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

Stephen E. Harris[1], Jason Munshi-South[2*], Craig Obergfell[3], & Rachel O'Neill[3]

[1]Program in Ecology, Evolutionary Biology, & Behavior, The Graduate Center, City University of New York (CUNY), New York, NY, USA

[2] Louis Calder Center-Biological Field Station & Department of Biology, Fordham University, Armonk, NY, USA

[3] Molecular & Cell Biology, University of Connecticut, Storrs, CT, USA

Corresponding author: Jason Munshi-South

E-mail: jason@NYCevolution.org

Tel: 1-646-318-6353

Fax: 1-646-660-6250

PeerJ PrePrints | https://peerj.com/preprints/13v2/ | v2 received: 21 Aug 2013, published: 21 Aug 2013, doi: 10.7287/peerj.preprints.13v2

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, we developed transcriptomic resources for urban *P. leucopus* to examine evolutionary changes in protein-coding regions for an exemplar 'urban adapter'. We 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, we 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.

PeerJ PrePrints | https://peerj.com/preprints/13v2/ | v2 received: 21 Aug 2013, published: 21 Aug 2013, doi: 10.7287/peerj.preprints.13v2

Introduction

Urbanization dramatically alters natural habitats [1], 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 [2]. 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 [3], 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 [4], and the development of next generation sequencing (NGS) has facilitated the generation of genomic resources for uncharacterized species in natural environments [5–7]. 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 [4,8,9].

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 [10–12]. 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 [1,13,14]. Introductions of new predators and competitors alter ecological interactions [15], and new or more abundant parasites or pathogens influence immune system processes [16]. 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 [17–19]. Recent studies provide some evidence of local adaptation and rapid evolution in urban patches. Using a candidate gene approach, Mueller et al. [20] 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 [21] found rapid adaptation to roadside breeding pond conditions in the salamander, *Ambystoma maculatum*, and Cheptou et al. [22] 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 [23].

Peromyscus spp. are an emerging model system for examining evolution in wild populations [24-26], 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 [27–31]. 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 [26] and adaptive variation in pelage color on light-colored soil substrates [25, 32, 33]. 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 [33,34]. The smallest patches in NYC often contain the highest population densities of white-footed mice [35], most likely due to ecological release and obstacles to dispersal [36,37]. Consistently elevated population density in urban patches compared to surrounding rural populations is predicted to result in densitydependent selective pressures on traits related to immunology, intraspecific competition, and male-male competition for mating opportunities, among others [38,39].

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 [40,41]. This genetic structure contrasts with weak differentiation reported for *Peromyscus* spp. at regional scales [42], or even between populations isolated on different islands for thousands of generations [43,44]. 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 we 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 [45,46]. 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 [47]. By analyzing SNPs within thousands of transcripts, it is feasible to identify candidate genes underlying rapid divergence of populations in novel environments [5, 47-49]. Certain statistical approaches, such as the ratio between non-synonymous and synonymous (p_N/p_S) substitutions, can be applied to pooled transcriptome data to identify potential signatures of selection between isolated populations [23,50,51]. If positive selection is acting on a codon, then non-synonymous mutations should be more common than under neutral expectations [52,53].

Here, we describe the results of *de novo* transcriptome sequencing, annotation, SNP discovery, and outlier scans for selection among urban and rural white-footed mouse populations. We 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. We then identified coding region SNPs between urban and rural populations, and scanned this dataset for signatures of positive selection using p_N/p_S between populations and McDonald-Kreitman tests between multiple species. We 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.

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 we 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, cilium (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 $p_N/p_S = 0.28 \pm 0.56$). 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. [23] found that 0.4% of genes in their *Pomacea canaliculata* dataset were positively selected, Renaut et al. [54] reported 0.5% in *Coregonus clupeaformis*, and Wang et al [55] 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 [56,57].

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. 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). We also identified *solute carrier organic anion transporter family member 1A5* (CP_contig01204, p_N/p_S = 1.55), a gene 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 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 *Translocation protein SEC62, Histone H1-like protein in spermatids 1, Aldo-keto reductase family 1 member C12, Fibrinogen alpha chain, Solute carrier organic anion transporter family member 1A5*, and *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. *P. leucopus* was compared to *R. norvegicus*, and *C. griseus* when *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 [58] and is especially useful for organisms without extensive genomic resources like *P. leucopus* [51,54,59–61]. We compared the relative merits of two established long-read assembly programs, CAP3 and Newbler, for assembling our transcriptomes [60, 61]. Despite the substantially fewer megabases per run generated by 454 FLX+ compared to Illumina or SOLiD sequencing [62], we still ran into computational limitations during assembly when using options for cDNA sequence. Similar to Cahais et al. [63], we 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 [64] 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. [65]; N50 = 1,333 bp in *Silene vulgaris*, Sloan et al. [51]; N50 = 1,588 bp in *Spalax galili*, Malik et al. [66]; and N50 = 854 bp in *Arctocephalus gazella,* Hoffman & Nichols [67].

We 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, we used a new protocol from Roche that excluded normalization of libraries. Surprisingly, we 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 [68,69], but normalization does not disproportionately influence gene discovery when enough sequencing coverage is achieved [70]. We found that read coverage per transcript increased for our non-normalized plates compared to the normalized pilot plate. However, Ekblom et al. [71] suggest that differences in technologies and sequencing effort may ultimately affect comparisons between normalized and non-normalized cDNA libraries, and any differences we 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 [72]. Deep divergence and high rates of chromosome evolution across these lineages [73] may have affected the percentage of identified homologous gene transcripts. Ramsdell et al. [74] 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 [75], 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* [76] and 79.7% in the plant *Silene vulgaris* [50]. These rates are much higher than some other organisms with few model relatives, such as 19.58% in a bat, *Artibeus jamaicensis*, [77], 18% in a butterfly, *Melitaea cinxia*, [59], and 29.2% in the gastropod, *Pomacea canaliculata*, [23]. Phylogenetic analyses support *Peromyscus* spp. and *Cricetulus* spp. as members of a monophyletic clade that diverged separately from *Mus* and *Rattus* [72], 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 [78], and a draft genome has also been sequenced [79]. 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 we sequenced in this study. Yang et al. [80] 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 [81]. 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. [80], and Chrast et al. [82] 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 [83] in *Mus*, and 454 sequencing of cDNA libraries from *C. griseus* identified 13,187 annotations in ovary [76]. UniGene [84] 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 [85]. 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 [76]. 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 [80,81].

SNP discovery and characterization

Without a reference genome, aligning reads to assembled transcripts and assigning mismatches as SNPs [86] is an acceptable substitute for generating sequence polymorphisms for non-model species [51,54,87]. Difficulties may persist in distinguishing true SNPs from false positives created by sequencing errors, misaligned reads, or alignment of reads to paralogous genes. Identifying true SNPs depends on assembly quality, filtering criteria of nucleotide mismatches during alignment, and statistical models used to call nucleotide variants [88]. Incorporating a probabilistic framework in SNP-calling algorithms greatly reduces false positives [89,90].

We used conservative filtering criteria when calling SNPs to minimize false positives. SAMtools [91] excels at SNP detection with low sequence coverage by comparing multiple samples simultaneously [89,90]. We 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 [92]. The thresholds we 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 [93]. 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*. We 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

We 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 [6,52,94]. This approach has recently been used to identify genes under positive or purifying selection between cichlid fish lineages in Nicaragua [56], between lake whitefish species pairs [54], and within an invasive gastropod [23]. Studies traditionally identify positive selection in genes with $p_N / p_S > 1.0$. We 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 unsampled [56,57]. The p_N / p_S index can also be used when samples have been pooled prior to sequencing [95], unlike summary statistics that rely on allele frequencies [51].

We 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. We calculated a neutrality index (NI) as $(p_N / p_S)/(d_N / d_S)$ to look at deviations from neutral expectations [96,97]. While we detected an excess of non-synonymous polymorphisms within *P. leucopus* in genes with functions including demethylation, xenobiotic metabolism, and innate immunity, we did not find evidence of positive selection between species. While these patterns could suggest purifying selection preventing the fixation of harmful mutations [98] or indicate balancing selection acting to maintain favorable alleles in different populations [26], 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 we 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). We 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 [41].

New predators, competitors, parasites, and pathogens can drive local adaptation of traits, especially those related to immunity, in novel urban environments [15,16]. We 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 [99], and positive selection is found to act on pathogen recognition genes within the complement activation pathway [100]. 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 [101], and white-footed mice in NYC may be evolving to efficiently recognize them and respond immunologically. We 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 [102], 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 [17–19]. One potential marker of increased exposure to pollutants is hypermethylation of regulatory regions of the genome [17,103–105]. 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. [106] 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* [18], and Wirgin et al. [107] also found rapid PCB resistance in tomcod from the Hudson River through positive selection. In urban to rural comparisons we 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 [108], 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 [109,110]. Patterns of divergence and positive selection have been robustly identified in *cytochrome p450* genes in natural populations of both *Mus musculus* when ingesting toxins through their diet and *Tetrahymena thermophila* exposed to toxic environments [110,111]. *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. We will differentiate between drift and selection in future work by examining genetic divergence between multiple urban and rural populations at these candidate loci and additional genes. We 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 [112]. However, current statistical tests address these issues and are generally robust in identifying positive selection [113]. In the case of a single population, $p_N/p_S > 1$ may not represent positive selection. Kryazhimskiy & Plotkin [114] 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 [114]. All samples used to calculate p_N/p_S for this study came from reproductively isolated and genetically structured populations [40]. We 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 [113].

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 we have missed many candidate genes. Additionally, such analyses do not identify adaptive variation in gene regulatory regions as opposed to transcribed cDNA [115]. Ratios such as p_N / p_S may also vary widely when there are relatively few mutations per gene [56, 108]. 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 [116,117]. 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.

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 [118]. 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 [40,41]. 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 [78] 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 [119]. However, the normalization process introduces additional steps and biases in library preparation [50], and resulted in a relatively low number of total high-quality 454 reads. 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). Individual RNA extracts were pooled by population and organ type and selected 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 [120] a long-read assembler that performs well in transcriptome assemblies [60], 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. We 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 [121]. 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, [122] and the trim-fastq.pl perl script implemented in Popoolation [93]. 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 analzyed. 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=rnax -t=dnax, [123]), 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 [124]. We 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, [89]) 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 [87-89]. In addition to the default parameters in SAMtools, we 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 [93,118,124–126].

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 [125,126]. 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_N/p_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^{-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 [93] 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_N) to synonymous (p_S) SNP substitutions (p_N/p_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 [127] 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 [128] 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) [129] to calculate the ratio of non-synonymous (p_N) to synonymous (p_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 ± 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 [130]. 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. We 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 norevegicus*, or *Mus musculus*. Orthologous sequences were codon-aligned using MACSE [131] and then scanned for evidence of recombination using a GARD analysis implemented in the Data Monkey webserver [132,133]. For McDonald-Kreitman tests, orthologous genes between *Peromyscus leucopus* and *Rattus norvegicus* or *Cricetulus griseus* were codon aligned with MACSE [131]. 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 [134]. 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: 10.5061/dryad.r8ns3).

Acknowledgements

We thank the New York State Department of Environmental Conservation, 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. We thank Paolo Cocco, Julie Sesina, and Diane Jacob for their assistance in the field and lab. Matthew

MacManes, three anonymous reviewers and the Associate Editor provided many comments that substantially improved the manuscript.

References

1. Shochat E, Warren PS, Faeth SH, McIntyre NE, Hope D (2006) From patterns to emerging processes in mechanistic urban ecology. Trends in Ecology and Evolution 21: 186–191. doi:10.1016/j.tree.2005.11.019.
2. United Nations (2011) World Urbanization Prospects The 2011 Revision. UN Department of Economic and Social Affairs'.
3. Martin LJ, Blossey B, Ellis E (2012) Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. Frontiers in Ecology and the Environment 10: 195–201. doi:10.1890/110154.
4. Pickett ST, Cadenasso ML, Grove JM, Boone CG, Groffman PM, et al. (2011) Urban ecological systems: scientific foundations and a decade of progress. Journal of Environmental Management 92: 331–362. doi:10.1016/j.jenvman.2010.08.022.
5. Rice AM, Rudh A, Ellegren H, Qvarnström A (2010) A guide to the genomics of ecological speciation in natural animal populations. Ecology Letters: 9–18. doi:10.1111/j.1461-0248.2010.01546.x.
6. Hohenlohe PA, Phillips PC, Cresko WA (2011) Using Population Genomics To Detect Selection in Natural Populations: Key Concepts and Methodological Considerations. International Journal of Plant Sciences 171: 1059–1071. doi:10.1086/656306.USING.
7. Storz J, Hoekstra H (2007) The study of adaptation and speciation in the genomic era. Journal of Mammalogy 88: 1–4.
8. White J, Antos M, Fitzsimons J, Palmer G (2005) Non-uniform bird assemblages in urban environments: the influence of streetscape vegetation. Landscape and Urban Planning 71: 123–135. doi:10.1016/j.landurbplan.2004.02.006.

9. Grimm NB, Faeth SH, Golubiewski NE, Redman CL, Wu J, et al. (2008) Global change and the ecology of cities. Science (New York, NY) 319: 756–760. doi:10.1126/science.1150195.

10. Blair RB (2001) Birds and butterflies along urban gradients in two ecoregions of the U.S. In: Biotic Homogenization. Lockwood JL, McKinney ML, editors Norwell, MA: Kluwer Academic Publishers.

11. McKinney ML (2002) Urbanization, biodiversity, and conservation. Bioscience 52: 883–890.

12. McKinney ML (2006) Urbanization as a major cause of biotic homogenization. Biological Conservation 127: 247–260. doi:10.1016/j.biocon.2005.09.005.

13. Bjorklund M, Ruiz I, Senar JC (2010) Genetic differentiation in the urban habitat: the great tits (Parus major) of the parks of Barcelona city. Biological Journal of the Linnean Society 99: 9–19. doi:10.1111/j.1095-8312.2009.01335.x.

14. Wandeler P, Funk SM, Largiadèr CR, Gloor S, Breitenmoser U (2003) The city-fox phenomenon: genetic consequences of a recent colonization of urban habitat. Molecular Ecology 12: 647–656.

15. Peluc SI, Sillett TS, Rotenberry JT, Ghalambor CK (2008) Adaptive phenotypic plasticity in an island songbird exposed to a novel predation risk. Behavioral Ecology 19: 830–835. doi:10.1093/beheco/arn033.

16. Sih A, Ferrari MCO, Harris DJ (2011) Evolution and behavioural responses to human-induced rapid environmental change. Evolutionary Applications 4: 367–387. doi:10.1111/j.1752-4571.2010.00166.x.

17. Yauk C, Polyzos A, Rowan-Carroll A, Somers CM, Godschalk RW, et al. (2008) Germ-line mutations, DNA damage, and global hypermethylation in mice exposed to particulate

air pollution in an urban/industrial location. Proceedings of the National Academy of Sciences of the United States of America 105: 605–610. doi:10.1073/pnas.0705896105.

18. Whitehead A, Triant D, Champlin D, Nacci D (2010) Comparative transcriptomics implicates mechanisms of evolved pollution tolerance in a killifish population. Molecular Ecology 19: 5186–5203. doi:10.1111/j.1365-294X.2010.04829.x.

19. Francis RA, Chadwick MA (2012) What makes a species synurbic? Applied Geography 32: 514–521. doi:10.1016/j.apgeog.2011.06.013.

20. Mueller JC, Partecke J, Hatchwell BJ, Gaston KJ, Evans KL (2013) Candidate gene polymorphisms for behavioural adaptations during urbanization in blackbirds. Molecular Ecology 22: 3629–3637. doi:10.1111/mec.12288.

21. Brady SP (2012) Road to evolution? Local adaptation to road adjacency in an amphibian (Ambystoma maculatum). Scientific Reports 2. doi:10.1038/srep00235.

22. Cheptou P-O, Carrue O, Rouifed S, Cantarel A (2008) Rapid evolution of seed dispersal in an urban environment in the weed Crepis sancta. Proceedings of the National Academy of Sciences of the United States of America 105: 3796–3799. doi:10.1073/pnas.0708446105.

23. Sun J, Wang M, Wang H, Zhang H, Zhang X, et al. (2012) De novo assembly of the transcriptome of an invasive snail and its multiple ecological applications. Molecular Ecology Resources 12: 1133–1144. doi:10.1111/1755-0998.12014.

24. Weber JN, Peterson BK, Hoekstra HE (2013) Discrete genetic modules are responsible for complex burrow evolution in Peromyscus mice. Nature 493: 402–405. doi:10.1038/nature11816.

25. Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. Science (New York, NY) 325: 1095–1098. doi:10.1126/science.1175826.

26. Storz JF, Sabatino SJ, Hoffmann FG, Gering EJ, Moriyama H, et al. (2007) The molecular basis of high-altitude adaptation in deer mice. PLoS Genetics 3: e45. doi:10.1371/journal.pgen.0030045.

27. Ungvari Z, Krasnikov BF, Csiszar A, Labinskyy N, Mukhopadhyay P, et al. (2008) Testing hypotheses of aging in long-lived mice of the genus Peromyscus: association between longevity and mitochondrial stress resistance, ROS detoxification pathways, and DNA repair efficiency. Age 30: 121–133. doi:10.1007/s11357-008-9059-y.

28. O'Neill R, Szalai G, Gibbs R, Weinstock G (1998) Sequencing the genome of Peromyscus. White paper proposal: 14.

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

30. Metzger LH (1971) Behavioral Population Regulation in the Woodmouse, Peromyscus leucopus. American Midland Naturalist 86: 434–448.

31. Wang G, Wolff JO, Vessey SH, Slade NA, Witham JW, et al. (2008) Comparative population dynamics of Peromyscus leucopus in North America: influences of climate, food, and density dependence. Population Ecology 51: 133–142. doi:10.1007/s10144-008-0094-4.

32. Linnen CR, Hoekstra HE (2009) Measuring natural selection on genotypes and phenotypes in the wild. Cold Spring Harbor Symposia on Quantitative Biology 74: 155–168. doi:10.1101/sqb.2009.74.045.

33. Mullen LM, Hoekstra HE (2008) Natural selection along an environmental gradient: a classic cline in mouse pigmentation. Evolution; International Journal of Organic Evolution 62: 1555–1570. doi:10.1111/j.1558-5646.2008.00425.x.

34. Puth LM, Burns CE (2009) New York's nature: a review of the status and trends in species richness across the metropolitan region. Diversity and Distributions 15: 12–21. doi:10.1111/j.1472-4642.2008.00499.x.

35. Ekernas LS, Mertes KJ (2007) The influence of urbanization, patch size, and habitat type on small mammal communities in the New York metropolitan region: a preliminary report. Transactions of the Linnaean Society of New York 10: 239–264.

36. Barko VA, Feldhamer GA, Nicholson MC, Davie DK (2003) Urban Habitat: a Determinant of White-Footed Mouse (Peromyscus Leucopus) Abundance in Southern Illinois. Southeastern Naturalist 2: 369–376. doi:10.1656/1528-7092(2003)002[0369:UHADOW]2.0.CO;2.

37. Nupp TE, Swihart RK (1996) Effect of forest patch area on population attributes of white-footed mice (Peromyscus leucopus) in fragmented landscapes. Canadian Journal of Zoology 74: 467–472.

38. Lankau R (2010) Rapid Evolution and Mechanisms of Species Coexistence. Annual Review of Ecology, Evolution, and Systematics 42: 335–354. doi:10.1146/annurev-ecolsys-102710-145100.

39. Lankau RA, Strauss SY (2011) Newly rare or newly common: evolutionary feedbacks through changes in population density and relative species abundance, and their management implications. Evolutionary Applications 4: 338–353. doi:10.1111/j.1752-4571.2010.00173.x.

40. Munshi-South J, Kharchenko K (2010) Rapid, pervasive genetic differentiation of urban white-footed mouse (Peromyscus leucopus) populations in New York City. Molecular Ecology 19: 4242–4254. doi:10.1111/j.1365-294X.2010.04816.x.

41. Munshi-South J (2012) Urban landscape genetics: canopy cover predicts gene flow between white-footed mouse (Peromyscus leucopus) populations in New York City. Molecular Ecology 21: 1360–1378. doi:10.1111/j.1365-294X.2012.05476.x.

42. Yang D-S, Kenagy GJ (2009) Nuclear and mitochondrial DNA reveal contrasting evolutionary processes in populations of deer mice (Peromyscus maniculatus). Molecular Ecology 18: 5115–5125. doi:10.1111/j.1365-294X.2009.04399.x.

43. Degner JF, Stout IJ, Roth JD, Parkinson CL (2007) Population genetics and conservation of the threatened southeastern beach mouse (Peromyscus polionotus niveiventris): subspecies and evolutionary units. Conservation Genetics 8: 1441–1452. doi:10.1007/s10592-007-9295-1.

44. Ozer F, Gellerman H, Ashley M V (2011) Genetic impacts of Anacapa deer mice reintroductions following rat eradication. Molecular Ecology: 3525–3539. doi:10.1111/j.1365-294X.2011.05165.x.

45. Boitard S, Schlötterer C, Nolte V, Pandey RV, Futschik A (2012) Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. Molecular Biology and Evolution 29: 2177–2186. doi:10.1093/molbev/mss090.

46. Futschik A, Schlötterer C (2010) Massively Parallel Sequencing of Pooled DNA Samples--The Next Generation of Molecular Markers. Genetics 218: 207–218. doi:10.1534/genetics.110.114397.

47. Gompert Z, Buerkle CA (2011) A hierarchical bayesian model for next-generation population genomics. Genetics 187: 903–917. doi:10.1534/genetics.110.124693.

48. Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100: 158–170. doi:10.1038/sj.hdy.6800937.

49. Bonin A (2008) Population genomics: a new generation of genome scans to bridge the gap with functional genomics. Molecular Ecology 17: 3583–3584. doi:10.1111/j.1365-294X.2008.03854.x.

50. Ungerer MC, Johnson LC, Herman MA (2008) Ecological genomics: understanding gene and genome function in the natural environment. Heredity 100: 178–183. doi:10.1038/sj.hdy.6800992.

51. Sloan DB, Keller SR, Berardi AE, Sanderson BJ, Karpovich JF, et al. (2012) De novo transcriptome assembly and polymorphism detection in the flowering plant Silene vulgaris (Caryophyllaceae). Molecular Ecology Resources 12: 333–343. doi:10.1111/j.1755-0998.2011.03079.x.

52. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–936.

53. Ellegren H (2008) Sequencing goes 454 and takes large-scale genomics into the wild. Molecular Ecology 17: 1629–1631. doi:10.1111/j.1365-294X.2008.03699.x.

54. Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae). Molecular Ecology 19 Suppl 1: 115–131. doi:10.1111/j.1365-294X.2009.04477.x.

55. Wang X-W, Zhao Q-Y, Luan J-B, Wang Y-J, Yan G-H, et al. (2012) Analysis of a native whitefly transcriptome and its sequence divergence with two invasive whitefly species. BMC Genomics 13: 529. doi:10.1186/1471-2164-13-529.

56. Elmer KR, Fan S, Gunter HM, Jones JC, Boekhoff S, et al. (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. Molecular Ecology 19: 197–211. doi:10.1111/j.1365-294X.2009.04488.x.

57. Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genetics 168: 1457–1465. doi:10.1534/genetics.104.030478.

58. Metzker ML (2010) Sequencing technologies - the next generation. Nature reviews Genetics 11: 31–46. doi:10.1038/nrg2626.

59. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, et al. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology 17: 1636–1647. doi:10.1111/j.1365-294X.2008.03666.x.

60. Meyer E, Aglyamova G V, Wang S, Buchanan-Carter J, Abrego D, et al. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics 10: 219. doi:10.1186/1471-2164-10-219.

61. Santure AW, Gratten J, Mossman JA, Sheldon BC, Slate J (2011) Characterisation of the transcriptome of a wild great tit Parus major population by next generation sequencing. BMC Genomics 12: 283. doi:10.1186/1471-2164-12-283.

62. Glenn TC (2011) Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11: 759–769. doi:10.1111/j.1755-0998.2011.03024.x.

63. Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, et al. (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Molecular Ecology Resources 12: 834–845. doi:10.1111/j.1755-0998.2012.03148.x.

64. Mundry M, Bornberg-Bauer E, Sammeth M, Feulner PGD (2012) Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. PloS One 7: e31410. doi:10.1371/journal.pone.0031410.

65. Ewen-Campen B, Shaner N, Panfilio KA, Suzuki Y, Roth S, et al. (2011) The maternal and early embryonic transcriptome of the milkweed bug Oncopeltus fasciatus. BMC Genomics 12: 61. doi:10.1186/1471-2164-12-61.

66. Malik A, Korol A, Hübner S, Hernandez AG, Thimmapuram J, et al. (2011) Transcriptome Sequencing of the Blind Subterranean Mole Rat, Spalax galili: Utility and Potential for the Discovery of Novel Evolutionary Patterns. PLoS ONE 6: e21227. doi:10.1371/journal.pone.0021227.

67. Hoffman J, Nichols H (2011) A novel approach for mining polymorphic microsatellite markers in silico. PloS ONE 6(8):e23283.

68. Christodoulou DC, Gorham JM, Herman DS (2011) Construction of normalized RNA-seq libraries for next- generation sequencing using the crab duplex-specific nuclease. Current Protocols in Molecular Biology: 1–14. doi:10.1002/0471142727.mb0412s94.Construction.

69. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12: 499–510. doi:10.1038/nrg3012.

70. Vijay N, Poelstra J, Künstner A, Wolf J (2012) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Molecular Ecology 46: 620–634. doi:10.1111/mec.12014.

71. Ekblom R, Slate J, Horsburgh GJ, Birkhead T, Burke T (2012) Comparison between Normalised and Unnormalised 454-Sequencing Libraries for Small-Scale RNA-Seq Studies. Comparative and Functional Genomics 2012: 281693. doi:10.1155/2012/281693.

72. Steppan S, Adkins R, Anderson J (2004) Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. Systematic Biology 53: 533–553. doi:10.1080/10635150490468701.

73. Mlynarski EE, Obergfell CJ, O'Neill MJ, O'Neill RJ (2010) Divergent patterns of breakpoint reuse in Muroid rodents. Mammalian genome : official journal of the International Mammalian Genome Society 21: 77–87. doi:10.1007/s00335-009-9242-1.

74. Ramsdell CM, Lewandowski A a, Glenn JLW, Vrana PB, O'Neill RJ, 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. doi:10.1186/1471-2148-8-65.

75. Ferreira de Carvalho J, Poulain J, Da Silva C, Wincker P, Michon-Coudouel S, et al. (2013) Transcriptome de novo assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species Spartina maritima and Spartina alterniflora (Poaceae). Heredity 110: 181–193. doi:10.1038/hdy.2012.76.

76. MacManes MD, Lacey EA (2012) The Social Brain: Transcriptome Assembly and Characterization of the Hippocampus from a Social Subterranean Rodent, the Colonial Tuco-Tuco (Ctenomys sociabilis). PLoS ONE 7: e45524. doi:10.1371/journal.pone.0045524.

77. Shaw TI, Srivastava A, Chou W-C, Liu L, Hawkinson A, et al. (2012) Transcriptome Sequencing and Annotation for the Jamaican Fruit Bat (Artibeus jamaicensis). PLoS ONE 7: e48472. doi:10.1371/journal.pone.0048472.

78. Becker J, Hackl M, Rupp O, Jakobi T, Schneider J, et al. (2011) Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. Journal of Biotechnology 156: 227–235. doi:10.1016/j.jbiotec.2011.09.014.

79. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, et al. (2011) The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nature Biotechnology 29: 735–741. doi:10.1038/nbt.1932.

80. Yang X, Schadt EE, Wang S, Wang H, Arnold AP, et al. (2006) Tissue-specific expression and regulation of sexually dimorphic genes in mice. Genome Research 16: 995–1004. doi:10.1101/gr.5217506.

81. Chapple R (2012) The developmental liver transcriptome of Rattus norvegicus University of Missouri.

82. Chrast R, Scott H, Papasavvas M (2000) The Mouse Brain Transcriptome by SAGE: Differences in Gene Expression between P30 Brains of the Partial Trisomy 16 Mouse Model of Down Syndrome (Ts65Dn) and Normals. Genome Research 10: 2006–2021. doi:10.1101/gr.158500.

83. Shima JE, McLean DJ, McCarrey JR, Griswold MD (2004) The murine testicular transcriptome: characterizing gene expression in the testis during the progression of spermatogenesis. Biology of Reproduction 71: 319–330. doi:10.1095/biolreprod.103.026880.

84. Pontius JU, Wagner L, Schuler GD (2003) UniGene: a unified view of the transcriptome. The NCBI Handbook: Bethesda (MD): National Center for Biotechnology Information.

85. Glenn JLW, Chen C-F, Lewandowski A, Cheng C-H, Ramsdell CM, et al. (2008) Expressed sequence tags from Peromyscus testis and placenta tissue: analysis, annotation, and utility for mapping. BMC Genomics 9: 300. doi:10.1186/1471-2164-9-300.

86. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. The Plant Journal : For Cell and Molecular Biology 51: 910–918. doi:10.1111/j.1365-313X.2007.03193.x.

87. Collins L, Biggs P, Voelckel C, Joly S (2008) An Approach to Transcriptome Analysis of Non-Model Organisms Using Short-Read Sequences. Genome Informatics 21: 3–14.

88. De Wit P, Pespeni MH, Ladner JT, Barshis DJ, Seneca F, et al. (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput

sequencing data analysis. Molecular Ecology Resources 12: 1058–1067. doi:10.1111/1755-0998.12003.

89. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12: 443–451. doi:10.1038/nrg2986.

90. Altmann A, Weber P, Bader D, Preuß M, Binder EB, et al. (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. Human Genetics: 1541–1554. doi:10.1007/s00439-012-1213-z.

91. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25: 2078–2079. doi:10.1093/bioinformatics/btp352.

92. McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2011) Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution. doi:10.1016/j.ympev.2011.12.007.

93. Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, et al. (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PloS One 6: e15925. doi:10.1371/journal.pone.0015925.

94. Oleksyk TK, Smith MW, O’Brien SJ (2010) Genome-wide scans for footprints of natural selection. Philosophical transactions of the Royal Society of London Series B, Biological sciences 365: 185–205. doi:10.1098/rstb.2009.0219.

95. Baldo L, Santos ME, Salzburger W (2011) Comparative transcriptomics of eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. Genome Biology and Evolution 3: 443–455. doi:10.1093/gbe/evr047.

96. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654. doi:10.1038/351652a0.

97. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415: 1022–1024. doi:10.1038/4151022a.

98. Andersen KG, Shylakhter I, Tabrizi S, Grossman SR, Happi CT, et al. (2012) Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. Philosophical Transactions of the Royal Society of London Series B, Biological Sciences 367: 868–877. doi:10.1098/rstb.2011.0299.

99. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, et al. (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genetics 4: e1000144. doi:10.1371/journal.pgen.1000144.

100. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, et al. (2007) Dynamic evolution of the innate immune system in Drosophila. Nature Genetics 39: 1461–1468. doi:10.1038/ng.2007.60.

101. Bradley CA, Altizer S (2007) Urbanization and the ecology of wildlife diseases. Trends in Ecology & Evolution 22: 95–102. doi:10.1016/j.tree.2006.11.001.

102. Turner LM, Chuong EB, Hoekstra HE (2008) Comparative analysis of testis protein evolution in rodents. Genetics 179: 2075–2089. doi:10.1534/genetics.107.085902.

103. Janssens TKS, Roelofs D, van Straalen NM (2009) Molecular mechanisms of heavy metal tolerance and evolution in invertebrates. Insect Science 16: 3–18. doi:10.1111/j.1744-7917.2009.00249.x.

104. Somers CM, Yauk CL, White P a, Parfett CLJ, Quinn JS (2002) Air pollution induces heritable DNA mutations. Proceedings of the National Academy of Sciences of the United States of America 99: 15904–15907. doi:10.1073/pnas.252499499.

105. Somers CM, Cooper DN (2009) Air pollution and mutations in the germline: are humans at risk? Human Genetics 125: 119–130. doi:10.1007/s00439-008-0613-6.

106. McGuire KL, Payne SG, Palmer MI, Gillikin CM, Keefe D, et al. (2013) Digging the New York City Skyline: Soil Fungal Communities in Green Roofs and City Parks. PLoS ONE 8: e58020. doi:10.1371/journal.pone.0058020.

107. Wirgin I, Roy NK, Loftus M, Chambers RC, Franks DG, et al. (2011) Mechanistic basis of resistance to PCBs in Atlantic tomcod from the Hudson River. Science (New York, NY) 331: 1322–1325. doi:10.1126/science.1197296.

108. Carroll MC (2004) The complement system in regulation of adaptive immunity. Nature immunology 5: 981–986. doi:10.1038/ni1113.

109. Su T, Ding X (2004) Regulation of the cytochrome P450 2A genes. Toxicology and Applied Pharmacology 199: 285–294. doi:10.1016/j.taap.2003.11.029.

110. Büntge A (2010) Tracing Signatures of Positive Selection in Natural Populations of the House Mouse Christian-Albrechts-Universität, Kiel.

111. Fu C, Xiong J, Miao W (2009) Genome-wide identification and characterization of cytochrome P450 monooxygenase genes in the ciliate Tetrahymena thermophila. BMC Genomics 10: 208. doi:10.1186/1471-2164-10-208.

112. Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99: 364–373. doi:10.1038/sj.hdy.6801031.

113. Zhai W, Nielsen R, Goldman N, Yang Z (2012) Looking for Darwin in Genomic Sequences--Validity and Success of Statistical Methods. Molecular Biology and Evolution 29: 2889–2893. doi:10.1093/molbev/mss104.

114. Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. PLoS Genetics 4: e1000304. doi:10.1371/journal.pgen.1000304.

115. Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. Proceedings of the National Academy of Sciences of the United States of America 104: 8605–8612. doi:10.1073/pnas.0700488104.

116. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science (New York, NY) 327: 883–886. doi:10.1126/science.1183863.

117. Li J, Li H, Jakobsson M, Li S, Sjödin P, et al. (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Molecular Ecology 28: 28–44. doi:10.1111/j.1365-294X.2011.05308.x.

118. Sikes RS, Gannon WL (2011) Guidelines of the American Society of Mammalogists for the use of wild mammals in research. Journal of Mammalogy 92: 235–253. doi:10.1644/10-MAMM-F-355.1.

119. Babik W, Stuglik M, Qi W, Kuenzli M, Kuduk K, et al. (2010) Heart transcriptome of the bank vole (Myodes glareolus): towards understanding the evolutionary variation in metabolic rate. BMC Genomics 11: 390. doi:10.1186/1471-2164-11-390.

120. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Research 9: 868–877.

121. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, et al. (2010) Galaxy: a web-based genome analysis tool for experimentalists. Current Protocols in Molecular Biology 89: 19.10.1–19.10.21. doi:10.1002/0471142727.mb1910s89.

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

123. Kent WJ (2002) BLAT---The BLAST-Like Alignment Tool. Genome Research 12: 656–664. doi:10.1101/gr.229202.

124. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 26: 589–595. doi:10.1093/bioinformatics/btp698.

125. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics (Oxford, England) 21: 3674–3676. doi:10.1093/bioinformatics/bti610.

126. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research 36: 3420–3435. doi:10.1093/nar/gkn176.

127. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) 26: 841–842. doi:10.1093/bioinformatics/btq033.

128. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 26: 2460–2461. doi:10.1093/bioinformatics/btq461.

129. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, et al. (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics, Proteomics & Bioinformatics 4: 259–263. doi:10.1016/S1672-0229(07)60007-2.

130. Montoya-Burgos JI (2011) Patterns of positive selection and neutral evolution in the protein-coding genes of Tetraodon and Takifugu. PLoS ONE 6: e24800. doi:10.1371/journal.pone.0024800.

131. Ranwez V, Harispe S, Delsuc F, Douzery EJP (2011) MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PloS One 6: e22594. doi:10.1371/journal.pone.0022594.

132. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Molecular Biology and Evolution 23: 1891–1901. doi:10.1093/molbev/msl051.

133. Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics (Oxford, England) 26: 2455–2457. doi:10.1093/bioinformatics/btq429.

134. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics (Oxford, England) 25: 1451–1452. doi:10.1093/bioinformatics/btp187.

Figure Legends

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.

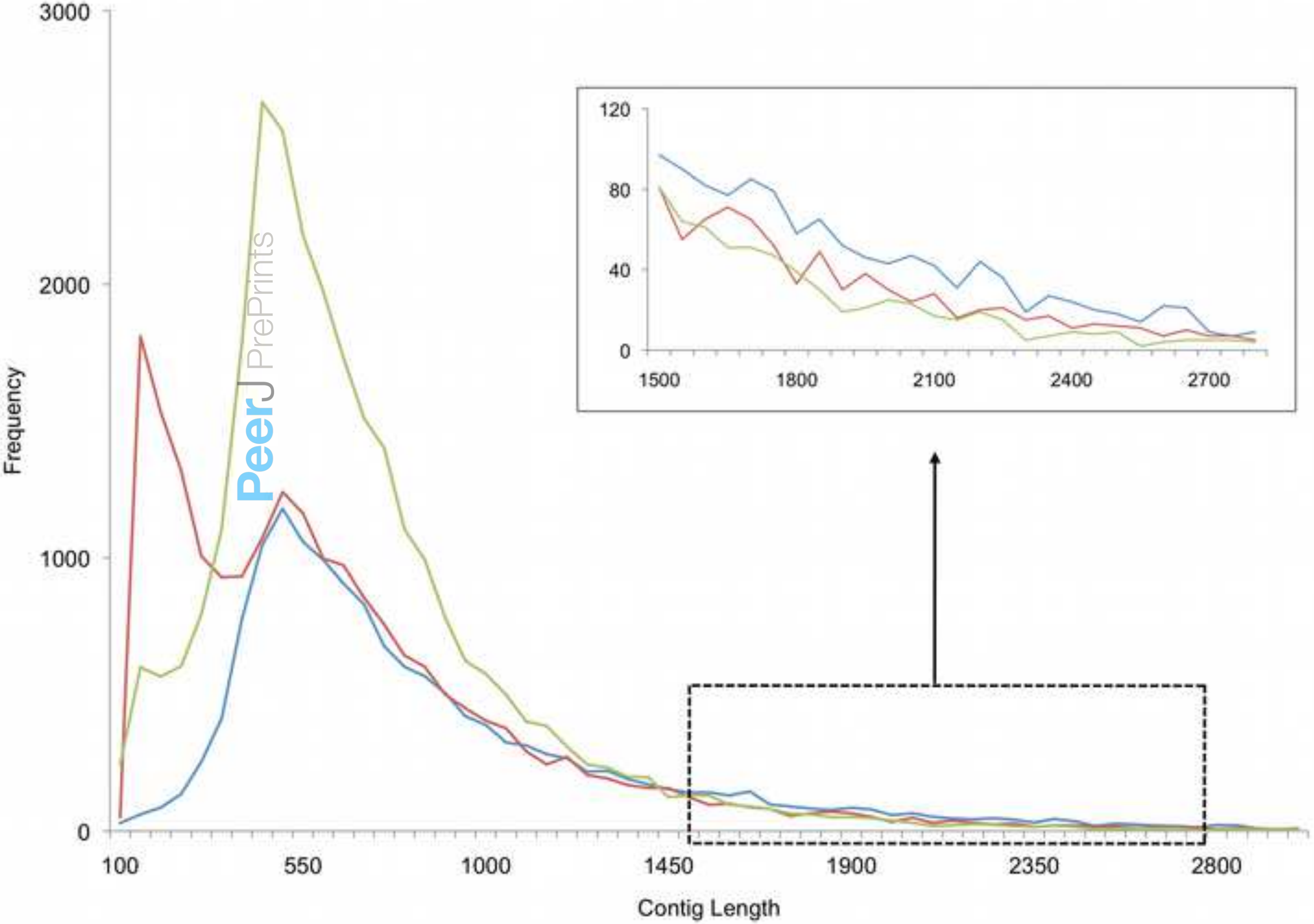
Figure 3. Annotation of final reference transcriptome. Number of assembled *P. leucopus* contigs from four different tissue types that had significant hits with known proteins on BLASTX, and GO term annotations from reference databases using Blast2Go; Blue = Total number of contigs, Red = BLASTX hits, 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.

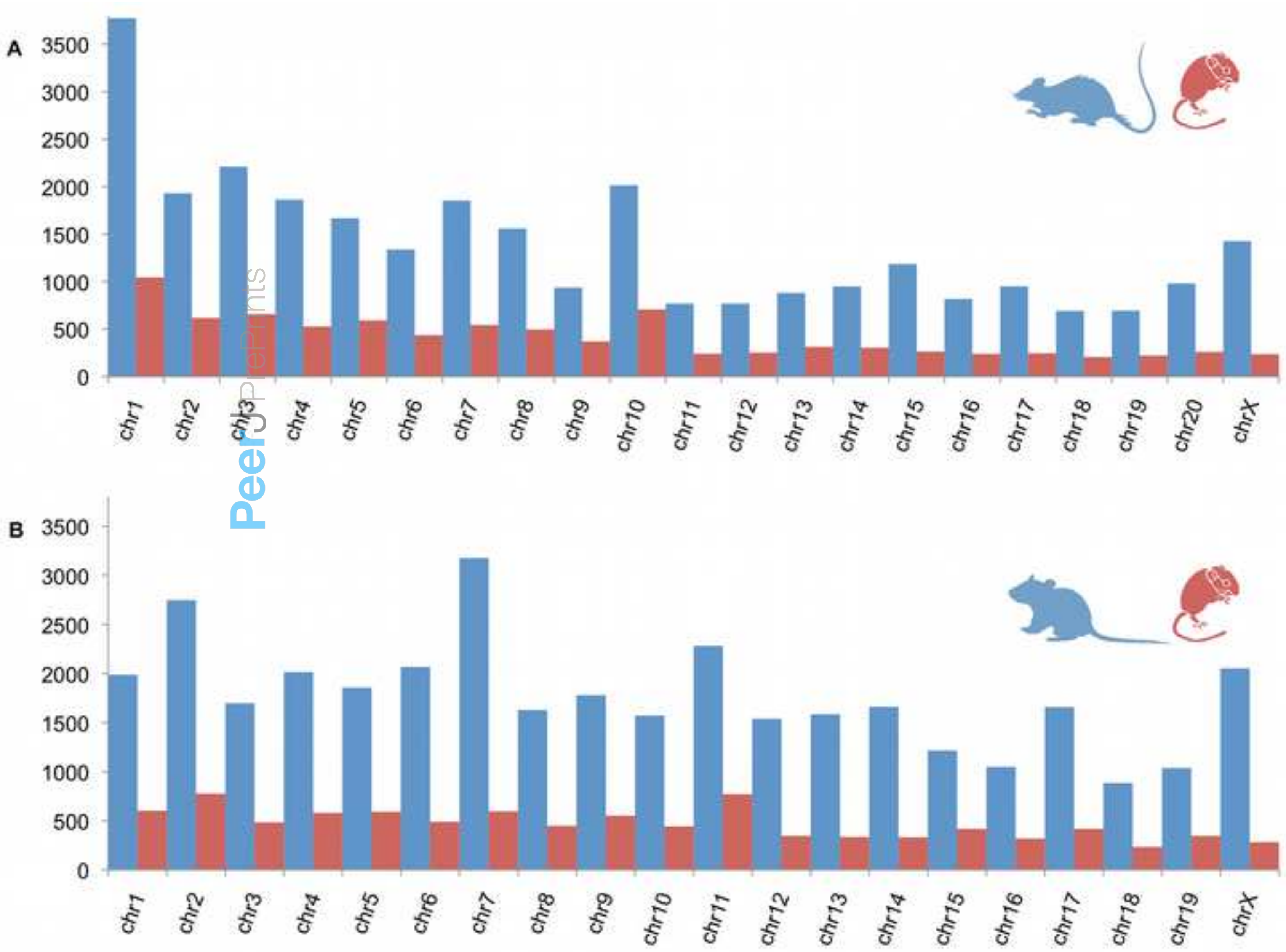
Figure 5. Non-synonymous (p_N) SNP substitutions plotted vs. synonymous (p_S) substitutions for 354 genes. Each circle represents one unique assembled contig. (a) Pairwise comparisons for all urban populations. (b) Pairwise comparisons for urban to rural populations. The dashed line

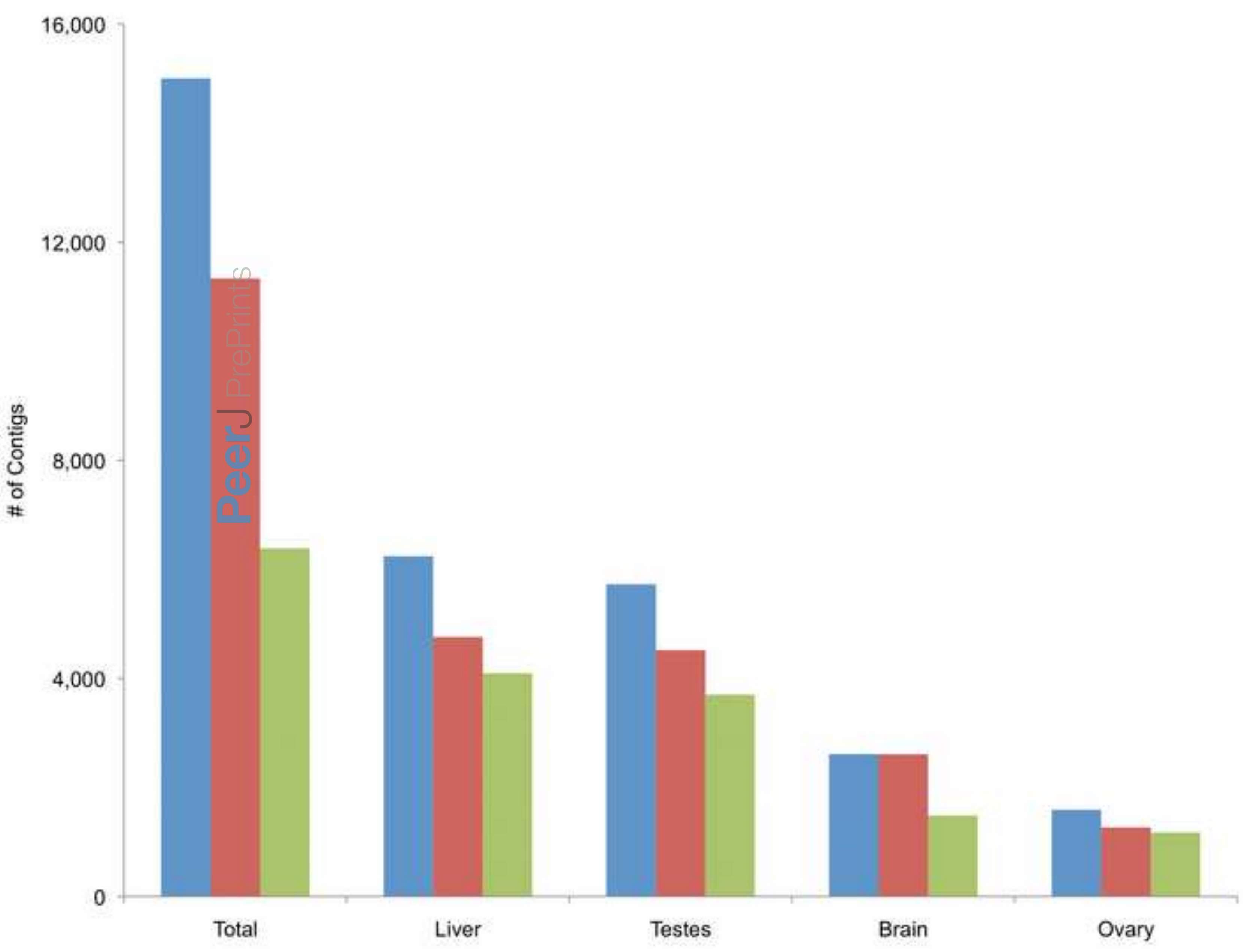
denotes $p_N / p_S = 1$, and circles above the line ($p_N / p_S > 1$) indicate candidates for positive selection. The solid line shows the slope for $p_N / p_S = 0.5$.

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. impermeable 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).

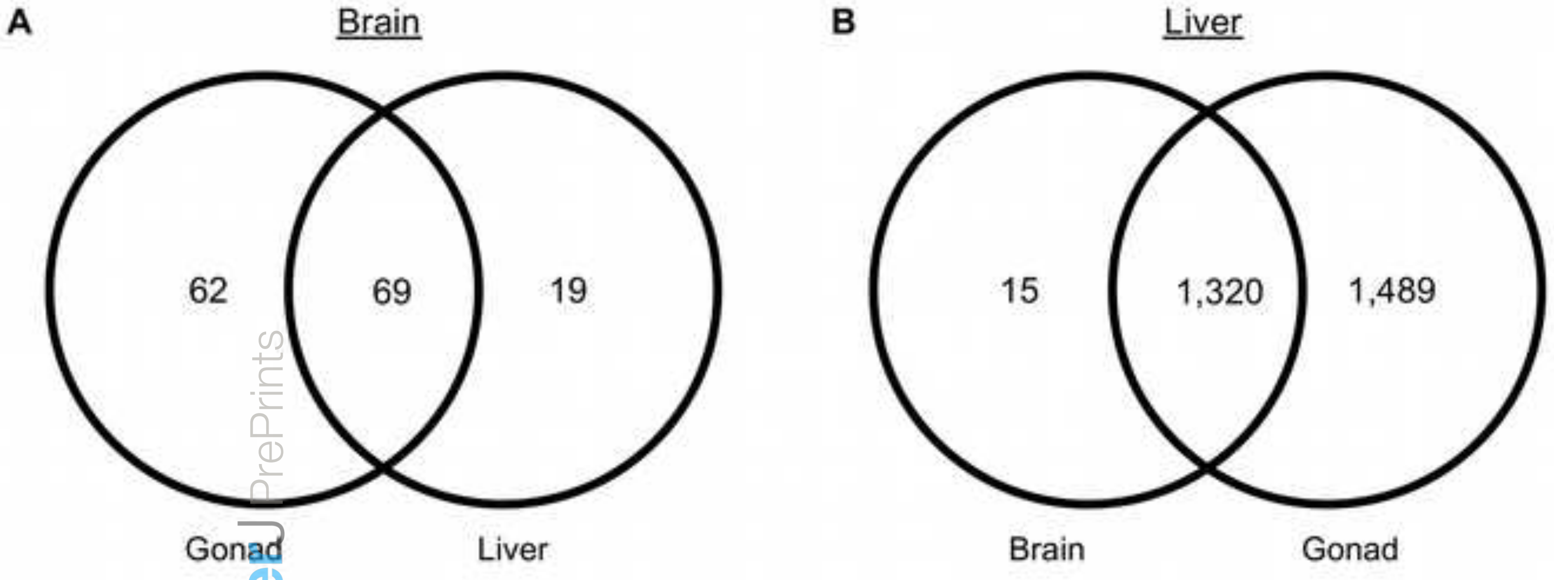


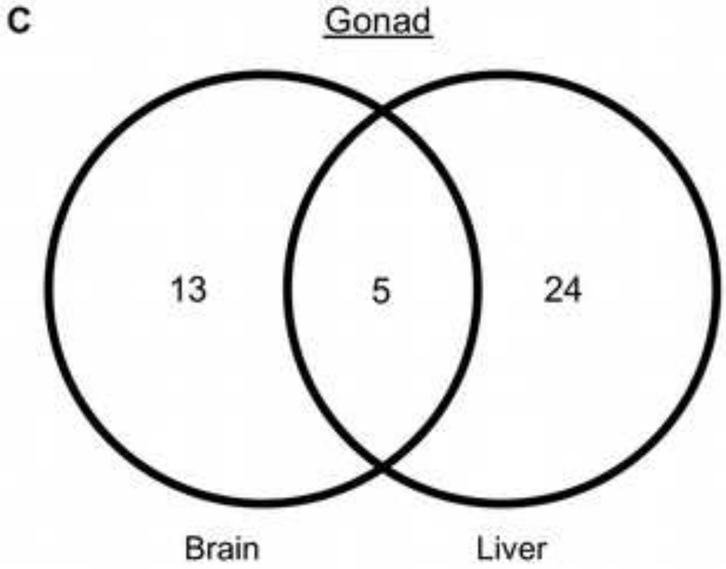
3000
2000
1000
0
Frequency
100
550
1000
1450
1900
2350
2800
Contig Length
120
80
40
0
1500
1800
2100
2400
2700


A
3500
3000
2500
2000
1500
1000
500
0
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chrX
B
3500
3000
2500
2000
1500
1000
500
0
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chrX

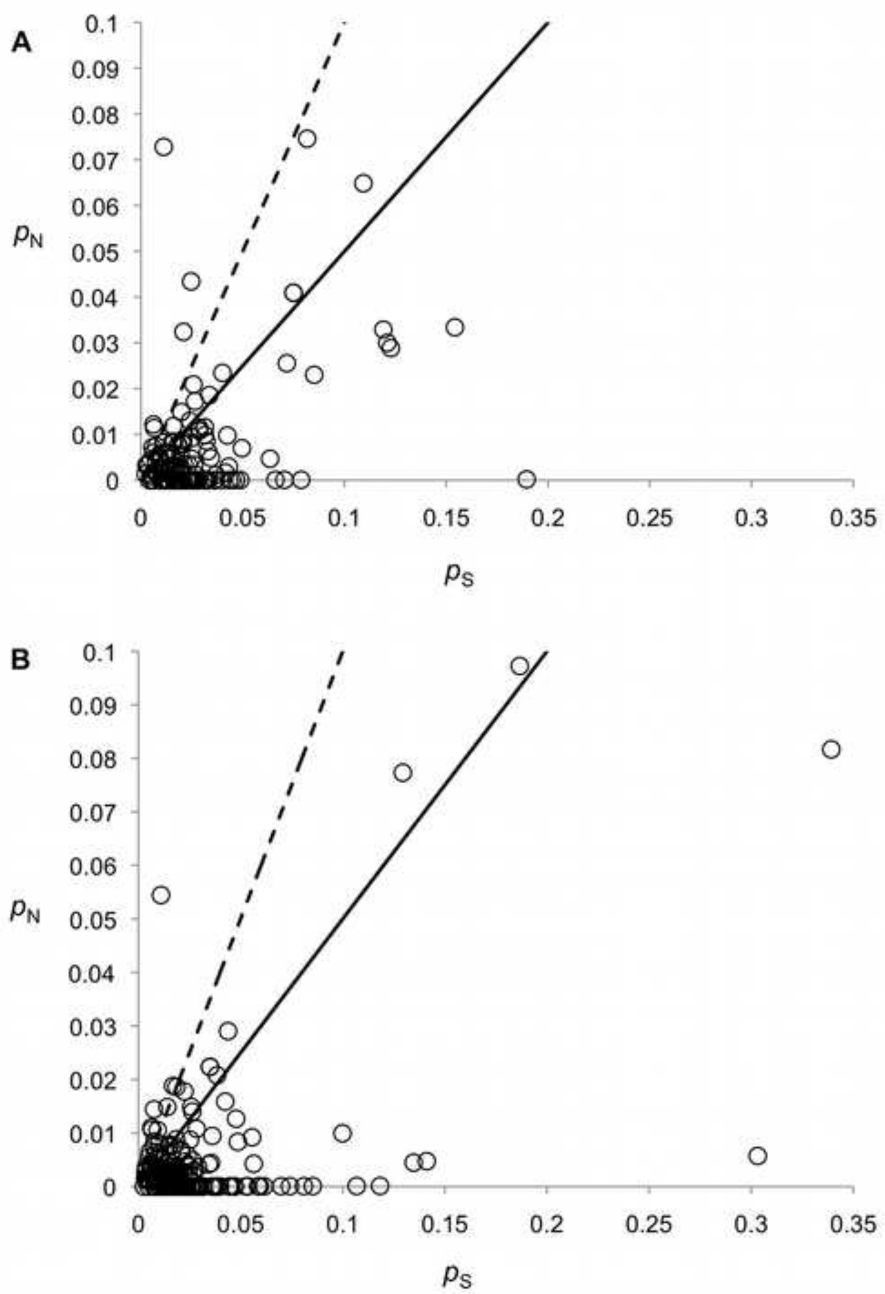


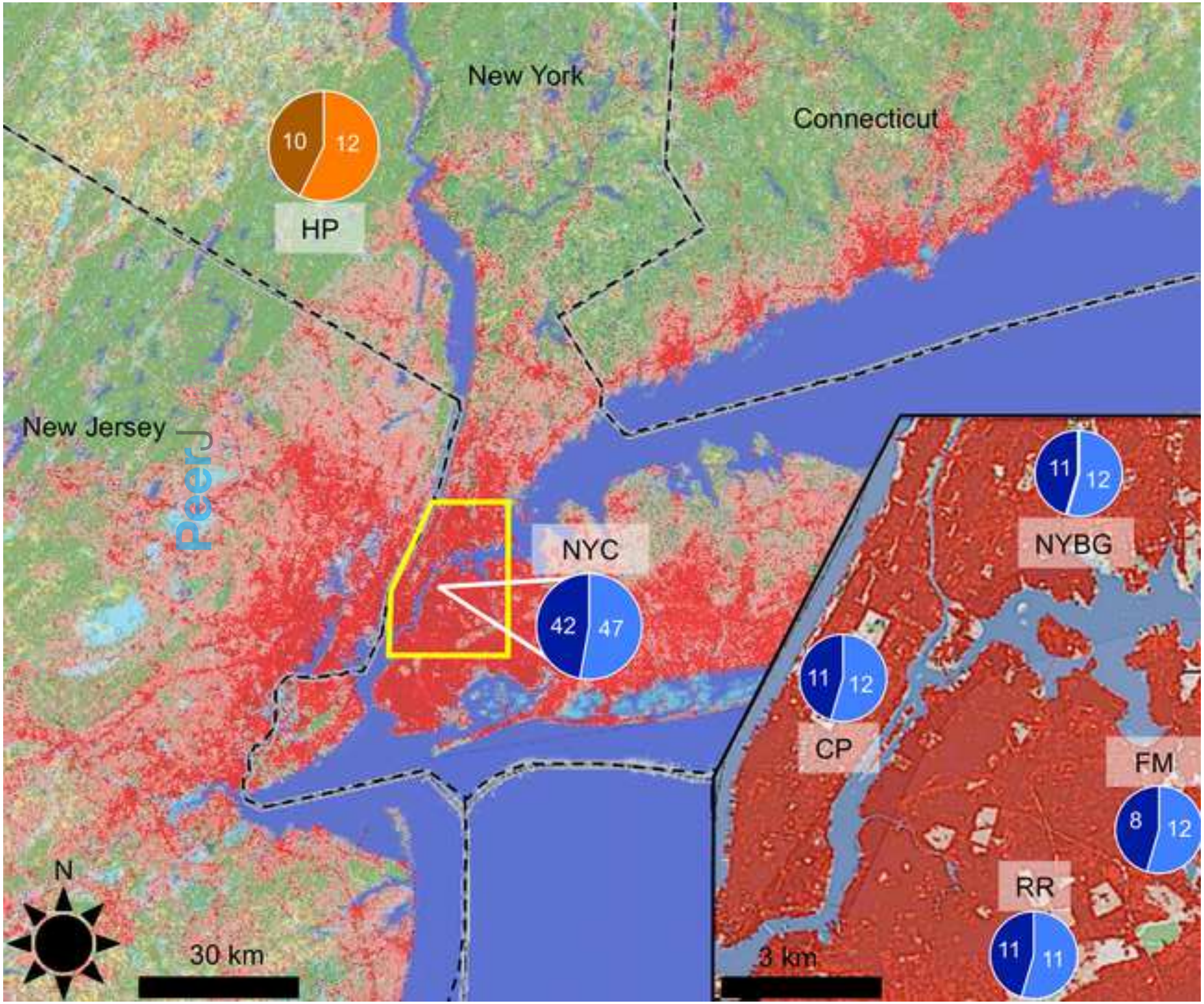
PeerJ PrePrints | https://peerj.com/preprints/13v2/ | v2 received: 21 Aug 2013, published: 21 Aug 2013, doi: 10.7287/peerj.preprints.13v2





PeerJ PrePrints | https://peerj.com/preprints/13v2/ | v2 received: 21 Aug 2013, published: 21 Aug 2013, doi: 10.7287/peerj.preprints.13v2


A
0.1
0.09
0.08
0.07
0.06
p_N
0.05
0.04
0.03
0.02
0.01
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
p_S
B
0.1
0.09
0.08
0.07
0.06
p_N
0.05
0.04
0.03
0.02
0.01
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
p_S


New York
Connecticut
10
12
HP
New Jersey
NYC
42
47
11
12
NYBG
11
12
CP
FM
8
12
RR
11
11
N
30 km
3 km

Tables

Table 1. Results of transcriptome assembly using three different approaches.

Assembly Method	No. Contigs	Mean Contig Length (bp)	Median Contig Length (bp)	N50[*]	Length (Mb)[**]
Newbler genome[a]	20,570	630 ± 504	516	830	12.95
Cap3[b]	27,497	653 ± 380	566	732	17.95
Newbler cDNA[c]	15,004 (Isotigs)	895 ± 752	683	1,039	13.42

[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.

PeerJ PrePrints | https://peerj.com/preprints/13v2/ | v2 received: 21 Aug 2013, published: 21 Aug 2013, doi: 10.7287/peerj.preprints.13v2

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

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

* = 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.

	GO term	FDR	# Sequences
Liver			
	ATP binding	5.31E-24	184
	zinc ion binding	5.93E-20	154
	transcription factor complex	3.91E-19	148
	electron carrier activity	8.53E-18	251
	structural constituent of ribosome	5.51E-15	117
	soluble fraction	2.35E-12	97
	microsome	1.53E-10	83
	protein homodimerization activity	2.75E-10	81
	oxygen binding	1.97E-09	93
	perinuclear region of cytoplasm	9.92E-09	69
	GTP binding	7.64E-08	62
	GTPase activity	2.82E-05	42
	ubiquitin-protein ligase activity	2.82E-05	42
	NADH dehydrogenase (ubiquinone) activity	5.01E-05	40
	drug binding	6.65E-05	39
	sequence-specific DNA binding	6.65E-05	39
	double-stranded DNA binding	8.90E-05	38
	mitochondrial respiratory chain complex I	1.18E-04	37
	transcription coactivator activity	1.18E-04	37
	catalytic step 2 spliceosome	1.58E-04	36
Brain			
	protein complex	1.27E-06	569
	plasma membrane	4.30E-92	567
	signal transduction	2.15E-39	525
	cytosol	1.79E-08	411
	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 $p_N / p_S > 1$.

	Sequence name	p_N / p_S	Gene name	Gene function
Pairwise Urban:Rural Comparisons				
	HP_contig01773	1.01	Translocation protein SEC62	Post-translational protein translocation into the endoplasmic reticulum; plasma membrane protein
	HP_contig02632	1.05	39S ribosomal protein L51	Part of mitochondrial ribosomal large subunit (39S); involved in protein translation
	HP_contig02656	1.07	Histone H1-like protein in spermatids 1	Transcriptional regulation and / or chromatin remodeling through DNA binding during spermatogenesis
	HP_contig01778	1.12	PHD finger protein 8	Removal of methyl groups from histones
	HP_contig01919	1.18	Aldo-keto reductase family 1, member C12	Xenobiotic metabolism; oxidation-reduction process
	HP_contig00870	1.74	Camello-like 1	Metabolic process; mitochondrial inner membrane protein
	HP_contig01783	1.89	Cytochrome P450 2A15	Metabolic process; testosterone 7a-hydroxylase activity
Pairwise Urban:Urban Comparisons				
	CP_contig00473	1.23	Fibrinogen alpha chain	Glycoprotein circulating in the blood; functions in blood coagulation and part of the most abundant component of blood clots
	CP_contig01204	1.55	Solute carrier organic anion transporter family member 1A5	Membrane protein; transports hormones; facilitates intestinal absorption of bile acids and renal uptake of indoxyl sulfate
	CP_contig00256	1.76	Serine protease inhibitor a3c	Bind to proteases and inhibit proteolysis; often involved in blood coagulation and inflammation
	CP_contig00748	1.97	Alpha-1-acid glycoprotein 1	Transport protein in the blood stream; binds and distributes synthetic drugs throughout body; modulates innate immune response

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.

	Polymorphisms			Divergence				
Gene Name	Non-synonymous (*P*n)	Synonymous (*Ps*)	Ratio (*Pn*/*Ps*)	Non-synonymous (*D*n)	Synonymous (*D*s)	Ratio (*D*n/*D*s)	Neutrality Index	P-value
Translocation protein SEC62	2	1	2	18	28	0.64	3.11	0.55
39S ribosomal protein L51	4	1	4	15	32	0.47	8.53	0.05
Histone H1-like protein in spermatids 1	2	1	2	20	12	1.67	1.20	1.00
PHD finger protein 8	9	3	3	36	51	0.71	4.25	0.03
Aldo-keto reductase family 1, member C12	3	1	3	18	37	0.49	2.67	0.08
Camello-like 1	4	1	4	41	23	1.78	2.24	0.65
Cytochrome P450 2A15*	6	1	6	13	28	0.46	12.92	0.01
Fibrinogen alpha chain	3	1	3	101	93	1.08	2.76	0.62
Solute carrier organic anion transporter 1A5	9	3	3	21	37	0.57	5.29	0.02
Serine protease inhibitor a3c*	4	1	4	25	19	1.32	1.27	0.65
Alpha-1-acid glycoprotein 1	4	1	4	68	44	1.55	2.59	0.65

* = McDonald Kreitman test used *Cricetulus griseus*

Supporting Information

Figure S1. 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.

Figure S2. Distribution of species with the most top-hit BLASTX results in Blast2Go using the Newbler cDNA assembly as the query.

Table S1. Sequencing and assembly statistics for Newbler cDNA transcriptome assembly by tissue type and 454 sequencing plate.

Table S2. 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 S3. Candidate loci with p_N / p_S between 0.5 and 1.

PeerJ PrePrints | https://peerj.com/preprints/13v2/ | v2 received: 21 Aug 2013, published: 21 Aug 2013, doi: 10.7287/peerj.preprints.13v2