Phylogenomic Insights into Mouse Evolution Using a Pseudoreference Approach

Brice A.J. Sarver[1], Sara Keeble[1], Ted Cosart[1], Priscilla K. Tucker[2], Matthew D. Dean[3], and Jeffrey M. Good[1,*]

[1]Division of Biological Sciences, University of Montana, Missoula, MT

[2]Department of Ecology and Evolutionary Biology and Museum of Zoology, University of Michigan, Ann Arbor, MI

[3]Molecular and Computational Biology, University of Southern California, Los Angeles, CA

*Corresponding author: E-mail: jeffrey.good@umontana.edu.

Accepted: February 20, 2017

Data deposition: This project has been deposited at the NCBI Sequence Read Archive under the accession PRJNA323493.

Abstract

Comparative genomic studies are now possible across a broad range of evolutionary timescales, but the generation and analysis of genomic data across many different species still present a number of challenges. The most sophisticated genotyping and downstream analytical frameworks are still predominantly based on comparisons to high-quality reference genomes. However, established genomic resources are often limited within a given group of species, necessitating comparisons to divergent reference genomes that could restrict or bias comparisons across a phylogenetic sample. Here, we develop a scalable pseudoreference approach to iteratively incorporate sample-specific variation into a genome reference and reduce the effects of systematic mapping bias in downstream analyses. To characterize this framework, we used targeted capture to sequence whole exomes (~54 Mbp) in 12 lineages (ten species) of mice spanning the *Mus* radiation. We generated whole exome pseudoreferences for all species and show that this iterative reference-based approach improved basic genomic analyses that depend on mapping accuracy while preserving the associated annotations of the mouse reference genome. We then use these pseudoreferences to resolve evolutionary relationships among these lineages while accounting for phylogenetic discordance across the genome, contributing an important resource for comparative studies in the mouse system. We also describe patterns of genomic introgression among lineages and compare our results to previous studies. Our general approach can be applied to whole or partitioned genomic data and is easily portable to any system with sufficient genomic resources, providing a useful framework for phylogenomic studies in mice and other taxa.

Key words: *Mus musculus*, bioinformatics, mapping bias, introgression, comparative genomics.

Introduction

The efficient generation and analysis of comparative genome-wide data sets remains a key challenge in evolutionary biology. Massively parallel sequencing has made it relatively easy to generate whole-genome sequencing (WGS) data sets, enabling comparative genomic studies across a broad range of evolutionary timescales. However, the empirical and analytical resources required to generate comparative WGS data sets are still somewhat limiting in species groups with large, complex genomes. As a partial solution, various partitioning approaches are often used to generate comparative genome-wide data across broader sets of species (e.g., restriction site-associated DNA sequencing, targeted capture, transcriptomics, etc.; reviewed in Ekblom and Galindo 2011; Jones and Good 2016). These approaches overcome the extra costs associated with WGS, but analyzing such data across a diverse sample of species still presents a number of challenges. In particular, the most sophisticated analytical frameworks often rely upon approaches developed for WGS and the numerous benefits afforded by a high quality reference genome (e.g., efficient genotyping, physical location, and associated functional annotation; Li and Durbin 2009; Li et al. 2009; McKenna et al. 2010; Yandell and Ence 2012). Thus, as with WGS, the types of analyses that can be conducted using genome-wide partitioned data can be limited by the existence, quality, and completeness of a reference genome.

One common solution in species lacking genomic resources is to use an established reference from another species. Most

genotyping approaches are reference-based at some level and thus depend on accurate sequence read mapping (DePristo et al. 2011), which decreases with increasing sequence divergence from the reference (Li et al. 2008). Although mapping algorithms allow for reference mismatches to account for some divergence, polymorphism, or sequencing error (Nielsen et al. 2011; Ruffalo et al. 2011; Liu et al. 2012), mapping to a divergent reference can generate a number of systematic biases that could compromise comparative evolutionary analyses. For example, sequences that show substantial divergence from a reference will map with lower quality and effectively hide corresponding sample-specific variation. Analyses relying on full sequence information, such as those often used in phylogenetics or molecular evolution, may be particularly sensitive to these issues because called genotypes may converge towards the reference, resulting in an overestimated similarity between subject and reference sequences in divergent regions. This phenomenon, generally referred to as reference (or mapping) bias, has been discussed most frequently with regard to its effect on detecting allele-specific expression in transcriptomic analyses (e.g., Satya et al. 2012; Stevenson et al. 2013; Panousis et al. 2014; Brandt et al. 2015), yet it impacts any comparative study where reads are mapped to a divergent reference. For example, reference bias could lead to the underestimation of rates of molecular evolution or the overestimation of phylogenetic discordance due to stochastic genealogical processes (i.e., incomplete lineage sorting) or hybridization. One approach that has shown some promise in alleviating these concerns is the generation of "pseudogenomes," or reference genomes that incorporate sample-specific variation (Holt et al. 2013; Huang et al. 2013, 2014). This allows annotation to be carried over from a reference while accounting for sequence divergence during the mapping stage. Here, we extend these previous works by developing a scalable pseudoreference approach to iteratively incorporate sample-specific variation into a reference and thereby reduce the effects of systematic mapping bias in downstream analyses.

The house mouse (*Mus musculus*) is an important model of mammalian biology and a compelling system in which to develop comparative genomic approaches and resources. In addition to extensive genetic and developmental resources, the mouse was the second mammal to be sequenced (Chinwalla et al. 2002), and the mouse reference (C57BL/6, a mosaic lab strain primarily of *M. musculus domesticus* origin; Yang et al. 2011) remains second in quality only to the human genome. House mice have also emerged as a powerful system to address fundamental questions in genome evolution, population genetics, and speciation (e.g., Good et al. 2010; Halligan et al. 2010; Kousathanas et al. 2014; Turner et al. 2014; Phifer-Rixey and Nachman 2015; Larson et al. 2016). Although most evolutionary genomic studies in this group have focused on a few closely related species and subspecies (e.g., Keane et al. 2011; Yang et al. 2011), house mice are embedded within a radiation of ~38 species that shared a common ancestor ~7.5 Ma (Schenk et al. 2013). Several of these species already have developed inbred laboratory strains, providing a unique combination of genetic and genomic resources that could be leveraged to address a wide array of evolutionary questions in mammals. However, aspects of the *Mus* phylogeny remain unresolved, including uncertainty in the evolutionary relationships among some key lineages that are relatively closely related to house mice (e.g., *M. spretus/spicilegus/macedonicus* and *M. caroli/cookii/cervicolor*; Hammer and Silver 1993; Lundrigan et al. 2002; Chevret et al. 2005; Tucker et al. 2005; Bryja et al. 2014). In addition to uncertainty in overall species relationships, it is also unclear how much phylogenetic discordance there is across the house mouse genome due to incomplete lineage sorting or gene flow between species (e.g., Keane et al. 2011, Song et al. 2011). Resolving these outstanding issues is an important step in developing the mouse system for comparative evolutionary studies.

In this study, we use targeted capture to generate whole exome data (54 Mb targeted, exons and flanking regions) across 10 species of mice (*Mus*). We use these data to evaluate the general performance of our pseudoreference approach in mitigating the effects of reference bias. We then use the pseudoreferences to resolve the phylogenetic relationships among these mouse species while assessing phylogenetic discordance at different genomic scales and the extent of introgression between some lineages. In addition to insights into the evolutionary history of these species, our study provides a foundation for future comparative studies in mice and a general framework for rapidly generating phylogenomic data sets in other groups of closely related species.

Materials and Methods

Exome Capture

Illumina sequencing libraries were generated using whole genomic DNA from ten species (*Mus caroli*, *M. cervicolor*, *M. cookii*, *M. macedonicus*, *M. minutoides*, *M. musculus*, *M. pahari*, *M. platythrix*, *M. spicilegus*, and *M. spretus*) including three wild-derived inbred strains of house mice (*M. musculus domesticus*: LEWES/EiJ, hereafter dom^{LEWES}; *M. m. musculus*: CZECHII/EiJ and PWK/PhJ, hereafter $mus^{CZECHII}$ and mus^{PWK}) (supplementary material table S1, Supplementary Material online). Libraries were individually indexed following Meyer and Kircher (2010), pooled (Pool 1: *M. caroli*, *M. cervicolor*, *M. cookii*, *M. minutoides*, *M. pahari*, *M. platythrix*; Pool 2: $mus^{CZECHII}$, dom^{LEWES}, *M. macedonicus*, mus^{PWK}, M. *spicilegus*, *M. spretus*), enriched with two NimbleGen SeqCap EZ Mouse exome capture reactions (Fairfield et al. 2011), and 100 bp paired-end sequenced on an Illumina HiSeq 2000. This in-solution enrichment platform targets 54.3 Mbp of exonic regions with the mouse genome (NCBI37/mm9).

Quality Assessment and Iterative Mapping

Raw reads were cleaned using the expHTS pipeline (available from https://github.com/msettles/expHTS; last accessed

February 28, 2017), which trims adapters and low-quality bases, merges overlapping reads, and removes identical reads (putative PCR duplicates). Initial capture performance statistics were calculated using CollectHsMetrics in Picard v2.5.0 (available from http://github.com/broadinstitute/picard; last accessed February 28, 2017). To mitigate reference bias, we employed an iterative mapping strategy to generate species-specific exomes embedded within the mouse reference genome (GRCm38). Cleaned reads were mapped to the reference genome using the MEM algorithm of BWA v0.7.15 (Li and Durbin 2009; Li 2013). Duplicate reads were identified postmapping using Picard v2.5.0. For multiply mapped reads, only the location with the best mapping quality was included in the analysis. Regions with insertions or deletions (indels) were identified and realigned, and single nucleotide variants (SNVs) were called using HaplotypeCaller within the Genome Analysis Toolkit (GATK) v3.6 (McKenna et al. 2010; DePristo et al. 2011). Resulting SNVs were filtered for a minimum quality of 30 and a minimum sequencing depth of at least five independent reads. These variants were injected back into the original reference using FastaAlternateReferenceMaker within the GATK. Additional processing of files, such as indexing, merging, and sorting, was accomplished using SAMtools v1.3.1 (Li et al. 2009) and Picard v2.5.0, as required. After each round, the modified reference was used as the starting point for additional iterations, starting with remapping of all reads and proceeding through variant calling. The early rounds of this iterative procedure should systematically introduce variants from the sample into the reference, increasing the number of sample reads that map and the number of variants that can be confidently called until the number of incorporated reads stabilizes across subsequent iterations. At this point, we inserted IUPAC ambiguity codes at putative heterozygous positions.

It was initially unclear how many iterations of mapping and reference generation ought to be performed to remove reference bias in our study. Preliminary evaluation (data not shown) suggested more than three iterations of mapping and genotyping would be required to incorporate most variation into a pseudoreference. We examined this empirically by identifying the number of iterations (5) at which read incorporation and per-site sequence divergence plateaued in the most divergent species in our sample, *M. pahari*. We then used this as the number of iterations necessary to produce a stable pseudoreference across all species in our sample. As a final step, each position with insufficient data to confidently call a sample genotype was excluded; an additional round of variant calling was performed with the EMIT_ALL_SITES argument set, producing a VCF with calls at each position. All remaining ambiguous positions (genotype quality <30, read depth <10 or >60) were hard masked (i.e., replaced with an "N") using GNU awk and bedtools v2.25 (Quinlan and Hall 2010). This produced a final consensus pseudoreference exome for each sample with the same coordinate system as the mouse reference. We also generated pseudoreferences without ambiguity codes for some downstream analyses. These are useful for bioinformatic analyses, including mapping and variant calling, which assume a haploid reference. All code necessary to replicate these procedures starting from cleaned reads is available as part of the *pseudo-it* project on GitHub (http://www.github.com/bricesarver/pseudo-it; last accessed February 28, 2017), and all pseudoreferences are available upon request.

Phylogenetic Inference

We used a two-tiered approach to resolve the phylogenetic relationships in our sample. First, we estimated the overall phylogeny from a concatenated alignment of gene sequences using the brown rat (*Rattus norvegicus*) as an outgroup. For each targeted protein-coding gene, we extracted the longest protein-coding transcript sequence based on the UCSC genes track (retrieved through the UCSC Genome Browser) from each iterated pseudoreference and from the whole genome reference sequence for *M. m. domesticus* strain C57BL/6. For each species, exons were extracted and assembled into transcripts using custom code and the Biostrings package (Pagès et al. 2016) in R v3.1.3 (R Core Team 2015) and then combined into a multispecies alignment. We then used BioMart (Smedley et al. 2015) to identify one-to-one orthologous transcripts in *R. norvegicus*. Each set of transcripts was translation aligned using TranslatorX (Abascal et al. 2010) with the Muscle progressive alignment algorithm (Edgar 2004). Alignments without a length evenly divisible by three or possessing internal stop codons were discarded (5702 genes). With this filtered gene set, we then performed concatenated analyses by chromosome to simplify data processing and to verify internal consistency of analyses. All transcript sets from each chromosome were combined into a supermatrix using Phyutility v2.2.6 (Smith and Dunn 2008). A tree was estimated for each chromosome with the MPI version of RAxML v8.2.3 (Stamatakis et al. 2005; Stamatakis 2014) using a simultaneous maximum likelihood (ML) search and rapid bootstrapping run under the GTR + Γ model of sequence evolution (autoMRE option). Trees were visualized using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree; last accessed February 28, 2017). Among-chromosome topological discordance was assessed by rooting trees with rat and estimating pairwise Robinson-Foulds distances (Robinson and Foulds 1981) using the ape library (Paradis et al. 2004) in R.

Second, we focused on finer-scale patterns of phylogenetic discordance. A phylogenetic tree assumes a series of bifurcating speciation events. However, the speciation process is not necessarily instantaneous and we expect some regions of the genome to show conflicting phylogenetic histories due to incomplete lineage sorting, hybridization, or undetected gene duplication. In phylogenetics, a distinction is made between the history of a locus (a "gene tree") and the true relationship

among lineages (a "species tree"; Maddison 1997). Several approaches have been developed to account for gene tree-species tree discordance under the multispecies coalescent (e.g., Edwards et al. 2007; Liu et al. 2009; Heled and Drummond 2010), yet many of these approaches are computationally intensive and thus less practical for genome-scale data sets. With these limitations in mind, we accounted for phylogenetic discordance in our data set using the computationally efficient species tree algorithm implemented in ASTRAL v4.10.11 (Mirarab et al. 2014; Mirarab and Warnow 2015; Sayyari and Mirarab 2016). Assuming sets of independent and accurately estimated gene trees, ASTRAL breaks each tree into its constituent quartets (i.e., four-taxa cases) and recovers a consistent estimate of the species tree.

Resolution of individual targets or transcript genealogies may be limited in our study, given the low overall levels of coding divergence between our focal species. To increase local phylogenetic signals, we expanded our working data set to include 5′- or 3′ untranslated regions (UTR) and all other regions targeted for capture. Though exome probes are usually contained within annotated exons, both the capture process itself and the iterative pseudoreference process allow for the discovery of variation in flanking regions. To incorporate this variation, we extended each target by 200 bp on both ends and merged regions that were up to 1 kbp apart, increasing the total data set from 54.3 to 163.4 Mbp. As above, we first used RAxML (GTR + Γ, 200 bootstrap replicates) to estimate an ML tree per chromosome by extracting extended targets using bedtools v2.2.5 and combining regions with AMAS (Borowiec 2016). For these data, no alignment is required because indel variation is not incorporated into the pseudoreference. We then repeated this procedure across autosomal windows of five different sizes (extended targets, 100 kbp, 500 kbp, 1 Mbp, and 5 Mbp), estimating ML phylogenies from each window using the fast hill-climbing algorithm in RAxML. Strong linkage disequilibrium typically extends 100 kbp or less within wild house mouse (*M. musculus*) populations (Laurie et al. 2007), suggesting that larger window sizes may combine regions with independent phylogenetic histories. Any window containing only missing data for at least one individual was discarded. For each window size, all trees were combined for species tree inference in ASTRAL. We also calculated among-locus phylogenetic discordance using the normalized quartet score, which quantifies the amount of quartet discordance relative to the species tree.

Testing for Introgression

Motivated by recent studies that identified introgression between mouse lineages (e.g., Teeter et al. 2008; Keane et al. 2011; Yang et al. 2011; Staubach et al. 2012; Janoušek et al. 2015; Liu et al. 2015) we tested for signatures of introgression within and between taxa from the *M. musculus* group (here, *M. m. musculus* and *M. m. domesticus*), the *M. spretus* group (*M. spretus*, *M. spicilegus*, and *M. macedonicus*), and between *M. cervicolor*, *M. cookii*, and *M. caroli*. We used the D-statistic (i.e., Patterson's D or the ABBA-BABA test) to characterize patterns among species (Green et al. 2010; Durand et al. 2011). Briefly, the D-statistic is a normalized difference of counts of two site patterns within a rooted four-taxa case: ABBA and BABA. ABBA counts indicate a sharing of alleles between the first taxon and a specified outgroup (A) and the second and third taxa (B), whereas the opposite is true for the BABA case. Significance was assessed using a chi-square test (see Pease and Hahn 2015), and 95% confidence intervals estimated using a nonparametric bootstrap with 10,000 replicates. Additionally, when our sampling allowed, we estimated the minimum proportion of genomic admixture ($\hat{f}$) following Durand et al. (2011). The *D*-statistic is relatively robust to genotyping error (Green et al. 2010; Durand et al. 2011), but could be sensitive to inherent differences in the source and quality of the exome data relative to the reference genome ($dom^{C57BL/6}$). Therefore, we limited our comparisons to sequenced exomes except when directly testing for differential introgression between *M. m. musculus* and the two available *M. m. domesticus* genotypes (dom^{LEWES}, $dom^{C57BL/6}$).

Results

Efficient Targeted Recovery of *Mus* Whole Exomes

Multiplex exome capture was successful across all samples. Sequencing efforts produced an average of ~22 million reads per sample with an average of 1.1% of targets showing no coverage. Given a combined target size of ~2% of the genome, this represents targeted recovery of 53.8 Mbp of sequence data (table 1) including most annotated genic regions in the mouse genome. Approximately 75% of raw reads were unique, resulting in average target coverage of 30× across samples (range: 20.6–39.3×) with ~80% of targeted bases sequenced to at least 10× coverage (table 1).

Evaluation of Iterative Pseudoreference Generation

To assess the performance of the iterative approach, we compared the same set of cleaned reads mapped to the mouse reference and to five-iteration pseudoreferences for each species (table 1). In all cases, mapping to a five-iteration pseudoreference resulted in minor increases in the coverage of targeted bases (e.g., 23.0–23.4× in *M. pahari*) and the percentage of targeted bases recovered at a given depth (e.g., +1.3% for targets with at least 10× coverage in *M. pahari*; table 1). In addition, reads were more confidently placed with each pseudoreference, resulting in an increase in usable bases and fewer reads discarded due to low mapping quality, as evidenced across iterations for the *M. pahari* exome (supplementary material table S2, Supplementary Material online).

Table 1
Exome Sequencing Coverage across Species

Sample	Total Reads	Bases on-Target	Target Coverage	% Low Quality Bases	% Target Bases ≥ 10×
M. caroli (ref)	18,923,169	1,423,165,223	26.2	8.1	78.2
M. caroli (5)	—	1,435,937,326	26.4	6.9	78.7
M. cervicolor (ref)	29,711,989	2,119,138,428	39.0	8.0	87.0
M. cervicolor (5)	—	2,133,497,561	39.3	6.9	87.5
M. cookii (ref)	29,089,576	2,119,862,202	39.0	8.1	86.5
M. cookii (5)	—	2,134,097,222	39.3	7.1	87.0
M. macedonicus (ref)	17,428,555	1,233,646,735	22.7	6.1	79.4
M. macedonicus (5)	—	1,239,060,658	22.8	5.4	79.7
M. minutoides (ref)	23,340,200	1,703,586,451	31.3	7.8	77.0
M. minutoides (5)	—	1,733,871,618	31.9	6.2	78.2
M. pahari (ref)	17,033,748	1,247,810,134	23.0	9.5	68.7
M. pahari (5)	—	1,273,913,560	23.4	7.7	70.0
M. platythrix (ref)	22,058,259	1,734,397,262	31.9	9.4	78.6
M. platythrix (5)	—	1,756,287,920	32.3	8.2	79.5
M. spicilegus (ref)	14,814,946	1,120,139,640	20.6	8.2	73.2
M. spicilegus (5)	—	1,124,893,538	20.7	7.6	73.5
M. spretus (ref)	16,200,749	1,157,756,368	21.3	6.9	73.6
M. spretus (5)	—	1,163,128,639	21.4	6.2	73.9
M. m. domesticus LEWES (ref)	25,565,922	1,920,922,154	35.3	6.2	87.8
M. m. domesticus LEWES (5)	—	1,921,986,302	35.4	6.1	87.8
M. m. musculus CZECHII (ref)	24,773,619	1,946,919,016	35.8	7.2	86.5
M. m. musculus CZECHII (5)	—	1,949,472,117	35.9	6.9	86.6
M. m. musculus PWK (ref)	22,785,276	1,652,768,831	30.4	6.3	85.4
M. m. musculus PWK (5)	—	1,655,627,360	30.5	6.0	85.5

NOTE.—Exome sequencing coverage for each species when mapped to the mouse reference genome (ref) or a species-specific pseudoreference after five rounds of iterative mapping (5). Shown are the total reads per library after cleaning (Total Reads), the number of bases in regions targeted by the capture (Bases On-Target), average coverage per target (Target Coverage), the percentage of bases in reads mapped with a MAPQ greater than zero (% Low Quality Bases), and the percentage of bases in targeted regions with at least 10× coverage.

In addition to modest increases in overall coverage, pseudoreference construction should also help mitigate systematic biases in standard descriptive statistics when mapping to a distantly related reference genome. To test this, we calculated the per-site divergence for targeted bases on Chromosome 1 (i.e., the number of homozygous alternative calls relative to the C57BL/6 reference divided by the total number of confidently genotyped sites) at each iteration for three samples—*M. m. domesticus* (dom^{LEWES}), *M. spretus*, and *M. pahari*—of increasing evolutionary distance from the reference. Divergence estimates were notably higher in all three species when using a five-iteration pseudoreference when compared with mapping straight to the mouse genome (fig. 1*A*). Increases in per-site divergence were lowest for *M. m. domesticus* (dom^{LEWES}, 0.19% vs. 0.22%; fig. 1*A*), and highest for the most distantly related lineage in our study, *M. pahari* (3.34% vs. 4.24%). In all cases, the most dramatic change was observed after mapping to the first estimated pseudoreference (i.e., iteration 2) and appeared to reach an asymptote by the fourth iteration. However, the relative magnitude of change scaled with divergence (fig. 1*B*), assuming that incremental increases reflect divergence estimates asymptotically approaching their true value. These results indicate that the number of iterations required to mitigate biases will be contingent on the divergence levels between sample(s) and reference(s) in a given study.

The impact of pseudoreference construction on estimates of sequence divergence should also be apparent within a genome, across sites that vary in levels of functional constraint, for example. To test this, we classified all confidently called sites in *M. pahari* as belonging to protein-coding exon sequence, 5′- or 3′-UTRs, or flanking regions (introns or intergenic). We observe the same trends, with the most dramatic changes in per-site divergence detected in the less constrained flanking regions, followed by UTRs and protein coding domains (fig. 1*C*).

Finally, we looked at the number and quality of variants called for *M. m. domesticus* (dom^{LEWES}), *M. spretus*, and *M. pahari* using the mouse reference and a five-iteration exome pseudoreference (Chromosome 1). We used HaplotypeCaller in the GATK (with –emitRefConfidence BP_RESOLUTION) to return genotype calls at each position and applied common quality filters to each set (as above). Although the number of confidently called sites decreased with divergence from the mouse reference genome, the total number of confidently called sites relative to the first iteration increased (table 2). Intuitively, we would also expect that genotype qualities should tend to increase in the

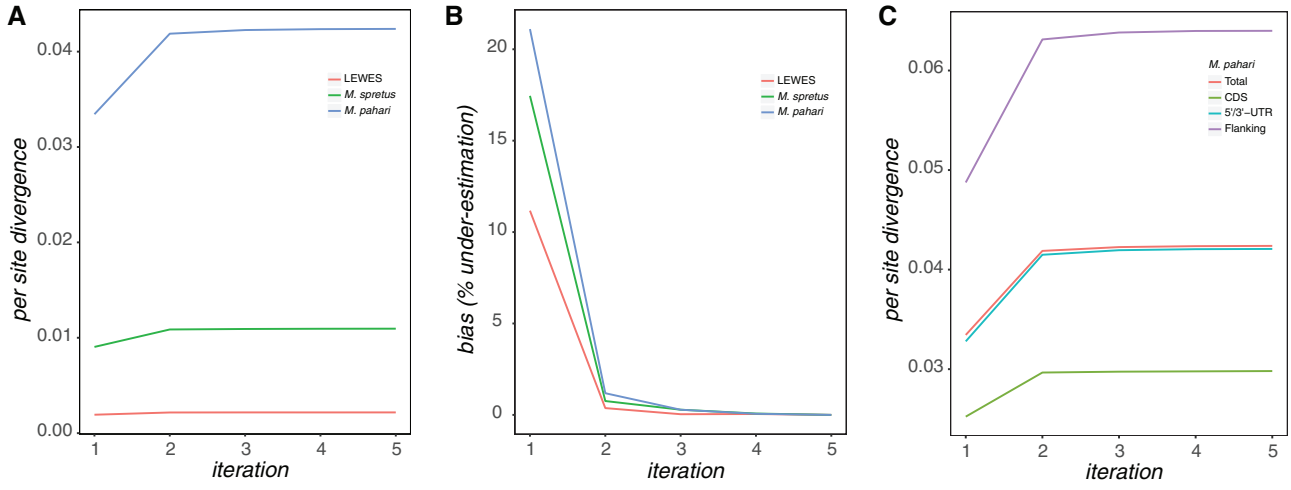

Fig. 1.—Reference bias and sequence divergence. (*A*) Per-site sequence divergence per iteration using confidently called positions on Chromosome 1 for *M. m. domesticus* (dom^{LEWES}), *M. spretus*, and *M. pahari*. (*B*) The bias in divergence estimates (% under-estimation) at each iteration relative to the per-site divergence of the sample's five iteration pseudoreference using the same data. (*C*) Per-site divergence for *M. pahari* partitioned by protein-coding sequence (CDS), untranslated exonic regions (5′/3′-UTR) and flanking sequences.

Table 2
Confidently Called Genotypes on Chromosome 1 Using the Mouse Reference Genome and a Five-Iteration Pseudoreference

Species	Genotypes, Mouse Reference	Genotypes, Five-Iteration Pseudoreference	Δ Genotypes Called	% Increase
M. m. domesticus (dom^{LEWES})	4,199,062	4,204,382	5,320	0.13
M. spretus	3,298,609	3,332,638	34,029	1.03
M. pahari	2,717,373	2,791,970	74,597	2.75

Note.—Confidently called genotypes on Chromosome 1 using the mouse reference genome and a five-iteration pseudoreference. *M. m. domesticus* is most closely related to the mouse reference genome (a mosaic lab strain primarily of *M. m. domesticus* origin), followed by *M. spretus* and *M. pahari*.

context of pseudoreferences. Consistent with this, we observed a positive skew in genotype qualities for all three species at positions that were confidently called relative to the mouse reference genome and the final pseudoreference (fig. 2). However, we also observed many sites where the genotype quality decreased, frequently reflecting the loss of reads at a position due to being more confidently placed elsewhere after iteration. We also observe cases where sites called as homozygous reference or alternative relative to the mouse reference are called heterozygous (and vice versa) due to the placement of reads with alternate alleles at a given site.

Resolving the *Mus* Phylogeny

We first estimated a phylogeny for each chromosome based on concatenation of protein-coding transcripts. After filtering, this data set consisted of 15,620 aligned transcripts (28.2 Mbp) with one-to-one orthologs in rat. RAxML produced the same fully resolved tree for all chromosomes with 100% bootstrap support for each bipartition (supplementary material fig. S1, Supplementary Material online). There was no topological discordance among chromosomes (Robinson-Foulds distances equal to zero). Additionally, there was no discordance among trees estimated using sets of transcripts without *Rattus* (26,624 transcripts with a total length of 43.7 Mbp, analysis not shown), and all trees were resolved with 100% bootstrap support. These analyses also confirmed that *M. pahari* is an outgroup relative to the other sequenced species based on the rooted phylogeny (supplementary material fig. S1, Supplementary Material online). We then repeated this procedure for an expanded data set including all targeted and flanking regions in mice (and excluding rats), and found the same general results of a fully resolved concatenated phylogeny with no discordance among chromosomes (fig. 3). Notably, these concatenated phylogenies resolve *M. spretus/spicilegus/macedonicus* and *M. caroli/cookii/cervicolor* as monophyletic groups with *M. spretus* and *M. caroli* placed as the basal lineages within each.

Using concatenation to resolve a phylogeny effectively averages over fine-scale discordance, which can inflate confidence in the overall tree and obscure important sources of incongruence (Hahn and Nakhleh 2016). Therefore, we also used a species-tree approach to quantify fine-scale topological discordance. To do this, we first estimated individual ML genealogies trees using all extended targets with data partitioned into five window sizes: 100,531 extended targets (mean alignment length = 1,622 bp; parsimony informative sites per target: mean = 26.8, median = 15.0), 13,628 100 kbp intervals (mean alignment length = 10,621 bp); 4,036 500 kbp intervals (37,322 bp); 2,665 1 Mbp intervals

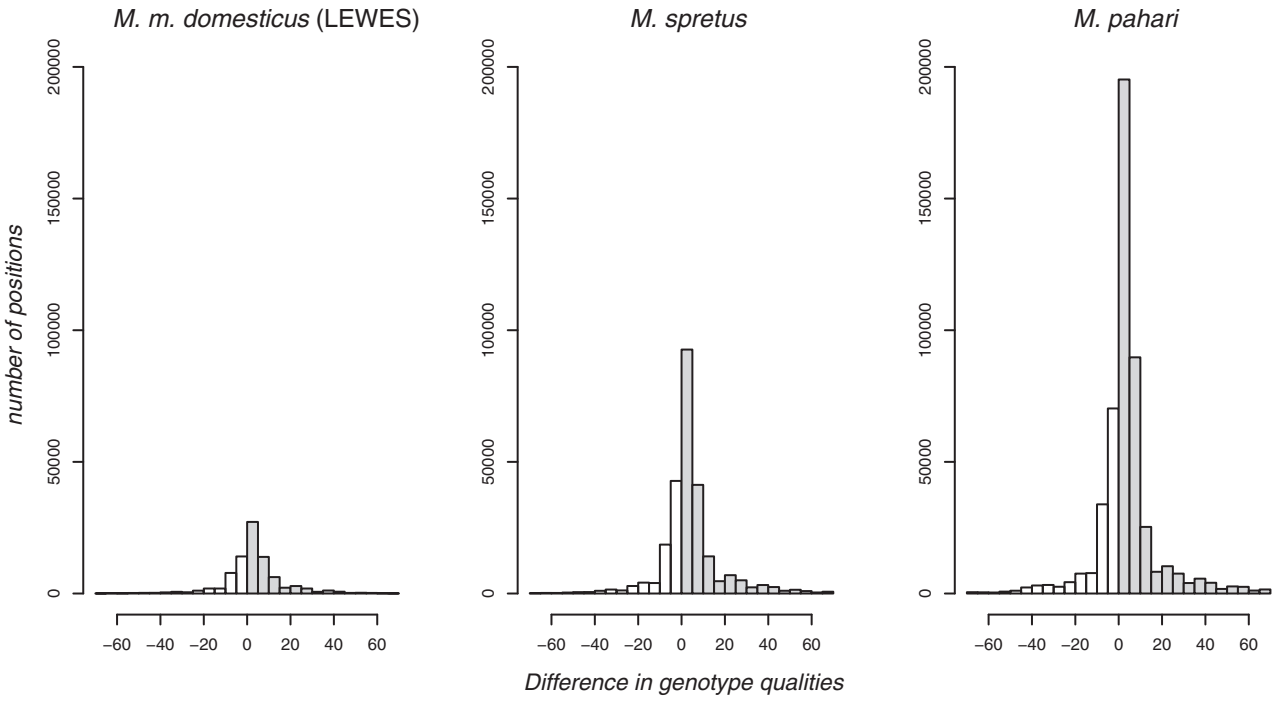

Fig. 2.—Differences in genotype qualities at shared positions called using a five-iteration pseudoreference or the mouse reference genome. Positive values reflect higher genotype qualities in the five-iteration pseudoreference (shown in gray). Positions with no change in genotype qualities are excluded to improve visualization. In each case, the distributions are skewed positively, indicating a trend towards more confident genotype calls in the pseudoreference.

(66,743 bp); and 511 5 Mbp intervals (291,249 bp). We then used these trees to estimate species trees while accounting for among-locus topological discordance. We detected no appreciable discordance in the point-estimate of the species tree (rooted on *M. pahari*; fig. 3) when compared with the per-chromosome concatenated trees at the 100 kbp, 500 kbp, 1 Mbp, and 5 Mbp scales (fig. 4). However, quartet support for some branches did vary by window size, and there was discordance at the target-level scale (fig. 4). The lowest support was found for branches defining the *M. pahari-platythrix-minutoides* group at the base of the tree, suggesting some uncertainty in the placement of these deep nodes. Indeed, *M. platythrix* and *M. pahari* share a common ancestor in the species tree estimated using the extended targets, contrary to all other analyses. Only 36% of quartets support this clade, and the branch is extremely short. We also observed some variation in support levels within other groups. For example, although the *M. spretus-spicilegus-macedonicus* clade itself was well supported across most analyses, only 43% of quartets support the species tree designation of this clade at the level of targets (fig. 4). Support steadily increased to 59% at the 100 kbp scale, 74% at the 500 kbp scale, 82% at the 1 Mbp scale, and 96% at the 5 Mbp scale. Thus, there is some fine-scale discordance in this group of interest, but the overall species tree generally shows more support than alternative phylogenies. Likewise, support for the *M. caroli-cookii-cervicolor* started at 60% at the target scale and reached 100% at the 5 Mbp scale. Normalized quartet scores suggest ~80% of all quartets support the species tree at the target scale, and this increased to ~99% at the 5 Mbp scale. Considering all analyses, the phylogeny for these taxa appears reasonably well resolved with relatively low levels of topological discordance, at least at the scales that can be reasonably evaluated with our exome data.

Introgression

We detected genotype asymmetries consistent with significant autosomal introgression between *M. m. domesticus* and *M. m. musculus*. We also detected some evidence for significant introgression between *M. cookii* and *M. caroli*. We did not detect autosomal introgression in other cases, including between lineages of the *M. spretus* group (*M. macedonicus*, *M. spicilegus*, and *M. spretus*) (fig. 5; supplementary material table S3, Supplementary Material online). Patterns of between-lineage allele sharing were variable among strains within *M. m. domesticus* and *M. m. musculus*, consistent with the notion of differential introgression due to recent gene flow (Yang et al. 2011). Our sampling allows us to estimate the minimum admixture proportion for a few of these instances. We estimated that ~7% of the genomes of mus^{PWK} ($\hat{f}$ =6.6%) and dom$^{C57BL/6}$ ($\hat{f}$ =6.8%) descend from introgression between *M. m. musculus* and *M. m. domesticus*.

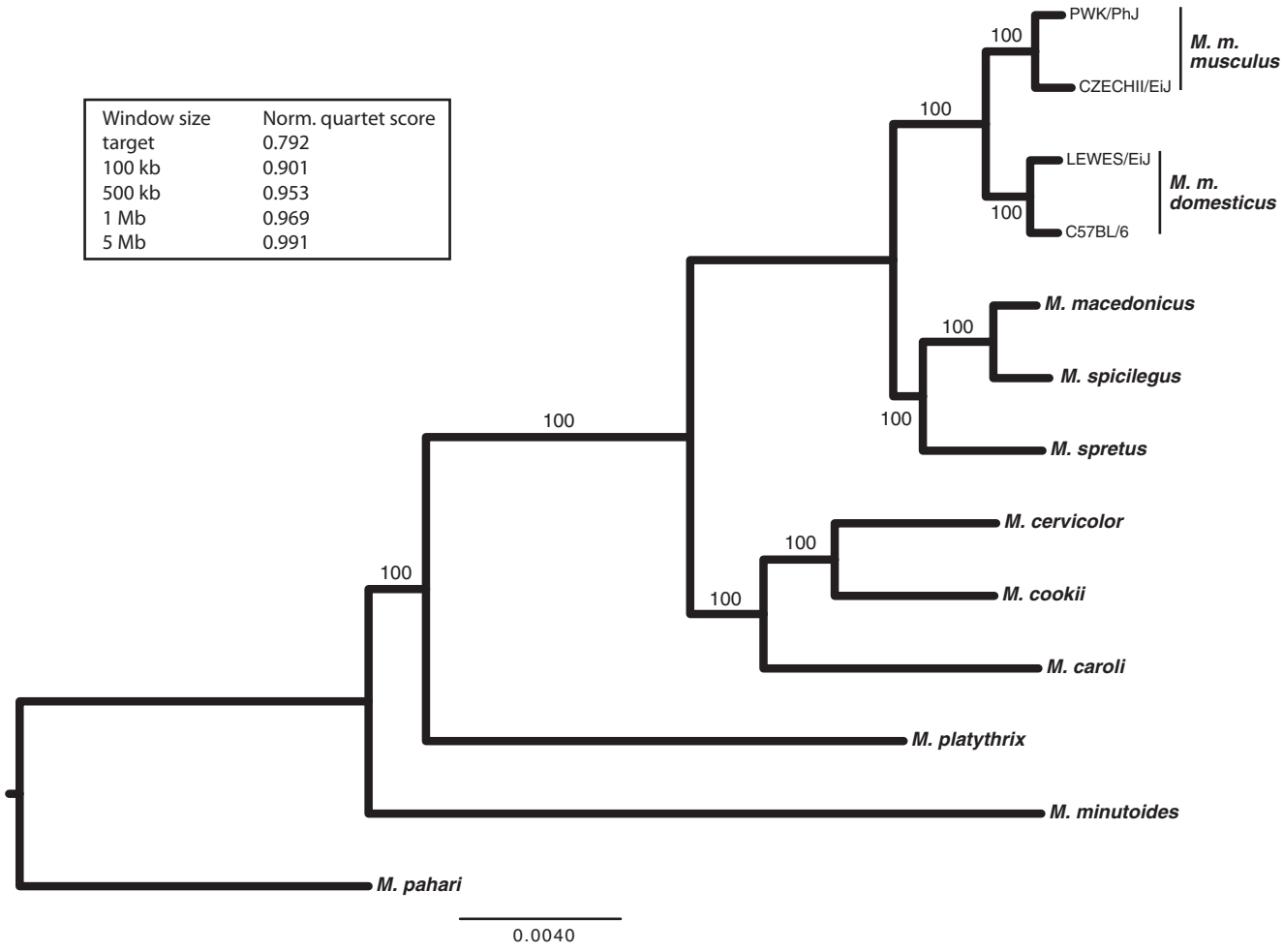

Fig. 3.—*Mus* phylogeny, rooted on *M. pahari*, estimated using all extended targets from Chromosome 1. ML bootstrap support values are listed above branches. There was no discordance between this tree and trees estimated from other chromosomes. The inset provides the normalized quartet scores calculated with ASTRAL from local genealogies estimated at five genomic scales.

Discussion

Genomic data sets are now commonplace in model and non-model systems. However, using a divergent reference genome to analyze genomic data sets can introduce reference biases that can affect biological inferences. To help address this outstanding issue, we developed a scalable pseudoreference approach to iteratively incorporate sample-specific variation into an established reference. Additionally, we describe the first targeted sequencing effort of complete exomes for approximately one-third of described *Mus* species diversity. Using these data, we resolve the phylogenetic relationships between these mouse species and describe patterns of introgression among lineages. Our analyses demonstrate that targeted exome sequencing is useful for both of these tasks and provides a proof-of-concept for similar analyses in other systems. More generally, our pseudoreference framework alleviates mapping biases that can lead to systematic underestimates in divergence and related statistics, providing a useful tool for comparative genomic analyses. Below, we discuss the general utility and limitations of our approach as well as the specific insights of our data to mouse evolution.

Exome Capture and Pseudoreference Construction

Ongoing work will continue to generate assembled and annotated reference genomes for many species of interest. However, high-quality reference genomes, which are critical to mapping reads generated from high throughput sequencing technologies, are still relatively scarce (Ellegren 2014). We were able to capture whole exomes across ten species spanning ~7.5 Myr of divergence (Schenk et al. 2013). Given the strong and comparable performance across all species, we anticipate that this capture approach would be effective over deeper evolutionary timescales. In addition to basic phylogenetic insights, our approach could also be used to generate comparative genomic data for in-depth analyses of molecular evolution over moderate timescales or as a

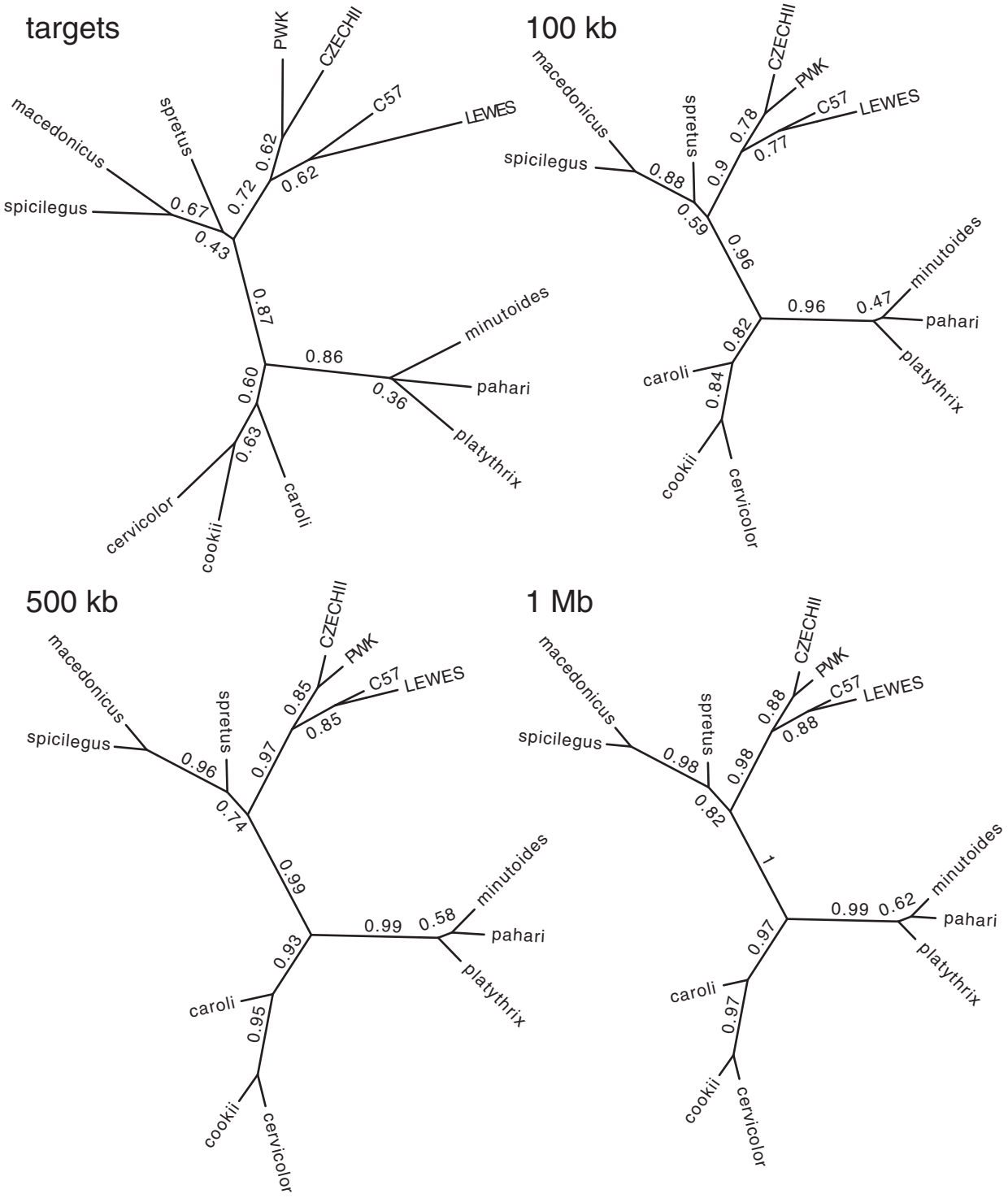

Fig. 4.—Unrooted species tree estimates from ASTRAL across four different window sizes (5 Mbp not shown). Branches are annotated with their local quartet scores.

supplement to lower-coverage whole genome data. Other studies have shown that targeted capture can be used to recover exome data over a broad range of evolutionary timescales (Vallender 2011; Bi et al. 2012; Jin et al. 2012; Hedtke et al. 2013), though the integration of such data into a well-annotated reference genome had not been explored.

Our transspecific capture and iterative pseudoreference approach leveraged the benefits of the mouse reference,

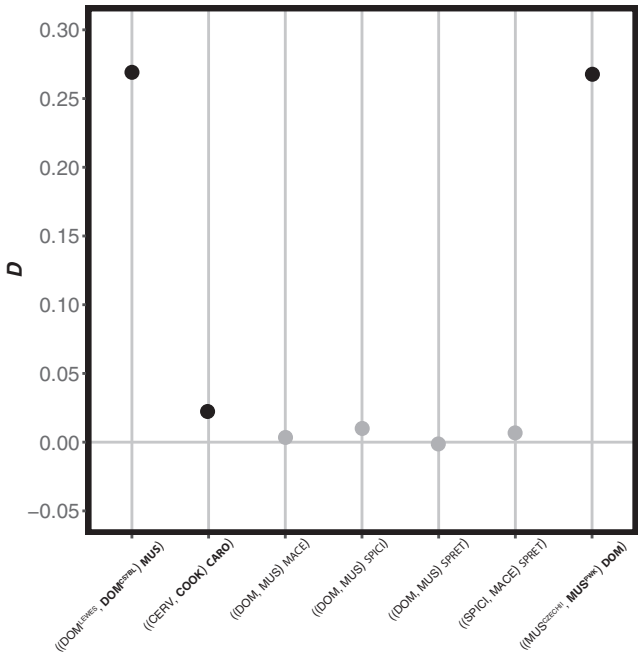

Fig. 5.—Introgression among taxa inferred from the ABBA-BABA test. The *x* axis identifies the three taxa examined for signatures of autosomal introgression using the D statistic where positive values reflect an excess of ABBA sites. CERV = *M. cervicolor*, COOK = *M. cookii*, CARO = *M. caroli*, MACE = *M. macedonicus*, SPICI = *M. spicilegus*, SPRET = *M. spretus*, DOM = *M. m. domesticus* (dom^{LEWES}), and MUS = *M. m. musculus* (mus^{CZECHII}) unless otherwise noted. CARO was used as the outgroup in all comparisons except for ((CERV, COOK) CARO), which used *M. pahari* (shown) or *M. m. musculus* (mus^{CZECHII}). Black circles and boldface taxa indicate significant deviations from zero (χ^2 test; corrected *P* value < 0.01; see details in supplementary material table S3, Supplementary Material online).

including position and annotation information, while mitigating the confounding effects of reference bias. Even among closely related species, we demonstrated that reference bias can have a strong impact on estimation of basic parameters, such as genetic divergence (fig. 1) and genotype quality (fig. 2). These simple comparisons illustrate that while reference-based genotyping is sensitive to divergence, the iterative pseudoreference approach reduces these biases over moderate levels of sequence divergence. Pseudoreferences, therefore, should generally increase the quality of and confidence in downstream analyses through incorporating additional reads and placing them with greater confidence. Implementation of our approach is straightforward (with the *pseudo-it* package), requires the same set of resources as standard mapping and variant calling, and preserves the coordinate system of the original reference. An alternative approach would be to *de novo* assemble targeted regions within each species (e.g., Bi et al. 2012). Whereas mapping to contigs assembled *de novo* is not expected to introduce reference bias, assembly requires substantially more computational power and results in a new coordinate system that needs to be linked between species. Any hard-earned empirical or computational annotation afforded by a reference would also need to be reestablished.

Several other approaches have been developed to combine sets of loci into workable references that can be used to call variants (e.g., PRGmatic; Hird et al. 2011), but are not iterative and cluster regions based on overall similarity. Recent studies using restriction-site associated DNA sequencing (RAD-seq) have shown that overall data quality is considerably higher when using a reference (Fountain et al. 2016; Shafer et al. 2016). Mapping of reads followed by *de novo* assembly would also be expected to reduce mapping bias and genotyping errors but consumes substantially more resources (Gan et al. 2011; Hunter et al. 2015). It is possible to obtain genomic coordinates of contig sets assembled *de novo* by aligning to a reference. However, such an approach is computationally demanding. For example, de Bruijn graph-based assembly would need to be performed under a range of k-mer values and clustered, and each assembly is both CPU and memory intensive. Furthermore, *de novo* assemblies from transcriptomic or capture data sets are often highly fragmented (e.g., Bi et al. 2012), leading to additional complications. We have shown that studies lacking species-specific references may benefit from an iterative approach, provided that a reference genome exists within a moderate evolutionary distance.

We also illustrated that the use of pseudoreferences can be combined with exome capture to resolve a species-level phylogeny (figs. 3 and 4) and inform about patterns of introgression (fig. 5). A resolved phylogeny is important for answering a variety of questions in evolutionary biology, including estimating speciation rates (e.g., Nee 2001), inferring rates of morphological evolution (e.g., Pennell and Harmon 2013), and characterizing patterns of molecular evolution (e.g., Zhang et al. 2005). In addition to exonic sequences, noncoding (e.g., introns and intergenic regions) may also be targeted for capture or recovered through anonymous partitioning approaches. Given that they tend to evolve more quickly than exons (fig. 1*C*), these noncoding regions would aid in resolving relationships among closely related taxa, inferring rates of evolution in concert with the phylogeny, or investigating finer-scale patterns of phylogenetic discordance. Because reference mapping biases scale with divergence (fig. 1), iterative mapping is likely to be particularly useful when analyzing more rapidly evolving nongenic regions.

Many of the questions listed above require that the tree be ultrametric (i.e., scaled relative to time), but it is still computationally intractable to estimate an ultrametric species tree with genome-scale data using Bayesian methods. To address this, others have recommended restricting analyses of whole genome data to the most informative regions or combining regions with similar underlying topologies (e.g., Jarvis et al. 2015; Mirarab et al. 2015). Given the need to subset WGS data, partitioned comparative data sets are obviously well suited for this general approach (though whole exome data

would still likely need to be subsampled). Using a reduced data set, it should be possible to fix the topology to the estimated species tree and use Bayesian approaches to estimate a substitution rate scaled relative to time (and scaled relative to absolute time if fossil calibrations are used). Fixing the tree eliminates one of the most computationally intensive parts of likelihood-based phylogenetic estimation, the recalculation of the likelihood after topological rearrangement, and would facilitate an analysis using many loci for a more accurate calculation of the substitution rate per unit time.

Our iterative approach is not without important limitations. For example, we did not take indel variation into account when iterating our pseudoreferences in order to maintain a consistent coordinate system across many species. Due to the deleterious effect of frameshift mutations, indels tend to be rare in protein coding regions and we chose to ignore them within our study. Others have incorporated indel information within pseudogenomes (Holt et al. 2013; Huang et al. 2013, 2014), though these studies were focused on pairwise contrasts between very closely related genomes and did not use iteration. Extending the pseudoreference approach to efficiently incorporate small-scale indels across a phylogenetic sample remains an important goal for future studies; however, this reference-based framework will always be limiting with respect to larger-scale structural variation (e.g., chromosomal translocations and inversions). Thus, the approach outlined here will be most useful for generating comparative evolutionary genomic data sets of orthologous loci that can be used for phylogenetic and population genomic inferences. The relevance of such reference-based comparative studies should continue to grow as high quality reference genomes become increasingly common across the tree of life.

Mus Phylogenomics

Previous works focused on *Mus* systematics lacked several lineages included in this study or were uncertain with respect to the branching order within certain clades. In particular, the relationships between *M. spretus/spicilegus/macedonicus* and *M. caroli/cookii/cervicolor* have remained unclear (e.g., Lundrigan et al. 2002; Tucker et al. 2005; Tucker 2008). For example, there was conflicting evidence about the relationships among *M. spretus*, *M. spicilegus*, and *M. macedonicus* and the placement of each relative to the *M. musculus* species group. Our analyses resolved this group as monophyletic as well as the phylogenetic relationships among all ten species (fig. 3). Though discordance among *M. spretus*, *M. spicilegus*, and *M. macedonicus* is appreciable at the scale of extended targets, a majority of quartets still support the species relationships inferred from all other data sets (fig. 4). Overall, the species phylogeny is relatively well supported even when accounting for among-locus phylogenetic discordance (fig. 4). This information is crucial for effectively designing genomic or functional genetic experiments in house mice that require comparisons to closely related species. Moreover, we note that the phylogenetic relationships that we recovered were robust across individual genealogies estimated at different local scales (fig. 4) and when considering targets from smaller subsets of the whole exome capture (e.g., by chromosome). However, in one case, using extended targets alone for species tree analysis transposed the relationships at the base of the tree on a short branch with low quartet support, presumably reflecting a lack of informative sites in the alignments. These patterns suggest that the same general phylogenetic conclusions would have been apparent using a much smaller set of targeted loci as long as enough phylogenetically informative sites are present to confidently resolve relationships.

We also detected introgression between some mouse lineages. These results were not unexpected and are in strong agreement with other studies investigating whole-genome ancestry among mouse strains. Classic inbred strains derive from early breeding efforts of mouse fanciers (Beck et al. 2000), which included some crosses between species and subspecies (Ferris et al. 1982; Tucker et al. 1992; Ideraabdullah et al. 2004). The mosaic nature of classic inbred strains of mice is well known (Bonhomme et al. 1987; Yalcin et al. 2010; Didion and Pardo-Manuel De Villena 2013), but the extent of introgression has been the subject of some debate. Using a genome-wide SNP genotyping platform, Yang et al. (2011) estimated that *M. m. domesticus* strain C57BL/6 has a genome composed of ~93% *M. m. domesticus*, and ~7% *M. m. musculus*. Our estimate of 6.8% is in close agreement with their inferences, indicating that levels of introgression within sequenced genic regions is similar to genome-wide patterns based on SNVs and that variable ascertainment schemes used to populate the Mouse Diversity Genotyping Array (Yang et al. 2009) do not appear to bias overall signatures of gene flow. Additionally, we detected a strong signature of *M. m. domesticus* introgression into the wild-derived *M. m. musculus* strain PWK/PhJ, consistent with Yang et al. (2011) (fig. 5, supplementary material table S3, Supplementary Material online). The context of introgression involving this and some other wild-derived strains remains unclear. For PWK/PhJ, this could reflect natural gene flow as this strain was derived from the Czech Republic near the natural hybrid zone between *M. m. domesticus* and *M. m. musculus*. However, it has also been suggested that the haplotype structures of introgressed regions in this and a few other wild-derived inbred stains are consistent with very recent gene flow, perhaps occurring in the laboratory subsequent to strain derivation (Yang et al. 2011).

Liu et al. (2015) found 0.02–0.8% *M. spretus* ancestry within *M. m. domesticus*, and other studies have described natural introgression between these taxa (Orth et al. 2002; Keane et al. 2011; Song et al. 2011). We did not detect appreciable introgression between *M. m. domesticus* (dom^{LEWES}) and *M. spretus*, suggesting that the extent of introgression between these species is variable among individuals.

As expected, we did not detect introgression between *M. spicilegus* or *M. macedonicus* and the currently allopatric *M. spretus*. However, we did detect introgression between *M. cookii* and *M. caroli*. These species are broadly distributed throughout Eastern and Southeastern Asia and can cooccur in the same localities along with *M. cervicolor* (Suzuki and Aplin 2012). Introgression between these lineages, therefore, is not unexpected. Interestingly, these findings further support the notion that association with humans may contribute to hybridization between *Mus* species. Evidence for natural introgression within *Mus* includes cases of secondary contact following human-associated range expansions of *M. spretus* and *M. musculus* and between various *M. musculus* lineages (Palomo et al. 2009; Jones et al. 2010; Gabriel et al. 2010; Bonhomme et al. 2011; Song et al. 2011; Suzuki et al. 2013). Additionally, while the historical relationships of *M. cookii* and *M. caroli* and humans are less clear, both species (along with *M. cervicolor*) primarily occur in rice fields and nearby areas (Suzuki and Aplin 2012). This suggests that contact between lineages, and subsequent hybridization, may have been facilitated by agricultural development. Collectively, our results suggest that exome capture approaches may provide a powerful tool to reliably investigate finer-scale patterns of introgression among *Mus* species.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to thank Colin Callahan, Ryan Bracewell, Emily Kopania, Matt Jones, Nathanael Herrera, Zak Claire-Salzler, Vanessa Stewart, Tom Brekke, Dan Vanderpool, and Erica Larson for helpful discussion. Jay Storz and two anonymous reviewers provided very useful feedback during the evaluation of this manuscript. This research was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01HD073439; J.M.G.), the National Science Foundation (1146525; M.D.D.), and the National Institute of General Medical Sciences (R01GM098536; M.D.D.). Genomics support instrumentation at the University of Montana was supported by a grant from the M. J. Murdock Charitable Trust.

Literature Cited

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38:7–13.

Beck JA, et al. 2000. Genealogies of mouse inbred strains. Nat Genet. 24:23–25.

Bi K, et al. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13:403.

Bonhomme F, et al. 2011. Genetic differentiation of the house mouse around the Mediterranean basin: matrilineal footprints of early and late colonization. Proc Biol Sci. 278:1034–1043.

Bonhomme F, Guenet J-L, Dod B, Moriwaki K, Bulfield G. 1987. The polyphyletic origin of laboratory inbred mice and their rate of evolution. Biol J Linn Soc. 30:51–58.

Borowiec ML. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ. 4:e1660.

Brandt DYC, et al. 2015. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. Genes|Genomes|Genetics 5:931–941.

Bryja J, et al. 2014. Pan-African phylogeny of *Mus* (subgenus *Nannomys*) reveals one of the most successful mammal radiations in Africa. BMC Evol Biol. 14:256.

Chevret P, Veyrunes F, Britton-Davidan J. 2005. Molecular phylogeny of the genus *Mus* (Rodentia: Murinae) based on mitochondrial and nuclear data. Biol J Linn Soc. 84:417–427.

Chinwalla AT, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43:491–498.

Didion JP, Pardo-Manuel De Villena F. 2013. Deconstructing *Mus gemischus*: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse. Mamm Genome. 24:1–20.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol. 28:2239–2252.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. Proc Natl Acad Sci U S A. 104:5936–5941.

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

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 29:51–63.

Fairfield H, et al. 2011. Mutation discovery in mice by whole exome sequencing. Genome Biol. 12:R86.

Ferris SD, Sage RD, Wilson AC. 1982. Evidence from mtDNA sequences that common laboratory strains of inbred mice are descended from a single female. Nature 295:163–165.

Fountain ED, Pauli JN, Reid BN, Palsbøll PJ, Peery MZ. 2016. Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. Mol Ecol Resour. 16:966–978.

Gabriel SI, Jóhannesdóttir F, Jones EP, Searle JB. 2010. Colonization, mouse-style. BMC Biol. 8:131.

Gan X, et al. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477:419–423.

Good JM, Giger T, Dean MD, Nachman MW. 2010. Widespread over-expression of the X chromosome in sterile F_1 hybrid mice. PLoS Genet. 6:e1001148.

Green RE, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.

Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution 70:7–17.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet. 6:e1000825.

Hammer MF, Silver LM. 1993. Phylogenetic analysis of the alpha-globin pseudogene-4 (Hba-ps4) locus in the house mouse species complex reveals a stepwise evolution of t haplotypes. Mol Biol Evol. 10:971–1001.

Hedtke SM, Morgan MJ, Cannatella DC, Hillis DM. 2013. Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. PLoS One 8:e67908.
Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27:570–580.
Hird SM, Brumfield RT, Carstens BC. 2011. PRGmatic: An efficient pipeline for collating genome-enriched second-generation sequencing data using a "provisional-reference genome." Mol Ecol Resour. 11:743–748.
Holt J, Huang S, McMillan L, Wang W. 2013. Read annotation pipeline for high-throughput sequencing data. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics—BCB'13. New York : ACM Press. p. 605–612.
Huang S, Holt J, Kao C-Y, McMillan L, Wang W. 2014. A novel multi-alignment pipeline for high-throughput sequencing data. Database 2014:bau057.
Huang S, Kao C-Y, McMillan L, Wang W. 2013. Transforming genomes using MOD files with applications. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics—BCB'13. New York, NY: ACM Press. p. 595–604.
Hunter SS, et al. 2015. Assembly by Reduced Complexity (ARC): a hybrid approach for targeted assembly of homologous sequences. bioRxiv. doi: 10.1101/014662.
Ideraabdullah FY, et al. 2004. Genetic and haplotype diversity among wild-derived mouse inbred strains. Genome Res. 14:1880–1887.
Janoušek V, Munclinger P, Wang L, Teeter KC, Tucker PK. 2015. Functional organization of the genome may shape the species boundary in the house mouse. Mol Biol Evol. 32:1208–1220.
Jarvis ED, et al. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience 4:4.
Jin X, et al. 2012. An effort to use human-based exome capture methods to analyze chimpanzee and macaque exomes. PLoS One 7:e40637.
Jones EP, Van Der Kooij J, Solheim R, Searle JB. 2010. Norwegian house mice (*Mus musculus musculus/domesticus*): distributions, routes of colonization and patterns of hybridization. Mol Ecol. 19:5252–5264.
Jones MR, Good JM. 2016. Targeted capture in evolutionary and ecological genomics. Mol Ecol. 25:185–202.
Keane TM, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289–294.
Kousathanas A, Halligan DL, Keightley PD. 2014. Faster-X adaptive protein evolution in house mice. Genetics 196:1131–1143.
Larson EL, et al. 2016. Contrasting levels of molecular evolution on the mouse X chromosome. Genetics 203:1841–1857.
Laurie CC, et al. 2007. Linkage disequilibrium in wild mice. PLoS Genet. 3:1487–1495.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997.
Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18:1851–1858.
Liu KJ, et al. 2015. Interspecific introgressive origin of genomic diversity in the house mouse. Proc Natl Acad Sci U S A. 112:196–201.
Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. Syst Biol. 58:468–477.
Liu Q, et al. 2012. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics 13 (Suppl 8):S8.
Lundrigan BL, Jansa S. a, Tucker PK. 2002. Phylogenetic relationships in the genus *Mus*, based on paternally, maternally, and biparentally inherited characters. Syst Biol. 51:410–431.
Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46:523–536.
McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.
Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010:pdb.prot5448.
Mirarab S, et al. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics 30:541–548.
Mirarab S, Bayzid MS, Boussau B, Warnow T. 2015. Response to comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree." Science 350:171.
Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.
Nee S. 2001. Inferring speciation rates from phylogenies. Evolution 55:661–668.
Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 12:443–451.
Orth A, Belkhir K, Britton-davidian J, Boursot P, Benazzou T. 2002. Hybridation naturelle entre deux espèces sympatriques de souris *Mus musculus domesticus* L. et *Mus spretus* Lataste. Comptus Rendes Biol. 325:89–97.
Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2016. Biostrings: string objects representing biological sequences, and matching algorithms. R package version 2.42.1.
Palomo LJ, Justo ER, Vargas JM. 2009. *Mus spretus* (Rodentia: Muridae). Mamm Species. 840:1–10.
Panousis NI, Gutierrez-Arcelus M, Dermitzakis ET, Lappalainen T. 2014. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. Genome Biol. 15:467.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.
Pease JB, Hahn MW. 2015. Detection and polarization of introgression in a five-taxon phylogeny. Syst Biol. 64:651–662.
Pennell MW, Harmon LJ. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann N Y Acad Sci U S A. 1289:90–105.
Phifer-Rixey M, Nachman MW. 2015. Insights into mammalian biology from the wild house mouse *Mus musculus*. Elife 4:e05959.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53:131–147.
Ruffalo M, LaFramboise T, Koyuturk M. 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 27:2790–2796.
Satya RV, Zavaljevski N, Reifman J. 2012. A new strategy to reduce allelic bias in RNA-Seq readmapping. Nucleic Acids Res. 40:e127.
Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol. 33:1654–1668.
Schenk JJ, Rowe KC, Steppan SJ. 2013. Ecological opportunity and incumbency in the diversification of repeated continental colonizations by muroid rodents. Syst Biol. 62:837–864.
Shafer ABA, et al. 2016. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol Evol. doi:10.1111/2041-210X.12700.
Smedley D, et al. 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 43:W589–W598.
Smith S. a, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–716.

Song Y, et al. 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. Curr Biol. 21:1296–1301.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.
Staubach F, et al. 2012. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). PLoS Genet. 8:e1002891.
Stevenson KR, Coolon JD, Wittkopp PJ. 2013. Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. BMC Genomics 14:536.
Suzuki H, et al. 2013. Evolutionary and dispersal history of Eurasian house mice *Mus musculus* clarified by more extensive geographic sampling of mitochondrial DNA. Heredity 111:375–390.
Suzuki H, Aplin KP. 2012. Phylogeny and biogeography of the genus *Mus* in Eurasia. In: Macholan, M, Baird, SJE, Munclinger, P, editors. Evolution of the house mouse. Cambridge: Cambridge University Press. p. 35–64.
Teeter KC, et al. 2008. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res. 18:67–76.
Tucker PK. 2008. Evolutionary history of the genus *Mus*. In: Chalupa, L, Williams, R, editors. Eye, retina, and the visual system of the mouse. Cambridge: The MIT Press. p. 7-11.
Tucker PK, Lee BK, Lundrigan BL, Eicher EM. 1992. Geographic origin of the Y chromosomes in "old" inbred strains of mice. Mamm Genome. 3:254–261.
Tucker PK, Sandstedt SA, Lundrigan BL. 2005. Phylogenetic relationships in the subgenus *Mus* (genus *Mus*, family Muridae, subfamily Murinae): examining gene trees and species trees. Biol J Linn Soc. 84:653–662.
Turner LM, White MA, Tautz D, Payseur BA. 2014. Genomic networks of hybrid sterility. PLoS Genet. 10:e1004162.
Vallender EJ. 2011. Expanding whole exome resequencing into non-human primates. Genome Biol. 12:R87.
Yalcin B, et al. 2010. Commercially available outbred mice for genome-wide association studies. PLoS Genet. 6:e1001085.
Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13:329–342.
Yang H, et al. 2009. A customized and versatile high-density genotyping array for the mouse. Nat Methods. 6:663–666.
Yang H, et al. 2011. Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet. 43:648–655.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.

Associate editor: Jay Storz