrounded by urban
infrastructure and impervious surface. For all sampling locations, I trapped individuals over a period of 1-3 nights each. At each site, I set between one and four 7x7 m transects with 49 3” x 3” x 9” Sherman live traps, depending on the total area of each sampling site. I weighed, sexed, and took morphological measurements for all individual mice. For Central Park, Flushing Meadow, New York Botanical Garden, Brook Haven Park & Wild Wood Park, High Point Park, and Clarence Fahnestock Park, I took tissue from already collected liver samples stored in RNALater (Ambion Inc., Austin, TX) at -80C. At all other sites, I collected tissue by taking 1 cm tail clips, placing in 80% ethanol, and storing at -20C in the laboratory. I extracted genomic DNA using standard extraction protocols, quantified the yield, and checked quality before genomic sequencing library preparation (see protocols described in (Munshi-South et al. 2015) for full details). All animal handling procedures were approved by the Institutional Animal Care and Use Committee at Brooklyn College, CUNY (Protocol Nos. 247 and 266) and by Fordham University’s Institutional Animal Care and Use Committee (Protocol No. JMS-13-03).

**RAD sequencing and SNP calling**

We filtered out relatives and low-quality samples, retaining 191 *P. leucopus* individuals from 23 sampling sites for the genome-wide SNP dataset. All samples came from a previous study to look at population genomics of *P. leucopus* from a neutral genome-wide SNP dataset generated from a double-digest RADseq protocol. As such, library preparation, sequencing, quality filtering, *de novo* assembly, and SNP calling are described in detail in Munshi-South et al. (2015). Briefly, we followed standard protocols for ddRADseq presented in Peterson et al. (2012), starting with DNA extraction using Qiagen DNEasy kits with an RNAsé treatment. Next we used a combination of the enzymes, SphI-HF and MluCI to generate similarly sized DNA
fragments. Using AMPure XP magnetic beads we cleaned the digested DNA then ligated barcodes and Illumina sequencing adapters to the fragments. We used a Pippin Prep for precise DNA fragment size excision from gels and then Phusion High-fidelity PCR reagents to add Illumina sequencing primers. The resulting fragments were sent to the NYU Center for Genomics and Systems Biology who sequenced the samples using three lanes of Illumina HiSeq 2000 2x100 bp paired-end sequencing. We checked initial quality of the raw reads using FastQC and subsequent primer removal, low-quality nucleotide trimming, and de novo SNP calling was conducted using the Stacks 1.21 pipeline (Catchen et al. 2013). Using default settings in Stacks, we called and filtered SNPs except for requiring that loci occur in 22 / 23 sampling sites, and within each site, occur in at least 50% of individuals. We chose a random SNP from each RAD tag to avoid linkage between loci. Additionally, we removed individuals if they had too few reads resulting in extremely small SNP datasets or if they showed high levels of relatedness to other white-footed mice sampled. These filters resulted in 14,990 SNPs in the final dataset I used for demographic modeling.

**Population structure and migration**

I investigated patterns of genetic diversity in order to define evolutionary clusters that could be used to inform demographic modeling of *P. leucopus* populations in the NYC region. I examined population structure and evidence of migration among all 23 sampling sites. Adegenet is a software package linking multivariate analyses with genetic marker data in order to concisely summarize genetic variability (Jombart 2008) using the R computing environment, version 3.1.2 (R Core Team 2013). Discriminant analysis of principle components (DAPC) is a model-free multivariate method implemented in adegenet to identify and describe clusters of genetically
related individuals from structured populations by maximizing the variance among groups and minimizing the variance within groups (Jombart et al. 2010). DAPC requires a priori knowledge of groups, and I assigned individual *P. leucopus* to the sampling location where they were caught. I performed DAPC on the data after calculating the *a*-score to determine the optimal number of principal components to retain and avoid over-fitting. The *a*-score maximizes the percent of individuals assigned to the correct group while minimizing the percent of assignments for completely randomized groups. I performed global and hierarchical analysis with subsets of the data to identify genetically structured populations. The main evolutionary clusters identified in the global analysis were split into two separate sub-sets of individuals and analyzed separately. This hierarchical analysis looked at structure within the sub-sets of the data to pick up fine scale patterns in populations structure. The compoplot function within adegenet allows results to be viewed in a STRUCTURE-like plot and is particularly useful for identifying individuals exhibiting substantial admixture. For all analyses, I used compoplot to identify genetic clusters that contained no more than 50 % admixed individuals. DAPC is an exploratory method, and I used it to putatively find the most isolated populations.

I also used sNMF version 0.5 (Frichot et al. 2014) to examine population structure. sNMF explores patterns of structure by assigning individual ancestry coefficients using sparse non-negative matrix factorization. sNMF does not make any model assumptions like requiring populations to be in Hardy-Weinberg and linkage equilibrium (Frichot et al. 2014), as opposed to other likelihood models like STRUCTURE (Pritchard et al. 2000). For the number of putative ancestral populations tested, I chose a range from $K = 1$ to $K = 11$ using default parameters, with 10 replicate runs for each value of $K$. I ran sNMF on the full 14,990 SNP dataset (≤ 50% of SNPs missing per population) and on a more conservative dataset with only ≤ 15% of SNPs
missing per population. In the case of missing data, sNMF imputes the missing genotype by resampling from the empirical frequency at each locus (Frichot et al. 2014), and using fewer missing data ensured any population structure patterns found were not due to incorrectly imputed genotypes (Fig. S1). To infer the most likely number of ancestral populations, each model run generates a cross-entropy calculation based on ancestry assignment error when using masked genotypes. The model with the smallest cross-entropy score implies it is the best prediction of the true number of K ancestral populations (Frichot et al. 2014).

**Demographic inference from genome-wide site frequency spectra**

In order to reduce model complexity for demographic inference, I attempted to group individuals into the minimum number of populations representing unique evolutionary clusters. Global analyses in DAPC and sNMF showed the highest support for two populations split by the East River and Long Island Sound, and hierarchical analyses showed support for isolated urban populations. Collectively, results suggested a minimum of seven putative populations could capture most of the genetic variation between populations (Mainland & Manhattan: MM, Long Island: LI, Central Park: CP, Van Cortlandt Park: VC, Inwood Hill Park: IP, Jamaica Bay: JB, Fort Tilden: FT, Fig. 1). Along with hierarchical population structure results, I chose several of the urban populations to include in demographic modeling based on the size of the park, the relative isolation of the park due to urbanization, and the population density of white-footed in the park. I generated the multi-population site frequency spectrum (MSFS) for subsets of populations to test specific demographic history scenarios. I used custom scripts (see supplemental materials) and the `dadi.Spectrum.from_data_dict` command implemented in `dadi` (Gutenkunst et al. 2009) to generate the MSFS. When I created the SNP dataset, I required a SNP to occur in $\geq 50\%$ of individuals from each population, so the MSFS was downprojected (a
process where a minimum number of individuals is required to have the SNP), to 50% to allow for missing data (Gutenkunst et al. 2009). Once the MSFSs were generated, I used the software program fastsimcoal2 (Excoffier et al. 2013) for demographic inference. Fastsimcoal2 (fsc2) uses a composite multinomial likelihood approach to infer demographic histories from the site frequency spectrum generated from genomic scale SNP datasets. The expected SFS under user defined demographic scenarios is obtained using coalescent simulations.

I tested demographic histories under a scenario of population isolation with migration (IM model). This included six hierarchical IM models and I compared inferred parameters between models (Fig. 2). There was one two-population IM model (seven free parameters) to test older divergence patterns between MM and LI suggested from the geologic record. The remaining five models were three-population IM models (15 free parameters) testing for recent urban population divergence. I chose to run separate models investigating each urban population separately in order to avoid inconsistencies from overparameterization. I considered an ancestral population that split at time $T_{\text{div}1}$ and then an urban population that split more recently at time $T_{\text{div}2}$. For $T_{\text{div}1}$ I included a range of parameters based on the LGM of the Wisconsian glacier, ~18,000 ybp. For $T_{\text{div}2}$ I considered a range of parameters incorporating the timeframe of urbanization in NYC, ~400 ybp. I allowed for migration between all populations, and tested occurrences of population bottlenecks directly after divergence when urban isolation was incorporated into the model (See Appendix 3.3 for demographic models and parameter ranges). During likelihood calculation, a conditional maximization algorithm (ECM) is used to maximize the likelihood of each parameter while keeping the others stabilized. This ECM procedure runs through 40 cycles where each likelihood is calculated using 100,000 coalescent simulations. While increasing the number of simulations can increase precision, accuracy does not
significantly increase past 100,000 simulations (Excoffier et al. 2013). Additionally, in order to avoid likelihood estimates that oversample parameter values at local maxima across the composite likelihood surface, I ran 50 replicates with each starting from different initial conditions. I chose the replicate with the highest estimated maximum likelihood score for each model. Using parametric bootstrapping, I generated confidence intervals for the most likely inferred demographic parameters generated. The SFS was simulated with the parameter values from the highest likelihood model and then new parameter values re-estimated from the simulated SFS. I ran 100 parametric bootstraps. In order to find consistent signals of divergence which could be attributed to urbanization, I compared parameter values and overlapping confidence intervals between models.

Results

Population genetic structure and admixture

Cross-validation of the DAPC analysis showed maximum discriminatory power without over-fitting by retaining 23 principal components. Sampling locations clustered according to geographic patterns along discriminant functions one and two (Fig. 3). Mainland & Manhattan sites formed one cluster while Long Island sites formed the other. Jamaica Bay and Fort Tilden on Long Island clustered separately from the two larger groups in this global analysis. I then split individuals into two subsets for separate DAPC analysis according to the global analysis result representing Mainland & Manhattan (MM) and Long Island (LI) individuals. For the MM groups, I retained 12 principal components for discriminant analysis. In this analysis, most sampling locations grouped together in one cluster, but discriminant function one separated three urban sampling locations into unique clusters. Central park (CP) formed its own geographic cluster while Van Cortlandt Park (VC) and Inwood Hill Park (IP) formed a separate cluster. Van
Cortland and Inwood Hill Parks were then further separated along discriminant function two (Fig. 3). I analyzed the genetic composition of these clusters using the *compoplot* function, which also found CP, VC, and IP, to contain individuals where ≤ 50% of their genetic ancestry was assigned to different sampling locations (Fig. S2). I kept seven PCs for the discriminant analysis of Long Island individuals. A similar pattern to the global analysis was seen, where all Long Island sampling locations formed one cluster except for Jamaica Bay (JB) and Fort Tilden (FT). CP, VC, IP, JB, and FT are sampling locations that all occur within the boundaries of New York City and are referred to as ‘urban’ populations from this point forward.

Individual ancestry assignment in sNMF supported the global analysis in DAPC. After ten replicates each testing a K value of 1 – 11, assignment of two ancestral populations always had the highest support (Cross Entropy = 0.199) with structure occurring between MM and LI sampling sites. There was evidence of admixture between many individuals, which could come from populations with low genome-wide variation, possibly due to a bottleneck event. Overall, however, the distinction between MM and LI was clear using both the full and ≤ 15% of SNPs missing dataset (Fig 4, Fig S2).

**Demographic inference**

Parameters were free to vary in demographic modeling using fastsimcoal2, but all six models converged on similar parameter values estimated from the observed MSFS. Parameter estimates with the highest likelihood generally fell within the upper and lower bounds generated from parametric bootstrapping (Fig. 5, Table 1). The first two-population model tested divergence time, effective population size, and migration rates between MM and LI populations (Model 1, Fig. 2). The divergence time for the MM and LI split was inferred to be 13,599 ybp
and the MM effective population size (Ne) was 50x larger than the LI Ne (Table 1). Divergence times are based on a generation time of 0.5 years for *Peromyscus leucopus*. Migration was also inferred to be low, < 1 individual per generation, between MM and LI (Table 1).

The inferred demography for the more complex three-population models generally supported results from the two-population model. The first two complex models both estimated the divergence between MM and LI, but one model tested for divergence of JB and LI after the MM and LI split (Model 5, Fig. 2) while the other model tested divergence between FT and LI after the MM and LI split (Model 6, Fig. 2). This model also tested the likelihood of a bottleneck event when FT and JB, both urban populations, diverged. I set up the other three complex models in an identical fashion, except I tested the urban populations of CP (Model 2, Fig. 2), VC (Model 3, Fig. 2), or IP (Model 4, Fig. 2) for divergence from MM after the MM and LI split. Point estimates for demographic parameters converged on similar values and generally fell within the 95% confidence limits from parametric bootstrapping (Fig. 5, Table 1). The average divergence time for MM and LI was 14,679 ybp SD = 956.19. Similar to the two-population model, the MM Ne was larger (at least 2x in each model) than the LI Ne. The individual urban populations all had Ne’s 10x smaller than MM, but often similar to LI. The divergence time for the five tested urban populations, even with variation in # of generations per year, was consistent with the timeframe of urbanization (mean divergence = 233 ybp SD = 164.5). There was evidence for bottlenecks when urban populations diverged in CP, IP, FT, and VC, while JB had a Ne 1.86x larger than present day estimates (Table 1, modeled as instantaneous population size change). There was further evidence of low gene flow between MM and LI, and evidence of asymmetrical gene flow from MM into all urban populations.
Discussion

Evidence for genetic structure and admixture from RADseq data

I estimated the post-glacial demographic history of white-footed mouse populations in the NYC metropolitan area using genome-wide SNP data, and divergence time estimates aligned with the known geologic record and contemporary habitat change. Urbanization in NYC is a recent phenomenon, and I provide evidence using genome-wide SNP data that habitat fragmentation from urbanization strongly influences the evolutionary history of urban fauna. sNMF is extremely fast when dealing with large genomic datasets, and the analyses supported assignment of individuals to two main groups. The clearest signal showed Long Island as a likely ancestral population while Mainland & Manhattan represented another. This older geographic split occurs along the East River and Long Island Sound. The geologic record supports the separation of Long Island from the Mainland from rising sea levels ~13,000 ybp, and results from sNMF likely represent this geographic split. There was evidence for admixture in both populations, though this is not unexpected as mice are known to be decent swimmers, and the number of bridges and high human traffic between MM and LI provides many opportunities for individual mice to move across the barrier. When the full SNP dataset with ≤ 50% of SNPs missing was used, the assignment of some individuals to MM or LI contradicted their sampling location (Fig. 4). For the final analysis, I used a smaller subset of SNPs with ≤ 15% missing data. The contradictory assignments were not observed with the ≤ 15% missing data set (Fig. S1). The signal from ≤ 50% missing data set is likely due to decreased accuracy of sNMF in assigning individuals to the correct ancestry in the presence of ≥ 20% of missing SNPs (Frichot et al. 2014).
DAPC also supported the two population MM and LI history. Compared to sNMF, however, DAPC was also able to distinguish more fine-scale genetic structure. After confirming the MM and LI populations, I looked at LI and MM individuals separately in DAPC. These analyses suggested that several urban parks were representative of distinct populations. Central Park, Van Cortlandt Park, and Inwood Hill Park are located within MM while Jamaica Bay and Fort Tilden occur on Long Island, but all parks have similar characteristics. These urban parks are all relatively large, have a high proportion of vegetative cover, and are surrounded by a dense urban matrix. There were no rural sampling locations that showed evidence of being distinct populations, suggesting that CP, VC, IP, JB, and FT have been isolated due to urbanization. One caveat of this approach is the *a priori* assignment of sampling locations as population units. DAPC maximizes inter-population genetic variation, and thus may tend to identify separate evolutionary clusters based on weak genetic structuring. However, I have extensively sampled this geographic area and the fragmented nature of urban and suburban green spaces creates natural barriers around parks (Munshi-South 2012; Munshi-South & Nagy 2014). I believe that our sampling coverage and knowledge of sampling site boundaries allowed us to make informed population unit assignments when using DAPC.

**P. leucopus** population history during recent urbanization in NYC

Inferred parameter estimates show a consistent signal of an older geographic split between Long Island and Mainland & Manhattan populations followed by recent divergence of populations within NYC parks. All the models estimated a large ancestral population size followed by divergence of MM and LI ~13,600 ybp, followed by population contraction. This estimate is in line with geologic records of the area, though the exact date *P. leucopus* initially
migrated into NY and Long Island likely took place over hundreds to thousands of years after glacial retreat. Ecological succession in this area progressed from tundra, to fir and spruce forests, and then the climax community of oak and hickory dominated forests, the preferred white-footed mouse habitat (Smith 1957).

Our other demographic models examined whether contemporary urban populations diverged from MM or LI within the historical timeframe of urbanization in NYC. The history of modern NYC’s changing landscape begins with the arrival of Europeans in 1609, when only 1% of the Manhattan landscape was urbanized in the first few years after European arrival. Over the next 400 years, they altered, built, and redefined the natural wetlands, grasslands, and forests of NYC until 97% of the land had been heavily altered for human use (Sanderson & Brown 2007). Across all five of our models that included urban populations, the inferred time of divergence was always within the 400 year window of European settlement. While 400 years, representative of ~800 generations, is relatively recent, detailed demographic parameter inference over very young time scales is possible using the SFS from large enough genomic datasets (Excoffier et al. 2013). *Fastsimcoal2* was used to identify divergence times in sunskinks across South East Asia using RADseq generated loci and found reliable estimates of divergence as young as 435 generations before present (Barley et al. 2015). When examining human population demographic history, Excoffier *et al.* (2013) were confidently able to make inferences about migration, divergence, and population size changes that occurred as recently as 140 generations before present. Translating genetic estimates of divergence time from generations to years is dependent on an estimate of generation time for white-footed mice. I used a conservative value of 0.5 years, but the true value likely fluctuated over time and may have been as low as 0.3 or as high as 1 in different time periods. Even given potential variation in
generation time, many point estimates for urban park divergence are in line with the founding of urban parks in NYC. These estimates indicate that divergence between white-footed mouse populations was driven by urbanization of the landscape. More broadly, these results indicate that isolation in urban fragments is a strong enough force to impact the evolutionary history of urban fauna.

Within urban populations, there was evidence of bottlenecks immediately after divergence. I inferred the rate of population contraction during the bottleneck event, and for CP, VC, IP, and FT, populations resized to a $N_e < 1$. One commonly used rule suggests that $N_e$ is one fifth the population census size (Mace & Lande 1992), and given stochasticity in parameter estimates, this low $N_e$ value suggests that a small remnant group of individual white-footed mice within these parks at the time of the bottleneck provided most of the genetic variation found within the parks today. The population size of JB at the time of divergence was much larger in the past, and has contracted to the current size inferred for the present day population. This is consistent with the historical record. All of the urban populations except for JB were managed land and cleared for various purposes. Jamaica Bay (JB) is an 18,000 acre wetland estuary ecosystem, and while it has been reshaped and its ecosystem goods and services utilized by both Native Americans and Europeans, it has never undergone the extensive development seen in the other urban parks (Black & Brown 2001). JB has likely maintained a *P. leucopus* population that has fluctuated due to storms, flooding and habitat degradation, but has only recently undergone severe contraction due to encroaching urbanization.

I also inferred migration rates between all populations. In general, migration was high and variable among the 23 sampling sites, but I saw consistent patterns of low migration between MM and LI, and asymmetrical migration of individual mice from MM into urban populations.
In this study, I do not rely on the exact migration rates when describing the population history of *P. leucopus* in NYC. Given the extremely recent divergence times I am reporting here, the high migration rates could be due to retained ancestral polymorphisms from incomplete lineage sorting or geographic structure, which is difficult to distinguish from admixture (Lohse & Frantz 2014). Instead I report general patterns in migration rates seen between populations.

*Fastsimcoal2* is one of many current methods to infer the demographic history of a population from genome-wide datasets by calculating likelihoods from coalescent simulations (Li & Durbin 2011; Harris & Nielsen 2013; Sheehan et al. 2013; Excoffier et al. 2013; Bhaskar et al. 2014) or diffusion approximation (Gutenkunst et al. 2009; Lukic & Hey 2012). Many of these methods require data that is not attainable for many study organisms. Reference genomes, high-coverage full genome sequences, or genome-wide haplotypes may be required. Many of these methods are also restricted by the complexity of demographic models that can be tested. Computational time is another large constraint for parameter inference from complex models. *Fastsimcoal2* is relatively fast at calculating the approximate likelihood from unlinked SNP loci, which can be readily generated using RADseq approaches. It is also able to take advantage of the full multi-population site frequency spectrum, avoiding some loss of information. Recently *fastsimcoal2* has been used on similar SNP datasets to infer relatively complex demographic histories. Originally, the authors of the program were able to confirm many previous demographic parameter estimates about human expansion and divergence in Africa using SNP chips and complex demographic models consisting of up to 10 independent populations (Excoffier et al. 2013). Ultimately, they were able to find ancient divergence (> 110,000 years) between the human Yoruba and San populations. In a non-model organism, Barley et al. (2015) were able to find the time of divergence in sun skinks between multiple islands in Asia with
relatively low migration between islands. Papadopoulou & Knowles (2015) went one step beyond simple divergence time estimation, and tested the species-pump hypothesis of cycles of connectivity and isolation as a driver of species divergence in Caribbean crickets. Using the geologic and bathymetric record, they narrowed divergence time between species to a period of repeated connectivity and isolation between islands, supporting the species-pump hypothesis. Similarly, I tie divergence times in urban parks to the start of intense urbanization in NYC.

One limitation is that the use of approximate likelihoods can negatively affect the quality of demographic parameter inference (Kim et al. 2015). To account for this, I ran multiple models testing urban population divergence and looked for convergence on similar demographic parameters (Fig. 5, Table 1). While point estimates converged on similar values across models, there are several instances where they fell outside the 95 % confidence intervals generated from parametric bootstrapping (Table 1). This discrepancy is not unexpected as our models are simplified representations of the true population demographic history. Additionally, all parameters are correlated and several combinations of parameter values can lead to similar expected SFSs (Excoffier et al. 2013). I cannot account for every demographic event since the last glacial maximum. When estimating divergence and migration between approximately delimited P. leucopus populations, I am specifying an overly simplistic model to uncover the general patterns that explain the true history. I benefitted from the detailed geologic and historical records for the NYC landscape that allowed us to independently verify that our estimates are biologically realistic.

Our results suggest a combination of geography, historical geologic events, and human driven habitat change has left a measureable genetic signal across the genome of P. leucopus in NYC. These findings are consistent with other studies on urban evolutionary history which also
show that urbanization may influence evolutionary processes. Urban populations are expected to be genetically isolated from rural populations due to barriers to dispersal from human infrastructure (Sih et al. 2011). Zurich, Switzerland represents an urban center with some similarities to NYC, including natural geographic barriers in the form of the Limmat River that runs through its center. It also contains an urban adapter, the red fox, *V. vulpes*, that colonized Zurich within the last 15 years. Similar to NYC white-footed mice, urban *V. vulpes* represent isolated populations with increasing fox density since colonization (Wandeler et al. 2003b). Even with this short time frame and evidence of migration, urban red foxes are genetically differentiated from rural counterparts and the highest $F_{ST}$ values are found between urban sites (Wandeler et al. 2003b). Despite the extremely short distances between many urban parks representative of suitable habitat for urban adapters, individual parks are often isolated from each other with complex population dynamics between them (Davis & Glick 2009). *Parus major* populations in Barcelona, Spain showed genetic differentiation between urban parks and rural sites, but some populations within the city acted as ecological sinks (Bjorklund et al. 2010). I did not find evidence of ecological sinks in NYC, but population histories were variable between parks even separated by only one or two kilometers.

**Conclusions**

I have used the composite-likelihood approach implemented in *fastsimcoal2* to infer demographic parameters from the SFS generated for NYC populations of *P. leucopus*. Patterns of genetic variation and population structure reflect past demographic processes (Li et al. 2012), and genome-wide neutral SNP data generated from ddRADseq provided enough information to show that urbanization is driving divergence between NYC populations of white footed mice.
The demographic models I tested suggest divergence times and migration patterns that are consistent with what I know about the history of NYC. Using the known geologic and historical record of the region gave support for our demographic parameter estimates. With these precautions I believe our study has inferred accurate estimates for divergence times, population sizes, and patterns of migration for urban populations of *Peromyscus leucopus* in NYC. Moving forward, this inferred demographic history could act as the null background model and increase accuracy when identifying signatures of selection from protein-coding sequences.
Table 1. Inferred demographic parameters with 95% confidence values from parametric bootstrapping for all models.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>LI_MM</th>
<th>LI_MM_CP</th>
<th>LI_MM_VC</th>
<th>LI_MM_IP</th>
<th>LI_JB_MM</th>
<th>LI_FT_MM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 3</td>
<td>Model 4</td>
<td>Model 5</td>
<td>Model 6</td>
</tr>
<tr>
<td></td>
<td>(Point Estimate) (95% CI)</td>
<td>(Point Estimate) (95% CI)</td>
<td>(Point Estimate) (95% CI)</td>
<td>(Point Estimate) (95% CI)</td>
<td>(Point Estimate) (95% CI)</td>
<td>(Point Estimate) (95% CI)</td>
</tr>
<tr>
<td>Ancestral Ne</td>
<td>-</td>
<td>68083</td>
<td>-</td>
<td>81275</td>
<td>-</td>
<td>90482</td>
</tr>
<tr>
<td></td>
<td>(53492- 111400)</td>
<td>(36559- 2310630)</td>
<td>(71280- 312469)</td>
<td>(53734- 283269)</td>
<td>(85828- 269490)</td>
<td>(68608- 284971)</td>
</tr>
<tr>
<td>Ne</td>
<td>1057</td>
<td>12075</td>
<td>8991</td>
<td>8723</td>
<td>7534</td>
<td>6515</td>
</tr>
<tr>
<td></td>
<td>(798- 2203)</td>
<td>(7286- 81320)</td>
<td>(8110- 74287)</td>
<td>(6639- 70959)</td>
<td>(9849- 74872)</td>
<td>(8325- 71939)</td>
</tr>
<tr>
<td>Long island Ne</td>
<td>(LI)</td>
<td>(LI)</td>
<td>(LI)</td>
<td>(LI)</td>
<td>(LI)</td>
<td>(LI)</td>
</tr>
<tr>
<td>Ne</td>
<td>54886</td>
<td>25140</td>
<td>16416</td>
<td>15138</td>
<td>17186</td>
<td>12584</td>
</tr>
<tr>
<td></td>
<td>(50557- 105507)</td>
<td>(13266- 106624)</td>
<td>(13984- 93388)</td>
<td>(12371- 80804)</td>
<td>(18832- 76218)</td>
<td>(12915- 77670)</td>
</tr>
<tr>
<td>Mainland &amp; Manhattan Ne</td>
<td>(MM)</td>
<td>(MM)</td>
<td>(MM)</td>
<td>(MM)</td>
<td>(MM)</td>
<td>(MM)</td>
</tr>
<tr>
<td>Ne</td>
<td>27198</td>
<td>29440</td>
<td>29669</td>
<td>28662</td>
<td>29666</td>
<td>29354</td>
</tr>
<tr>
<td></td>
<td>(6110- 34227)</td>
<td>(24111- 29637)</td>
<td>(17400- 29663)</td>
<td>(23600- 29647)</td>
<td>(18312- 29636)</td>
<td>(12662- 29600)</td>
</tr>
<tr>
<td>Time of divergence (LI_MM)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_MM)</td>
</tr>
<tr>
<td>Ne</td>
<td>746</td>
<td>373</td>
<td>462</td>
<td>423</td>
<td>423</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(512- 13682)</td>
<td>(459- 8586)</td>
<td>(545- 14379)</td>
<td>(667- 7749)</td>
<td>(638- 6298)</td>
<td></td>
</tr>
<tr>
<td>Time of divergence (Urban LI</td>
<td>(CP_M)</td>
<td>(CP_M)</td>
<td>(IP_M)</td>
<td>(IP)</td>
<td>(IP)</td>
<td>(IP)</td>
</tr>
<tr>
<td>or MM)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_MM)</td>
</tr>
<tr>
<td>Ne</td>
<td>3.2</td>
<td>5.5</td>
<td>4.8</td>
<td>4.8</td>
<td>4.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.8- 4.6)</td>
<td>(2.9- 7.6)</td>
<td>(2.5- 7.2)</td>
<td>(3.2- 7.2)</td>
<td>(3.2- 7.1)</td>
<td></td>
</tr>
<tr>
<td>Ancestral recombination rate</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_M)</td>
<td>(LI_MM)</td>
</tr>
<tr>
<td>Ne</td>
<td>1.22</td>
<td>3.2</td>
<td>4.8</td>
<td>4.8</td>
<td>4.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.9- 1.140)</td>
<td>(1.8- 4.6)</td>
<td>(2.5- 7.2)</td>
<td>(3.2- 7.2)</td>
<td>(3.2- 7.1)</td>
<td></td>
</tr>
<tr>
<td>Urban population</td>
<td>(CP)</td>
<td>(CP)</td>
<td>(VC)</td>
<td>(VC)</td>
<td>(IP)</td>
<td>(IP)</td>
</tr>
<tr>
<td>Ne</td>
<td>9.1x10^-4</td>
<td>1.8x10^-5</td>
<td>1.4x10^-5</td>
<td>4.8x10^-5</td>
<td>1.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5.2x10^-6- 2.5x10^-3)</td>
<td>(3.8x10^-4- 2.7x10^-3)</td>
<td>(2.2x10^-6- 7.8x10^-6)</td>
<td>(1.8x10^-5- 0.83)</td>
<td>(2.2x10^-5- 1.6x10^-4)</td>
<td></td>
</tr>
<tr>
<td>Mig,(X)_to_MM</td>
<td>(LI)</td>
<td>(LI)</td>
<td>(LI)</td>
<td>(LI)</td>
<td>(LI)</td>
<td>(LI)</td>
</tr>
<tr>
<td>Ne</td>
<td>1.5x10^-3</td>
<td>1.0x10^-3</td>
<td>1.4x10^-3</td>
<td>2.1x10^-3</td>
<td>1.3x10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.1x10^-3- 3.1x10^-3)</td>
<td>(7.5x10^-3- 4.1x10^-3)</td>
<td>(7.9x10^-3- 3.7x10^-3)</td>
<td>(7.9x10^-3- 2.9x10^-3)</td>
<td>(5.3x10^-4- 9.5x10^-4)</td>
<td></td>
</tr>
<tr>
<td>Mig,(X)_to_LI</td>
<td>(MM)</td>
<td>(MM)</td>
<td>(MM)</td>
<td>(MM)</td>
<td>(MM)</td>
<td>(MM)</td>
</tr>
<tr>
<td>Ne</td>
<td>2.7x10^-3</td>
<td>0.16</td>
<td>5.1x10^-3</td>
<td>3.8x10^-3</td>
<td>2.3x10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.1x10^-3- 6.1x10^-3)</td>
<td>(6.5x10^-4- 0.39)</td>
<td>(4.4x10^-4- 0.22)</td>
<td>(2.5x10^-4- 1.4x10^-4)</td>
<td>(1.4x10^-4- 6.1x10^-4)</td>
<td></td>
</tr>
<tr>
<td>Mig,(X)_to_MM</td>
<td>(CP)</td>
<td>(CP)</td>
<td>(VC)</td>
<td>(VC)</td>
<td>(IP)</td>
<td>(IP)</td>
</tr>
<tr>
<td>Ne</td>
<td>5.1x10^-3</td>
<td>0.07</td>
<td>1.4x10^-3</td>
<td>1.2x10^-3</td>
<td>1.1x10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.0x10^-3- 6.1x10^-3)</td>
<td>(3.5x10^-3- 2.3x10^-3)</td>
<td>(5.6x10^-4- 1.9x10^-4)</td>
<td>(1.9x10^-4- 5.9x10^-4)</td>
<td>(1.3x10^-4- 8.9x10^-4)</td>
<td></td>
</tr>
<tr>
<td>Mig,(X)_to_LI</td>
<td>(CP)</td>
<td>(CP)</td>
<td>(VC)</td>
<td>(VC)</td>
<td>(IP)</td>
<td>(IP)</td>
</tr>
<tr>
<td>Ne</td>
<td>2.9x10^-3</td>
<td>1.4x10^-3</td>
<td>5.4x10^-3</td>
<td>5.3x10^-3</td>
<td>1.0x10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(9.3x10^-3- 6.3x10^-3)</td>
<td>(1.5x10^-3- 3.3x10^-3)</td>
<td>(8.6x10^-3- 2.7x10^-3)</td>
<td>(5.9x10^-3- 4.8x10^-3)</td>
<td>(8.7x10^-3- 6.9x10^-3)</td>
<td></td>
</tr>
<tr>
<td>Mig,(X)_to_MM</td>
<td>(CP)</td>
<td>(CP)</td>
<td>(VC)</td>
<td>(VC)</td>
<td>(IP)</td>
<td>(IP)</td>
</tr>
<tr>
<td>Ne</td>
<td>3.6x10^-3</td>
<td>5.1x10^-3</td>
<td>5.9x10^-3</td>
<td>4.4x10^-3</td>
<td>3.2x10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.7x10^-3- 5.9x10^-3)</td>
<td>(7.5x10^-3- 3.5x10^-3)</td>
<td>(5.0x10^-3- 4.5x10^-3)</td>
<td>(1.7x10^-3- 2.4x10^-3)</td>
<td>(1.3x10^-3- 2.1x10^-3)</td>
<td></td>
</tr>
</tbody>
</table>
Ne = effective population size. Time of divergence is in generations. Migration is reported as the coalescent $m$,
proportion of individuals that move from one population to another per generation.
Figure 1. Map of NYC region with Long Island (LI), Mainland (MM), Inwood Hill Park (IP), Central Park (CP), Van Cortlandt Park (VC), Fort Tilden (FT), and Jamaica Bay (JB) shaded. Each point represents a sampling location and is shaded to correspond to the population assignment.
Figure 2. Three main demographic histories tested in \textit{fastsimcoal2}. MM = Mainland & Manhattan. LI = Long Island. CP = Central Park. VC = Van Cortlandt Park. IP = Inwood Hill Park. JB = Jamaica Bay. FT = Fort Tilden. TDIV = Time of Divergence. N = Effective Population Size. M = migration rate. (A) Model 1 tests divergence between MM and LI. (B) Models 2, 3, and 4 test divergence of CP, VC, and IP from MM, respectively, after the MM and LI split. (C) Models 5 and 6 test for divergence of JB and FT from LI, respectively, after the MM and LI split.
**Figure 3.** Scatterplot generated after discriminant analysis of principle components. Population coordinates are plotted onto discriminant functions 1 and 2. Ellipses are drawn around points from the same *a priori* assigned cluster. (A) Global analysis including all 23 sampling locations. Groups of clusters are shaded according to geographic patterns. LI = Long Island. MM = Mainland and Manhattan. (B) Scatterplot generated for just Long Island populations. (C) Scatterplot generated for just Mainland and Manhattan plots.
Figure 4. Population structure analysis of 14,990 SNPs analyzed in sNMF. Vertical bars represent individuals and are assigned to populations by color. Results are shown for K=2 (highest support). Red = assigned to Long Island. Blue = assigned to Mainland & Manhattan. Mixed colors represent admixture.
Figure 5. Box plots of parametric bootstrap results for inferred demographic parameters. Results are from all six tested models. (A) $N_e$ of LI inferred from each model. X-axis labeled with the urban population being tested for divergence. (B) $N_e$ of MM inferred from each model. X-axis labeled with the urban population being tested for divergence. (C) $N_e$ for each diverged urban population. (D) Time of divergence for urban populations from LI or MM. (E) Time of divergence for LI and MM inferred from each model. (F) # of migrants per generation from urban populations into MM and MM into urban populations.
Chapter 4

Scans for positive selection reveal candidate genes involved in local adaptation to urbanization

Abstract

Urbanization significantly alters natural ecosystems, and its rate is only expected to increase globally as more humans move into urban centers. Urbanized landscapes are often highly fragmented. A common occurrence in human-altered landscapes is habitat fragmentation. Isolated populations within these fragments may adapt in response to novel urban ecosystems, but few studies have found strong evidence of evolutionary responses in urban environments. I used several genome scan methods and genotype-environment association (GEA) approaches to examine signatures of selection in transcriptomes from urban white-footed mice (Peromyscus leucopus) in New York City. I scanned transcriptomes from 48 P. leucopus individuals (eight per population) from six environmentally heterogeneous locations (three urban and three rural) for evidence of rapid local adaption in isolated urban habitats. I analyzed 154,770 SNPs from the transcriptomes, and identified patterns of allele frequency differentiation between urban and rural sites. Patterns in the site frequency spectrum (SFS) for urban sites also indicated selective sweeps in 95 contigs. High quality gene annotations were available for 85 contigs, and 54 of these outlier genes were associated with environmental measures of urbanization. These measures included percent impervious surface and human population density surrounding urban sites. The majority of candidate genes were involved in metabolic functions, especially dietary specialization. A subset of these genes have well-established roles in metabolizing lipids and carbohydrates, including transport of cholesterol and desaturation of fatty acids. Our results
reveal clear genetic differentiation between rural and urban sites that likely resulted from rapid local adaptation in urbanizing habitats. The specific candidate loci that I identified suggest that populations of *P. leucopus* are using novel food resources in urban habitats or metabolizing nutrients in different ways. Ultimately, our data support the small, but growing, body of evidence indicating that cities represent novel ecosystems.

**Keywords**: transcriptome, white-footed mouse, *Peromyscus leucopus*, genotype-environment association, $F_{ST}$ outlier, selective sweep
Introduction

Traits are adaptive when they increase an organism’s fitness in a specific environment (Barrett & Hoekstra 2011), and identification of specific genotypes underlying adaptive traits is a major goal in evolutionary biology. Many studies have identified the genetic basis underlying adaptation, but they often focus on a small number of well-known, conspicuous traits (Nachman et al. 2003; Pool & Aquadro 2007; Linnen et al. 2009; Storz et al. 2009). In the current era of high-throughput DNA sequencing, where sequencing costs continue to drop by orders of magnitude (De Wit et al. 2015), it is now feasible to generate genomic datasets for natural populations of non-model organisms. Researchers can use a reverse-ecology approach where candidate genes behind ecologically relevant, but non-conspicuous, phenotypes are identified based on patterns of variation and signatures of selection in protein-coding sequences (Li et al. 2008). Here I examine local adaptation in isolated urban populations of white-footed mice, *Peromyscus leucopus*, in NYC. I scan *P. leucopus* transcriptomes and identify regions and genes with divergent and skewed allele frequencies indicative of positive selection. I incorporate demographic history directly into our null model using parameter estimates previously inferred with genome-wide SNP datasets. I then examine the statistical association between allele frequencies and environmental measures of urbanization.

Natural selection varies in strength across environmentally heterogeneous landscapes, and populations of a species exposed to environmental variation often become locally adapted to a suite of biotic and abiotic factors (Tiffin & Ross-Ibarra 2014). Local adaptation occurs when a population has higher fitness within its home environment compared to an environmentally heterogeneous location or when the fitness at home is higher than any other introduced population would have (Kawecki & Ebert 2004). The relative fitness differences are often
Driven by different genotypes in each population (Williams 1966). This scenario contrasts with early views on adaptation that assumed evolution occurred too slowly for direct observation (Reznick & Ghalambor 2001). However, evidence has accumulated that shows selection coefficients can be high in nature (Endler 1986), and examples of local adaptation have increasingly been identified. For example, island populations of Mus musculus (Berry, 1964), heavy metal tolerance in local populations of plants (Antonovics et al. 1971), body size variation in isolated population in Drosophila (Huey et al. 2000) and adaptation to new hosts in Prodoxus moths (Groman & Pellmyr 2000) are all documented cases of local adaptation.

Traditional approaches for identifying local adaptation involve reciprocal transplant or common garden experiments (Merila & Hendry 2014), but local adaptation also leaves a predictable pattern of genetic variation and differentiation along environmental gradients across the genome (Savolainen et al. 2013). Measuring changes in the site frequency spectrum (SFS), the distribution of allele frequencies across sites, from genomic data can be an efficient method of detecting past selection (Merila & Hendry 2014). Positive directional selection increases interspecific variation at selected loci compared to the genomic background (Beaumont 2005), decreases nucleotide diversity around the selected locus through genetic hitchhiking (Hermisson 2009), and skews the SFS towards excess low and high frequency variants (Nielsen 2005). Balancing selection leaves a generally opposite pattern with decreased intraspecific genetic diversity (Nielsen 2005), low genetic differentiation between sites (Foll & Gaggiotti 2008), and an excess of intermediate frequency alleles (Nielsen 2005). Negative, or purifying selection reduces genetic diversity and differentiation, and only low frequency variants increase in the SFS (Nielsen 2005).
Local adaptation has increasingly been shown to occur in many taxa (Stinchcombe & Hoekstra 2008; Bonin 2008; Linnen et al. 2009; Hohenlohe et al. 2010a; Turner et al. 2010; Ellison et al. 2011; De Wit & Palumbi 2013). Uncovering the genetic basis of local adaptation has provided insight into a variety of evolutionary processes including speciation, maintenance of genetic diversity, range expansion, and species response to changing environments (Savolainen et al. 2013; Tiffin & Ross-Ibarra 2014). Cities represent one of the fastest growing and most rapidly changing environments around the world. Urbanization leads to habitat loss and fragmentation, changes in resource availability, novel species interactions, altered community composition, and increased exposure to pollutants (McKinney 2002; Chace & Walsh 2004; Shochat et al. 2006; Sih et al. 2011). Each of these ecological consequences may exert strong selective pressure, and there is mounting evidence that rapid adaptation occurs in many urban organisms. Another cause of rapidly changing environments is global climate change, where increasing temperatures and altered precipitation patterns strongly influence the life history traits of many species (Franks & Hoffmann 2011). These two processes, urbanization and climate change, are not mutually exclusive, however. Few studies have convincingly provided evidence of local adaptation, as opposed to plastic phenotypic responses, to human induced climate change (Boutin & Lane 2014; Merilä & Hendry 2014; Franks et al. 2014). Understanding local adaptation in urban habitats may lead to general insights about local adaptation to future climate change threats, both of which represent cases of general rapid evolution in changing environments. What traits are most likely involved in local adaptation? How quickly do populations respond to selective pressures and adapt locally? What environmental variables have the largest impact on populations and drive local adaptation?
the same genes and alleles involved in local adaptation also involved in similarly changing environments, i.e. is there evidence of convergent local adaptation?

I examine species in novel urban ecosystem as case studies of local adaptation. White-footed mice are good candidates for local adaptation because they are widespread and are one of the few native mammals that thrive in extremely small, fragmented urban forests (Pergams & Lacy 2007; Rogic et al. 2013; Munshi-South & Nagy 2014). *P. leucopus* tend to be found at higher densities in urban patches due to a thick understory and fewer predators and competitors (Rytwinski & Fahrig 2007). Increased density may also be due to limited *P. leucopus* dispersal between urban sites. Munshi-South (2012) found barriers to dispersal between isolated NYC parks, with migrants only moving along significantly vegetated corridors throughout the city. There is also substantial genetic structure between NYC parks as measured by microsatellites (Munshi-South & Kharchenko 2010), genome-wide SNPs (Munshi-South et al. 2015) and demographic modeling (Harris et al. 2015). I have also previously found evidence of divergence and selection in urban populations of NYC white-footed mice (Harris et al. 2013), though I used much smaller datasets and less sophisticated approaches than presented here. Collectively, strong selective pressures from urbanization, lack of gene flow between NYC parks, genetic structure found between geographically close urban sites, and evidence of urbanization driving neutral allele frequency patterns in urban populations (Munshi-South et al. 2015) makes it likely that populations of urban white-footed mice are adapting to strong selective pressures in spite of the influence of genetic drift.

Urbanization and global climate change are relatively recent disturbances that rapidly change native ecosystems. Over short timescales, standing genetic variation, as opposed to novel mutations in organisms, often underlies adaptation (Barrett & Schluter 2008; Stapley et al.
As these pre-existing mutations spread to fixation they produce a detectable signal in the form of ‘hard’ or ‘soft’ selective sweeps (Herisson & Pennings 2005; Messer & Petrov 2013). Additionally, ecologically important traits involved in local adaptation are often quantitative traits with many genes of small effect involved in producing the desired phenotype (Orr 2005; Rockman 2012). In order to distinguish these more subtle signatures of selection, I use multiple tests that provide greater statistical power and higher resolution at identifying types and age of selection when used together (Grossman et al. 2010; Hohenlohe et al. 2011).

I use transcriptomes sequenced from urban and rural populations of *P. leucopus* to produce estimates of nucleotide diversity \( \pi \) (Tajima 1983), Tajima’s \( D \) (Tajima 1989), and \( F_{ST} \) (Wright 1951) and make inferences about the evolutionary processes at work in these populations. Several studies have used this suite of population genetic statistics to detect candidate genes that are the target of selection (Stajich & Hahn 2005; Hohenlohe et al. 2010a; Tennessen et al. 2010; Nadeau et al. 2012). Major challenges in using \( \pi \) or Tajima’s \( D \) to are distinguishing between types of selection, and then disentangling demographic processes from selection (Biswa & Akey 2006). The difficulty comes when neutral demographic processes, like population bottlenecks, produce signatures of variation in the genome similar to those produced by selection (Oleksy et al. 2010; Li et al. 2012). For example, a population bottleneck followed by an expansion will create genomic regions with low genetic diversity that resemble signatures from selection. Alleles present in the few breeding individuals during the bottleneck will become widespread during the expansion (Pavlidis et al. 2010). There has been much discussion on how to deal with the confounding effects of demographic history on identifying selection (Excoffier et al. 2009; Li et al. 2012; Vitti et al. 2013; Lotterhos & Whitlock 2015). The prevailing approach is to produce genome-wide data and assume selection
acts on one or a few loci while demographic processes act across the genome. Outlier tests for loci under selection generate a null distribution, usually based on an island model of population differentiation (Excoffier et al. 2009), and then identify candidate genes with genetic differentiation beyond the null model’s limits. The true demographic history of most organisms is much more complex, and computational approaches have been developed to robustly infer demographic parameters (Gutenkunst et al. 2009; Excoffier et al. 2013). The inferred demographic history can then be used to construct a more realistic null model, reducing the rate of false positives in outlier based tests of selection (Excoffier et al. 2009; Yoder et al. 2014).

I use the inferred demographic history of urban populations of P. leucopus (see Chapter 3) to simulate comparable SNP datasets to our observed sequence data. I then use two genome scan tests that identify outlier loci based on population differentiation and the SFS, respectively. Bayescan uses a Bayesian approach to identify SNPs that show extreme allele frequency divergence between populations (Foll & Gaggiotti 2008). SweeD is a likelihood based test that finds evidence of selective sweeps by looking for regions with a SFS that deviates from neutral expectations (Pavlidis et al. 2013). I also use an emerging approach for identifying loci underlying local adaptation by examining associations between allele frequencies and environmental variables. Several tests have been developed based on the relationship between genotypes and environmental variables, falling under the general category of genotype-environment association (GEA) tests (Joost et al. 2007; Coop et al. 2010; Frichot et al. 2013; Lotterhos & Whitlock 2015). GEA tests perform better than genome scan based outlier tests under complex demographic scenarios (Lotterhos & Whitlock 2015) but can suffer from a high rate of false positives. Analyses suggest that using genome scan-based outlier tests in conjunction with GEA tests leads to reliable outlier loci identification (De Villemereuil et al.
GEA tests also identify local adaptation in polygenic phenotypes where each polymorphism has a relatively weak effect (Frichot et al. 2013), because correlations between alleles and environmental variables do not rely on the strength of genetic differentiation or SFS skew between populations. I examine the association between quantitative metrics of urbanization (percent impervious surface and human population density) and polymorphisms between rural and urban populations.

In this study, I examine transcriptomes generated from RNAseq for 48 *Peromyscus leucopus* individuals from three urban sites in NYC and three rural sites from the surrounding area. Including population pairs that are near each other and genetically similar, but occur in different environments (urban versus rural), increases the power to identify candidate genes under selection (Lotterhos & Whitlock 2015). I use traditional population genetic summary statistics to generate per-site estimates and find loci with patterns of genetic variation that deviate from neutral expectations. Next, I use several tests of selection that use our transcriptome-wide SNP datasets to determine whether these deviations are due to recent selection in urban populations of white-footed mice. To increase power, reduce false positives, identify more subtle signals of selection from standing genetic variation, and find candidate genes involved in polygenic phenotypic traits, I simulate a null background model from the inferred demographic history for NYC populations of *P. leucopus*. I also identify the candidate genes experiencing selection from ecological pressures in urban habitats by directly associating alleles with environmental measurements of urbanization. I use overlapping results from multiple tests and environmental associations in order to generate a reliable list of candidate genes involved in the local adaptation of *P. leucopus* populations to the urban environment. This study is the first to use transcriptome-wide patterns of genetic variation for analyses of local adaptation in cities.
Evidence of local adaptation in urban populations reveals how urbanization acts as an evolutionary force, gives insights into important traits for local adaptation, and provides an example of the speed of evolution in rapidly changing environments.

Materials and Methods

Sampling, library preparation, and transcriptome assembly

*Peromyscus leucopus* is one of the most abundant small mammals in North America, preferring the typical oak-hickory forest commonly found in the eastern U. S. (Wang *et al.* 2008). They are generalists that burrow in a variety of habitats (Metzger 1971; Vessey & Vessey 2007), and feed on a wide-range of invertebrates, nuts, fruit, vegetation, and fungus (Ostfeld *et al.* 1996). There is also evidence that *Peromyscus* spp. can adapt to environmental change (Storz *et al.* 2007a, 2009, 2010; Mullen & Hoekstra 2008a; Linnen *et al.* 2009; Weber *et al.* 2013; Natarajan *et al.* 2013). I provide a brief overview of methods here, but see Chapters 1 and 2 for full details about individual sampling and transcriptome sequencing and annotation.

White-footed mice were sampled from 2009 - 2013. Eight individual white-footed mice (equal numbers of males and females) were randomly chosen from six sampling locations representative of urban and rural habitats (Fig. 1, see Chapters 1,2). Three sampling sites occurred within NYC parks: Central Park in Manhattan (CP), New York Botanical Gardens in the Bronx (NYBG), and Flushing Meadow—Willow Lake in Queens (FM). These sites represented urban habitats surrounded by high-volume roads and dense human infrastructure. The remaining three sites occurred ~100 km outside of NYC in rural, undisturbed habitat representative natural environments for *Peromyscus leucopus*. High Point State Park is in the Kittatinny Mountains in New Jersey (HIP), Clarence Fahnestock State Park is located in the
Hudson Highlands in New York (CFP), and Brookhaven and Wilde Wood State Parks and neighboring sites on the northeastern end of Long Island, New York (BHvwp). Mice were sacrificed and liver, gonad, and brain tissue were harvested in the field for immediate storage in RNAlater (Ambion). Total RNA was extracted and ribosomal RNA was removed during library preparation. The reverse transcribed cDNA was sequenced using the 454 GS FLX+ and SOLiD 5500 xl systems using standard RNAseq protocols. SNPs were called with the Genome Analysis Toolkit pipeline using a Bayesian genotype likelihood model (GATK version 2.8, DePristo et al. 2011).

Summary statistics

SNP information was stored in a VCF (variant call format) file and summary statistics were calculated using vcftools (Danecek et al. 2011). These analyses were used for general estimates of diversity for each population and were calculated for each site. I calculated per-site nucleotide diversity ($\pi$), Tajima’s $D$, and $F_{ST}$. I also calculated the statistics for each contig (per-site statistic summed across all SNPs per contig divided by total sites) and found the average estimate for each population, including all pairwise population comparisons for $F_{ST}$.

Sans for positive selection based on population differentiation

Population structure analyses for protein coding sequences show that the three urban sites and three rural sites comprise two distinct groups, but there was also hierarchical structure indicating that urban sites represent unique evolutionary clusters (Harris et al. 2015). I used the $F_{ST}$ based analysis implemented in Bayescan v. 2.1 (Foll & Gaggiotti 2008) to compare population-specific allele frequencies with global averages and identify outlier SNPs with
divergence patterns better explained by models that include positive selection parameters. Bayescan identifies markers that differ between groups more than would be expected under neutral genetic processes. Based on a set of neutral allele frequencies under a Dirichlet distribution, Bayescan uses a Bayesian model to estimate the probability that a given locus is under the effect of selection. To generate more realistic allele frequency distributions, I used Bayescan to analyze coalescent simulations of SNP datasets based on the neutral demographic history inferred specifically for *P. leucopus* populations in Chapter 3. I generated 100 sets of 100,000 SNPs each from a three population, isolation with migration model using the previously inferred parameter estimates for divergence time, effective population size, migration rate, and population size change in the coalescent based software program, fastsimcoal2 (Excoffier *et al.* 2013). In short, the model represented a deep split between an ancestral population into Long Island, NY and the mainland (including Manhattan) 29,440 generations before present (GBP). Migration was asymmetrical from the mainland into Long Island and an urban population later became isolated 746 GBP. Urban populations were also modeled to include a bottleneck event at the time of divergence. Finally, I allowed migration to occur between all three populations (Chapter 3). Bayescan was run independently on each simulated dataset using default parameters. I also ran Bayescan multiple times on the observed SNP data. I performed a global analysis, one Bayescan run where all individuals were partitioned into Urban and Rural groups, and finally analyses on all individual pairwise population comparisons. Outlier SNPs were retained if they had a false discovery rate (FDR) value ≤ 0.1 and if the calculated $F_{ST}$ and posterior odds probability were higher than for any value calculated from the simulated dataset.

**Analysis for selective sweeps**
I also scanned the transcriptome to look for contigs where the observed SFS showed an excess of low frequency and high frequency minor alleles, a signal indicative of a recent selective sweep in the region. The composite likelihood ratio (CLR) statistic is used to identify regions where the observed SFS matches the expected SFS generated from a selective sweep (Kim & Stephan 2002; Nielsen et al. 2005; Pavlidis et al. 2010). I calculated the CLR along sliding windows across the transcriptome using the software program SweeD (Pavlidis et al. 2013). SweeD is an extension of the popular Sweepfinder (Nielsen et al. 2005) and is optimized for large next generation sequencing (NGS) datasets. SweeD was run separately for each population and on individual contigs directly from vcf files using default parameters except for setting a sliding window size of 200 bp and using the folded SFS, as I lack an outgroup to infer the ancestral state. The window within each contig with the highest CLR score is the likely location of a selective sweep, and statistical significance was chosen based on neutral simulations. SweeD does not inherently identify outlier regions, but rather, the CLR statistic is computed using a selective sweep model on the observed dataset and needs to be compared to a neutral model calibrated with the background SFS generated from simulations. I used 100 datasets with 100,000 SNPs each, simulated under the inferred neutral demographic history for urban and rural populations of white-footed mice in NYC. The CLR was calculated using SweeD for all simulated datasets and the resulting distribution was used to set a significance cutoff. For the observed dataset, SweeD was run separately on each contig because I lacked a genome to provide clear linkage information. Outlier regions and their associated contigs were chosen as candidates if their CLR statistic was greater than any produced when calculated for neutral simulations and if they fell within the top 0.01% of the CLR distribution for the observed SNPs. Choosing outliers within the top 0.01% of the distribution is a conservative cutoff value.
When looking for regions with genetic patterns of a selective sweep, Wilches (2014) filtered regions within the top 5% of the distribution. Selective sweeps from artificial selection in rice, *Oryza glaberrima*, were identified with a cutoff value of 0.5% (Chen *et al.* 2014) and regions within the Gorilla genome were identified as significant if CLR scores were in the top 0.5% (McManus *et al.* 2014). I chose an even more stringent filter of 0.01% because I lacked a reference genome and analyses were restricted to relatively short individual contigs.

**Genotype-environment association tests for environmental selection**

In order to associate outlier SNPs and candidate loci identified above with potential environmental selection pressures, I used LFMM (Frichot *et al.* 2013), a software program that is one of the recently emerging genotype-environment association (GEA) approaches to identifying selection (Hedrick *et al.* 1976; Joost *et al.* 2007; Coop *et al.* 2010; Frichot *et al.* 2013; Lotterhos & Whitlock 2015). Latent Fixed Mixed Modeling (LFMM) tests for correlations between environmental and genetic variation while accounting for the neutral genetic background and structure between populations (Frichot *et al.* 2013). I tested three environmental variables associated with urbanization, the percent impervious surface within a two kilometer buffer around each sampling site, human density within a two kilometer buffer around each sampling site, and a simple designation of urban or rural for each site. The genotype file contained information for all individuals and included the outlier SNPs detected in Bayescan and SweeD. An important first step in using the LFMM algorithm is to define the number of latent factors, K, that can be used to define population structure in the genetic background. To identify the appropriate number of K latent factors in our dataset, I used default parameters and performed a PCA followed by a recommended Tracy-Widom test to find the number of eigenvalues with
significant p values \( \leq 0.01 \) (Patterson et al. 2006; Frichot & Fran 2014). Results suggested the use of six latent factors. Thus, I ran LFMM with default parameters except for a \( K = 6 \), an increased number of MCMC cycles = 100,000, and a burn-in = 50,000. Using author recommendations, I combined 10 replicate runs and readjusted the p values to increase the power of the test. LFMM uses \(|z|\) scores to report the probability of a SNP’s association with an environmental variable. After correcting for multiple testing, I used a cutoff value of \( q \leq 0.1 \).

**Functional annotation of candidate gene**

The contigs containing outlier SNPs identified using the tests for selection above were obtained from the *P. leucopus* transcriptome. The gene annotation pipeline implemented in Blast2GO (Conesa et al. 2005; Götz et al. 2008) was used to find homologous sequences from the NCBI non-redundant protein database using BLASTX, and associated gene ontology (GO) terms were retrieved. Gene ontology (GO) terms are a standardized method of ascribing functions to genes. Blast2GO retrieves GO terms associated with BLASTX hits and also uses the KEGG database to describe biochemical pathways linking different enzymes (Ogata et al. 1999; Kanehisa et al. 2014).

**Results**

**Genetic diversity statistics**

There were 154,770 total SNPs retained for use in looking at patterns of genetic variation and performing tests of selection. For each population I obtained estimates of nucleotide diversity, Tajima’s \( D \), and pairwise \( F_{ST} \). There were differences in genetic diversity between urban and rural populations greater than one standard deviation. Urban populations had a two-
fold decrease in nucleotide diversity compared to the rural populations (Table 1). For instance, the average nucleotide diversity for all three rural populations was 0.224 ± 0.034, while the average for urban populations was only 0.112 ± 0.019. The average Tajima’s $D$ calculation within populations did not show substantial differences between populations (Table 1). For all populations, Tajima’s $D$ was slightly positive, with rural populations only slightly more positive than urban populations, though not significantly different. Given the large standard deviations for average population calculations, however, there is a wide range of Tajima’s $D$ values for individual loci but not a clear signal of directional selection for the entire dataset. Average pairwise $F_{ST}$ calculated using vcftools ranged from a low of 0.018 ± 0.364 between two rural populations (CFP_HIP) to a high of 0.110 ± 0.520 between two urban populations (CP_FM, Table 2). $F_{ST}$ calculations between populations using protein coding sequences were very similar to calculations made for neutral genome-wide SNP datasets from the same $P. leucopus$ populations (Munshi-South et al. 2015), and they supported findings that these populations lack an isolation-by-distance pattern. Comparisons between rural populations had the lowest $F_{ST}$ values, urban to rural populations had the second lowest, and urban to urban population comparisons had the highest overall $F_{ST}$ values despite being less than 5 km apart (Table 2).

**Outlier detection**

The test for positive or balancing selection implemented in Bayescan for the global analysis revealed 309 (0.19%) SNPs potentially under the influence of divergent selection. To investigate divergent selection due to urbanization, sampling sites were grouped and classified as urban or rural, and genome scans using Bayescan on this dataset uncovered 40 (0.025%) SNPs with signatures of positive selection (Fig. 2A, Table 3). Eight of these SNPs were found in the
global analysis. Individual urban to rural population comparisons did not find any outlier SNPs, and zero SNPs were revealed to be under balancing selection. $F_{ST}$ for outlier SNPs ranged from 0.21 - 0.33, much higher than the population average. When Bayescan was run on the simulated neutral dataset, which included bottlenecks during urban population divergence, there were zero identified outlier SNPs. I did, however, only include outlier SNPs from the observed dataset with FDR and posterior odds values that were smaller and larger, respectively, than the most extreme values for the simulated data ($FDR \leq 0.6$ and $\log_{10}(PO) \geq -0.196$).

Outlier regions showing signatures of selective sweeps from the SweeD analysis were identified using comparisons to neutral expectations. To generate the null distribution of the CLR statistic I tested the 100 SNP datasets simulated under the inferred demographic history for NYC populations of *P. leucopus*. I found that CLR scores in the top 5% of the distribution were generally 2x - 3x lower than for the top 5% of the observed dataset. I ran SweeD runs on observed SNPs within individual contigs and identified outliers by filtering for a CLR score $\geq 3.53$ (the maximum CLR from simulated data). I also chose regions that fell within the top 0.01% of the observed distribution (Fig. 2B). SweeD identified regions with SFS patterns that fit a selective sweep model in 55 contigs (40,908 contigs in *P. leucopus* transcriptome, 0.13%) within urban populations (Table 4). Contig 35790-44, which codes for the lipid transporter *Apolipoprotein B100*, had the highest CLR score, CLR = 8.56, and all outliers had CLR scores $\geq 4.97$. There was no overlap of outliers between Bayescan and SweeD.

**Environmental associations**

I used LFMM to examine statistical associations of outlier SNPs with environmental measures of urbanization. Thirty of 40 outliers identified from Bayescan could be associated
with at least one of the three environmental variables tested, which clearly delineate urban and rural sampling locations (Fig. 3A, Table 3). All 30 of the identified SNPs were associated with whether a site was classified as urban or rural. Only seven of the outlier SNPs were associated with percent impervious surface surrounding the sampling site and five were associated with human density. Twenty-six of the 55 outlier contigs in urban populations containing selective sweep regions as identified in SweeD could be associated with one of the environmental variables (Table 4). Again, all 26 significant associations involved classification of a site as either urban or rural. Fourteen outliers from SweeD were associated with percent impervious surface and eight were associated with human density surrounding the sampling location. Some contigs containing outlier SNPs associated with environmental variables were unique to individual urban populations, possibly indicating local adaptation within parks or selection on a polygenic trait where mutations in different genes can lead to the same phenotypic effect.

**Functional annotation**

The full contig sequences containing the outlier SNPs were obtained from the *P. leucopus* transcriptome (See Chapter 2) and used to identify functional annotations. Of the 40 contigs identified by Bayescan as divergent between urban and rural populations, 36 could be annotated with gene names and functional information (Table 3). Of these, 29 were also associated with the urban environmental variables. For the Bayescan outlier sequences, the ten most frequent gene ontology terms attributed to the DNA sequences involved organismal metabolism (Appendix 4.1). Some outliers occurred within well-studied genes with known functions and biochemical pathways. These included a farnesoid-x-receptor (FXR, Contig 25795-154) gene, the protein ABCC8 (Contig 26183-148), a Hermansky-Pudlak syndrome gene (Hps1, Contig
36706-36), KDM8, a histone demethylase (Contig 7750-426), a myosin light chain kinase (MYLK, Contig 7975-4180), and the gene SORBS2 (Contig 37967-26). These genes were identified as likely experiencing divergent selection between urban and rural populations and showed environmental associations with urbanization.

The SweeD analysis found regions within 55 contigs that showed a signature of a selective sweep (Table 4). Forty-nine could be annotated with gene names and gene ontology terms, and 25 were also associated with urbanization. Overall, sequences were associated with metabolic processes, similar to the outliers found in Bayescan, and many genes were involved with basic metabolic functions such as glycolysis and ATP production (Appendix 4.1). A few contigs were annotated with well-studied genes and clearly understood functions. Contig 35790-44 was annotated as the gene APOB, an apolipoprotein, and Contig 10636-348 was an aflatoxin reductase gene AKR7A1. There was also the gene FADS1, part of the fatty acid denaturase family (Contig 342-1776), a heat-shock protein (Hsp90, Contig 3964-627), and a hepatocyte growth factor activator gene (Contig 8960-388). Most gene annotations did not have known phenotypic traits related to their function, but KEGG analysis revealed several contigs involved in the same biochemical pathways: galactose metabolism, fructose metabolism, and mannose metabolism (Appendix 4.2).

Discussion

The results of this study provide insight into the genetic basis of local adaptation, which is key for understanding the ecological and evolutionary processes that affect biodiversity and how organisms respond to changing environments. I hypothesized that populations of *P. leucopus* in urban habitat fragments within NYC adapt in response to selective pressures from
urbanization. Previous work supports this claim. Clear evidence from neutral non-coding (see Chapter 3) and protein coding datasets (see Chapter 2) of population structure between urban and rural sampling sites suggests NYC populations of white-footed mice are genetically isolated. Urbanization also impacts genetic diversity across the genome (Munshi-South et al. 2012, Harris et al. 2015, Chapter 3). *P. leucopus* populations along an urban-to-rural gradient in NYC had reduced nucleotide diversity and heterozygosity in urban populations (Munshi-South et al. 2015). I also found lower overall nucleotide diversity in urban populations, but this does not preclude the action of natural selection as an evolutionary process. Despite reduced genetic diversity, strong selection can drive local adaptation in bottlenecked populations (Oliver & Piertney 2012). Genome scan results here suggest this might be the case for *P. leucopus* in NYC. Additionally, demographic inference indicates that NYC populations became isolated within the timeframe of urban settlement (see Chapter 3).

I have also found evidence for older occurrences of divergent selection in NYC white-footed mice by investigating non-synonymous polymorphisms between pooled transcriptome samples (Harris et al. 2013). The pooled transcriptome dataset (Chapter 1) compared three urban sites to one rural site and found 11 candidate genes. Two of the eleven were direct matches to outliers in these current analyses (Serine protease inhibitor a3c and Solute carrier organic anion transporter 1A5). Three other genes are from the same gene families or involved in the same biological processes as those described here. One gene was an aldo-keto reductase protein, part of the same gene family as our SweeD identified aflatoxin reductase gene (Contig 10636-348). The aldo-keto reductase gene family comprises a large group essential for metabolizing various natural and foreign substances (Hyndman et al. 2003). Two others, camello-like 1 and a cytochrome P450 (CYP1A1) gene, are involved in metabolism of drugs and lipids. In
Peromyscus spp., CYPA1A is directly expressed along with Hsp90 (outlier from current SweeD analysis) when exposed to environmental toxins (Settachan 2001). Here, I increased the numbers of individuals and sites studied compared to Harris et al. 2013 and are finding similar gene annotations in candidate genes. Collectively, these findings suggest that urban populations of *P. leucopus* may be adapting in response to selective pressures from urbanization.

In this study, I indeed observe patterns of divergent positive selection between urban and rural populations of *P. leucopus*. Additionally, I were able to associate outlier SNPs with environmental variables representative of urbanization, and annotate the bulk of outliers with full gene annotations. The majority of candidate genes deal with organismal metabolism, particularly diet-related breakdown of lipids and carbohydrates. I discuss what these finding mean for organisms as they are exposed to novel urban ecosystems, and for understanding the ecological processes and time frame of recent local adaptation in general.

**The utility of using genome scan methods to test for selection**

Over the past decade, genome scan methods have become a feasible and common way for investigating polymorphisms across the genome in order to detect and disentangle neutral (demographic) and adaptive (selection) evolutionary processes (De Villemereuil et al. 2014). One of the most popular approaches looks at locus specific allele frequency differentiation between sampling locations as measured by $F_{ST}$ (Lewontin & Krakauer 1973; Weir & Cockerham 1984). Sites with extremely high allele frequency differences may be subjects of positive directional selection. Bayescan (Foll & Gaggiotti 2008) builds on this idea and identifies outliers using a Bayesian approach. It calculates the posterior probability of a site being under the influence of selection by testing two models, one that includes selection and one
that does not. The model that does not invoke selection is based on a theorized neutral
distribution of allele frequencies.

While Bayescan has been shown to be the most robust differentiation method with
respect to confounding demographic processes (Pérez-Figueroa et al. 2010; De Villemereuil et
al. 2014), population bottlenecks, hierarchical structure, recent migration, or variable times to
most-recent-common-ancestor (MRCA) between populations can artificially inflate $F_{ST}$ values
(Hermisson 2009; Lotterhos & Whitlock 2014). Population structure and specific demographic
history must be built into the null distribution of $F_{ST}$ in order to avoid false positives. I dealt with
the issue of type I errors by running Bayescan on simulated SNP datasets generated under the
neutral inferred demographic history for urban populations of $P. leucopus$ in NYC. I only chose
outliers in our observed dataset if their posterior probability was greater than any found from
simulations. The outliers captured when comparing urban to rural sites made up 0.025% of the
total number of loci analyzed from the transcriptome. This number is in line with candidates
uncovered from a similar study (0.05%) that looked at high and low altitude populations of the
plant $S. chrysanthemifolius$. Many studies find higher percentages of outlier loci using
Bayescan, 4.5% in the American pika across its range in British Colombia (Henry & Russello
2013), and 5.7% in Atlantic herring across their range (Limborg et al. 2012). Our lower overall
percentage of outliers may be because I included the known demographic history in our tests,
because of the relatively recent isolation of urban populations of $P. leucopus$, or due to the fact
that I did not have complete transcriptome sequences for our populations.

A second genome scan approach looks at patterns in the SFS within a population as
opposed to allele differentiation between populations. The statistics developed around the SFS
are used to look at genetic hitchhiking around a selected locus that produces a pattern
characteristic of a selective sweep (Schlötterer 2003; Pavlidis *et al.* 2008). The main footprint that selective sweeps leave on the SFS is an excess of rare low frequency and high frequency variants (Nielsen 2005). The SweepFinder method (Nielsen *et al.* 2005), recently upgraded to the NGS compatible version SweeD (Pavlidis *et al.* 2013), uses a composite likelihood ratio test based on the ratio between the likelihood of a null (neutral evolution model) and the alternative (selective sweep) hypothesis. Like differentiation based methods, the weakness of hitchhiking methods is the confounding effect certain demographic processes have on the SFS. A strong population bottleneck can lead to variances in the genealogical history so that some loci have decreased genetic diversity and an excess of low frequency variants (Hermisson 2009). Again, however, building the known demographic history into the null model readily reduces false positive rates (Pavlidis *et al.* 2013).

I included the *P. leucopus* demographic history into our analysis, and found 0.04% of the transcriptome to contain regions with SFS patterns indicative of selective sweeps. This rate is in line with other studies that found 0.5% of regions in domesticated rice to show evidence of selective sweeps, though this might be unusually high due to artificial selection (Wang *et al.* 2014), 0.02% of loci in black cottonwood experiencing selective sweeps across geographic regions (Zhou *et al.* 2014), and 0.02% of regions across the entire Gorilla genome to show hitchhiking patterns (McManus *et al.* 2014).

Individual genome scan approaches look at different aspects of genomic structure and by themselves can miss true outliers, type II errors, or identify false positive, type I errors. Several studies have shown that a general principle to follow in order to avoid these errors is to perform multiple tests looking at various aspects of the genome (Nielsen 2005; Grossman *et al.* 2010; Hohenlohe *et al.* 2010b). I used Bayescan and SweeD to identify outliers experiencing positive
selection, but did not find any overlapping candidate genes between them. This finding is not necessarily unexpected as the two tests look at different selection scenarios, divergent local selection versus population-wide positive selection in the form of selective sweeps (Hermisson 2009). $F_{ST}$-based methods can pick up on divergence between alleles relatively quickly, while models for selective sweeps typically require nearly-fixed derived alleles (Hohenlohe et al. 2010b). Given the recent time frame of urbanization in NYC, not enough generations may have passed since NYC white-footed mice have become isolated to find complete selective sweeps in loci that overlap with outliers from Bayescan. In the case of NYC populations of $P. leucopus$, it is likely that adaptation is occurring from standing genetic variation in the form of soft sweeps (Hermisson & Pennings 2005), which are not readily identified by programs like SweeD (De Villemereuil et al. 2014). To give further support to this idea, I found several outliers across the various tests I ran that are unique to specific urban populations, which is characteristic of soft sweeps, as they and polygenic traits can lead to outlier SNPs unique to populations (Messer & Petrov 2013). Despite the lack of overlapping outlier SNPs between the two tests, further evidence that positive selection is acting in urban populations of $P. leucopus$ was found with an additional approach. Independent confirmation of candidate genes came from correlating genotypes and environmental variables, a method that may be more powerful than the genome scans above for identifying SNPs under selection (Savolainen et al. 2013).

**Environmental associations strengthen evidence of local adaptation to urbanization**

Genotype-environment association tests are a growing class of methods that provide fine scale detail about the ecological processes driving selection by identifying loci with allele frequencies that are correlated with environmental factors. Several have recently been developed
(Joost et al. 2007; Coop et al. 2010; Frichot et al. 2013), and here I used LFMM (Frichot et al. 2013) to associate outlier SNPs with environmental measurements that capture the effects of urbanization. LFMM is uniquely suited for our dataset as it has been found to perform better than other methods in the presence of hierarchical structure and when polygenic selection is acting on many loci with small effect (De Villemereuil et al. 2014). In our dataset, there are many layers of structure including urban and rural differentiation (Harris et al. 2015, Chapter 2), patterns of geographic structure between mainland mice and Long Island, NY (Chapter 3), and population structure between individual urban parks (Munshi-South & Kharchenko 2010). It also has more power when the sampling size is less than 10 individuals per populations, there is no evidence of IBD, and sampling design of the experiment involves pairs in environmentally heterogeneous habitats (Lotterhos & Whitlock 2015). I sampled eight white-footed mice per population, find no evidence of IBD (Munshi-South et al. 2015), and sampled environmentally heterogeneous rural and urban locations.

Using LFMM, I found that 75 % and 47 % of outliers from Bayescan and SweeD, respectively, could also be associated with one or more environmental variables. These results complement our findings that positive selection is acting on urban populations of white-footed mice. I acknowledge that impervious surface, human density, or classification as urban may be correlated with a different environmental selection force, but our results ultimately support an evolutionary scenario where isolated urban populations are experiencing divergent positive selection that is strongly affected by one or more environmental variables associated with urbanization. These results are also consistent with other studies combining genome scan methods and GEA tests. Limborg et al. (2012) found 62.5 % of the outliers identified in Bayescan to be correlated with temperature or salinity changes in Atlantic herring, and 26.3 % of
genome scan outliers could be associated with temperature or latitude in the tree species, *A. glutinosa* (De Kort *et al.* 2014).

The percent impervious surface and human density around a park, or the classification of sites as urban or rural, are efficient metrics for determining whether a sampling location has been affected by urbanization (Munshi-South *et al.* 2015). Once I can define urban sites, there are several predications I can make about how ecological processes are changing within them. One of the most obvious consequences of human altered environments is habitat loss and fragmentation (McKinney 2002; Sih *et al.* 2011). The act of fragmentation and the building of infrastructure invariably changes the net primary productivity due to increasing percentages of impervious surface or artificial landscapes, parks and yards (Shochat *et al.* 2006). Additionally, species interactions change as organisms are forced into smaller areas or separated by infrastructure (Shochat *et al.* 2006). This includes impediments to migration across the urbanized landscape. Humans often introduce invasive species into habitats (Sih *et al.* 2011) leading to increased competition or novel predator-prey interactions. Urbanization also changes the types and availability of resources available in the altered habitat (McKinney 2002; Sih *et al.* 2011). Pollution is also a major consequence of urbanization (Donihue & Lambert 2014), and can include chemical, noise, or light pollution (Sih *et al.* 2011).

Given the rapid alteration of environments during urbanization, behavioral flexibility and phenotypic plasticity are thought to play an important role in a species’ response to novel urban ecosystems (Sih *et al.* 2011). Climate change, another form of human-induced rapid environmental change, is often used as a model for understanding plastic and evolutionary responses in organisms. Franks *et al.* (2014), in a comprehensive review of phenotypic changes in plants in response to climate change, reported that the majority of studies showed evidence of
plastic responses. They also found many studies showed evidence of adaptation, though not always conclusively. Looking at animal responses to climate, Boutin & Lane (2014) found similar findings but even less conclusive evidence of adaptation versus plasticity, maybe due to the motility of animals and difficulty in establishing common garden or reciprocal transplant experiments. While it is likely that *P. leucopus* in NYC are displaying some plastic phenotypic responses in urban ecosystems, our results provide evidence of heritable evolutionary responses as well.

Between divergent allele frequencies, a skewed SFS, and environmental associations, I find several overlapping lines of evidence that support rapid divergent positive selection in white-footed mice. Urban ecologists are increasingly finding evidence of selection acting in urban environments (Donihue & Lambert 2014), and our results are in line with other studies that have found rapid local adaptation to ecological pressures from urbanization. Yeh (2004) found sexually selected tail coloration in juncos was rapidly evolving in urban populations compared to rural ones. European blackbirds show reduced migratory behavior in cities, and there is also evidence of selection on genes underlying anxiety behavior across multiple urban areas (Partecke et al. 2006; Mueller et al. 2013b). Cheptou et al. (2008) found weeds in urban vegetation plots surrounded by paved surfaces had a higher percentage of non-dispersing seeds and that this trait was genetically based. In marine species living in the polluted waters around urban areas, rapid adaptation for PCB resistance occurred in both killifish and tomcod (Whitehead et al. 2010; Wirgin et al. 2011). The realization that a diverse range of taxa may adapt to human induced landscape change suggests rapid adaptation to anthropogenic driven environmental change may be pervasive in nature.
Functional roles and ecological relevance of candidate genes

The model rodent species *Mus musculus*, *Rattus norvegicus*, and *Cricetulus griseus*, all have deeply sequenced, assembled and annotated reference genomes. These resources allowed us to annotate 89.5% of contigs containing outlier SNPs and genomic regions with high quality gene information. These annotations provided us with information about the traits affected by candidate genes. Urban *P. leucopus* specifically exhibited genetic patterns that suggest positive selection in genes from the mitochondria, a potentially significant finding considering mitochondrial genes are often used for demographic inference (Munshi-South & Nagy 2014). Tests for selection also identified genes that protect cellular health in stressful environments, modulate melanism throughout the body, genes that are involved in epigenetic control of gene expression, or involved in digestion and metabolism of lipids and carbohydrates.

Gene ontology vocabulary assigns gene function according to biological process, molecular function, and cellular component. Across all candidate genes and gene ontology terms, involvement with mitochondria was one of the most common assignments (Appendix 4.1). Whether genes were involved in energy production through metabolism of food or were actual mitochondrial proteins, it appears evolution in mitochondria and metabolic processes is extremely important for *P. leucopus* living in urban parks. Mitochondrial genes were traditionally used as neutrally evolving markers, but researchers are finding evidence of selection on mitochondrial DNA across taxa (Oliveira et al. 2008; Balloux 2010). One example includes mitochondrial haplotypes associated with more efficient non-shivering thermogenesis and higher fitness in over-wintering shrews (Fontanillas et al. 2005). In *Peromyscus leucopus*, Pergams & Lacy (2007) found complete mitochondrial haplotype replacement in present-day white-footed mice living in the urban Chicago environment compared to haplotypes found in museum skins.
collected from before urbanization. The agent of selection is not clear, but independent research found evidence of negative selection acting on the mitochondrial D-loop gene in NYC P. leucopus (Munshi-South & Nagy 2014). These findings are not surprising. Many mitochondria-related metabolic functions are affected by the same environmental variables that change in response to urbanization like temperature (Urban = heat island effect) (Balloux 2010), population density (Urban = barriers to dispersal around parks) (Lankau & Strauss 2011; Munshi-South 2012), or resource availability (Urban = increased non-native prey) (Burcelin et al. 2002). In novel urban ecosystems, P. leucopus may be experiencing different energy requirements than rural counterparts.

One example of uniquely urban energy requirements comes from the signature of a selective sweep and a strong correlation with urban site classification found in the heat-shock protein Hsp90. Heat shock proteins are a gene family that have repeatedly been found to play a pivotal role in adaptation to environmental stress (Limborg et al. 2012). In a landmark study, cryptic variation in Hsp90 specifically, was found to act as a capacitor for the loss of eyes in cavefish (Rohner et al. 2013). Essentially, under normal environmental conditions, Hsp90 masks phenotypic variation in eye size, but under high stress conditions, Hsp90 is effectively inhibited allowing for eye size variation and eventual selection for unmasked phenotypic traits. In Peromyscus spp., Hsp90 acts a chaperone for many proteins, including a suite of metabolizing receptors activated by dioxin-like industrial toxins often found in polluted soil samples (Settachan 2001). When P. maniculatus was exposed to soils inundated with the toxin, 2,3,7,8 TCDD, maintenance of their circadian rhythm was affected and mice became active 3 hours earlier than under normal conditions (Settachan 2001). The aldo-keto reductase gene, aflatoxin aldehyde reductase (AKR7), was also an outlier in our analyses and is also important for
metabolizing environmental toxins (Hyndman et al. 2003). Aflatoxin is a natural carcinogen often found in cereals and nuts contaminated with the fungus, A. flavus and is metabolically activated by cytochrome P450 (an outlier from chapter 1) (Jin & Penning 2007). In experiments on Rattus norvegicus, researchers found AKR7 is upregulated in the liver when exposed to various classes of toxins and quickly acts to metabolize them, protecting cellular health (Ellis et al. 2003). I found P. leucopus caught in NYC had more enlarged, scarred, and fatty livers than those from rural populations (Harris, personal observation), and this may be directly related to ecological conditions in urban environments that promote environmental toxin accumulation. Due to proximity to human infrastructure, urban soils consistently show increased levels of heavy metal contamination (McDonnell et al. 1997). Urban ecosystems also experience the heat island effect with higher temperatures than rural locations (McDonnell et al. 1997), leaf litter that quickly decomposes but is of poor quality (Pouyat et al. 1997), and NYC in particular experiences high humidity in warmer months (National Oceanic and Atmospheric Administration, NOAA). The combination of constantly decaying vegetation, high temperatures, and high humidity is ideal for healthy communities of the fungus A. flavus, the primary producer of aflatoxins. Hsp90, AKR7, and cytochrome P450 (outlier from chapter 1) may be under selective pressures in NYC to efficiently metabolize higher concentrations of toxins in P. leucopus exposed to polluted urban soils or food sources in NYC.

Energy requirements may also be different in in urban populations because of dietary shifts. I found a surprising number of candidate genes with functions related to the metabolism and transport of lipids and carbohydrates. These genes were strongly correlated with environmental measures of urbanization, with clearly divergent allele frequencies between urban and rural sites (Fig. 3B). APOB-100 is the primary apolipoprotein that binds and transports
lipids, including both forms of cholesterol (HDL and LDL), and *Mus musculus* knock-out models result in hyperglycemia and obesity (Lloyd *et al.* 2008). FADS1, a farnesoid-x-receptor, is a nuclear receptor antagonist that is involved in bile synthesis and modulates high fat diets, with variation in expression affecting rates of obesity in mice (Li *et al.* 2013). Manually curated protein annotations show MYLK and SORBS2 are both directly involved in the gastrointestinal system, involved in smooth muscle contractions and absorption of water and sodium in the intestine, respectively (Magrane & Consortium 2011; Consortium 2014). ABCC8 is an ATP-binding cassette transporter, and knock-out mice models lack insulin secretion in response to glucose (Seghers *et al.* 2000). Finally, KEGG analysis found that two contigs (10636-348 and 27546-129) represent proteins that are both directly involved in Galactose, Fructose and Mannose metabolism (Ogata *et al.* 1999).

These candidate genes suggest that white-footed mice in isolated urban parks are responding to resource differences between urban and rural habitats. One prediction would be that urban *P. leucopus* consume a diet with higher overall fat content. The typical diet of *P. leucopus* across its range consists of arthropods, fruits, nuts, various green vegetation, and fungus (Wolff *et al.* 1985). They are especially reliant on oak mast cycles and an important predator of gypsy moths (Ostfeld *et al.* 1996). They are generalists and opportunistic in the food they eat, and thus many different food resources could drive diet differences in urban versus rural systems. Urbanization in NYC has lead to relatively small green patches that are surrounded by a dense urban matrix. The high percent of impervious surface is detrimental to the persistence of white-tailed deer in urban parks, leading to their exclusion throughout the majority of NYC. An overabundance of deer, like what occurs in our rural sampling sites, leads to the removal of the vegetative understory and inhibits regeneration of many plants (Stewart 2001). In these heavily
browsed habitats lacking a thick vegetative understory, there is direct correlation with length of
deer browsing in the area and invertebrate species diversity and abundance (Stewart 2001;
Allombert et al. 2005). As the understory is cleared by deer there are fewer food resources and
habitats for woodland invertebrates.

This is not the case for urban parks that often have extremely thick and healthy
understories (Leston & Rodewald 2006). Although the understory of urban forest fragments is
typically composed of invasive plants, such an understory can produce a number of novel seed
and fruit resources (McKinney 2008), as well as support a high abundance, if not diversity, of
invertebrate prey (McDonnell et al. 1997). *P. leucopus* in NYC are likely so successful in urban
ecosystems because they take advantage of the new food sources in urban habitats, including
seeds and other plant parts from an invasive understory layer, as well as invertebrates that may
thrive in urban fragments. There has been much research on adaptation to diet specialization,
especially in human populations. One well known case involves mutations in the human lactase
gene that lead to lactase persistence, most likely in response to a cattle domestication event
(Enattah et al. 2008). Another study that looked at more subtle shifts in allele frequencies across
human populations found outlier SNPs within genes that more efficiently metabolize proteins
found in the root and tuber based diets that humans switched to as they moved into polar
coregions (Hancock et al. 2010). There is also growing evidence of adaptation in native
predators in order to consume exotic or toxic prey species (Carlsson et al. 2009), for example,
larger mouthparts in the Australian soapberry bug to increase foraging on invasive balloon vines
(Carroll et al. 2005).

I hypothesize that urban *P. leucopus* have much higher fat content in their diets due to
increased seed or invertebrate abundance, and local adaption is occurring to more efficiently
metabolize the increased lipids and carbohydrates. There is strong genetic evidence that divergent positive selection is occurring between urban and rural mice, but in order to confirm hypotheses, it would be worth performing common garden experiments to measure metabolic rates when mice from different habitats are fed a consistent diet, or sequencing these same candidate genes across a broader range of urban and rural sites to look for similar signatures of selection. Next steps could also attempt to associate outlier SNPs with more fine scale ecological measurements like temperature, environmental pollutant level, or vegetative understory cover. Diet analyses between sites can also be undertaken and with the use of a metabarcoding approach using next generation sequencing, the entire diet can easily be identified from *P. leucopus* waste (Pompanon *et al.* 2012; Soininen *et al.* 2013).

**Conclusion**

Results strongly suggest that populations of *Peromyscus leucopus* within urban parks in NYC are adapting to the effects of urbanization. Focusing on protein-coding regions of the genome, using multiple tests of selection that analyze different parts of genomic structure, and associating outliers with environmental variables that capture the ecological changes imposed by urbanization allowed us to narrow in on specific genes underlying recent adaptation in urban habitats. In line with the definition of an ‘urban adapter’ (McKinney 2002), the generalist *P. leucopus* is successful in urban parks, and our results suggest this may be because they can adapt to changing dietary resources in urban ecosystems and potentially metabolize increased chemical pollutants in their environment. While I find definitive evidence of genetic variation between urban and rural sampling sites, further work needs to be done to look at specific polymorphisms and their impact on translation and protein folding.
Next steps should include SNP assays or full sequencing of outlier genes in more individuals from an increased number of sites across the urban - rural gradient. With this further confirmation, ecological based studies of diet can be pursued. Humans are increasingly altering the natural landscape through urbanization and indirectly through global climate change. Despite this, there are few studies with clear evidence of adaptation in novel urban ecosystems. Our study begins to address this issue using the statistical power of genomic datasets and finds that rapid adaptation is possible in recently disturbed ecosystems. By providing further understanding of contemporary evolution in response to urbanization, I have begun to answer important questions about the traits involved in adaptation to human modified landscapes and what environmental variables most likely drive this adaptation. Hopefully, these insights can be used for urban ecosystem management as global biodiversity continues to deal with unprecedented environmental change in the new era of the Anthropocene.
Table 1. Summary statistic averages for six *P. leucopus* populations

<table>
<thead>
<tr>
<th>Population</th>
<th>Nucleotide Diversity, $\pi$ (mean ± SD)</th>
<th>Tajima’s $D$ (mean ± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP</td>
<td>0.131 ±0.173</td>
<td>0.318 ±0.522</td>
</tr>
<tr>
<td>FM</td>
<td>0.112 ±0.166</td>
<td>0.301 ±0.522</td>
</tr>
<tr>
<td>NYBG</td>
<td>0.094 ±0.153</td>
<td>0.280 ±0.500</td>
</tr>
<tr>
<td>BHwwp</td>
<td>0.198 ±0.186</td>
<td>0.350 ±0.549</td>
</tr>
<tr>
<td>CFP</td>
<td>0.211 ±0.184</td>
<td>0.336 ±0.543</td>
</tr>
<tr>
<td>HIP</td>
<td>0.263 ±0.182</td>
<td>0.349 ±0.569</td>
</tr>
</tbody>
</table>
Table 2. Average pairwise $F_{ST}$ among six *P. leucopus* populations

<table>
<thead>
<tr>
<th></th>
<th>BHwwp</th>
<th>HIP</th>
<th>CFP</th>
<th>CP</th>
<th>NYBG</th>
<th>FM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHwwp</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIP</td>
<td>0.042 ±0.376</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFP</td>
<td>0.034 ±0.400</td>
<td>0.018 ±0.364</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP</td>
<td>0.089 ±0.458</td>
<td>0.060 ±0.417</td>
<td>0.063 ±0.447</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYBG</td>
<td>0.070 ±0.477</td>
<td>0.054 ±0.428</td>
<td>0.043 ±0.462</td>
<td>0.092 ±0.536</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>FM</td>
<td>0.056 ±0.468</td>
<td>0.061 ±0.420</td>
<td>0.057 ±0.456</td>
<td>0.109 ±0.520</td>
<td>0.085 ±0.535</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3. Results for selection from Bayescan and associations with environmental variables across urban (CP, FM, NYBG) and rural (BHwwp, CFP, HIP) populations. I = percent impervious surface, D = human density, C = Urban or Rural Classification

<table>
<thead>
<tr>
<th>Urban to Rural LFMM results</th>
<th>Outliers</th>
<th>Gene</th>
<th>( I )</th>
<th>( D )</th>
<th>( C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>27691-127</td>
<td>retroviral nucleocapsid protein gag containing protein</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>25795-154</td>
<td>af478441_1farnesoid-x-receptor alpha splice variant 1</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>37015-34</td>
<td>tubulin folding cofactor e-like isoform x6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>902-1236</td>
<td>alkyldihydroxyacetonephosphate peroxisomal</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3135-709</td>
<td>transmembrane 9 superfamily member 1 isoform 2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>27707-127</td>
<td>autophagy-related protein 2 homolog a isoform x2</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>38397-23</td>
<td>--</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3567-665</td>
<td>gram domain-containing protein 3</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>2482-790</td>
<td>protein diaphanous homolog 1 isoform x1</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>37967-26</td>
<td>sorbin and sh3 domain-containing protein 2 isoform x3</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>17974-242</td>
<td>40s ribosomal protein s15a-like protein</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>36437-38</td>
<td>jnk sapk-inhibitory isoform cra_a</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>7975-418</td>
<td>myosin light chain smooth muscle</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>12107-321</td>
<td>--</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>5754-511</td>
<td>otu domain-containing protein 3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>35973-42</td>
<td>isoform cra_a</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>36706-36</td>
<td>hermansky-pudlak syndrome 1 protein homolog</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>27887-125</td>
<td>26s proteasome non-ATPase regulatory subunit 9</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1749-927</td>
<td>utrophin isoform x2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>29218-108</td>
<td>n-alpha-acetyltransferase 50 isoform x1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>31201-85</td>
<td>transmembrane protein 115</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>22365-204</td>
<td>transmembrane protein 19 isoform x1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>36701-36</td>
<td>isoform cra_b</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7690-428</td>
<td>casp8-associated protein 2</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>2260-821</td>
<td>a kinase anchor protein isoform cra_a</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1371-1036</td>
<td>signal recognition particle 9 kDa protein</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>19-4220</td>
<td>cytoplasmic dynein 1 heavy chain 1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>20787-217</td>
<td>adp-ribosylation factor-like protein 1</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>36491-37</td>
<td>5-oxoprolinase isoform x1</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>low molecular weight phosphotyrosine protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23896-185</td>
<td>phosphatase-like</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>38691-21</td>
<td>protein mdm4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1396-1029</td>
<td>proteasome activator complex subunit 1</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>7750-426</td>
<td>lysine-specific demethylase 8</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>GeneID</td>
<td>Description</td>
<td>11279-335</td>
<td>26257-147</td>
<td>31894-78</td>
<td>26183-148</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>mitochondrial ribosomal protein l37</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PREDICTED: uncharacterized protein C1orf167</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>homolog</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>--</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>atp-binding cassette sub-family c member 8-like</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>succinate dehydrogenase</td>
<td></td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>adaptin ear-binding coat-associated protein 1</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Results for selection from SweeD and associations with environmental variables across urban (CP, FM, NYBG) and rural (BHwwp, CFP, HIP) populations. Columns to the left of the outliers show the population where the SweeD identified outlier was found. I = percent impervious surface, D = human density, C = Urban or Rural Classification

<table>
<thead>
<tr>
<th>Population</th>
<th>SweeD</th>
<th>LFMM results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>I</td>
</tr>
<tr>
<td>CP FM NYBG Combined</td>
<td>Outliers</td>
<td></td>
</tr>
<tr>
<td>- - - +</td>
<td>10099-359</td>
<td>--</td>
</tr>
<tr>
<td>+ - - -</td>
<td>10636-348</td>
<td>aflatoxin b1 aldehyde reductase member 2</td>
</tr>
<tr>
<td>+ - - -</td>
<td>1115-1128</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>113-2629</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>11470-332</td>
<td>--</td>
</tr>
<tr>
<td>+ - - -</td>
<td>1156-1114</td>
<td>--</td>
</tr>
<tr>
<td>+ - - +</td>
<td>124-2491</td>
<td>--</td>
</tr>
<tr>
<td>+ - - -</td>
<td>12718-311</td>
<td>--</td>
</tr>
<tr>
<td>- + - +</td>
<td>13665-297</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>14528-283</td>
<td>solute carrier family 22 (organic anion transporter) member 7</td>
</tr>
<tr>
<td>- - - +</td>
<td>148-2324</td>
<td>--</td>
</tr>
<tr>
<td>- - + -</td>
<td>1583-971</td>
<td>isoform cra_a</td>
</tr>
<tr>
<td>- - + +</td>
<td>1596-968</td>
<td>small ubiquitin-related modifier 2 isoform 2</td>
</tr>
<tr>
<td>- - - +</td>
<td>17779-244</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>1782-919</td>
<td>solute carrier family 39 (zinc transporter) member 1</td>
</tr>
<tr>
<td>- + + -</td>
<td>17856-243</td>
<td>serine protease inhibitor a3c-like</td>
</tr>
<tr>
<td>- - + +</td>
<td>2034-860</td>
<td>dehydrogenase reductase (sdr family) member 3</td>
</tr>
<tr>
<td>- - - +</td>
<td>20378-220</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>21270-213</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>22908-200</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>23358-193</td>
<td>--</td>
</tr>
<tr>
<td>- + - -</td>
<td>243-1951</td>
<td>solute carrier family member 13</td>
</tr>
<tr>
<td>+ + + +</td>
<td>25500-158</td>
<td>pentatricopeptide repeat domain-containing protein mitochondrial isoform x2</td>
</tr>
<tr>
<td>- + - -</td>
<td>26488-144</td>
<td>--</td>
</tr>
<tr>
<td>- - + -</td>
<td>2736-755</td>
<td>--</td>
</tr>
<tr>
<td>- - + -</td>
<td>27546-129</td>
<td>6- liver type</td>
</tr>
<tr>
<td>- - - +</td>
<td>280-1900</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>28127-122</td>
<td>sarcosine mitochondrial</td>
</tr>
<tr>
<td>- - - +</td>
<td>28528-117</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>289-1886</td>
<td>--</td>
</tr>
<tr>
<td>- - - +</td>
<td>29117-109</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 1. Map of NYC and surrounding area showing included sample localities. Sites in Red are urban parks within New York City. CP = Central Park; NYBG = New York Botanical Gardens; FM = Flushing Meadow/Willow Lake; BHwwp = Brookhaven and Wilde Wood State Park; CFP = Clarence Fahnestock State Park; HIP = High Point State Park
Figure 2. (a) Bayescan results; Red line is FDR = 0.1. (b) Sweed results, Red line corresponds to p value $\leq 0.0001$
Figure 3. (a) Population values for three environmental variables. (b) Allele frequencies for three contigs found as outliers in a genome scan and GEA test.
Appendix 1

Appendix 1.1

Figure 1.1. Frequency distribution of depth of coverage (reads / contig). (a) The Newbler cDNA assembly. Red line indicates median coverage = 4.9 reads, Interquartile range (IQR) = 4.1. (b) The Newbler genomic assembly, median = 4.7 reads, IQR = 4.6. (c) The Cap3 assembly, median = 5.0 reads, IQR = 7.0.
Appendix 1.2

**Figure 1.2.** Distribution of species with the most top-hit BLASTX results in Blast2Go using the Newbler cDNA assembly as the query.
Appendix 1.3

Table 1.3. Sequencing and assembly statistics for Newbler cDNA transcriptome assembly by tissue type and 454 sequencing plate.

<table>
<thead>
<tr>
<th></th>
<th>Liver</th>
<th>Brain</th>
<th>Ovaries</th>
<th>Testis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Plate 1</td>
<td>Plate 3</td>
<td>Plate 4</td>
<td>Plate 2</td>
</tr>
<tr>
<td>Total # reads</td>
<td>191,244</td>
<td>775,896</td>
<td>701,974</td>
<td>722,086</td>
</tr>
<tr>
<td># Contigs containing reads</td>
<td>8,177</td>
<td>8,564</td>
<td>6,260</td>
<td>9,858</td>
</tr>
</tbody>
</table>
Appendix 1.4

Table 1.4. Full list of over represented GO terms for all tissue pairwise comparisons from Fisher’s Exact Test (FDR ≤ 0.5). (a) Liver. (b) Brain. (c) Gonads.

<table>
<thead>
<tr>
<th>Brain to Liver</th>
<th>GO term</th>
<th>FDR</th>
<th># Sequences</th>
<th>Brain to Gonads</th>
<th>GO term</th>
<th>FDR</th>
<th># Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>plasma membrane</td>
<td>2.41E-44</td>
<td>567</td>
<td></td>
<td>plasma membrane</td>
<td>4.30E-92</td>
<td>567</td>
</tr>
<tr>
<td></td>
<td>cell-cell signaling</td>
<td>3.86E-40</td>
<td>232</td>
<td></td>
<td>cell-cell signaling</td>
<td>2.79E-61</td>
<td>232</td>
</tr>
<tr>
<td></td>
<td>cell differentiation</td>
<td>1.11E-38</td>
<td>372</td>
<td></td>
<td>signal transduction</td>
<td>2.15E-39</td>
<td>525</td>
</tr>
<tr>
<td></td>
<td>cytoskeleton organization</td>
<td>1.48E-24</td>
<td>145</td>
<td></td>
<td>cytoskeleton</td>
<td>9.39E-34</td>
<td>369</td>
</tr>
<tr>
<td></td>
<td>multicellular organismal development</td>
<td>2.94E-24</td>
<td>555</td>
<td></td>
<td>anatomical structure morphogenesis</td>
<td>1.89E-30</td>
<td>291</td>
</tr>
<tr>
<td></td>
<td>signal transduction</td>
<td>1.62E-23</td>
<td>525</td>
<td></td>
<td>cell differentiation</td>
<td>5.07E-28</td>
<td>372</td>
</tr>
<tr>
<td></td>
<td>anatomical structure morphogenesis</td>
<td>1.77E-19</td>
<td>291</td>
<td></td>
<td>behavior</td>
<td>6.72E-28</td>
<td>133</td>
</tr>
<tr>
<td></td>
<td>actin binding</td>
<td>1.77E-19</td>
<td>93</td>
<td></td>
<td>cytoplasmic membrane-bound vesicle</td>
<td>1.33E-22</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td>behavior</td>
<td>2.81E-15</td>
<td>133</td>
<td></td>
<td>ion transport</td>
<td>3.13E-17</td>
<td>209</td>
</tr>
<tr>
<td></td>
<td>ion channel activity</td>
<td>6.50E-14</td>
<td>62</td>
<td></td>
<td>ion channel activity</td>
<td>5.21E-17</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>cytoplasmic membrane-bound vesicle</td>
<td>1.72E-11</td>
<td>197</td>
<td></td>
<td>cellular homeostasis</td>
<td>9.82E-16</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>motor activity</td>
<td>2.08E-11</td>
<td>48</td>
<td></td>
<td>actin binding</td>
<td>3.54E-15</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>protein kinase activity</td>
<td>5.16E-09</td>
<td>77</td>
<td></td>
<td>response to external stimulus</td>
<td>9.26E-15</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>protein complex</td>
<td>6.06E-09</td>
<td>569</td>
<td></td>
<td>receptor activity</td>
<td>1.97E-14</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>cellular homeostasis</td>
<td>1.52E-08</td>
<td>134</td>
<td></td>
<td>receptor binding</td>
<td>2.79E-14</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>calcium ion binding</td>
<td>3.14E-07</td>
<td>109</td>
<td></td>
<td>calcium ion binding</td>
<td>7.69E-13</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>cilium</td>
<td>8.46E-07</td>
<td>27</td>
<td></td>
<td>cytoskeleton organization</td>
<td>9.13E-13</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>ion transport</td>
<td>4.08E-06</td>
<td>209</td>
<td></td>
<td>enzyme regulator activity</td>
<td>1.28E-11</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>structural molecule activity</td>
<td>1.93E-05</td>
<td>141</td>
<td></td>
<td>Golgi apparatus</td>
<td>1.51E-10</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td>transcription regulator activity</td>
<td>2.73E-05</td>
<td>59</td>
<td></td>
<td>cytosol</td>
<td>1.79E-08</td>
<td>411</td>
</tr>
<tr>
<td></td>
<td>cell cycle</td>
<td>6.17E-05</td>
<td>166</td>
<td></td>
<td>response to abiotic stimulus</td>
<td>4.97E-08</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>cytosol</td>
<td>9.77E-05</td>
<td>411</td>
<td></td>
<td>lipid binding</td>
<td>5.13E-07</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>chromosome microtubule organizing center</td>
<td>1.20E-03</td>
<td>91</td>
<td></td>
<td>response to stress</td>
<td>1.13E-06</td>
<td>308</td>
</tr>
<tr>
<td></td>
<td>nucleus</td>
<td>2.09E-03</td>
<td>514</td>
<td></td>
<td>protein complex</td>
<td>1.27E-06</td>
<td>569</td>
</tr>
<tr>
<td></td>
<td>cell death</td>
<td>3.06E-03</td>
<td>247</td>
<td></td>
<td>cell death</td>
<td>1.78E-06</td>
<td>247</td>
</tr>
<tr>
<td></td>
<td>response to abiotic stimulus</td>
<td>3.12E-03</td>
<td>88</td>
<td></td>
<td>motor activity</td>
<td>8.38E-06</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>chromatin binding</td>
<td>2.90E-02</td>
<td>29</td>
<td></td>
<td>endosome</td>
<td>8.46E-06</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Golgi apparatus</td>
<td>3.96E-02</td>
<td>168</td>
<td></td>
<td>response to endogenous stimulus</td>
<td>1.04E-05</td>
<td>122</td>
</tr>
</tbody>
</table>

<p>|                  | embryonic development            | 5.12E-05 | 92          | proteinaceous extracellular matrix | 1.90E-04 | 34          |
|                  | proteinaceous extracellular matrix | 1.92E-04 | 331         | protein transport                | 3.24E-04 | 152         |
|                  | signal transducer activity       | 8.24E-04 | 48          |</p>
<table>
<thead>
<tr>
<th>Function</th>
<th>Value</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>protein kinase activity</td>
<td>1.61E-03</td>
<td>77</td>
</tr>
<tr>
<td>cell envelope</td>
<td>1.88E-03</td>
<td>6</td>
</tr>
<tr>
<td>carbohydrate binding</td>
<td>8.60E-03</td>
<td>43</td>
</tr>
<tr>
<td>external encapsulating structure</td>
<td>1.38E-02</td>
<td>6</td>
</tr>
<tr>
<td>structural molecule activity</td>
<td>2.03E-02</td>
<td>141</td>
</tr>
<tr>
<td>neurotransmitter transporter activity</td>
<td>2.73E-02</td>
<td>6</td>
</tr>
<tr>
<td>extracellular space</td>
<td>3.68E-02</td>
<td>68</td>
</tr>
<tr>
<td>generation of precursor metabolites and energy</td>
<td>4.94E-02</td>
<td>134</td>
</tr>
<tr>
<td>GO term</td>
<td>FDR</td>
<td># Sequences</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>Gonads to Liver</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cilia</td>
<td>1.26E-14</td>
<td>104</td>
</tr>
<tr>
<td>cell cycle</td>
<td>8.88E-12</td>
<td>571</td>
</tr>
<tr>
<td>transcription regulator activity</td>
<td>3.57E-08</td>
<td>183</td>
</tr>
<tr>
<td>nucleolus</td>
<td>2.16E-07</td>
<td>498</td>
</tr>
<tr>
<td>reproduction</td>
<td>1.92E-06</td>
<td>680</td>
</tr>
<tr>
<td>nuclear chromosome</td>
<td>9.86E-06</td>
<td>119</td>
</tr>
<tr>
<td>cytoskeleton organization</td>
<td>1.75E-04</td>
<td>218</td>
</tr>
<tr>
<td>nucleoplasm</td>
<td>1.77E-04</td>
<td>687</td>
</tr>
<tr>
<td>regulation of gene expression, epigenetic</td>
<td>1.91E-04</td>
<td>41</td>
</tr>
<tr>
<td>microtubule organizing center</td>
<td>6.18E-04</td>
<td>150</td>
</tr>
<tr>
<td>protein kinase activity</td>
<td>6.52E-04</td>
<td>161</td>
</tr>
<tr>
<td>RNA binding</td>
<td>6.70E-04</td>
<td>637</td>
</tr>
<tr>
<td>structural molecule activity</td>
<td>1.13E-02</td>
<td>369</td>
</tr>
<tr>
<td>motor activity</td>
<td>1.71E-02</td>
<td>64</td>
</tr>
<tr>
<td>viral reproduction</td>
<td>1.75E-02</td>
<td>339</td>
</tr>
<tr>
<td>cell differentiation</td>
<td>3.89E-02</td>
<td>613</td>
</tr>
<tr>
<td>chromatin binding</td>
<td>4.61E-02</td>
<td>78</td>
</tr>
</tbody>
</table>
**Appendix 1.5**

**Table 1.5.** Candidate loci with $p_N/p_S$ between 0.5 and 1.

<table>
<thead>
<tr>
<th>Sequence name</th>
<th>$p_N/p_S$</th>
<th>Gene name</th>
<th>Gene function</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP_contig02521</td>
<td>0.51</td>
<td>Apoptogenic protein mitochondrial</td>
<td>Regulation of mitochondrial induced apoptosis</td>
</tr>
<tr>
<td>HP_contig01711</td>
<td>0.51</td>
<td>Fibrinogen alpha chain</td>
<td>Glycoprotein circulating in the blood; functions in blood coagulation and part of the most abundant component of blood clots</td>
</tr>
<tr>
<td>HP_contig02065</td>
<td>0.52</td>
<td>Complement factor h</td>
<td>Glycoprotein circulating in plasma; regulation of complement activation</td>
</tr>
<tr>
<td>HP_contig00430</td>
<td>0.53</td>
<td>Murinoglobulin-1 precursor</td>
<td>Protease activity; acute phase response</td>
</tr>
<tr>
<td>HP_contig01727</td>
<td>0.54</td>
<td>Ornithine mitochondrial</td>
<td>Transfer of ornithine across inner mitochondrial membrane</td>
</tr>
<tr>
<td>HP_contig00783</td>
<td>0.54</td>
<td>Carboxymethylenebutenolide homolog</td>
<td>Cysteine hydrolase, protein binding</td>
</tr>
<tr>
<td>HP_contig00807*</td>
<td>0.54</td>
<td>Isoform cra_a</td>
<td>Uncharacterized cellular membrane protein</td>
</tr>
<tr>
<td>HP_contig01783</td>
<td>0.58</td>
<td>Cytochrome p450 2a15</td>
<td>Metabolic process; testosterone 7a-hydroxylase activity</td>
</tr>
<tr>
<td>HP_contig00737</td>
<td>0.58</td>
<td>Catechol o-methyltransferase</td>
<td>Catalyzes methylation for degradation of neurotransmitters and catecholic xenobiotics</td>
</tr>
<tr>
<td>HP_contig05051</td>
<td>0.60</td>
<td>Hypothetical protein I79_019498</td>
<td>Uncharacterized</td>
</tr>
<tr>
<td>HP_contig02710</td>
<td>0.63</td>
<td>L-xylulose reductase</td>
<td>Metabolic processes; catalyzes NADPH-dependent reduction</td>
</tr>
<tr>
<td>HP_contig01169</td>
<td>0.64</td>
<td>Alpha-1-acid glycoprotein precursor</td>
<td>Transport protein in the blood stream; binds and distributes synthetic drugs throughout body; modulates innate immune response</td>
</tr>
<tr>
<td>HP_contig02231</td>
<td>0.66</td>
<td>Isoform cra_b</td>
<td>Transmembrane transport protein</td>
</tr>
<tr>
<td>HP_contig02824</td>
<td>0.67</td>
<td>Nadh dehydrogenase</td>
<td>Mitochondrial respiratory chain complex; electron transport</td>
</tr>
<tr>
<td>HP_contig03468</td>
<td>0.73</td>
<td>Coatomer subunit beta</td>
<td>Protein transportation between Golgi body and ER; required for budding from Golgi body</td>
</tr>
<tr>
<td>HP_contig01714</td>
<td>0.74</td>
<td>Complement factor I</td>
<td>Serine protease; regulation of complement activation by cleavage of complement system components</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Expr.</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>-------</td>
<td>---------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>HP_contig01785</td>
<td>0.77</td>
<td>Kininogen-1 isoform 2 precursor</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inflammatory response; involved in blood coagulation - negative regulation</td>
<td></td>
</tr>
<tr>
<td>HP_contig03812</td>
<td>0.79</td>
<td>Biorientation of chromosomes in cell division protein 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Required for proper orientation of chromosomes during cell division</td>
<td></td>
</tr>
<tr>
<td>HP_contig01991</td>
<td>0.80</td>
<td>Calcium binding and coiled-coil domain 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Innate immune response; receptor protein for bacteria; mediate macroautophagy</td>
<td></td>
</tr>
<tr>
<td>HP_contig01688</td>
<td>0.83</td>
<td>Polymeric immunoglobulin receptor</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transports immunoglobulins across cell to apical surface for secretion; adaptive immune system</td>
<td></td>
</tr>
<tr>
<td>HP_contig02433</td>
<td>0.99</td>
<td>40s ribosomal protein s3a</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNA binding; translational initiation</td>
<td></td>
</tr>
<tr>
<td>HP_contig03921</td>
<td>1.00</td>
<td>Bile salt sulfotransferase-like</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transferase activity</td>
<td></td>
</tr>
<tr>
<td>RR_contig00497</td>
<td>0.50</td>
<td>Arginase-1 like</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arginine metabolism</td>
<td></td>
</tr>
<tr>
<td>RR_contig00554</td>
<td>0.51</td>
<td>Ornithine mitochondrial</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Metabolic processes; mitochondrial membrane</td>
<td></td>
</tr>
<tr>
<td>NYBG_contig00478</td>
<td>0.54</td>
<td>Fibrinogen alpha chain</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Glycoprotein circulating in the blood; functions in blood coagulation and part of the most abundant component of blood clots</td>
<td></td>
</tr>
<tr>
<td>NYBG_contig00650</td>
<td>0.54</td>
<td>Leucine-rich repeat and wd repeat-containing protein 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Involved with DNA replication initiation and silencing; Binds to methylated histones and restricts transcription</td>
<td></td>
</tr>
<tr>
<td>CP_contig00326</td>
<td>0.55</td>
<td>---NA---</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Uncharacterized</td>
<td></td>
</tr>
<tr>
<td>CP_contig00764</td>
<td>0.55</td>
<td>Murinoglobulin-2 precursor</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Protease activity; acute phase response</td>
<td></td>
</tr>
<tr>
<td>RR_contig00859</td>
<td>0.59</td>
<td>Liver carboxylesterase b-1-like</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Xenobiotic metabolism, detoxification through hydrolysis of ester and amide bonds</td>
<td></td>
</tr>
<tr>
<td>NYBG_contig00447</td>
<td>0.59</td>
<td>Complement c3 precursor</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Innate immune response; major role in activation of complement pathway</td>
<td></td>
</tr>
<tr>
<td>RR_contig00239</td>
<td>0.59</td>
<td>Cytochrome p450 family 2 subfamily b</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Xenobiotic metabolism</td>
<td></td>
</tr>
<tr>
<td>RR_contig00587</td>
<td>0.63</td>
<td>Estradiol 17-beta-dehydrogenase 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Steroid biosynthetic process; oxidation-reduction</td>
<td></td>
</tr>
<tr>
<td>CP_contig00527</td>
<td>0.65</td>
<td>Cytochrome p450 2d27-like</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Xenobiotic metabolism; oxidation-reduction</td>
<td></td>
</tr>
<tr>
<td>NYBG_contig00664</td>
<td>0.73</td>
<td>Protein maelstrom homolog isoform 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reproductive process; active in spermatogenesis; repress transposable elements to retain germline integrity</td>
<td></td>
</tr>
<tr>
<td>NYBG_contig00513</td>
<td>0.74</td>
<td>Cytochrome family subfamily polypeptide 13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Xenobiotic metabolism; oxidation-reduction</td>
<td></td>
</tr>
<tr>
<td>NYBG_contig00284</td>
<td>0.81</td>
<td>Catechol o-methyltransferase</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Methylation activity; introduces methyl groups to inactivate neurotransmitters</td>
<td></td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>pN/pS</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>CP_contig00544</td>
<td>0.91</td>
<td>Gametogenin-binding protein 1</td>
<td></td>
</tr>
<tr>
<td>RR_contig01212</td>
<td>0.91</td>
<td>Isoform cra_a</td>
<td></td>
</tr>
<tr>
<td>NYBG_contig00790*</td>
<td>0.98</td>
<td>Orosomucoid 1</td>
<td></td>
</tr>
<tr>
<td>CP_contig00444*</td>
<td>0.99</td>
<td>Vitamin d-binding protein</td>
<td></td>
</tr>
<tr>
<td>CP_contig00449</td>
<td>0.99</td>
<td>Kininogen-1 isoform 2 precursor</td>
<td></td>
</tr>
</tbody>
</table>

*= Gene contained $p_N/p_S$ between 0.5 and 1 in two independent population pairwise comparisons
Appendix 2

Appendix 2.1

Table 2.1 Microsatellite repeats identified

<table>
<thead>
<tr>
<th>Microsatellite Type (Min. Repeat Number)</th>
<th>Contig Name</th>
<th>Motif</th>
<th>Left Primer Sequence</th>
<th>Right Primer Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Di(10)-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14094-290</td>
<td>AG</td>
<td></td>
<td>ACTGGGAGTA</td>
<td>TGCTCTCCTTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGCTTTGAGG</td>
<td>CATGTTCACC</td>
</tr>
<tr>
<td>15039-276</td>
<td>AG</td>
<td></td>
<td>CGCATGCACAC</td>
<td>ACCACATATCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCTGTCAC</td>
<td>AGCCTGTCC</td>
</tr>
<tr>
<td>20426-220</td>
<td>AG</td>
<td></td>
<td>TTCAAAGGGGA</td>
<td>CTTACTGAGAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ACGACAGCAG</td>
<td>AGGCCAGGC</td>
</tr>
<tr>
<td>21021-215</td>
<td>AT</td>
<td></td>
<td>TTTGGCCTGGA</td>
<td>TCTTTTGGTGTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TGTGAGAGGG</td>
<td>CTTCTCGGG</td>
</tr>
<tr>
<td>21826-208</td>
<td>AG</td>
<td></td>
<td>TCCTTCCTTGCA</td>
<td>AGGTGTGATTCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AAGCTGAAG</td>
<td>GACTCCTGTG</td>
</tr>
<tr>
<td>22550-203</td>
<td>AC</td>
<td></td>
<td>GCAACATTCCG</td>
<td>GGGCTAGGGGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGTCAGCAG</td>
<td>GACAGAGATG</td>
</tr>
<tr>
<td>24636-173</td>
<td>AC</td>
<td></td>
<td>CGGCCTCCATA</td>
<td>TCACGATGGCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TACACATGCA</td>
<td>CTTTCTCTTG</td>
</tr>
<tr>
<td>25938-152</td>
<td>AC</td>
<td></td>
<td>TCTGTAACTCC</td>
<td>AGGGCTACAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ACTGAGCAG</td>
<td>AGTGAACCC</td>
</tr>
<tr>
<td>25957-152</td>
<td>AC</td>
<td></td>
<td>TCTGTAACTCC</td>
<td>TCTCAAGCAGCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ACTGAGCAG</td>
<td>AAGAGGCAG</td>
</tr>
<tr>
<td>Start</td>
<td>End</td>
<td>Type</td>
<td>Sequence 1</td>
<td>Sequence 2</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>28787-114</td>
<td>AC</td>
<td>Tri(8)-</td>
<td>GGGCTTTGTGA</td>
<td>GCAGTCAATA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GTTTATCGGG</td>
<td>AAGGGCCGAG</td>
</tr>
<tr>
<td>32874-69</td>
<td>AC</td>
<td></td>
<td>TCTCTGACCAG</td>
<td>ATGCTCCCAAGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CAATCACCC</td>
<td>TCACAGATG</td>
</tr>
<tr>
<td>32920-68</td>
<td>AG</td>
<td></td>
<td>CCGAGTTTACG</td>
<td>TGGCAAGACCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCCAGCTTG</td>
<td>AATGGTGGAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tetra(4)-</td>
<td>CCGGACAACA</td>
<td>TCGCTCATGAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ACATAGCTGG</td>
<td>TACGGCTTTC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Type</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-4012</td>
<td>CCCG</td>
<td></td>
<td>CAGCACATTCA</td>
<td>AGGACGTATGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGGGCTTCC</td>
<td>TCTGCCCTC</td>
</tr>
<tr>
<td>102-2754</td>
<td>AGGG</td>
<td></td>
<td>TCATCAAGAGC</td>
<td>CGGGACACTTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ATCCGAGGG</td>
<td>TTGCACTTC</td>
</tr>
<tr>
<td>457-1608</td>
<td>ACCC</td>
<td></td>
<td>GCCAAGATGA</td>
<td>GAAACTGGCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AACAGGAGCC</td>
<td>GTGTGGAGTC</td>
</tr>
<tr>
<td>530-1517</td>
<td>CCCG</td>
<td></td>
<td>TGGAAGATCA</td>
<td>CTGGACAGAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GACACGCTCG</td>
<td>GCTCAGGAG</td>
</tr>
<tr>
<td>653-1394</td>
<td>AAAC</td>
<td></td>
<td>CACCTCAGAG</td>
<td>CATGATTTGAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ATGGACCCAG</td>
<td>CTGCCACAG</td>
</tr>
<tr>
<td>1520-988</td>
<td>ACAG</td>
<td></td>
<td>CCCTACAGGCC</td>
<td>GTTGCTTGCTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CATGTCTC</td>
<td>CACCTCTTG</td>
</tr>
<tr>
<td>1617-962</td>
<td>AGCG</td>
<td></td>
<td>GGCCGCAGAA</td>
<td>GTCAAGAGGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GATGATGAAC</td>
<td>GGCAGAGTC</td>
</tr>
<tr>
<td>1737-930</td>
<td>AGGC</td>
<td></td>
<td>GTTCAGCCTT</td>
<td>GAATGGATGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CCAGTCTTC</td>
<td>ACAGCTGCAC</td>
</tr>
<tr>
<td>3410-680</td>
<td>AAGG</td>
<td></td>
<td>GATATCCAGAC</td>
<td>ACCTTCTGCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CAACGGCTC</td>
<td>TTACACCTG</td>
</tr>
<tr>
<td>Start</td>
<td>End</td>
<td>Sequence</td>
<td>Start</td>
<td>End</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>----------</td>
<td>-------</td>
<td>-----</td>
</tr>
<tr>
<td>4523</td>
<td>582</td>
<td>CCCG AGACAGTGCC CTCTCATTTG ACCCTCCACAA GTCCAACTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4691</td>
<td>572</td>
<td>AGCG GGGAAAGTTTCC ACGCGAGAG TCTTTTCTTCT TCCTAGCCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7208</td>
<td>445</td>
<td>AGCG GAAATCCAGA GAGTATTGAC GGGCACGACG CGTCCAGCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7523</td>
<td>434</td>
<td>CCCG GCCTTTAAGC ACCTCACTC AGGAATCAGA GAGACGGAGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8490</td>
<td>402</td>
<td>ACAG TAGCCGTTTCC TCTGCTGG TTCTCTTCCTC TTCCAGCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8731</td>
<td>395</td>
<td>AGGG GCATCAGTGT TTCAGCTCC TTCTCCGCTCC ACATCCTAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8828</td>
<td>391</td>
<td>AAGG GTCTGTAGGGA GAGTGCAGG GCAGAACACTG ACAAGGTACG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9059</td>
<td>385</td>
<td>CCCG CTTTAAGGAGC CGCTCTGC AACCTGAAAG ACGGCAAAGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11563</td>
<td>330</td>
<td>AAGC CAAGCGCTCTA TCRACTGAC TGGGCTGGCTT TCTCTGATG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11896</td>
<td>324</td>
<td>CCCG CCAGAGTCTAG GCAGAAGTTT GCAGAAGAGG CTGGAAGGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12177</td>
<td>309</td>
<td>AGCC GAAGAGCACC GAGGCGAG GATTCCACTGC CCACAATCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12826</td>
<td>309</td>
<td>ACAG AACAAGCAGC TGTTGTGAGC CATCAAAGAG CCTGGCCATG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13035</td>
<td>306</td>
<td>AGGC CCGCATCTTTA GTCAGCAGC GTCTCTAACC TGGGTGTGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13611</td>
<td>298</td>
<td>ACAG CCATGGGCCAT ACTCGGAAG CCAGGGCTATG CAAAGAAACC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position</td>
<td>Sequence</td>
<td>Position</td>
<td>Sequence</td>
<td>Position</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td>13698-297</td>
<td>CCGG</td>
<td>GACCACGCGT</td>
<td>AGACCTCGAC</td>
<td>TGAAAGCTGG</td>
</tr>
<tr>
<td>15970-264</td>
<td>AAAG</td>
<td>CACTGATGTCT</td>
<td>TCGCTGCAG</td>
<td>TCCAGGACAC</td>
</tr>
<tr>
<td>19855-225</td>
<td>ATCC</td>
<td>GGATGGATAG</td>
<td>TGGGATACATG</td>
<td>GAGTCTTTGTGT</td>
</tr>
<tr>
<td>21617-210</td>
<td>ACTG</td>
<td>GTGCCCTCCCAT</td>
<td>GGTGTCTG</td>
<td>CTGAAGGTGG</td>
</tr>
<tr>
<td>23071-199</td>
<td>AGGG</td>
<td>CCCATTTCCTG</td>
<td>TGAGAGCAG</td>
<td>GGCAACATCTC</td>
</tr>
<tr>
<td>26575-142</td>
<td>AAGC</td>
<td>TGAAAAGGTGG</td>
<td>AGACTGAGCC</td>
<td>CACCACTGCTT</td>
</tr>
<tr>
<td>26642-142</td>
<td>AGGC</td>
<td>CTCTACCCTG</td>
<td>AGTGAGGTC</td>
<td>GAGTCCTCAGG</td>
</tr>
<tr>
<td>27675-127</td>
<td>AAAC</td>
<td>ATGCAGCTGG</td>
<td>AGAGGTATCG</td>
<td>ACAGGCTGCTT</td>
</tr>
<tr>
<td>27827-126</td>
<td>AAAC</td>
<td>CAACTCCTCCC</td>
<td>AGATCCTCC</td>
<td>GCAAGGCCTTC</td>
</tr>
<tr>
<td>28127-122</td>
<td>AGGC</td>
<td>GTCTCCAATCC</td>
<td>ACCCTAGGC</td>
<td>GAGGTTCTCGG</td>
</tr>
<tr>
<td>30409-94</td>
<td>AAAC</td>
<td>CGAAAGCATAA</td>
<td>GGACAGAGCAG</td>
<td>AGCAATGCCA</td>
</tr>
<tr>
<td>33022-67</td>
<td>ACAT</td>
<td>AAAACCTCTCTC</td>
<td>TGGCCTCTG</td>
<td>TTCCCGGCACTC</td>
</tr>
<tr>
<td>35222-48</td>
<td>ATCC</td>
<td>TGGATGATTTGG</td>
<td>ATGGTGTTGG</td>
<td>CCCACACACTT</td>
</tr>
<tr>
<td>36493-37</td>
<td>ACAG</td>
<td>GGTGCAGTCTCT</td>
<td>TAAGAGGCC</td>
<td>AGAAGAACCAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Penta(4)-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10565-350</td>
<td>AAAAC</td>
<td>GCTGTGTACAT</td>
<td>ACCACTGCCCCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GTCCGCTTC</td>
<td>GATCTAAGAG</td>
<td></td>
</tr>
<tr>
<td>23608-189</td>
<td>ACCCC</td>
<td>GTGCCATCTCA</td>
<td>GTACCGGTAAT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCAGCTCC</td>
<td>GTTAGGACTG</td>
<td></td>
</tr>
<tr>
<td>31814-79</td>
<td>CCCGG</td>
<td>ACATACTCTTC</td>
<td>ACCCTTACACT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACGCTCCAC</td>
<td>GCAGAGCTC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hexa(4)-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3565-665</td>
<td>AGGATG</td>
<td>AAACACGTTCTT</td>
<td>GGGAATTACAT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCCATGGGT</td>
<td>TGGGACCAGAGC</td>
<td></td>
</tr>
<tr>
<td>4543-581</td>
<td>AAGTCC</td>
<td>TCCCAGAAGC</td>
<td>GAATGAGAAA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACAGTCTTCC</td>
<td>GGCAGGGCTTC</td>
<td></td>
</tr>
<tr>
<td>10198-357</td>
<td>AAGAGG</td>
<td>TGTAAGAAAGA</td>
<td>CTCTGCAGAAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCTGGGAAGG</td>
<td>CCAATTGCC</td>
<td></td>
</tr>
<tr>
<td>11279-335</td>
<td>AGGCC</td>
<td>CTTCCAGTTCC</td>
<td>GAGGAGTAGG</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TCGTGTTGC</td>
<td>AGGCTTGGCC</td>
<td></td>
</tr>
<tr>
<td>29344-106</td>
<td>ATGCC</td>
<td>CACACTGCAGT</td>
<td>CACAGTCGCTC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTCACCAGAG</td>
<td>TTGAGTCAG</td>
<td></td>
</tr>
<tr>
<td>29390-106</td>
<td>ATGCC</td>
<td>CACACTGCAGT</td>
<td>GGCACAGCTGT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTCACCAGAG</td>
<td>CACTAGTTTC</td>
<td></td>
</tr>
<tr>
<td>34061-58</td>
<td>ACCCTG</td>
<td>TGGTCTGAGCA</td>
<td>TGCTCTCTTCC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CACTTCCTC</td>
<td>CTGCTGTAC</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 3

Appendix 3.1

Figure 3.1 sNMF results for K=2 (highest support) of ≤ 15% missing SNP dataset (3,664 SNPs).

Red = assigned to Long Island. Blue = assigned to Mainland & Manhattan.
Appendix 3.2

**Figure 3.2** Membership probability plots generated from *compoplot* for global, MM, and LI groups of individuals. Each vertical bar represents an individual, and solid colors indicate 100% probability of assigning the individual to the correct sampling location.
Appendix 3.3

Figure 3.3 (a) TPL file and (b) EST file (input files for fastsimcoal2) showing setup for demographic modeling and parameter size ranges.

a. TPL

```plaintext
// Number of population samples (demes)
3
// Population effective sizes (number of genes)
NP1
NP2
NP3
// Samples sizes and samples age
9
66
116
// Growth rates: negative growth implies population expansion
0
0
0
// Number of migration matrices: 0 implies no migration between demes
3
// Migration matrix 0
0 MIG1 MIG2
MIG3 0 MIG4
MIG5 MIG6 0
// Migration matrix 1
0 0 0
0 0 MIG4
0 MIG6 0
// Migration matrix 2
0 0 0
0 0 0
0 0 0
// Historic event: time, source, sink, migrants, new deme size, new growth rate, migration matrix index
3 historic events
TISLAND1 0 0 0 RESIZE1 0 0
TISLAND2 0 2 1 1 0 1
TISLIV 2 1 1 RESIZE2 0 2
// Number of independent loci [chromosome]
10
// Per chromosome: Number of contiguous linkage block: a block is a set of contiguous loci
1
// Per block: data type, number of loci, per gen recomb and mut rates
FREQ 1 0 2.0e-8
```
b. EST

// Priors and rules file
// -------------------

[PARAMETERS]
// #isInt? #name #dist.#min #max
// all N are in number of haploid individuals
1 ANCSIZE unif 100 100000 output
1 NPOP1 unif 20 100000 output
1 NPOP2 unif 20 100000 output
1 NPOP3 unif 20 100000 output
 0 NIM1 logunif 0.01 20 hide
 0 NIM2 logunif 0.01 20 hide
 0 NIM3 logunif 0.01 20 hide
 0 NIM4 logunif 0.01 20 hide
 0 NIM5 logunif 0.01 20 hide
 0 NIM6 logunif 0.01 20 hide
1 TISLAND1 unif 4 1000 output
1 TISLAND2 unif 4 1000 output
1 TDIV unif 4 30000 output bounded
 0 RESIZE1 logunif 1e-10 1 output

[RULES]

TISLAND2 > TISLAND1
TISLAND2 < TDIV
NPOP3 > NPOP2
NPOP3 > NPOP1
NPOP2 > NPOP1

[COMPLEX PARAMETERS]

0 ZNM1 = 2*NIM1 hide
0 ZNM2 = 2*NIM2 hide
0 ZNM3 = 2*NIM3 hide
0 ZNM4 = 2*NIM4 hide
0 ZNM5 = 2*NIM5 hide
0 ZNM6 = 2*NIM6 hide
0 MIG1 = ZNM1/NPOP1 output
0 MIG2 = ZNM2/NPOP1 output
0 MIG3 = ZNM3/NPOP2 output
0 MIG4 = ZNM4/NPOP2 output
0 MIG5 = ZNM5/NPOP3 output
0 MIG6 = ZNM6/NPOP3 output
0 RESIZE2 = ANCSIZE/NPOP3 output
## Appendix 4

### Appendix 4.1

**Table 4.1** Blast2GO table with BLASTX hits and top three supported Gene Ontology terms for outlier genes from Bayescan and SweeD

<table>
<thead>
<tr>
<th>Seq. Name</th>
<th>Seq. Description</th>
<th>Seq. Length</th>
<th>#Hits</th>
<th>min. eValue</th>
<th>mean Similarity</th>
<th>#GOs</th>
<th>GOs</th>
</tr>
</thead>
<tbody>
<tr>
<td>27887-125</td>
<td>26s proteasome non-ATPase regulatory subunit 9</td>
<td>572</td>
<td>5</td>
<td>1.02E-43</td>
<td>92.80%</td>
<td>1</td>
<td><strong>C:</strong> proteasome complex</td>
</tr>
<tr>
<td>17974-242</td>
<td>40s ribosomal protein s15a-like protein</td>
<td>243</td>
<td>5</td>
<td>4.92E-08</td>
<td>59.20%</td>
<td>9</td>
<td><strong>F:</strong> L-serine transmembrane transporter activity; <strong>C:</strong> integral to membrane; <strong>P:</strong> phosphatidylycerine metabolic process;</td>
</tr>
<tr>
<td>36491-37</td>
<td>5-oxoprolinase isoform x1</td>
<td>3504</td>
<td>5</td>
<td>0</td>
<td>96.60%</td>
<td>2</td>
<td><strong>P:</strong> metabolic process; <strong>F:</strong> hydrolase activity</td>
</tr>
<tr>
<td>26257-147</td>
<td>PREDICTED: uncharacterized protein C1orf167 homolog</td>
<td>1635</td>
<td>4</td>
<td>1.29E-09</td>
<td>81.00%</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>2260-821</td>
<td>a kinase anchor protein isoform cra_a</td>
<td>1645</td>
<td>5</td>
<td>1.01E-150</td>
<td>90.20%</td>
<td>13</td>
<td><strong>C:</strong> nuclear matrix; <strong>F:</strong> double-stranded DNA binding; <strong>F:</strong> zinc ion binding;</td>
</tr>
<tr>
<td>40819-1</td>
<td>adaptin ear-binding coat-associated protein 1</td>
<td>1043</td>
<td>5</td>
<td>3.68E-09</td>
<td>94.80%</td>
<td>5</td>
<td><strong>C:</strong> coated pit; <strong>C:</strong> clathrin vesicle coat; <strong>C:</strong> plasma membrane; <strong>C:</strong> trans-Golgi network; <strong>C:</strong> Golgi membrane; <strong>P:</strong> protein localization to Golgi apparatus;</td>
</tr>
<tr>
<td>20787-217</td>
<td>adp-ribosylation factor-like protein 1</td>
<td>1599</td>
<td>5</td>
<td>2.02E-112</td>
<td>100.00%</td>
<td>15</td>
<td><strong>F:</strong> retinoid X receptor binding; <strong>F:</strong> RNA polymerase II distal enhancer sequence-specific DNA binding; <strong>P:</strong> response to glucose stimulus</td>
</tr>
<tr>
<td>25795-154</td>
<td>af478441_1farnesoid-x-receptor alpha splice variant 1</td>
<td>1289</td>
<td>5</td>
<td>3.09E-160</td>
<td>96.20%</td>
<td>25</td>
<td><strong>F:</strong> flavin adenine dinucleotide binding; <strong>F:</strong> UDP-N-acetylmuramate dehydrogenase activity; <strong>P:</strong> lipid biosynthetic process;</td>
</tr>
<tr>
<td>902-1236</td>
<td>alkyldihydroxyacetone phosphate peroxisomal</td>
<td>1579</td>
<td>5</td>
<td>4.57E-140</td>
<td>99.80%</td>
<td>5</td>
<td><strong>F:</strong> sulfonylurea receptor activity; <strong>F:</strong> ATPase activity, coupled to transmembrane movement of substances; <strong>P:</strong> potassium ion import;</td>
</tr>
<tr>
<td>26183-148</td>
<td>atp-binding cassette subfamily c member 8-like</td>
<td>366</td>
<td>5</td>
<td>6.87E-59</td>
<td>99.40%</td>
<td>13</td>
<td>-</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene Name</td>
<td>Log2 Fold Change</td>
<td>p-Value</td>
<td>% Change</td>
<td>Canonical Pathways</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------</td>
<td>------------------</td>
<td>----------</td>
<td>----------</td>
<td>-----------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27707-127</td>
<td>autophagy-related protein 2 homolog a isoform x2</td>
<td>2.89E-12</td>
<td>96.00%</td>
<td>3</td>
<td>C:lipid particle; C:membrane; P:autophagy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7690-428</td>
<td>casp8-associated protein 2 cytoplasmic dynein 1 heavy chain 1 isoform x2</td>
<td>5.87E-51</td>
<td>90.80%</td>
<td>14</td>
<td>P:Fas signaling pathway; P:cell cycle; C:mitochondrion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19-4220</td>
<td>autophagy-related protein 2 homolog a isoform x2</td>
<td>8.59E-54</td>
<td>91.60%</td>
<td>4</td>
<td>F:microtubule motor activity; P:transport; C:dynein complex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3567-665</td>
<td>hermansky-pudlak syndrome 1 protein homolog</td>
<td>5.92E-71</td>
<td>95.40%</td>
<td>9</td>
<td>P:organelle organization; P:melanocyte differentiation; C:cytoplasmic membrane-bounded vesicle; F:nucleosome assembly; F:DNA binding; F:nucleic acid binding; C:cytoplasm; C:plasma membrane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36706-36</td>
<td>isoform cra_a</td>
<td>8.85E-12</td>
<td>80.80%</td>
<td>5</td>
<td>C:cytoplasm; F:non-membrane spanning protein tyrosine phosphatase activity; F:acid phosphatase activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35973-42</td>
<td>isoform cra_a</td>
<td>8.59E-54</td>
<td>91.60%</td>
<td>4</td>
<td>C:cytoplasm; F:non-membrane spanning protein tyrosine phosphatase activity; F:acid phosphatase activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36701-36</td>
<td>isoform cra_b jnk sakp-inhibitory isoform cra_a</td>
<td>8.93E-91</td>
<td>95.20%</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36437-38</td>
<td>low molecular weight phosphotyrosine protein phosphatase-like</td>
<td>7.90E-111</td>
<td>96.00%</td>
<td>9</td>
<td>C:cytoplasm; F:non-membrane spanning protein tyrosine phosphatase activity; F:acid phosphatase activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23896-185</td>
<td>lysine-specific demethylase 8</td>
<td>8.93E-91</td>
<td>95.20%</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11279-335</td>
<td>mitochondrial ribosomal protein l37</td>
<td>9.57E-12</td>
<td>99.40%</td>
<td>4</td>
<td>C:cytoplasm; F:N-acetyltransferase activity; F:protein binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7975-418</td>
<td>myosin light chain smooth muscle</td>
<td>7.42E-67</td>
<td>86.20%</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29218-108</td>
<td>n-alpha-acetyltransferase 50 isoform x1</td>
<td>9.57E-12</td>
<td>99.40%</td>
<td>4</td>
<td>C:cytoplasm; F:N-acetyltransferase activity; F:protein binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5754-511</td>
<td>3</td>
<td>7.42E-67</td>
<td>86.20%</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1396-1029</td>
<td>proteasome activator complex subunit 1</td>
<td>1.22E-154</td>
<td>96.80%</td>
<td>5</td>
<td>F:endopeptidase activator activity; P:positive regulation of endopeptidase activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2482-790</td>
<td>protein diaphanosus homolog 1 isoform x1</td>
<td>4.28E-21</td>
<td>81.25%</td>
<td>15</td>
<td>F:actin binding; P:actin filament polymerization; F:Rho GTPase binding P:G0 to G1 transition; P:negative regulation of apoptotic process; P:protein stabilization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38691-21</td>
<td>3</td>
<td>4.28E-21</td>
<td>81.25%</td>
<td>15</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seq. Name</td>
<td>Seq. Description</td>
<td>Seq. Length</td>
<td>#Hits</td>
<td>min. eValue</td>
<td>mean Similarity</td>
<td>#GOs</td>
<td>GOs</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------</td>
<td>-------------</td>
<td>-------</td>
<td>-------------</td>
<td>----------------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>27691-127</td>
<td>retroviral nucleocapsid protein gag containing protein</td>
<td>1224 5 2.99E-31</td>
<td>60.20%</td>
<td>3</td>
<td>F:nucleic acid binding; P:viral process; F:zinc ion binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1371-1036</td>
<td>signal recognition particle 9 kda protein</td>
<td>1381 5 4.23E-51</td>
<td>99.60%</td>
<td>4</td>
<td>C:signal recognition particle, endoplasmic reticulum targeting; P:negative regulation of translational elongation; F:7S RNA binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37967-26</td>
<td>sorbin and sh3 domain-containing protein 2 isoform x3</td>
<td>1472 5 0</td>
<td>96.80%</td>
<td>10</td>
<td>C:Z disc; F:structural constituent of cytoskeleton; F:structural constituent of muscle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14102-290</td>
<td>succinate dehydrogenase transmembrane 9 superfamily member 1 isoform 2</td>
<td>1250 5 0</td>
<td>98.20%</td>
<td>6</td>
<td>C:autophagic vacuole membrane; C:lysosomal membrane; C:integral to membrane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31201-85</td>
<td>transmembrane protein 115</td>
<td>688 5 2.73E-47</td>
<td>97.00%</td>
<td>4</td>
<td>C:integral to membrane; P:replicative cell aging; C:Golgi apparatus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22365-204</td>
<td>transmembrane protein 19 isoform x1</td>
<td>2106 5 8.76E-169</td>
<td>96.20%</td>
<td>1</td>
<td>C:integral to membrane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37015-34</td>
<td>tubulin folding cofactor e-like isoform x6 utrophin isoform x2</td>
<td>954 5 0</td>
<td>99.00%</td>
<td>2</td>
<td>C:cytoskeleton; C:cytoplasm F:integrin binding; C:membrane raft; F:zinc ion binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1749-927</td>
<td>SweeD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene ID</td>
<td>Description</td>
<td>Gene ID</td>
<td>Description</td>
<td>Count</td>
<td>P Value</td>
<td>Log2 Fold Change</td>
<td>Gene ID</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------</td>
<td>---------</td>
<td>------------------</td>
<td>---------</td>
</tr>
<tr>
<td>35790</td>
<td>apolipoprotein b-partial</td>
<td>35790</td>
<td>apolipoprotein b-partial</td>
<td>13706</td>
<td>5</td>
<td>82.40%</td>
<td>2034</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26488</td>
<td>pentatricopeptide repeat domain-containing protein mitochondrial isoform x2</td>
<td>26488</td>
<td>pentatricopeptide repeat domain-containing protein mitochondrial isoform x2</td>
<td>1179</td>
<td>5</td>
<td>94.00%</td>
<td>8466</td>
</tr>
<tr>
<td>17856</td>
<td>small ubiquitin-related modifier 2 isoform 2</td>
<td>17856</td>
<td>small ubiquitin-related modifier 2 isoform 2</td>
<td>3172</td>
<td>5</td>
<td>86.80%</td>
<td>1596</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Name</td>
<td>Accession</td>
<td>Score</td>
<td>FDR</td>
<td>Function Description</td>
<td>Coordinates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
<td>------------------------------------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>solute carrier family 22 (organic anion transporter) member 7</td>
<td>14528-283</td>
<td>2144</td>
<td>5</td>
<td>C:integral to plasma membrane; C:basolateral plasma membrane; P:ion transmembrane transport</td>
<td>90.60%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>solute carrier family 39 (zinc transporter) member 1</td>
<td>1782-919</td>
<td>1615</td>
<td>5</td>
<td>C:plasma membrane; P:zinc ion transmembrane transport; F:receptor binding</td>
<td>7.37E-170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>solute carrier family member 13</td>
<td>243-1951</td>
<td>2914</td>
<td>5</td>
<td>P:gluconeogenesis; P:response to calcium ion; F:calcium ion binding</td>
<td>98.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>udp-glucuronosyltransferase 3a1-like isoform x1</td>
<td>6315-484</td>
<td>2572</td>
<td>5</td>
<td>F:glucuronosyltransferase activity; P:cellular response to genistein; C:integral to membrane</td>
<td>83.80%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.2 KEGG analysis for biochemical pathways containing more than one outlier contig.

Colored boxes represent outlier genes
Bibliography


R Core Team (2013) R: A Language and Environment for Statistical Computing.

Ramsdell CM, Lewandowski A a, 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.


Settachan D (2001) Mechanistic and molecular studies into the effects of 2,3,7,8-tetrachlorodibenzop-dioxin and similar compounds in the deer mouse, Peromyscus maniculatus. Texas Tech University.


Trucchi E, Gratton P, Whittington JD *et al.* (2014) King penguin demography since the last glaciation inferred from genome-wide data King penguin demography since the last glaciation inferred from genome-wide data. *Proceedings of the Royal Society B: Biological Sciences, 281*, 20140528.


Wright S (1931) Evolution in mendelian populations. *Genetics, 97–159.*


