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[1,2*], Craig Obergfell[3], & Rachel O'Neill[3]

[1]Program in Ecology, Evolutionary Biology, & Behavior, The Graduate Center, City University of New York (CUNY), 365 Fifth Avenue, New York, NY 10016 USA

[2]Department of Natural Sciences, Baruch College, City University of New York (CUNY), 17 Lexington Avenue, New York, NY 10010 USA

[3]University of Connecticut, Molecular & Cell Biology, 354 Mansfield Road, Storrs, CT 06269 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/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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 the first to report candidate genes exhibiting signatures of directional selection in divergent urban ecosystems.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

Introduction

Urbanization dramatically alters natural habitats, and its speed and intensity will increase as nearly two-thirds of the world's human population is predicted to live in urban areas by 2030 (Ibanez-Alamo et al. 2010). Understanding how natural populations adapt to ecologically divergent urban habitats is thus an important and immediate goal for urban ecologists and evolutionary biologists. Few ecological and evolutionary studies are conducted in urban environments (Martin et al. 2012), but recent attitude shifts and technological advancements have removed many of the obstacles to working on urban wildlife. Multiple studies have demonstrated that urban areas are biologically diverse, productive, and viable (Pickett *et al.* 2011), and the development of next generation sequencing (NGS) has facilitated the generation of genomic resources for uncharacterized species in natural environments (Hohenlohe et al. 2010; Rice et al. 2010; Storz et al. 2007). 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 (Grimm et al. 2008; Pickett et al. 2011; White et al. 2005).

Cities typically experience a substantial decrease in biodiversity of many taxonomic groups as urban 'avoiders' disappear, accompanied by a rise in urban 'exploiters' that are primarily non-native human commensals such as pigeons or rats. Urban 'adapters' are native species that favor disturbed edge habitats such as urban forest fragments, relying on a combination of wild-growing and human-derived resources (Blair 2001; McKinney 2002, 2006). This last group is of primary interest for examining genetic signatures of recent evolutionary change in novel urban environments. Severe habitat fragmentation is one of the primary impacts of urbanization and often leads to genetic differentiation between populations (Wandeler et al. 2003; Shochat et al. 2006; Bjorklund et al. 2010). Introductions of new predators and

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

64 competitors alter ecological interactions (Peluc et al. 2008), and new or more abundant parasites
65 or pathogens influence immune system processes (Sih et al. 2011). Air, water, and soil pollution
66 typically increase in local urban ecosystems, and selection may favor previously-rare genetic
67 variants that more efficiently process these toxins (Francis & Chadwick, 2012; Whitehead et al.
68 2010; Yauk et al. 2008). Recent studies provide some evidence of local adaptation and rapid
69 evolution in urban patches. Brady (2012) found rapid adaptation to roadside breeding pond
70 conditions in the salamander, *Ambystoma maculatum*, and the weed, *Crepis sancta*, exhibited a
71 heritable increase in production of non-dispersing seeds over only 5-12 generations in extremely
72 fragmented urban tree pits (Cheptou *et al.* 2008). The genetic architecture of the phenotypes
73 under selection has not been described for either of these urban 'adapters', but outlier scans of
74 transcriptome sequence datasets are one promising approach (Sun *et al.* 2012).
75 *Peromyscus* spp. are an emerging model system for examining evolution in wild
76 populations (Storz et al. 2007; Linnen et al. 2009; Weber et al. 2013), but large-scale genomic
77 resources are not yet widely available. The genus contains the most widespread and abundant
78 small mammals in North America, and *Peromyscus* research on population ecology, adaptation,
79 aging, and disease has a long, productive history (Metzger 1971; O'Neill *et al.* 1998; Vessey &
80 Vessey 2007; Wang *et al.* 2008). An increasing number of studies have demonstrated that
81 *Peromyscus* spp. rapidly (i.e. in several hundreds to thousands of generations) adapt to divergent
82 environments. These examples include adaptation to hypoxia in high altitude environments
83 (Storz et al. 2007) and adaptive variation in pelage color on light-colored soil substrates (Linnen
84 & Hoekstra, 2009; Linnen et al. 2009; Mullen & Hoekstra, 2008). *P. leucopus* is the sole
85 *Peromyscus* spp. in New York City (J. Munshi-South, unpublished data), where it occupies most
86 small patches of secondary forest, shrublands, and meadows within NYC parklands (Puth &
87 Burns 2009; Munshi-South & Kharchenko 2010). The smallest patches in NYC often contain the
88 highest population densities of white-footed mice (Ekernas & Mertes 2007), most likely due to

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

ecological release and obstacles to dispersal (Nupp & Swihart 1996; Barko *et al.* 2003). Consistently elevated population density in urban patches compared to surrounding rural populations is predicted to result in density-dependent selective pressures on traits related to immunology, intraspecific competition, and male-male competition for mating opportunities, among others (Lankau 2010; Lankau & Strauss 2011).

White-footed mouse populations in NYC exhibit high levels of heterozygosity and allelic diversity at neutral loci within populations, but genetic differentiation and low migration rates between populations (Munshi-South & Kharchenko 2010; Munshi-South 2012). This genetic structure contrasts with weak differentiation reported for *Peromyscus* spp. at regional scales (Yang & Kenagy, 2009), or even between populations isolated on different islands for thousands of generations (Degner et al. 2007; Ozer et al. 2011). High genetic diversity within and low to nonexistent migration between NYC populations suggests that selection can operate efficiently within these 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 (Boitard et al. 2012; Futschik & Schlötterer, 2010). While pooling results in the loss of genetic information from individuals, the ability to identify SNPs in a population increases due to the inclusion of multiple individuals in the pool (Gompert & Buerkle 2011). By analyzing SNPs within thousands of transcripts, it is feasible to identify candidate genes underlying rapid divergence of populations in novel environments (Bonin, 2008; Rice et al. 2010; Stinchcombe & Hoekstra, 2008; Ungerer et al. 2008). 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 populations (Sloan et al. 2012; Sun et al.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

2012). If positive selection is acting on a codon, then non-synonymous mutations should be more common than under neutral expectations (Ellegren, 2008; Nielsen & Yang, 1998).

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. 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.

Materials and Methods

Ethics statement

All animal procedures were approved by the Institutional Animal Care and Use Committee at Brooklyn College, CUNY (Protocol No. 247), and adhered to the Guidelines of the American Society of Mammalogists for the Use of Wild Mammals in Research (Sikes & Gannon 2011).

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. 1). 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

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

(Munshi-South & Kharchenko 2010; Munshi-South 2012). The rural site, Harriman State Park located ~68 km north of Central Park, is one of the largest contiguous protected areas nearby and the most likely representative of a non-urban population of mice in proximity to NYC. Mice were trapped over a period of 1-3 nights at each site using four 7x7 transects of 3"x3"x9" Sherman live traps. Mice were killed by cervical dislocation and immediately dissected in the field. Livers, gonads and brains were extracted, rinsed with PBS to remove any debris from the surface of the tissue, and immediately placed in RNALater® (Ambion Inc., Austin, TX) on ice before transport and storage at -80°C. These tissue types were chosen for initial analysis due to their wide range of expressed gene transcripts (Yang et al 2006) and potential roles in adaptation to urban conditions.

RNA extraction and cDNA library preparation

Total RNA was extracted and cDNA libraries were pooled for all five populations for four multiplexed plates of 454 sequencing. The first plate of sequencing was normalized to produce equalized concentrations of all transcripts present, potentially allowing enhanced gene discovery and greater overall coverage of the transcriptome (Babik *et al.* 2010). However, the normalization process introduces additional steps and biases in library preparation (Sloan et al. 2012), 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

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

using the Trimmer-Direct cDNA normalization kit (Evrogen, Moscow, Russia). Then, normalized cDNA pools were sequenced with multiplex identifiers using standard 454 FLX Titanium protocols. This pilot plate contained cDNA pools for Harriman State Park and Flushing Meadows-Willow Lake.

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

Transcriptome assembly

Two methods were used to assemble the best transcriptome from all four 454 plates: Cap3 (Huang & Madan, 1999) a long-read assembler that performs well in transcriptome assemblies (Cahais et al. 2012), and Roche's proprietary software, Newbler (Version 2.5.3), that was designed specifically for assembling 454 sequencing reads with additional features for cDNA sequence. Newbler's cDNA options assemble reads into contigs, followed by assembly into larger 'isotigs' representing alternatively-spliced transcripts. Isotigs are then clustered into larger 'isogroups' representing full-length genes. Transcriptome assembly was attempted with the full

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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 (Blankenberg *et al.* 2010). Before the cDNA assembly, nucleotides with poor quality scores, primer sequences, and long poly(A) tails were removed using cutadapt (Version 1.2.1 2012, Martin, 2011) and the trim-fastq.pl perl script implemented in Popoolation (Kofler *et al.* 2011). The filtered fastq files were then used as input for Cap3 or Newbler with the cDNA assembly option, using default parameters for both assemblies. These assemblies (1. genome assembly with Newbler, 2. cDNA assembly with Newbler, and 3. cDNA assembly with Cap3) were compared to identify the best full reference transcriptome for downstream analysis.

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

Alignment to model rodent genomes

Peromyscus assemblies were initially characterized and annotated by performing two separate analyses using *Mus musculus* and *Rattus norvegicus* genomic resources. The first analysis was used to determine the number of likely genes in each assembly. BLASTN searches were

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

214 performed against *Mus musculus* (NCBI Annotation Release 103) and *Rattus norvegicus* (NCBI
215 build 5.1) cDNA reference libraries downloaded from NCBI. BLASTN matches were considered
216 significant when sequence identity was greater than 80%, alignment length was at least 50% of
217 the total length of either the query or subject sequence, and the *e*-value was less than 10^{-5}. While
218 significant, these hits may not be ideal for population genomic analyses due to inclusion of
219 paralogous gene matches, matches between multi-gene families, and false positive orthologous
220 gene matches. In order to identify individual isotigs representing a single gene with known
221 function useful for statistical analysis, BLASTN results were further filtered by including query
222 hits that matched only one subject ID (i.e. gene) and *vice versa*. These contigs were annotated as
223 'Gene Candidates'.
224 The distribution of *P. leucopus* isotigs across model rodent genomes was analzyed. All *P.*
225 *leucopus* isotigs were mapped to chromosomes in the *Mus* (GRCm38) and *Rattus* (RGSC 5.0)
226 reference genomes. Default BLAT parameters were used with an exception for aligning mRNA
227 to genomes across species (-q=rnax -t=dnax, Kent, 2002), and best BLAT hits were parsed based
228 on percent identity and score (# match − # mismatch).
229
230 *Mapping and SNP discovery*
231 To generate a SNP library for downstream population genomic analysis, 454 reads were first
232 mapped to the Newbler cDNA assembly using the BWA-SW (http://bio-bwa.sourceforge.net/)
233 alignment algorithm for long read mapping (Li & Durbin 2010). We only used trimmed reads
234 from the final assembly, removed singletons before mapping to reduce false positive SNP calls
235 from sequencing errors or duplicate reads, and included reads with a mapping quality > 20 in
236 SAMtools. The SAM file from BWA-SW was used in the SAMtools package (v. 0.1.17, Li et al.
237 2009) to call SNPs. The SNP calling pipeline implemented in SAMtools uses base alignment
238 quality (BAQ) calculations to generate likelihoods of genotypes, can overcome low coverage by

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

using sequence information from multiple samples to call variants, and uses Bayesian inference to make SNP calls with high confidence (Altmann et al. 2012; Li et al. 2009; Nielsen et al. 2011). 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 (Kofler et al. 2011; Rubin et al. 2010; Altmann et al. 2012).

Functional annotation of transcriptomes

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

Prediction of Open Reading Frames (ORFs) and p_N/p_S calculations

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

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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 (Quinlan & Hall 2010) was used to filter contigs and generate a fasta file of putative ORFs (identified above) for each population assembly. The USEARCH (http://www.drive5.com/usearch/) clustering and alignment software for genomic datasets (Edgar 2010) was used to create pairwise alignments between all population ORFs using an *e*-value ≤ 0.001. Signatures of selection between aligned ORFs were identified using KaKs_Calculator1.2 (Zhang *et al.* 2006) 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 were included. 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. 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: XXXXXXX).

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

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

assembly (Fig. 2). 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). The Newbler cDNA assembly had the highest proportion (47%) of BLASTN hits that were characterized as '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 13,390,740 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, 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

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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. 3). 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.

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. 4). 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

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

transcriptome (2,613 isotigs) included a lower number of assembled contigs and percent GO annotation (56.8%; Fig. 4).

One-tailed Fisher's Exact tests (FDR ≤ 0.5) indicated that liver had the most GO terms that were significantly over-represented compared to the other tissue types (Table 3). 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). The full list of GO terms was examined between the tissue types and 1,320 annotations in liver were overrepresented when compared to brain and gonads, and 69 annotations in brain when compared to gonad and liver (Fig. 5). Gonads had the least number of annotations (five) commonly overrepresented when compared to brain and liver (Fig. 5). 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 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 with at least three SNPs were used to calculate p_N/p_S between each

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

pair of populations. The majority of the ORFs did not exhibit statistical signatures of positive selection (overall mean ± SE $p_N/p_S = 0.28 \pm 0.56$). From the 2,307 pairs of homologous cDNA sequences between populations that contained predicted ORFs and ≥ three SNPs, p_N/p_S values for 19 (0.8%) contigs exceeded 1.0 (Table 4, Fig. 6). Nine contigs (0.4%) exhibited p_N/p_S values > 1 in urban to urban comparisons and 11 contigs (0.5%) 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. 6); p_N/p_S greater than 0.5 is a less conservative filter for detecting positive selection, especially when using truncated ORFs (Swanson *et al.* 2004; Elmer *et al.* 2010).

Different genes showed strong ($p_N/p_S > 1$) signatures of selection when urban populations were compared to other urban populations than when urban and rural populations were compared. Candidate genes identified from the ORF pairs (i.e. $p_N/p_S > 1$) in urban to rural comparisons were related to metabolic processes (including xenobiotic metabolism), the immune system, 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, $p_N/p_S = 1.18$). Another candidate gene was found in two independent pairwise population comparisons and is involved in the alternative pathway of the innate immune response: *complement Factor B* (HP_contig01699, $p_N/p_S = 1.08$). 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. Two of these genes are involved in regulating the innate immune response, *complement factor H* (RR_contig00157, $p_N/p_S = 6.50$) and *alpha-1-acid glycoprotein*

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

1 (CP_contig00748, $p_N/p_S = 1.97$), through regulation of alternative pathway activation and modulating innate immune response while circulating in the blood, respectively. The other immune system genes are involved in blood coagulation and inflammation, *serine protease inhibitor a3c* (CP_contig00256, $p_N/p_S = 1.76$) and *fibrinogen alpha chain* (CP_contig00473, $p_N/p_S = 1.23$). We also identified *solute carrier organic anion transporter family member 1A5* (CP_contig01204, $p_N/p_S = 1.55$) that facilitates intestinal absorption of bile acids and renal uptake and excretion of uremic toxins.

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

Discussion

De novo transcriptome assembly and characterization

Compared to other NGS technologies, 454 transcriptome sequencing provides longer read lengths ideal for *de novo* assembly (Metzker 2010) and is especially useful for organisms without extensive genomic resources like *P. leucopus* (Vera *et al.* 2008; Meyer *et al.* 2009; Renaut *et al.* 2010; Santure *et al.* 2011; Sloan *et al.* 2012). We compared the relative merits of two established long-read assembly programs, CAP3 and Newbler, for assembling our transcriptomes (Mundry et al. 2012; Cahais et al. 2012). Despite the substantially fewer megabases per run generated by 454 FLX+ compared to Illumina or SOLiD sequencing (Glenn 2011), we still ran into computational limitations during assembly when using options for cDNA sequence. Similar to Cahais et al. (2012), we had the most success after compressing the raw reads into a smaller

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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 (number of long contigs, average contig length, and proportion of assembled contigs representing a single gene). Newbler performed well at assembling full-length cDNA contigs, and our results are in line with Mundry et al's (2012) findings that Newbler outperformed other assembly programs in simulated experiments. The N50 value reported here is comparable to *de novo* Newbler cDNA assemblies for other organisms: N50 = 1,735 bp in *Oncopeltus fasciatus*, Ewen-Campen et al. (2011); N50 = 1,333 bp in *Silene vulgaris*, Sloan et al. (2012); N50 = 1,588 bp in *Spalax galili*, Malik et al. (2011); and N50 = 854 bp in *Arctocephalus gazella,* Hoffman & Nichols (2011).

We sequenced samples using normalized and non-normalized cDNA pools and examined the influence each protocol had on gene discovery. At the time the libraries were prepared for the first pilot plate, Roche had not yet provided a protocol for cDNA library preparation. Following sequencing of the first normalized plate, the company released a preferred protocol excluding normalization of libraries, and we followed this preferred method for subsequent sequencing. 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 (Christodoulou *et al.* 2011; Davey *et al.* 2011), but normalization does not disproportionately influence gene discovery when enough sequencing coverage is achieved (Vijay *et al.* 2012). Normalization also reduces the read coverage per transcript and may lead to fragmented assemblies (Cahais *et al.* 2012), thus reducing the number of correctly-assembled transcripts and informative sites (SNPs) for downstream population genomic studies. We found that read coverage per transcript increased for our non-normalized plates compared to the normalized pilot plate. However, Ekblom et al.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

437 (2012) suggest that differences in technologies and sequencing effort may ultimately affect
438 comparisons between normalized and non-normalized cDNA libraries, and any differences we
439 identify may be due to different protocols used to extract RNA and prepare pooled libraries.
440
441 *Mapping to rodent genomes*
442 The mammalian laboratory models *Mus* and *Rattus* have extensively annotated genomes that
443 provide a good substitute reference for other rodent sequencing projects. We compared our
444 *Peromyscus* transcriptome to both genomes and found 9,418 (62.8% of assembled transcriptome)
445 and 8,786 (58.6%) putative homologous genes in *Mus* and *Rattus*, respectively. The New World
446 *Peromyscus* and Old World *Mus* and *Rattus* lineages last shared a common ancestor ~25 million
447 years ago (Steppan *et al.* 2004). Deep divergence and high rates of chromosome evolution
448 across these lineages (Mlynarski *et al.* 2010) may have affected the percentage of identified
449 homologous gene transcripts. Ramsdell et al. (Ramsdell *et al.* 2008) found the *Peromyscus*
450 genome to be more similar to *Rattus* than *Mus* due to an enhanced level of genome
451 rearrangement in *Mus* compared to ancestral muroids. Our results support these findings given
452 that most *Peromyscus* transcripts mapped to different chromosomes (96.1%) between *Mus* and
453 *Rattus*. Our homologous gene matches between *Peromyscus* and *Rattus* also represented a
454 higher proportion (30.1%) of total *Rattus* genes than homologous gene matches between
455 *Peromyscus* and *Mus* (25.7%). Non-homologous hits and mapping differences between
456 reference genomes may also be due to highly variable or alternatively spliced transcripts,
457 contamination by genomic DNA, or inclusion of low-quality data (Ferreira de Carvalho *et al.*
458 2013), although our assembly methods included measures to limit the influence of these artifacts.
459
460 *Functional annotation and tissue comparisons*

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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

Transcriptome studies in model rodents provide useful context for understanding how much of each tissue-specific transcriptome we sequenced in this study. Yang et al. (2006) used microarray analysis to identify 12,845 active genes in *Mus* liver, and RNA-Seq using an Illumina HiSeq 2000 on *Rattus* liver identified 7,514 known genes (Chapple 2012). Our gene discovery was between 40-60% of these previously reported liver transcriptomes, and thus we may have detected nearly half the genes expressed in the liver in white-footed mice. In brain tissue, 4,508 genes were identified in *Mus* by Yang et al. (2006), and Chrast et al. (2000) report ~4,000 genes identified by SAGE analysis 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.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

Microarray analysis of testis RNA identified up to 13,812 known genes (Shima *et al.* 2004) in *Mus*, and 454 sequencing of cDNA libraries from the biopharmaceutical CHO cell line in the closely related *C. griseus* identified 13,187 annotations in ovary (Becker et al. 2011). UniGene, a database of transcribed sequences by organism and cell type (Pontius *et al.* 2003), includes 8,946 genes for *Mus* testis, 5,285 for *Mus* ovaries, 4,355 for *Rattus* testis, and 5,093 for *Rattus* ovaries. The only cDNA library established in UniGene for *Peromyscus* spp. includes 635 putative genes from testis (Glenn *et al.* 2008). Our assembled libraries from gonad tissue fall within these ranges, and non-annotated transcripts could represent *Peromyscus*-specific genes. The total number of assembled isotigs for each tissue directly parallels the number of cDNA libraries sequenced for this study (Fig. 4). 454 sequencing produces less sequencing output, and genes transcribed at low levels might have been missed during library preparation and sequencing. Higher coverage from individual resequencing using short-read, high-throughput platforms should recover these rare transcripts. 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 (FDR ≤ 0.05) allowed us to identify annotated transcripts over-represented in one tissue compared to the others. The brain transcriptome of the social rodent, *C. sociabilis*, exhibited highly expressed genes involved with behavior and signal transduction (MacManes & Lacey 2012). Over-represented GO terms in *P. leucopus* brain tissue were related to similar major functions in the brain, including regulation of behavior, cellular signaling, actin binding, ion transport and channel activity, motor activity, and calcium ion binding. In liver, over-represented GO terms were largely dedicated to metabolic processes including ATP binding, GTP binding, NADH dehydrogenase, and electron carrier activity. There were also several GO terms related to the immune response, hematopoietic processes, and nutrient binding;

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

these annotations are supported by microarray and RNA-seq analyses of liver in mouse and rat, respectively (Yang *et al.* 2006; Chapple 2012).

SNP discovery and characterization

Without a reference genome for the study species, aligning reads to assembled transcripts and assigning mismatches as SNPs (Barbazuk *et al.* 2007) is an acceptable substitute for generating sequence polymorphisms for non-model species (Collins *et al.* 2008; Renaut *et al.* 2010; Sloan *et al.* 2012). Difficulties may persist in distinguishing true SNPs from false positives created by sequencing errors or misaligned reads. Alignment of reads to paralogous genes can also generate false SNPs, thus affecting downstream population genomic analyses. Identifying true SNPs depends on assembly quality, filtering criteria of nucleotide mismatches during alignment, and statistical models used to call nucleotide variants (De Wit *et al.* 2012). Incorporating a probabilistic framework in SNP-calling algorithms greatly reduces false positives (Nielsen *et al.* 2011; Altmann *et al.* 2012).

We could not validate our SNP calls using reference genome resources for *Peromyscus*, but used conservative filtering criteria when calling SNPs to minimize false positives. SAMtools (Li *et al.* 2009) excels at SNP detection with low sequence coverage by incorporating prior information about the probability of a SNP occurring based on simultaneous comparisons of multiple samples (Nielsen *et al.* 2011; Altmann *et al.* 2012). 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 (McCormack *et al.* 2011). 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 (Kofler *et al.* 2011) . Our SNP library represents highly confident variant calls and will serve as

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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 several candidate genes that may have experienced positive selection in urban populations of *P. leucopus*. Identifying ORFs in assembled transcriptomes and using SNPs to calculate the ratio of non-synonymous to synonymous substitutions (p_N/p_S) between populations can be a successful method for identifying the operation of natural selection on individual loci (Nielsen & Yang 1998; Oleksyk *et al.* 2010; Hohenlohe *et al.* 2011). This approach has recently been used to identify genes under positive or purifying selection between cichlid fish lineages in Nicaragua (Elmer et al. 2010), between lake whitefish species pairs (Renaut et al. 2010), and within *Pomacea canaliculata*, an invasive gastropod (Sun *et al.* 2012). 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 non-synonymous substitutions are not sampled from non-sequenced codons (Swanson *et al.* 2004; Elmer *et al.* 2010). The p_N/p_S index can also be used when samples have been pooled prior to sequencing (Baldo *et al.* 2011), unlike summary statistics that rely on minor allele frequency spectra. Pooling cDNA from multiple individuals introduces biases in allele frequencies because of variability in transcript presence and abundance (Sloan *et al.* 2012).

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

559 Many ecological changes arising from urbanization may drive local adaption to novel
560 conditions in fragmented urban populations. Based on the existing urban ecology literature, we
561 made several predictions about the types of adaptive traits present in urban habitats. Genes
562 involved in divergence of urban and rural populations of white-footed mice are likely associated
563 with quantitative traits affected by crowded (i.e. high population density) and polluted urban
564 environments (life history, longevity, reproduction, immunity, metabolism, thermoregulatory and
565 / or toxicological traits). We identified candidate genes showing strong positive selection (p_N/p_S
566 >1) that supported these predictions between urban and rural populations of mice, but also
567 between individual urban populations. The urban matrix is a strong enough barrier to dispersal
568 that white-footed mouse populations in individual city parks may experience highly localized
569 selective pressures in addition to selective pressures that are general to urban environments
570 (Munshi-South 2012).
571 New predators, competitors, parasites, and pathogens can drive local adaptation of traits,
572 especially those related to immunity, in novel urban environments (Peluc *et al.* 2008; Sih *et al.*
573 2011). We identified candidate genes involved in the innate immune system and activation of the
574 complement pathway to identify pathogens. Additionally, two candidate genes were identified in
575 comparisons of urban populations that function in blood coagulation and inflammation. The
576 innate immune system is a biochemical pathway that removes pathogens by identifying and
577 killing target cells (Kosiol *et al.* 2008), and positive selection is found to act on pathogen
578 recognition genes within the complement activation pathway (Sackton *et al.* 2007). The
579 introduction of invasive species, population growth of 'urban exploiters', and increased traffic,
580 trade, and transportation within cities can introduce large numbers of novel pathogens (Bradley
581 & Altizer 2007). Our results suggest that white-footed mice in NYC may be evolving to
582 efficiently recognize and respond immunologically to such urban pathogens. We also identified
583 several genes involved in metabolism that were divergent between populations, and a gene

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

584 expressed during spermatogenesis that was divergent between urban and rural populations.
585 Rapid evolution has been identified in reproductive proteins between *Peromyscus* spp. affecting
586 spermatogenesis, sperm competition, and sperm-egg interactions (Turner *et al.* 2008), and the
587 intensity of sperm competition and reproductive conflict may be increasing in dense *P. leucopus*
588 populations in NYC.

589 Increasing air, water, and soil pollution are all typical impacts of urbanization (Yauk *et*
590 *al.* 2008; Whitehead *et al.* 2010; Francis & Chadwick 2012). One potential marker of increased
591 exposure to pollutants is hypermethylation of regulatory regions of the genome (Somers *et al.*
592 2002; Yauk *et al.* 2008; Janssens *et al.* 2009; Somers & Cooper 2009). Positive selection may
593 also be acting on genes involved in xenobiotic metabolism. Heavy metals including mercury,
594 lead, and arsenic occur at increased concentrations within NYC park soils (S. Harris,
595 unpublished data), and McGuire et al. (2013) found lower pH and higher concentrations of heavy
596 metals in NYC parks compared to green roofs. In *Fundulus heteroclitus* from multiple polluted
597 estuaries along the Atlantic coast, genetic mechanisms of recent PCB resistance were identified
598 in multiple populations (Whitehead *et al.* 2010). Wirgin et al. (2011) also found that tomcod in
599 the Hudson River rapidly adapted to increased PCB concentrations through positive selection for
600 a two amino acid deletion that reduces binding affinity. In urban to rural comparisons we found
601 two potential toxicological candidate genes: one gene involved in metabolizing foreign chemical
602 compounds (i.e. xenobiotics), and a demethylase that removes methyl groups from histone
603 lysines.

604 Comparing candidate genes from all pairwise analyses with p_N / p_S between 0.5 and 1
605 reveals several additional patterns. Four complement factor proteins involved in the innate
606 immune system were identified: *complement factor H*, *I*, *B*, and *c3 precursor*. All proteins
607 function in the alternative pathway, which acts continuously in an organism without antibody
608 activation to clear foreign pathogens (Carroll 2004). The alternative pathway for complement-

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

609 mediated immunity has been found to be rich in positively selected genes across several
610 mammalian genomes, including *Mus* and *Rattus* (Kosiol *et al.* 2008). Cell signaling and protein
611 recognition genes within the innate immune system's response to pathogens are likely under
612 similar selective pressures as the adaptive immune response (Sackton *et al.* 2007). We also
613 identified proteins that interact with the complement system in multiple population pair analyses
614 in both urban-to-urban and urban-to-rural population comparisons. *Alpha1-acid glycoprotein*
615 was identified in four separate population comparisons and likely modulates the innate immune
616 system while circulating in the blood (Fournier *et al.* 2000). Four *cytochrome p450* genes, *2d27-*
617 *like*, *family 2 subfamily B*, *subfamily polypeptide 13*, and *2a15,* exhibited p_N/p_S between 0.5 and
618 1 in urban populations and between urban and rural populations. The *cytochrome p450* family of
619 genes plays a major role in xenobiotic metabolism, including detoxification in variable
620 environments (Su & Ding 2004; Büntge 2010). Patterns of divergence and positive selection
621 have been robustly identified in *cytochrome p450* genes in natural populations of both *Mus*
622 *musculus* ingesting toxins through their diet and *Tetrahymena thermophila* exposed to toxic
623 environments (Fu *et al.* 2009; Büntge 2010). *P. leucopus* in NYC populations may be
624 experiencing different dietary demands and exposure to pollutants, leading to selective pressures
625 on detoxifying genes like the cytochrome p450 gene family.
626 We have identified several transcriptome-wide trends of divergence between urban and
627 rural populations of white-footed mice, as well as between isolated urban populations within
628 NYC. Identification of candidates from the same gene families indicates that positive selection
629 may be acting directly on protein pathways of white-footed mice in urban populations. However,
630 inferring the function of candidate genes and phenotypes influenced by selection should be done
631 with an excess of caution after identifying statistical signatures of positive selection.
632 Purifying selection can sometimes lead to $p_N/p_S > 1$, and individual codons within a gene
633 can have an excess of non-synonymous substitutions due to random biological processes

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

(Hughes 2007). However, current statistical tests address these issues and are generally robust in identifying positive selection (Zhai *et al.* 2012). In the case of a single population, $p_N/p_S > 1$ may not represent positive selection. Kryazhimskiy & Plotkin (2008) demonstrated that the relationship between p_N/p_S and selection is radically different when samples are being compared that originated from the same population; p_N/p_S actually decreases in response to positive selection. To make inferences about selection between two samples using p_N/p_S, samples must come from reproductively isolated populations with fixed substitution differences (Kryazhimskiy & Plotkin 2008). All samples used to calculate p_N/p_S for this study came from reproductively isolated and genetically structured populations (Munshi-South & Kharchenko 2010). 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 (Zhai *et al.* 2012).

The ability of p_N/p_S to detect genes under positive selection is limited in some situations, so it is likely that we have missed many candidate genes by using this single statistic. Additionally, such analyses do not identify adaptive variation in gene regulatory regions as opposed to transcribed cDNA (Prud'homme et al 2007). Amino acid changes may also be more representative of past rather than recent or ongoing selection (Elmer et al. 2010; Hohenlohe et al 2011), and p_N/p_S ratios can vary widely when there are relatively few mutations per gene (Hughes 2007; Renaut et al 2010). 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 multiple tests of selection that provide more statistical power and higher resolution when identifying types and age of selection in single candidate genes (Grossman et al 2010; Li et al.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

2012). Our ongoing work in this system uses transcriptomic and genomic libraries from individual mice from several populations, and 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.

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.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

References

Altmann A, Weber P, Bader D *et al*. (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human Genetics*, 1541–1554.

Babik W, Stuglik M, Qi W *et al*. (2010) Heart transcriptome of the bank vole (Myodes glareolus): towards understanding the evolutionary variation in metabolic rate. *BMC Genomics*, **11**, 390.

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.

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**, 910–8.

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.

Becker J, Hackl M, Rupp O *et al*. (2011) Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. *Journal of Biotechnology*, **156**, 227–35.

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.

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

Blankenberg D, Von Kuster G, Coraor N *et al*. (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, **89**, 19.10.1–19.10.21.

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.

Bonin A (2008) Population genomics: a new generation of genome scans to bridge the gap with functional genomics. *Molecular Ecology*, **17**, 3583–4.

Bradley CA, Altizer S (2007) Urbanization and the ecology of wildlife diseases. *Trends in Ecology & Evolution*, **22**, 95–102.

Brady SP (2012) Road to evolution? Local adaptation to road adjacency in an amphibian (Ambystoma maculatum). *Scientific Reports*, **2**.

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

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

Carroll MC (2004) The complement system in regulation of adaptive immunity. *Nature Immunology*, **5**, 981–6.

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

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**, 3796–9.

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.

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.

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.

Conesa A, Götz S, García-Gómez JM *et al*. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, **21**, 3674–6.

Davey JW, Hohenlohe P a, Etter PD *et al*. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics*, **12**, 499–510.

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.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, **26**, 2460–1.

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.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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.
Ellegren H (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology*, **17**, 1629–31.
Elmer KR, Fan S, Gunter HM *et al*. (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Molecular Ecology*, **19**, 197–211.
Ewen-Campen B, Shaner N, Panfilio KA *et al*. (2011) The maternal and early embryonic transcriptome of the milkweed bug Oncopeltus fasciatus. *BMC Genomics*, **12**, 61.
Ferreira de Carvalho J, Poulain J, Da Silva C *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.
Fournier T, Medjoubi-N N, Porquet D (2000) Alpha-1-acid glycoprotein. *Biochimica et Biophysica Acta*, **1482**, 157–171.
Francis RA, Chadwick MA (2012) What makes a species synurbic? *Applied Geography*, **32**, 514–521.
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.
Futschik A, Schlötterer C (2010) Massively Parallel Sequencing of Pooled DNA Samples--The Next Generation of Molecular Markers. *Genetics*, **218**, 207–218.
Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular ecology resources*, 759–769.
Glenn JLW, Chen C-F, Lewandowski A *et al*. (2008) Expressed sequence tags from Peromyscus testis and placenta tissue: analysis, annotation, and utility for mapping. *BMC Genomics*, **9**, 300.
Gompert Z, Buerkle CA (2011) A hierarchical bayesian model for next-generation population genomics. *Genetics*, **187**, 903–17.
Götz S, García-Gómez JM, Terol J *et al*. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, **36**, 3420–35.
Grimm NB, Faeth SH, Golubiewski NE *et al*. (2008) Global change and the ecology of cities. *Science (New York, N.Y.)*, **319**, 756–60.
Grossman SR, Shylakhter I, Karlsson EK *et al*. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science (New York, N.Y.)*, **327**, 883–6.
Hoffman J, Nichols H (2011) A novel approach for mining polymorphic microsatellite markers in silico. *PloS One*, **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.
Huang X, Madan a (1999) CAP3: A DNA sequence assembly program. *Genome research*, **9**, 868–77.
Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity*, **99**, 364–73.
Ibanez-Álamo J, Soler M (2010) Does urbanization affect selective pressures and life history strategies in the common blackbird (Turdus merula L.)? *Biological Journal of the Linnean Society*, **101**, 759–766.
Janssens TKS, Roelofs D, Van Straalen NM (2009) Molecular mechanisms of heavy metal tolerance and evolution in invertebrates. *Insect Science*, **16**, 3–18.
Kent WJ (2002) BLAT---The BLAST-Like Alignment Tool. *Genome Research*, **12**, 656–664.
Kofler R, Orozco-terWengel P, De Maio N *et al*. (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS One*, **6**, e15925.
Kosiol C, Vinar T, Da Fonseca RR *et al*. (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genetics*, **4**, e1000144.
Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. *PLoS Genetics*, **4**, e1000304.
Lankau R (2010) Rapid Evolution and Mechanisms of Species Coexistence. *Annual Review of Ecology, Evolution, and Systematics*, **42**, 335–354.
Lankau R a., 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.
Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **26**, 589–95.
Li H, Handsaker B, Wysoker A *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078–9.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

Li J, Li H, Jakobsson M *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.
Linnen CR, Hoekstra HE (2009) Measuring natural selection on genotypes and phenotypes in the wild. *Cold Spring Harbor Symposia on Quantitative Biology*, **74**, 155–68.
Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. *Science (New York, N.Y.)*, **325**, 1095–8.
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). (GG de Polavieja, Ed,). *PLoS One*, **7**, e45524.
Malik A, Korol A, Hübner S *et al*. (2011) Transcriptome Sequencing of the Blind Subterranean Mole Rat, Spalax galili: Utility and Potential for the Discovery of Novel Evolutionary Patterns (P Michalak, Ed,). *PLoS One*, **6**, e21227.
Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet*, **17**, 10–12.
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.
McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2011) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*.
McGuire KL, Payne SG, Palmer MI *et al*. (2013) Digging the New York City Skyline: Soil Fungal Communities in Green Roofs and City Parks (JA Gilbert, Ed,). *PLoS One*, **8**, e58020.
McKinney ML (2002) Urbanization, biodiversity, and conservation. *Bioscience*, **52**, 883–890.
McKinney ML (2006) Urbanization as a major cause of biotic homogenization. *Biological Conservation*, **127**, 247–260.
Metzger LH (1971) Behavioral Population Regulation in the Woodmouse, Peromyscus leucopus. *American Midland Naturalist*, **86**, 434–448.
Metzker ML (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics*, **11**, 31–46.
Meyer E, Aglyamova G V, Wang S *et al*. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, **10**, 219.
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.
Mullen LM, Hoekstra HE (2008) Natural selection along an environmental gradient: a classic cline in mouse pigmentation. *Evolution; International Journal of Organic Evolution*, **62**, 1555–70.
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.
Munshi-South J (2012) Urban landscape genetics: canopy cover predicts gene flow between white-footed mouse (Peromyscus leucopus) populations in New York City. *Molecular Ecology*, **21**, 1360–1378.
Munshi-South J, Kharchenko K (2010) Rapid, pervasive genetic differentiation of urban white-footed mouse (Peromyscus leucopus) populations in New York City. *Molecular Ecology*, **19**, 4242–4254.
Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, **12**, 443–51.
Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–36.
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.
O'Neill R, Szalai G, Gibbs R, Weinstock G (1998) Sequencing the genome of Peromyscus. *White paper proposal*, 14.
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.
Ozer F, Gellerman H, Ashley M V (2011) Genetic impacts of Anacapa deer mice reintroductions following rat eradication. *Molecular Ecology*, 3525–3539.
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.
Pickett ST a, Cadenasso ML, Grove JM *et al*. (2011) Urban ecological systems: scientific foundations and a decade of progress. *Journal of Environmental Management*, **92**, 331–62.
Pontius JU, Wagner L, Schuler GD (2003) UniGene: a unified view of the transcriptome. *The NCBI Handbook: Bethesda (MD): National Center for Biotechnology Information*.
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**, 8605–12.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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.
Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, **26**, 841–2.
Ramsdell CM, Lewandowski A a, Glenn JLW *et al*. (2008) Comparative genome mapping of the deer mouse (Peromyscus maniculatus) reveals greater similarity to rat (Rattus norvegicus) than to the lab mouse (Mus musculus). *BMC Evolutionary Biology*, **8**, 65.
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**, 115–31.
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.
Sackton TB, Lazzaro BP, Schlenke TA *et al*. (2007) Dynamic evolution of the innate immune system in Drosophila. *Nature Genetics*, **39**, 1461–8.
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.
Shaw TI, Srivastava A, Chou W-C *et al*. (2012) Transcriptome Sequencing and Annotation for the Jamaican Fruit Bat (Artibeus jamaicensis) (ML Baker, Ed,). *PLoS One*, **7**, e48472.
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**, 319–30.
Shochat E, Warren PS, Faeth SH, McIntyre NE, Hope D (2006) From patterns to emerging processes in mechanistic urban ecology. *Trends in Ecology & Evolution*, **21**, 186–91.
Sih A, Ferrari MCO, Harris DJ (2011) Evolution and behavioural responses to human-induced rapid environmental change. *Evolutionary Applications*, **4**, 367–387.
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.
Sloan DB, Keller SR, Berardi AE *et al*. (2012) De novo transcriptome assembly and polymorphism detection in the flowering plant Silene vulgaris (Caryophyllaceae). *Molecular Ecology Resources*, **12**, 333–43.
Somers CM, Cooper DN (2009) Air pollution and mutations in the germline: are humans at risk? *Human Genetics*, **125**, 119–30.
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**, 15904–7.
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**, 533–53.
Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–70.
Storz J, Hoekstra H (2007) The study of adaptation and speciation in the genomic era. *Journal of Mammalogy*, **88**, 1–4.
Storz JF, Sabatino SJ, Hoffmann FG *et al*. (2007) The molecular basis of high-altitude adaptation in deer mice. *PLoS Genetics*, **3**, e45.
Su T, Ding X (2004) Regulation of the cytochrome P450 2A genes. *Toxicology and Applied Pharmacology*, **199**, 285–94.
Sun J, Wang M, Wang H *et al*. (2012) De novo assembly of the transcriptome of an invasive snail and its multiple ecological applications. *Molecular Ecology Resources*, **12**, 1133–1144.
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**, 1457–65.
Turner LM, Chuong EB, Hoekstra HE (2008) Comparative analysis of testis protein evolution in rodents. *Genetics*, **179**, 2075–89.
Ungerer MC, Johnson LC, Herman MA (2008) Ecological genomics: understanding gene and genome function in the natural environment. *Heredity*, **100**, 178–83.
Ungvari Z, Krasnikov BF, Csiszar A *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**, 121–33.
Vera JC, Wheat CW, Fescemyer HW *et al*. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–47.
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.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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.

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**, 647–56.

Wang G, Wolff JO, Vessey SH *et al*. (2008) Comparative population dynamics of Peromyscus leucopus in North America: influences of climate, food, and density dependence. *Population Ecology*, **51**, 133–142.

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

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.

Whitehead a, Triant D a, Champlin D, Nacci D (2010) Comparative transcriptomics implicates mechanisms of evolved pollution tolerance in a killifish population. *Molecular Ecology*, **19**, 5186–5203.

Wirgin I, Roy NK, Loftus M *et al*. (2011) Mechanistic basis of resistance to PCBs in Atlantic tomcod from the Hudson River. *Science (New York, N.Y.)*, **331**, 1322–5.

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

Xu X, Nagarajan H, Lewis NE *et al*. (2011) The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nature Biotechnology*, **29**, 735–41.

Yang D-S, Kenagy GJ (2009) Nuclear and mitochondrial DNA reveal contrasting evolutionary processes in populations of deer mice (Peromyscus maniculatus). *Molecular Ecology*, **18**, 5115–25.

Yang X, Schadt EE, Wang S *et al*. (2006) Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Research*, **16**, 995–1004.

Yauk C, Polyzos A, Rowan-Carroll A *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**, 605–10.

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.

Zhang Z, Li J, Zhao X-Q *et al*. (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics*, **4**, 259–63.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

Figure Legends

Figure 1. 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).

Figure 2. 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 3. 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) Blue = total number of genes per chromosome for *Mus*. Red = number of aligned *Peromyscus* isotigs per *Mus* chromosome.

Figure 4. 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.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

954

955 **Figure 5. Over-represented GO terms from pairwise tissue comparisons (FDR ≤0.05).** (a)
956 Comparison of brain transcriptome to liver and gonad. (b) Comparison of liver to brain and
957 gonad. (c) Comparison of gonad to liver and brain.

958

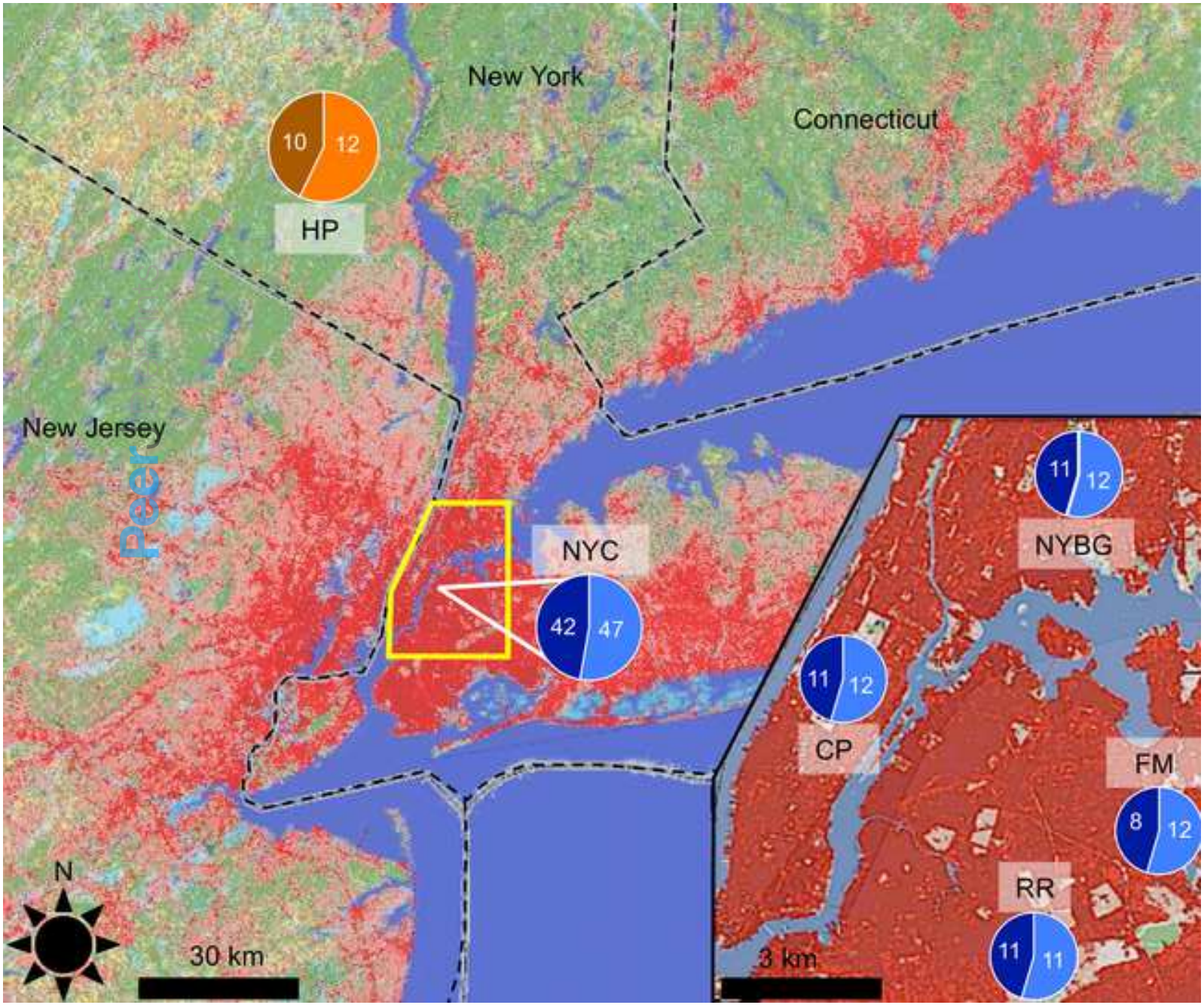
959 **Figure 6. Non-synonymous (p_N) SNP substitutions plotted vs. synonymous (p_S) substitutions**
960 **for 354 genes.** Each circle represents one unique assembled contig. (a) Pairwise comparisons for
961 all urban populations. (b) Pairwise comparisons for urban to rural populations. The dashed line
962 denotes $p_N / p_S = 1$, and circles above the line ($p_N / p_S > 1$) indicate candidates for positive
963 selection. The solid line shows the slope for $p_N / p_S = 0.5$.

964

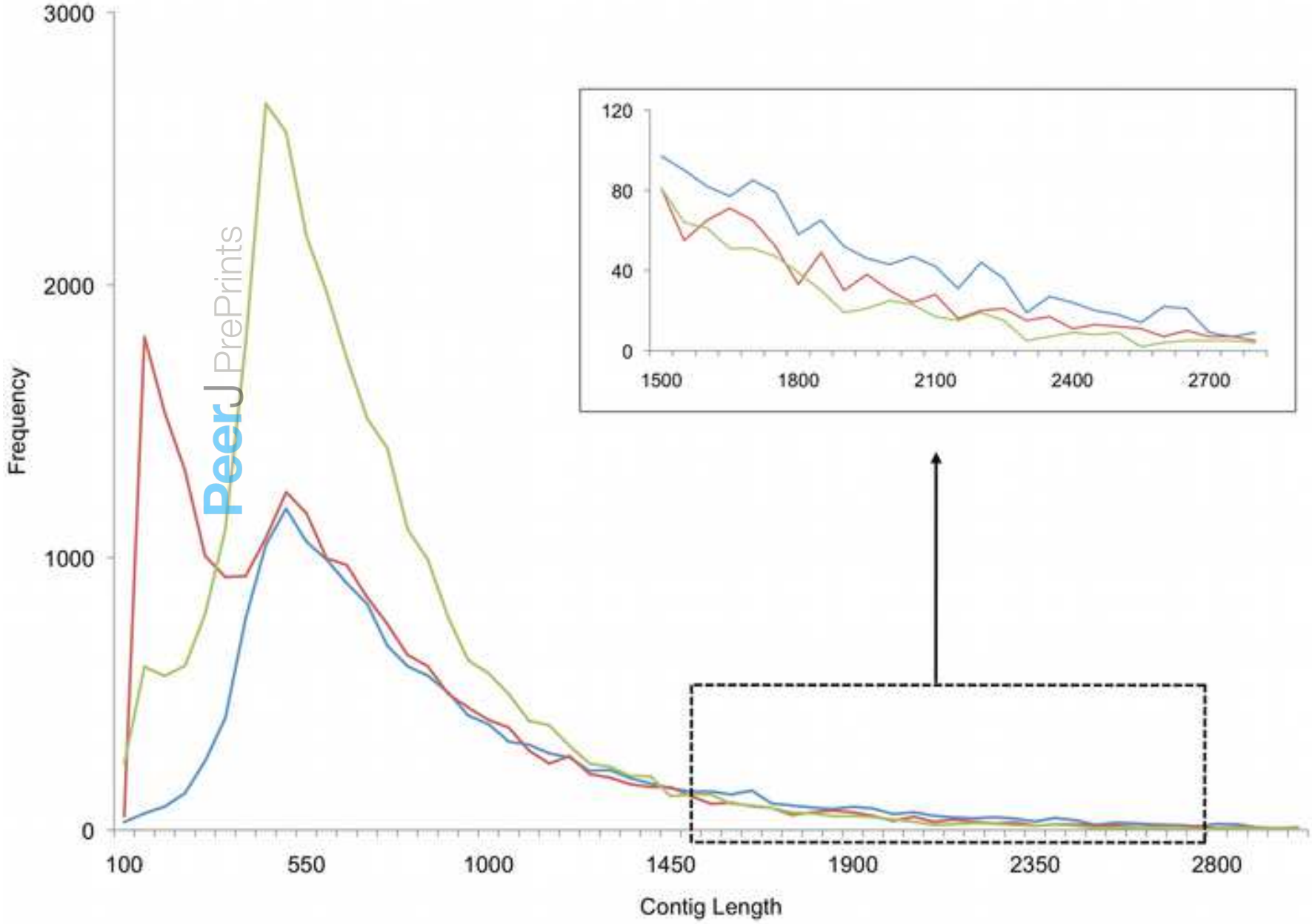
965

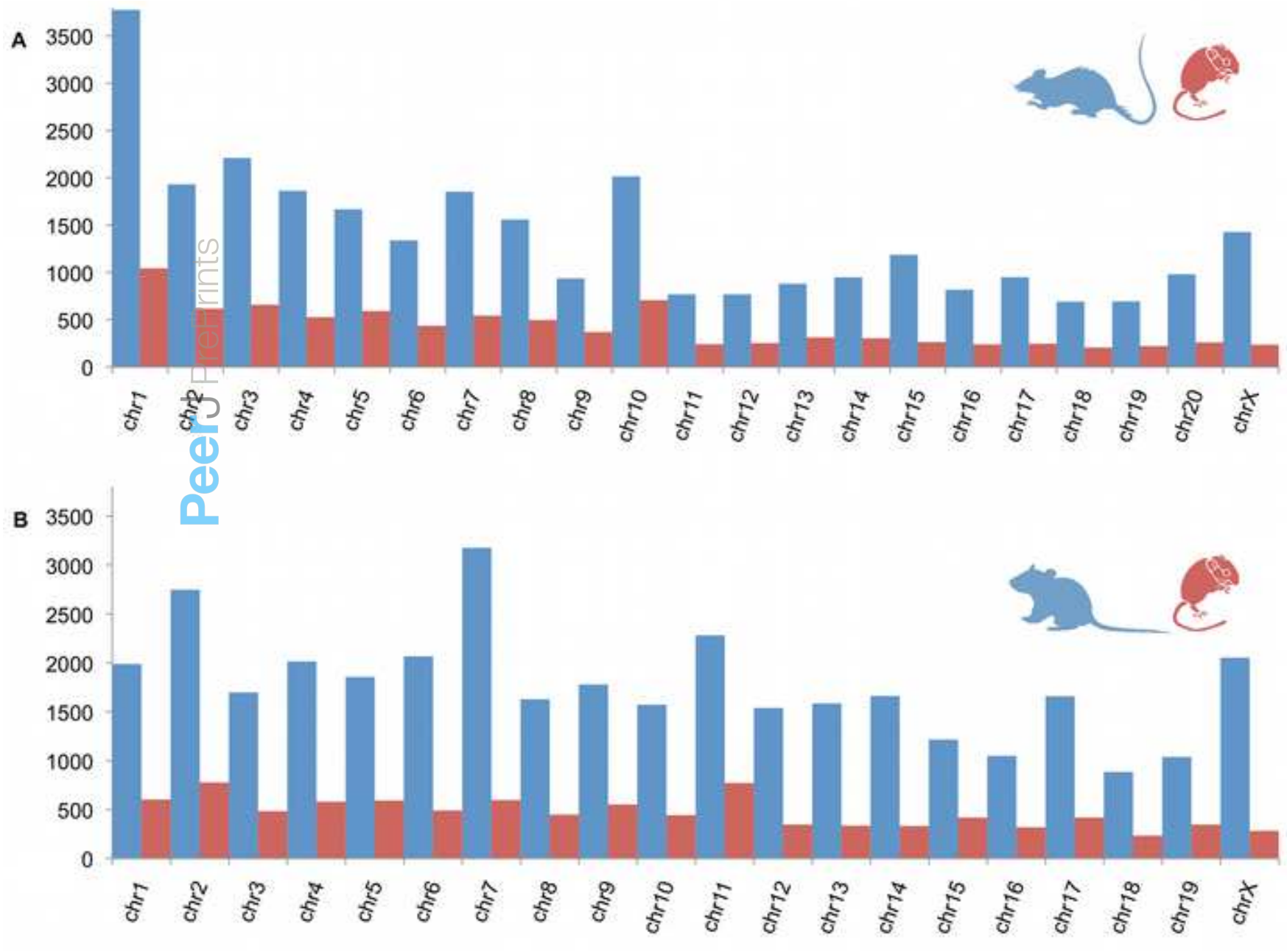
966

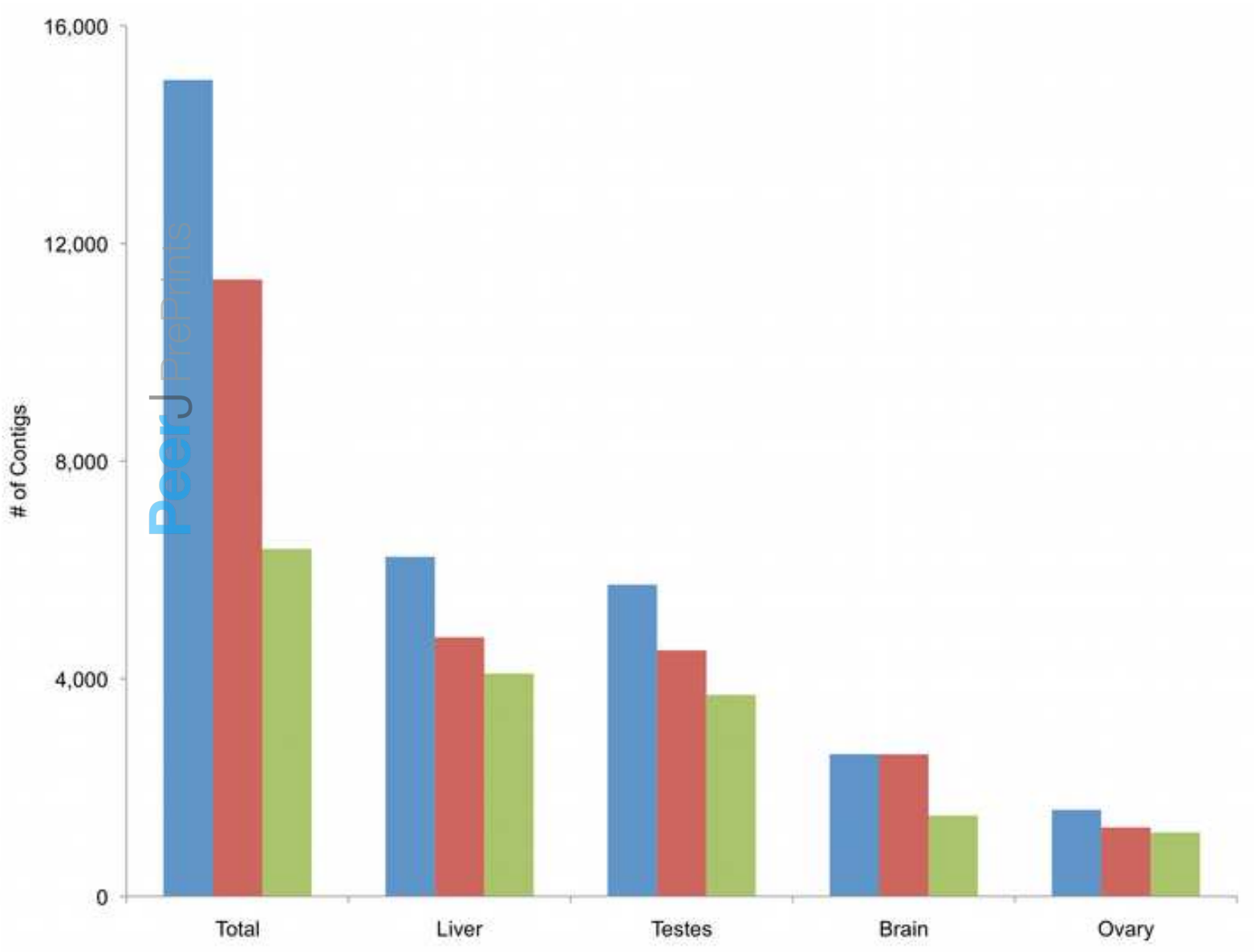
PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

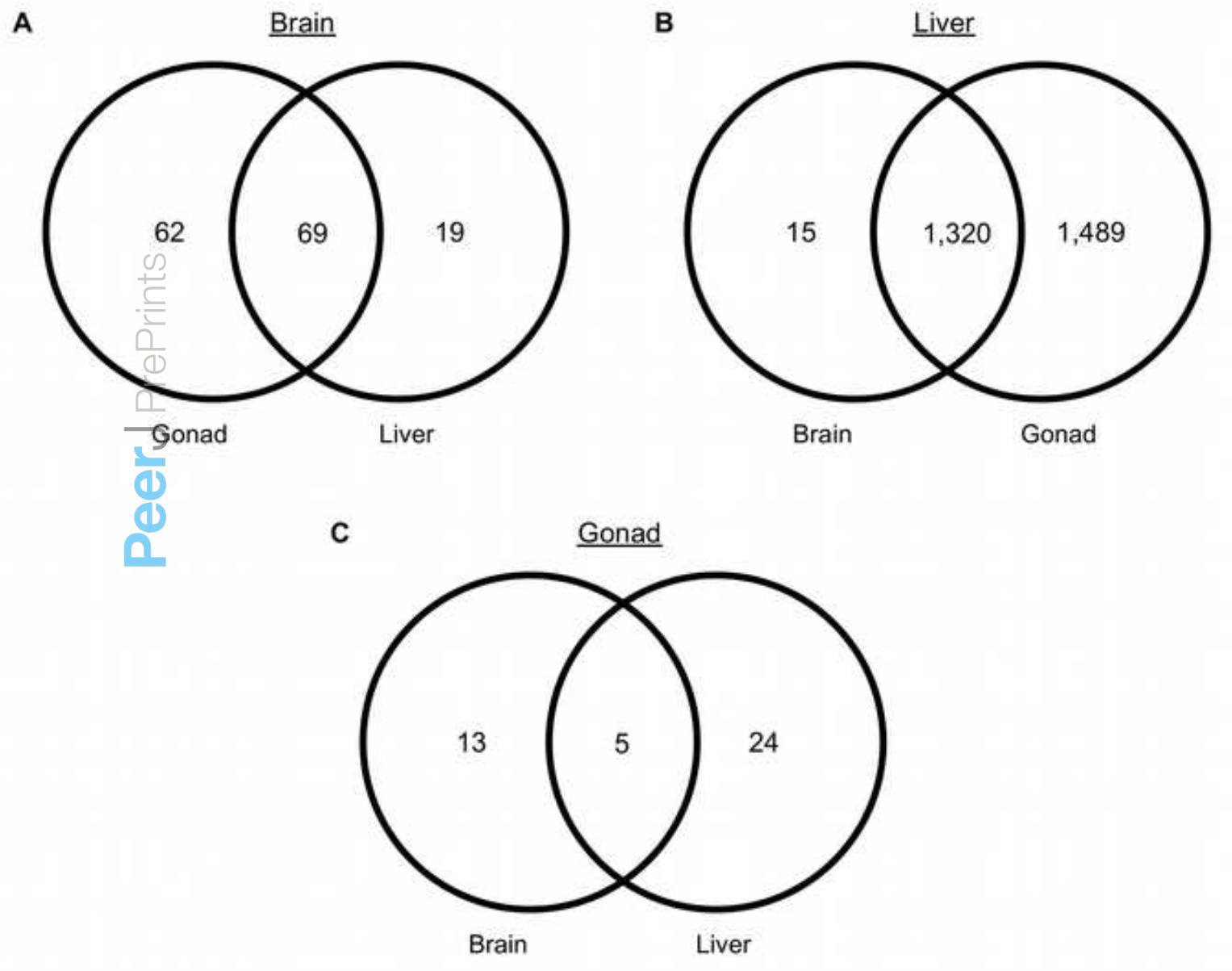


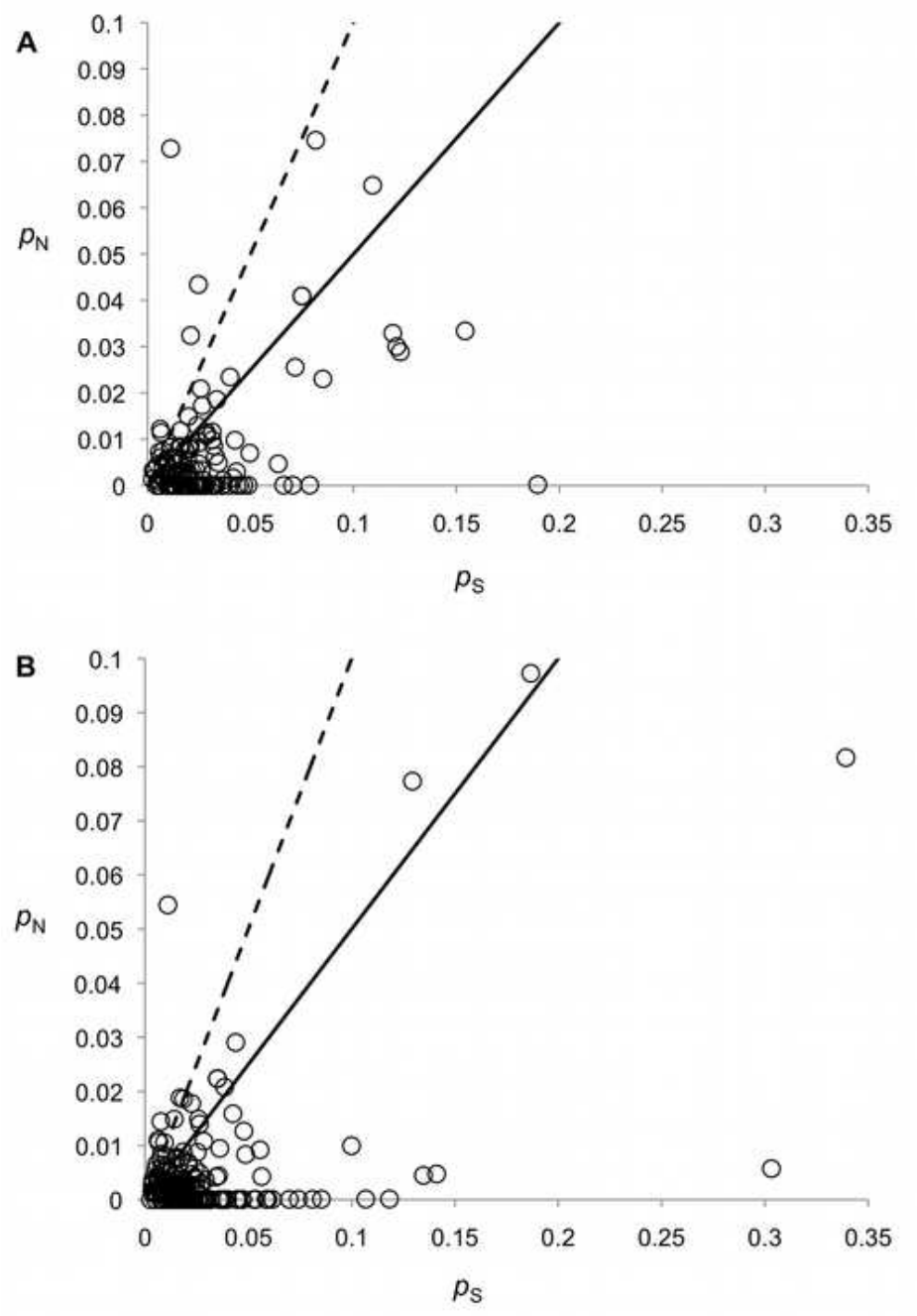
PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13











967 **Tables**

968 **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

969

970 [a]Newbler v. 2.5.3 large genomic assembly of total set of raw sequencing reads

971 [b]Cap3 assembly using 'assembled' or 'partially assembled' reads from Newbler genome

972 assembly

973 [c]Newbler v. 2.5.3 cDNA assembly using 'assembled' or 'partially assembled' reads from

974 Newbler genome assembly

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

976 **Total assembly length in Megabases.

977

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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*.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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.

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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_contig01699*	1.07	Complement factor B	Circulates in the blood; functions in the alternative pathway of the complement system during innate immune responses
	HP_contig01778	1.12	PHD finger protein 8	Removal of methyl groups from histones
	HP_contig03615	1.12	NA(+) / H(+) exchange regulatory cofactor NHE-RF3	Scaffold protein that connects regulatory elements with plasma membrane proteins; regulation of ion transport
	HP_contig01919	1.18	Aldo-keto reductase family 1, member C12	Xenobiotic metabolism; oxidation-reduction process
	HP_contig03408	1.62	Orm1-like 3	Endoplasmic reticulum plasma membrane protein; may regulate ER mediated signaling
	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
	RR_contig01212*	1.26	Isoform cra_a	Uncharacterized cellular membrane protein
	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
	NYBG_contig00118*	1.76	Orf 2	Contains reverse transcriptase domain
	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
	RR_contig00157	6.50	Complement Factor H-like	Regulates activation of the alternative complement pathway in the innate immune response

*= Gene contained $p_N / p_S > 1$ in two independent population pairwise comparisons

PeerJ PrePrints | https://peerj.com/preprints/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13

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/13 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.13