MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 403–416 doi: 10.1111/1755-0998.12077

De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set

SONAL SINGHAL*†
Museum of Vertebrate Zoology, University of California, Berkeley, 3101 Valley Life Sciences Building, Berkeley, CA 94720-3160, USA, †Department of Integrative Biology, University of California, Berkeley, 1005 Valley Life Sciences Building, Berkeley, CA 94720-3140, USA

Abstract

High-throughput sequencing (HTS) is revolutionizing biological research by enabling scientists to quickly and cheaply query variation at a genomic scale. Despite the increasing ease of obtaining such data, using these data effectively still poses notable challenges, especially for those working with organisms without a high-quality reference genome. For every stage of analysis – from assembly to annotation to variant discovery – researchers have to distinguish technical artefacts from the biological realities of their data before they can make inference. In this work, I explore these challenges by generating a large *de novo* comparative transcriptomic data set data for a clade of lizards and constructing a pipeline to analyse these data. Then, using a combination of novel metrics and an externally validated variant data set, I test the efficacy of my approach, identify areas of improvement, and propose ways to minimize these errors. I find that with careful data curation, HTS can be a powerful tool for generating genomic data for non-model organisms.

Keywords: annotation, *de novo* assembly, suture zones, transcriptomes, variant discovery

Received 7 November 2012; revision received 13 December 2012; accepted 22 December 2012

Introduction

High-throughput sequencing (HTS) is poised to revolutionize the field of evolutionary genetics by enabling researchers to assay thousands of loci for organisms across the tree of life. Already, HTS data sets have facilitated a wide range of studies, including identification of genes under natural selection (Yi *et al.* 2010), reconstructions of demographic history (Luca *et al.* 2011) and broad scale inference of phylogeny (Smith *et al.* 2011). Daily, sequencing technologies and the corresponding bioinformatics tools improve, making these approaches even more accessible to a wide range of researchers. Still, acquiring HTS data for non-model organisms is nontrivial, especially as most applications were designed and tested using data for organisms with high-quality reference genomes. Assembly, annotation, variant discovery and homolog identification are challenging propositions in any genomics study (Nielsen *et al.* 2011; Baker 2012); doing the same *de novo* for non-model organisms adds an additional layer of complexity. Already, many studies have collected HTS data sets for organisms of evolutionary and ecological interest (Hohenlohe *et al.* 2010; Ellegren *et al.* 2012; Keller *et al.* 2012) and have developed associated pipelines. Some have published these pipelines to share with other researchers (Catchen *et al.* 2011; Hird *et al.* 2011; de Wit *et al.* 2012); such programs make HTS more accessible to a wider audience and serve as an excellent launching pad for beginning data analysis. However, because each HTS data set likely poses its own challenges and idiosyncrasies, researchers must evaluate the efficacy and accuracy of any pipeline for their data sets before they are used for biological inference. Evaluating pipeline success is easier for model organisms, where reference genomes and single nucleotide polymorphism (SNP) sets are more common; however, for most non-model organisms, we often lack easy metrics for gauging pipeline efficacy.

In this study, I generate a large HTS data set for five individuals each from seven phylogeographic lineages in three species of Australian skinks (family: *Scincidae*; Fig. S2), for which the closest assembled genome (*Anolis carolinesis*) is highly divergent [most recent common ancestor (MRCA), 150 Ma, Alföldi *et al.* (2011)]. These seven lineages are closely related; they shared a MRCA about 25 Ma (Skinner *et al.* 2011). This clade is the focus of a set of studies looking at introgression across lineage boundaries (Singhal & Moritz 2012), and to set the foundation

Correspondence: Sonal Singhal, E-mail: singhal@berkeley.edu

for this work, I generate and analyse transcriptomic data for lineages meeting in four of these contact zones, two of which are between sister lineages exhibiting deep divergence (*Carlia rubrigularis* N/S, *Lampropholis coggeri* C/S) and two which show shallow divergence (*Saproscincus basiliscus* C/S, *Lampropholis coggeri* N/C) (Fig. S2). I use these data to develop a bioinformatics pipeline to assemble and annotate contigs, and then, to define variants within and between lineages and identify homologs between lineages. Using both novel and existing metrics and an externally validated SNP data set, I am able to test the effectiveness of this pipeline across all seven lineages. In doing so, I refine my pipeline, identify remaining challenges and evaluate the consequences of these challenges for downstream inferences. My work makes suggestions to other researchers conducting genomics research with non-model organisms, offers ideas on how to evaluate the efficacy of pipelines, and discusses how the technical aspects of HTS sequencing can affect biological inference.

Methods

All bioinformatic pipelines are available as Perl scripts on https://sites.google.com/site/mvzseq/original-scripts-and-pipelines/pipelines, and they are summarized graphically in Figs 1a and S1. I have also shared R scripts (R Development Core Team 2011) that use GGPLOT2 to do the statistical analyses and graphing presented in this article (Wickham 2009).

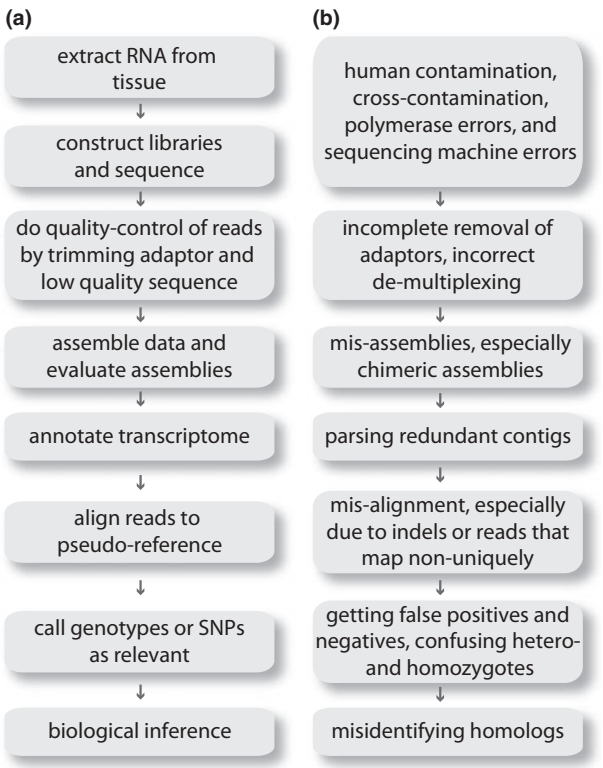

Fig. 1 (a) Pipeline for handling transcriptome data for *de novo* population genomic analyses, as presented in this study. (b) Errors introduced at each stage in the pipeline.

Library preparation and sequencing

Even though costs of sequencing continue to drop and assembly methods improve (Schatz *et al.* 2010; Glenn 2011), whole-genome *de novo* sequencing remains inaccessible for researchers interested in organisms with large genomes (*i.e.*, over 500 Mb) and for researchers who wish to sample variation at the population level. Thus, most *de novo* sequencing projects must still use some form of complexity reduction (*i.e.*, target-based capture or restriction-based approaches) to interrogate a manageable portion of the genome. Here, I chose to sequence the transcriptome, because it is appropriately sized to ensure high coverage and successful *de novo* assembly, I will surely obtain homologous contigs across taxa, I can capture both functional and noncoding variation, and assembly can be validated by comparing to known protein-coding genes.

Liver and, where appropriate, testes samples were collected from adult male and female lizards during a field trip to Australia in fall 2010 (Table S1); tissues and specimens are accessioned at the Museum of Vertebrate Zoology, UC-Berkeley. I extracted total RNA from RNA-later preserved liver tissues using the Promega Total RNA SV Isolation kit. After checking RNA quality and quantity with a Agilent Bioanalyser, I used the Illumina mRNA TruSeq kit to prepare individually barcoded libraries. Final libraries were quantified using qPCR, pooled at equimolar concentrations, and sequenced using four lanes of 100 bp paired-end technology on the Illumina HiSeq2000.

Data quality and filtration

I evaluated raw data quality by using the FASTQC v0.10.0 module (Andrews 2012) and in-house Perl scripts that calculate sequencing error rate. Sequencing error rates for Illumina reads have been reported to be as high as 1% (Minoche *et al.* 2011); such high rates can both lead to poor assembly quality and false positive calls for SNPs. To compare with these reported values, I derived an empirical estimate of sequencing error rate. To do so, I aligned a random subsample of overlapping forward–reverse reads (N = 100 000) using the local aligner BLAT v34 (Kent 2002), identified mismatches and gaps, and calculated error rates as the total number of errors divided by double the length of aligned regions. Data were then cleaned: exact duplicates due to PCR amplification were removed, low-complexity reads (e.g., reads that

consisted of homopolymer tracts or more than 20% 'N's) were removed, reads were trimmed for adaptor sequence and for quality using a sliding window approach implemented in TRIMMOMATIC v0.16 (Lohse *et al.* 2012), reads matching contaminant sources (e.g., ribosomal RNA and human and bacterial sources) were removed via alignment to reference genomes with BOWTIE2 v2.0.0-beta5 using default settings (Langmead & Salzberg 2012), and overlapping paired reads were merged using FLASH v1.0.2 (Magoč & Salzberg 2011). Following data filtration, but prior to read merging, I again estimated sequencing error rates using the method described above.

de novo assembly

Determining what kmer, or nucmer length, to use is key in *de novo* assembly of genomic data (Earl *et al.* 2011). In assembling data with even coverage, researchers typically use just one kmer (Earl *et al.* 2011); however, with transcriptome data, contigs have uneven coverage because of gene expression differences (Martin & Wang 2011). Thus, some have shown the ideal strategy for transcriptomes is to assemble data at multiple kmers and then assemble across the assemblies to reduce redundancy (Surget-Groba & Montoya-Burgos 2010). To assemble across assemblies, I first identify similar contigs using clustering algorithms [CD-HIT-EST v4.5.7; (Li & Godzik 2006)] and local alignments [BLAT v34; (Kent 2002)] and then assemble similar contigs using a lightweight *de novo* assembler [cap3; (Huang & Madan 1999)]. I used this custom multi-kmer approach along with other existing approaches, including:

1. A single kmer approach implemented in the program TRINITY R2012-01-25 [a *de novo* RNA transcript assembler, after which I used my clustering script (Grabherr *et al.* 2011)].
2. A single kmer approach implemented in ABYSS v1.3.2 [a *de novo* genomic assembler; (Simpson *et al.* 2009)], VELVET v1.1 [a *de novo* genomic assembler; (Zerbino & Birney 2008)], and SOAPDENOVO-TRANS v1.01 [a *de novo* RNA transcript assembler; (SOAP 2012)], which I implemented as a multi-kmer approach using my custom multi-kmer script.
3. A multi-kmer approach implemented in the program OASES v0.2 (Schulz *et al.* 2012).

I explore a wide range of assembly methods because generating a high-quality and complete assembly is key for almost all downstream applications. Particularly with genome assembly, which is both an art and a science, researchers should try multiple approaches and evaluate their efficacy before further analyses (Earl *et al.* 2011). However, without a reference genome, evaluating the quality of a *de novo* assembly is challenging. Here, I implement novel metrics for evaluating *de novo* transcriptome assemblies. In addition to existing metrics in the literature (N50, mean contig length, total assembly length) (Martin & Wang 2011), I determined which proportion of reads were used in the assembly, measured putative levels of chimerism in transcripts due to misassemblies, determined the proportion of assembled transcripts that could be annotated and the accuracy of these transcripts (as determined by the number of nonsense mutations or premature stop codons), and calculated the completeness and contiguity of the assembly (Martin & Wang 2011.

Here, I assembled across all individuals in a lineage rather than assembling each individual separately. Although this introduced additional polymorphism into the data which can reduce assembly efficiency (Vinson *et al.* 2005), previous work suggests the additional data lead to more complete assemblies (Singhal, unpublished).

Annotation

Following evaluation of my final assemblies, I chose the best assembly (here, Trinity-generated assemblies) for annotation to protein databases. Determining the most appropriate database for annotation is important, so I tested multiple options, including using a single-species database, whether from a distantly related but well-annotated genome or closely related but poorly annotated genome, using a multi-species database, or using a curated protein set, such as UniRef90 (Suzek *et al.* 2007). For one randomly selected lineage, I tested the efficiency and accuracy of five different reference databases:

1. The nonredundant Ensembl protein database (Flicek et al. 2012) for the lizard *Anolis carolinensis*; with a most recent common ancestor to my lineages of about ≈150 Ma, it is the closest available genome (Alföldi *et al.* 2011).
2. The nonredundant Ensembl protein data set for *Gallus gallus*, whose genome is of higher quality than the *Anolis* genome, but is more distantly related (≈250 Ma).
3. A nonredundant, curated data set (UniRef90) of proteins from a wide range of organisms, whose genes have been clustered at 90% similarity.
4. A highly redundant Ensembl protein data set for eight vertebrates sequenced to high quality (human, dog, rat, mouse, platypus, opossum, dog, chicken).
5. A highly redundant Ensembl protein data set for the 54 vertebrates whose genomes have been annotated.

I evaluated the number of matching contigs, and for the nonredundant data sets, the number of uniquely matching contigs. Distinguishing between contigs that

match and contigs that match uniquely is important, as despite my clustering during assembly, many contigs in the assembly appear redundant. These highly similar contigs likely result from misassemblies, allelic variants, alternative splicing isoforms or recently duplicated paralogs. Parsing these categories is challenging without a reference genome and when expected coverage across contigs is uneven. Especially for projects interested in functional genomics, annotation of redundant contigs remains an important and unresolved issue. Here, I try to mitigate these errors by using reciprocal BLAST best matching to annotate contigs and selecting the best match. In doing so, I likely failed to annotate recently evolved paralogs, but I should not have multiple copies of the same gene in my downstream analyses.

Once I determined the best database both with respect to efficacy and efficiency, I used a custom script to annotate the contigs using a reciprocal best-match strategy via BLASTX v2.2.24 and TBLASTX with an e-value cut-off of 1e-20 (Altschul *et al.* 1997) and defined the untranslated regions and coding sequence of the transcript using EXONERATE v2.1 (Slater & Birney 2005). Furthermore, initial tests of the annotation pipeline uncovered two challenges: first, many contigs were chimeric and consisted of multiple, combined transcripts, and second, many of the predicted open reading frames (ORFs) had nonsense mutations, largely due to frameshift mutations. To correct for chimeric contigs, I identified contigs that had two or more nonoverlapping and high-quality matches to different genes using BLASTX and split these contigs accordingly. Furthermore, I used the program FRAMEDP v1.2 to identify and correct for frameshift mutations (Gouzy *et al.* 2009).

Finally, I searched unannotated contigs against the NCBI 'nr' database using BLASTN to determine these contigs' identity. As described in the Results, these unannotated contigs largely went unidentified. Thus, although some of these unannotated transcripts have viable ORFs and/or had homologs in other lineages, and therefore, might be genes, I will be conservative and only use annotated transcripts in all downstream analyses.

Finally, to describe the putative biological functions of my annotated contigs, I determined gene ontology using BLAST2GO (Conesa *et al.* 2005).

Alignment

The first step in identifying variants or estimating gene expression levels is to align the sequencing reads to one's reference genome. Here, I use my annotated transcripts as a pseudo-reference genome (Wiedmann *et al.* 2008), thus aligning the reads used to generate the assembly to the assembly itself. Here, I tested seven different aligners [BOWTIE v0.12.7, BOWTIE2 v2.0.0-beta5, BWA v0.6.1, NOVOALIGN v2.07.07, SMALT v0.5.8, SOAPALIGNER v2.21, STAMPY v1.0.14; Langmead *et al.* (2009); Langmead & Salzberg (2012); Li & Durbin (2009); Lunter & Goodson (2011); Li *et al.* (2008)] to determine their efficacy and accuracy. These programs run the gamut of being fast, but less sensitive to being slower, and more sensitive. Here, sensitivity is defined as the aligner's ability to align reads with multiple mismatches. Previous results have shown (Li 2011) that alignment error is a common cause of miscalled SNPs, particularly alignment errors around indel sites. To evaluate these programs, I inferred genotypes from the alignments with SAMTOOLS v0.1.18 (Li *et al.* 2009). I then compared these genotypes to a small data set of known genotypes from one of the contact zones, *C. rubrigularis* N/S. In another study, I had Sanger sequenced 200–400 bp of sequence from 10 to 15 genes for the same individuals sequenced here (Singhal, unpublished). Importantly, all these genes were represented at high coverage ($\geq 20\times$ in this data set; thus, coverage is sufficiently great to ensure accurate genotype calling (Nielsen *et al.* 2012). I used these validated genotypes to determine the number of false positives (or variation called at a nonpolymorphic site) and negatives (or variation not called at a polymorphic site) in my inferred genotypes. Furthermore, I evaluated these programs based on the proportion of reads and read pairs they aligned and the concordance of SNP calls across data sets.

Variant discovery

Two major types of variant discovery are SNP identification and genotype calling. Many researchers are interested only in identifying SNPs or determining which nucleotide positions are variable in a sample of individuals. SNP-containing regions are then resequenced or genotyped for further analysis (Wiedmann *et al.* 2008). Increasingly, researchers are both identifying variable sites, and then, summarizing variation at these sites using the site frequency spectrum (SFS) or calling genotype likelihoods for each individual for subsequent population genomics analyses. SNP identification has become an easier exercise as sequencing costs dropped and coverage has increased. However, genotype calling remains a challenging proposition, particularly in diploid and polyploid individuals, as distinguishing heterozygosity, homozygosity and sequencing errors at variable sites is difficult unless there is high coverage [$\geq$ 20 $\times$, (Nielsen *et al.* (2012)]. Thus, I focus on genotype calling and its use in characterizing variation for population genomics analyses. Importantly, I assume in my approach and discussion that both alleles are expressed in each individual. Although there are some data to suggest that expression can be allele-biased,

properly controlling and testing for this issue requires having previously identified variants or genomic data (Skelly *et al.* 2011).

My results indicated that BOWTIE2 was the most effective and efficient aligner (see *Results*); thus, I used it for all downstream analyses. When identifying variants from alignment data, there are several approaches:

1 Brute strength methods, in which the read counts for given alleles at a site are calculated, and variants are determined by an arbitrary cut-off (Yang *et al.* 2011).
2 Maximum-likelihood (VARSCAN v2.2) and Bayesian methods (SAMTOOLS v0.1.18)(Koboldt *et al.* 2009; Li *et al.* 2009), in which algorithms consider strand bias, alignment quality, base quality and depth to call genotype likelihoods for individuals. These methods have been developed further to account for Hardy–Weinberg disequilibrium and linkage disequilibrium in calling and filtering variants (Li *et al.* 2009; DePristo *et al.* 2011), to use machine learning with a set of validated SNPs to improve algorithms (DePristo *et al.* 2011), and to realign reads near indel areas to ensure that inaccurate alignments do not lead to false SNPs.
3 Bayesian methods (ANGSD v0.3) which infer the SFS for all the variants in the data set, which is, in turn, used as a prior to estimate genotype likelihoods for individuals (Nielsen *et al.* 2012). This method is particularly useful for data sets with large population samples.

Here, I test these three general types of SNP and genotype discovery, using read counting, VARSCAN, SAMTOOLS and ANGSD in two sister lineage pairs for which I have validated genotypes (*C. rubrigularis* N/S and *L. coggeri* N/C). I both looked at concordance of SNP and genotype calls across methods and calculated the number of false positives and negatives.

Homolog discovery

Homologs between lineages must be identified for any comparative genomics analyses. In this study, my lineages are all closely related, so homology identification is less challenging than in many other comparative studies. However, ensuring I am identifying orthologs across lineages and not paralogs is challenging, particularly as my annotation pipeline could not conclusively distinguish orthologs and paralogs in the absence of a reference genome. With that caveat, I test three different methods for identifying homology:

1 Defining homologs by their annotation; *i.e.*, contigs that share the same annotation are assumed to be homologs.
2 Defining homologs by reciprocal best-hit BLAST, as is most commonly done in other studies (Moreno-Hagelsieb & Latimer 2008).
3 The SNP method, or defining homologs by mapping reads from one lineage to the other lineages' assembly, identifying variants and thus determining homologous sequence.

I evaluated these methods by the number of homologs found, the percent of aligned sequence between homologs and the raw number of differences between homologous sequence. I looked at homology discovery both between sister lineages and nonsister lineages, as I expect discovery across nonsister lineages will be harder.

Biological inference

Finally, I determined how robust biological inference is to the analysis method used. First, to determine how genotype calling affects downstream inference, I inferred the SFS and associated summary statistics (Tajima's D, θ, π) for one lineage across different genotype calling methods and different coverage levels using DADI 1.6.2 (Gutenkunst *et al.* 2009). Second, to determine how homology identification affects downstream inference; I determined dN/dS ratios using PAML 4.4 (Yang 2007) and raw sequence divergence for each gene across different methods of homology.

Results

Data quality and filtration

Library preparation and sequencing were successful for all individuals. On average, I generated 3.5 $\pm$ 0.5 Gb per individual. Duplication rates, low-complexity sequences and contamination levels were low (Table S2). However, aggressive filtering and merging significantly reduced the raw data set; I lost 27.1% $\pm$ 3.8% of raw base pairs per individual. As seen in Fig. S3, this strategy significantly improved the per-base quality of my data. Indeed, I was able to reduce sequencing error rates in my final data set five-fold (initial error rates: 0.3 $\pm$ 0.1%, final error rates: 0.06 $\pm$ 0.01%). These error rates are likely over-estimates, because I used the lower quality portion of the read (the tail end) to identify sequencing errors. Despite this reduction in error rates, profiling of mismatches across the reads showed that both the head and tail of the read still harbour a higher number of mismatches compared with the rest of the read. This pattern persisted even when the first and last five base pairs of each read were trimmed prior to alignment (Fig. S4). Possibly, as others have found residual adaptor sequence in their data sets despite using rigorous adaptor trimming (K. Bi, unpublished), these heightened error rates could be due to adaptor sequences leading to misalignments and spurious SNPs.

de novo assembly

To assemble my data, I tested five different programs, which employed different strategies (e.g., single kmer, built-in multi-kmer approach, my custom multi-kmer approach). I evaluated the assemblies on many metrics; here, I show data for four of these metrics. With respect to the percentage of paired reads that aligned to the assembly, SOAPDENOVO and TRINITY performed far better than the rest of the assemblers (Fig. 2a), suggesting their assemblies were more contiguous. The same two assemblers and Velvet also recovered the greatest number of annotated transcripts, measured here by the number of core eukaryotic genes found in these assemblies [core eukaryotic genes mapping approach; Parra *et al.* (2007); Fig. 2b]. OASES and TRINITY appeared to be the most accurate, as they contained the fewest number of nonsense mutations in annotated ORFs (Fig. 2c). Finally, OASES, TRINITY and SOAPDENOVO assemblies had the fewest number of putative chimeric transcripts (Fig. 2d). Looking across all these metrics, TRINITY emerges as the best assembler. Furthermore, TRINITY did a good job assembling most of the data; on average, just 8.1 ± 4.3% of contigs from other assemblies were unique to that assembly compared with TRINITY. As such, I used TRINITY assemblies for all downstream analyses. As seen in Table 1, the basic metrics of these assemblies (e.g., number of contigs, total length of assembly and N50) were fairly constant across all lineages. Unlike other studies (Comeault *et al.* 2012), I find no correlation between contig length and coverage, suggesting my assembly is not data-limited (Fig. S5). I do find a weak, but significant negative correlation between polymorphism levels and contig length ($r^2 = -0.169$, P-value <0.05; Fig. S6), suggesting that, for more variable contigs, combining across individuals negatively impacts assembly contiguity.

Annotation

After assembling the data, I annotated the assemblies to identify uniquely annotated contigs for downstream analyses and to refine the assemblies further. First, because my focal lineages are evolutionarily distant from the nearest genome (MRCA ≈150 Ma to *Anolis carolinensis*), I wanted to test the efficacy of different databases to annotate my contigs. While more complete databases did lead more annotated contigs (Table S3), the increase was marginal. Furthermore, larger databases consume significantly more computing time; here, annotating to the UniProt90 database took nearly 100 times the processor hours as annotating to *A. carolinensis*. Thus, I used the *A. carolinensis* database for all further annotations. Importantly, I could annotate these genomes to more distant relatives (*G. gallus* and *T. guttata*; MRCA ≈ 300 Ma), without seeing a significant decrease in annotation success (Table S3). This result suggests that such an

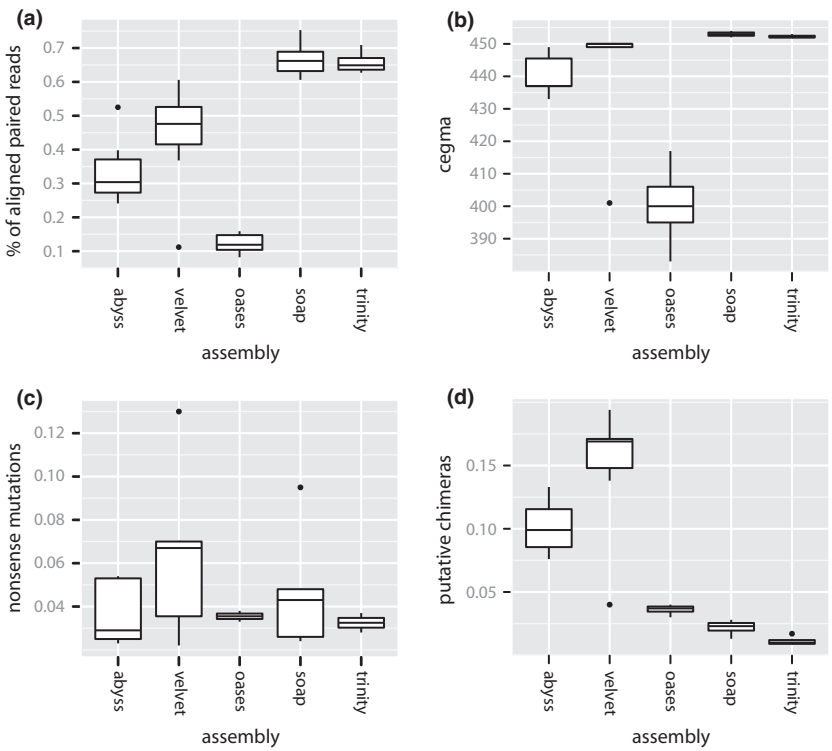

Fig. 2 Evaluation of assemblies across the seven sequenced lineages according to (a) percentage of paired reads that aligned to reference, (b) number of CEGMA genes that are found in assembly, (c) percentage of annotated coding sequences that had nonsense mutations, and (d) percentage of contigs that were putative chimeras.

Table 1 Summary of assemblies and their annotation. Complete annotated contigs are those with some 5′ and 3′ UTR sequence, as well as the full coding sequence

Assembly	Number contigs	Total length	n_{50}	Annotated contigs	Annotated contigs (unique)	Complete annotated contigs
Carlia rubrigularis, N	104648	89.1e6	1806	25198	12063	8179
C. rubrigularis, S	98280	84.3e6	1780	24323	11558	7697
Lampropholis. coggeri, N	96798	87.5e6	1972	22760	11457	7344
L. coggeri, C	106937	92.7e6	1845	23852	10894	7796
L. coggeri, S	112935	89.6e6	1549	23774	11029	7258
Saproscincus. basiliscus, C	84756	77.7e6	1951	21584	11221	7586
S. basiliscus, S	98685	83.5e6	1749	22031	11340	7696

annotation approach could work for organisms in even more genomically depauperate clades.

While annotating contigs, I identified a low percentage of chimeric contigs (≈4%), which I resolved by splitting these contigs into individual genes (Table S4). Inspecting alignments of sequencing reads to these chimeric contigs suggested that these contigs form during assembly and not due to technical errors during library preparation, as chimeric junctions generally had significantly reduced coverage. Furthermore, a small portion of the predicted ORFs of annotated contigs (≈3%) had premature stop codons. Although it is possible that these ORFs are pseudogenes (Kalyana-Sundaram *et al.* 2012), it seems more likely that they are due to assembly errors, as these contigs were generally highly expressed. Using FRAMEDP, I was able to identify and fix many of these likely frameshift errors (Table S4).

Through this pipeline, I annotated an average of 23 360 contigs per lineage, which matched to an average of 11 366 unique genes in the *A. carolinensis* genome (Table 1). I also recovered the full coding sequence for many genes; 67% of unique annotated contigs encompassed the entire coding sequence for a gene, including portions of the 5′ and 3′ UTRs. These numbers appear reasonable – the annotation for the *A. carolinensis* genome currently includes 19 K proteins, and liver tissue does not express all genes at a sufficiently high level to be represented here (Ramsköld *et al.* 2009). These genes contribute to a diversity of biological processes and serve a wide range of molecular functions, suggesting I assayed a varied portion of the transcriptome (Fig. S7).

Furthermore, my pipeline appears to be robust; almost all unannotated contigs failed to find a good match in the NCBI 'nr' database (Fig. S8). Approximately 9% of unannotated contigs matched to genes; however, further analysis of these matches showed that almost all of them matched with such low-quality to prevent annotation.

In addition, by annotating contigs rigorously to limit the number of putative duplicate contigs, I significantly reduced the redundancy of my data set. When I aligned sequencing reads to my initial, unannotated assembly, I found that ≈10% of mapped reads aligned to multiple places in the assembly. Some of these multiple alignments might be because of biological redundancy – perhaps these reads are aligning across recently duplicated genes or across common motifs in genes – but it is likely a good portion of them are aligning multiply because the initial assembly had many redundant contigs (≈50% of annotated contigs were not unique). After annotating the genome and removing redundant contigs, the percentage of mapped reads that aligned nonuniquely was reduced to approximately 2%. However, removing redundant contigs also lead to an average 8% decline in overall mapping efficiency. Thus, it seems likely these redundant contigs are 'biologically real', but we do not yet have the tools to parse such contigs properly (Vijay *et al.* 2012).

Alignment

Identifying variants and quantifying gene expression first require that sequencing reads are aligned to the reference genome. Here, I tested the efficacy of seven different alignment programs, which employ different algorithms over a range of sensitivity and speed. I evaluated these programs in three ways. First, I used my externally validated set of genotypes to see how many genotypes were inferred correctly. Almost all of the aligners performed well and led to the correct genotype at $\geq$90% of the sites. Although the false negative rate was moderately high (≈5% for most aligners), the false positive rate was low (Table 2). BOWTIE2 clearly outperformed the rest of the aligners and was thus used for all downstream analyses. Second, I evaluated how many read pairs and reads the programs could align. Although NOVOALIGN, SMALT and STAMPY are generally considered to be more sensitive aligners, I found little variation in the percentage of reads aligned across programs (Fig. 3). BOWTIE2 and STAMPY were able to align the most paired reads, which is useful as aligning paired reads reduces the likelihood of errant matches and nonunique matches (Bao *et al.* 2011). Finally, I looked at overlap in SNPs

Table 2 Accuracy of genotype inference following the use of different programs for alignment; all genotypes were inferred using samtools postalignment. Parenthetical percentages show the relative proportions of genotype types

Genotype	BOWTIE	BOWTIE2	Bwa	Novoalign	Smalt	SOAPaligner	Stampy
Right genotype	379 (89.8%)	419 (99.2%)	381 (90.3%)	383 (90.8%)	393 (93.1%)	207 (49.0%)	391 (92.7%)
Wrong genotype	29 (6.9%)	3 (0.7%)	7 (1.7%)	9 (2.1%)	6 (1.4%)	52 (12.3%)	8 (1.9%)
False negative	12 (2.8%)	0 (0%)	34 (8.1%)	30 (7.1%)	23 (5.5%)	163 (38.6%)	23 (5.5%)
False positive	3	1	1	1	1	1	5

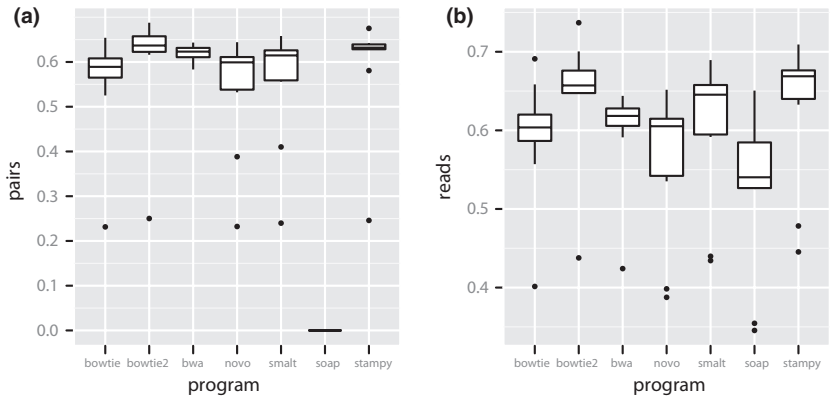

Fig. 3 Evaluation of different alignment software across three randomly selected lineages with respect to two metrics, (a) number of paired reads aligned and (b) number of reads aligned.

inferred across programs. Problematically, although all programs were fed the same reference genome and sequencing reads, I saw only moderate overlap – on average, only 77% ± 9% of SNPs were shared. Checking the raw alignments suggested that these discrepancies often arose from differences in alignment rather than differences in SNP inference postalignment. These results suggest that alignment is probably a major source of error in *de novo* HTS analyses, as has been suggested by other studies (Li 2011; Kleinman & Majewski 2012; Lin *et al.* 2012). Furthermore, although the common set of SNPs found across these programs is likely to be high-quality, considering only these SNPs is likely to lead to many false negatives. That said, when the same SNPs were called across programs, genotype inference was highly concordant; 94 ± 2% of genotype calls were the same across alignment methods, and inferred allele frequency at these SNPs was highly correlated ($r = 0.94 \pm 0.01$).

Variant discovery

After alignment, programs for variant inference are used to call SNPs and genotypes. In the previous tests, I used the variant discovery program SAMTOOLS for all analyses; here, I test a few approaches: a brute strength approach, in which I call SNPs and genotypes based solely on count data, two probabilistic methods (SAMTOOLS and VARSCAN), and a probabilistic method that uses the allele frequency spectrum (ANGSD). I first assessed accuracy of genotype calls by using my externally validated genotype set. In general, I found that all methods performed fairly well – particularly, when a SNP was identified, all programs inferred the correct genotype with high accuracy ($\geq$98%; Table 3). However, the count method of identifying variation led to many false positives, an unsurprising result given its failure to account for sequence error or alignment score. ANGSD had a high false negative rate, the reason for which remains unclear, though is possibly due to the small sample sizes used here. But, as shown by other work, ANGSD is best suited for correctly inferring the shape of the SFS (Nielsen *et al.* 2012). Comparing across all SNPs found across all programs, I found that concordance across all SNPs was moderate, similar to

Table 3 Accuracy of genotype inference across different programs for genotype inference; for all, BOWTIE2 was used for alignment. Parenthetical percentages show the relative proportions of genotype types

Genotype	ANGSD	Count data	SAMTOOLS	VARSCAN
Right genotype	520 (68.4%)	745 (98.0%)	750 (98.7%)	745 (98.0%)
Wrong genotype	3 (0.3%)	15 (2.0%)	10 (1.3%)	15 (2.0%)
False negative	230 (30.2%)	0 (0%)	0 (0%)	0 (0%)
False positive	6	134	1	12

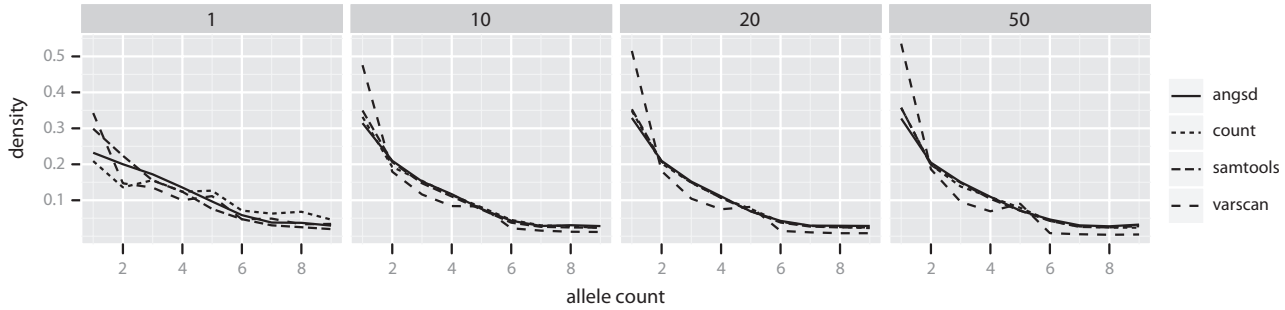

Fig. 4 Unfolded allele frequency spectrum for variants within a randomly selected lineage for sites represented at 1×, 10×, 20× and 50× coverage per individual, across different methods for genotype inference.

my comparative alignment results. On average, only 83% of SNP calls are shared across programs; this lack of concordance was largely driven by SNPs inferred from count data. More promisingly, when a site is inferred as a SNP, 98% of the genotype calls are shared across programs. Overall, these results suggested that SAMTOOLS performed the best, so I used it for all downstream analyses.

Upon defining SNPs and then genotypes for each individual, I explored how different variant discovery methods affect biological inference by constructing the SFS. Despite the only moderate levels of concordance in SNP calls, I find that the SFS is nearly identical across all the different approaches, but VARSCAN (Fig. 4). Importantly, this result only holds true when I restrict analysis to higher coverage contigs ($\geq$ 10 ×); low-coverage contigs show aberrant patterns. Although the SFS is similar across all approaches, estimates of key population genetic summary statistics (*i.e.*,θ_ω, π) vary depending on the approach – an unsurprising result given that the total number of SNPs inferred differs across approaches. Thus, prior to using these data for population genetic analyses, ascertainment bias must be factored into any downstream inference (Nielsen 2004). Finally, to look at these SNPs in greater detail, I annotated the SNPs I found in two sister lineages, with respect to how they are segregating, their location relative to the gene and their coding type (Fig. S10). Not only are the patterns of polymorphism and nonsynonymous/synonymous mutations reasonable (Begun *et al.* 2011), but there are many types of variants (*i.e.*, coding vs. noncoding, nonsynonymous vs. synonymous, fixed vs. polymorphic), which will allow the data to be used to identify adaptive signatures of molecular evolution, infer demographic history and develop markers.

Homolog discovery

To identify homologs between lineages, I tested three different methods and then evaluated their effectiveness. All three methods performed well, identifying more than 8000 homologous pairs between lineages within-genera and between-genera for a significant portion of the contig length (Fig. 5). However, with the SNP method for homology, alignment efficiency dropped off significantly in between-genera comparisons, leading to identified homologs being shorter. I chose to use reciprocal BLAST matching to identify homologs for all downstream analyses as it was able to identify more homologs than the two other methods and it worked well across evolutionary distances (Fig. 5). This approach identified 8800 homologous contigs across all seven lineages for use in comparative analyses.

Estimation of the summary statistics (sequence divergence and *dN/dS* ratios between homologs from lineage pairs) is affected by how homologs are defined (Fig. S11). Defining homologs via annotation or via reciprocal BLAST matching gives very similar results for both sequence divergence and *dN/dS*. However, using SNPs to reconstruct the homolog results in a fuzzier pattern. When I restrict the analysis to homologs with higher coverage (> 10 ×) for which there is greater confidence in SNP inference (see *Results: Variant Discovery*), all three methods are highly correlated. Thus, this method for homolog identification should account for differences in coverage, where appropriate.

Discussion

In creating and implementing a pipeline for high-throughput sequence data, I noted several possible sources of error (Fig. 1B):

1 Errors introduced during library preparation, which can include human contamination, errors introduced during PCR amplification of the library, and cross-contamination between samples.

2 Errors introduced during sequencing, the frequency and type of which are dependent on the chemistry of sequencing platform, and subsequent de-multiplexing.

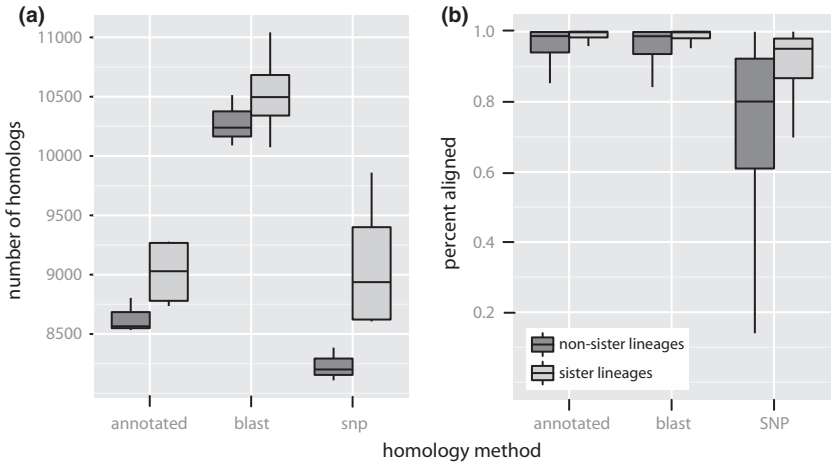

Fig. 5 Summary of different methods for homolog discovery between all lineage comparisons of interest, considering (a) number of homologs for which 75% of sequence was aligned and (b) percent of homolog aligned.

3 Errors introduced during assembly (Baker 2012), such as misassembly of reads to create chimeric contigs.
4 Errors due to misalignment of reads to assembly during variant discovery, particularly caused by indels in alignments and reads that map to multiple locations.
5 Errors in SNP and genotype calling, such as not sampling both alleles and thus mistakenly calling a homozygote.

To this, I add two additional sources of uncertainty that every study in evolutionary genomics faces – have contigs been annotated correctly and have orthologs between compared genomes been identified correctly (Chen *et al.* 2007)? Errors can arise at any stage in the process; such errors percolate through subsequent steps, likely affecting all downstream inference (Kleinman & Majewski 2012; Lin *et al.* 2012; Vijay *et al.* 2012). Whether using their own pipeline or a pre-existing pipeline, researchers will want to incorporate some of the checks suggested here to ensure that the pipeline is working well for their data and that incidence of errors is low. Moving forward, the questions become how to limit these errors and how to mitigate their effects.

All these sources of error are nontrivial, but with careful data checking and willingness to discard low-quality data, it is possible to mitigate the effects of these errors. First, as has now become standard, scrubbing reads for low-quality bases and adaptors is a must – as shown here, read cleaning can reduce error rates noticeably. When possible, merging reads from paired-end reads can further decrease error rates and will lead to more accurate estimates of coverage for expression studies (Magoč & Salzberg 2011). Second, having a high-quality assembly is crucial both for accurate annotation and variant discovery. Inferring the quality of *de novo* assemblies is challenging, as there are no clear metrics or comparisons to use (Martin & Wang 2011). However, I propose a few metrics, which can be used with transcriptome data – primarily, looking for assemblies that minimize chimerism and nonsense mutations that are contiguous, and that capture a significant portion of known key genes. Undoubtedly, errors remain in the final assemblies, but these metrics helped me select the most accurate assembly for downstream analyses. In addition, contig redundancy in final assemblies remains a pressing challenge. By using a strict reciprocal-BLAST annotation strategy, I removed many of these apparently redundant contigs. However, this approach certainly removed some biologically real contigs that were recent duplicates and alternative splicing isoforms of interest to those interested in expression differences between biological groups (Vijay *et al.* 2012). Researchers should continue to explore better methods to identify orthologs and paralogs. Until better methods are developed, using strict criteria for identifying nonredundant gene sets is a must, as most biological inference programs assume that each contig offered for analysis is a unique evolutionary unit.

Alignment and variant discovery remain notable challenges. In part, a poor-quality assembly genome truly can affect variant discovery – alignments across misassemblies can lead to errant SNP calls, particularly when misassemblies introduce indels (Li 2011). Furthermore, unless some sort of redundancy reduction is used, many contigs will be nearly identical in an assembly, leading to a high rate of nonunique alignments and miscalled SNPs. I was able to remove most redundant contigs, and thus, I reduced the proportion of nonunique alignments. I still see evidence for errors in alignment as (i) discrepancies between our externally validated SNP set and genotype calls from these alignments and (ii) the only moderate level of congruence between different approaches fuelled by the same data. The same patterns hold for SNP inference after alignment. Some of these errors are probably driven by the quality of the assembly – by removing alternative splicing isoforms and recently

duplicated genes, some of the reads probably misaligned to retained contigs although they were derived from another, rejected contig. However, many of these errors disappear at higher coverage, thus, given these data, the best approach is to rely on contigs with higher coverage −10 to 20×, at least – and to account for this ascertainment bias in any biological inference. Importantly, however, by relying on high-coverage contigs in transcriptome analyses one is biased to more slowly evolving genes, as there is a strong negative correlation between expression levels and rate of molecular evolution (Drummond & Wilke 2008).

Furthermore, to ensure the vagaries of variant discovery do not unduly influence our biological inference, we should use the genotype likelihoods and not genotype calls for downstream work. Ideally, researchers would conduct subsequent inference that use the SFS or genotype likelihoods as input, such as BAMOVA (Gompert & Buerkle 2011) or DADI (Gutenkunst *et al.* 2009), thus ensuring uncertainty in SNP and genotype calling is incorporated into model fitting. However, many analyses, particularly those used by most biodiversity researchers (*i.e.*, coalescent-based demography and phylogeny programs), require known genotypes or haplotypes. Until uncertainty is incorporated into such programs, researchers will have to arbitrarily choose cutoffs to determine most likely genotypes. In such cases, researchers might want to restrict their analyses to regions with high coverage, where calls are likely more certain (Nielsen *et al.* 2012).

Moving forward, how can we reduce the sources of errors stemming from alignment errors and genotype inference? Improved assemblies, facilitated by new long-read sequencing technologies, will certainly help. As researchers collect externally validated SNP data sets, they can use programs like GATK to recalibrate variant calling and to realign around indels (DePristo *et al.* 2011). Researchers will also increasingly sequence more individuals in a population, which will better take advantage of multi-sample methods like SAMTOOLS and ANGSD (Li *et al.* 2009; Nielsen *et al.* 2012). Finally, programs like CORTEX, which assemble across individuals to provide both a reference assembly and individual assemblies, are promising (Iqbal *et al.* 2012). Simulations suggest that this method can also better handle data with indel polymorphism.

Finally, homolog discovery is a challenge in any genome project (Chen *et al.* 2007), and this project was no exception. All three methods I tested for homolog discovery worked well, but I recommend only using a SNP-based approach between lineages that are closely related and for contigs with high coverage. Moving forward, as we acquire more comparative genomic data across the tree of life, homolog discovery should become an easier problem, as fuelled by comparative clustering programs like ORTHOMCL (Chen *et al.* 2005).

Given this, other researchers should carefully consider the benefits and challenges of working with transcriptomic data before embarking on similar studies. For researchers interested in obtaining variation data for non-model organisms and who do not require expression data, they might consider using restriction-based methods like RADtags or reduced-representation libraries (Hohenlohe *et al.* 2010) or collecting target-based capture data (Bi *et al.* 2012). Restriction-based methods are cheaper than transcriptome methods, and they do not require that genetic samples have been preserved to maintain RNA quality. However, finding homologous contigs across phylogenetic depths can be challenging, and such contigs typically cannot be annotated. Target-based capture methods can be used with low-quality DNA and have the same benefits of transcriptome data (*i.e.*, homologous contigs can be identified across phylogenetic depths and contigs can be annotated) without its disadvantages (*i.e.*, coverage is expected to even across contigs and redundancy in assemblies can be more easily handled) (Bi *et al.* 2012). However, exome-capture is more expensive than restriction-based methods and designing probes requires previously acquired genomic data. Thus, determining which approach is ideal for a given study depends on the number and quality of samples to be assayed, the amount of money available and the phylogenetic span of the samples.

Despite the challenges of HTS data generally and transcriptome data specifically, through this work I collated a large data set of over 12 K annotated contigs, spanning a wide range of biological functions, and over 100 K SNPs between lineage pairs, spanning a wide range of locations and coding types. Notably, I was able to do all these analyses using existing, open-source software and, but for assembly, using a low-end desktop machine. Genomic analyses are not just for those working with humans or mice anymore. With careful and thoughtful data curation, HTS can enable researchers to use genomic approaches to explore all the branches in the tree of life.

Acknowledgements

I gratefully acknowledge M. Chung, J. Penalba and L. Smith for technical support, and the Seqanswers.com community for providing timely and thoughtful advice. Three anonymous reviewers, R. Bell, K. Bi, J. Bragg, C.A. Buerkle, T. Linderoth, M. MacManes, C. Moritz, R. Nielsen, S. Ramirez, F. Zapata, and members of the Moritz Lab Group provided comments and suggestions during this work and on this manuscript that greatly improved its quality. Financial support for this work was provided by National Science Foundation (Graduate Research

Fellowship and Doctoral Dissertation Improvement Grant), the Museum of Vertebrate Zoology Wolff Fund, and a Rosemary Grant Award from the Society of the Study of Evolution. This work was made possible by the supercomputing resources provided by NSF XSEDE, in particular the clusters at Texas Advanced Computing Center and Pittsburgh Supercomputing Center.

References

Alföldi J, Palma FD, Grabherr M *et al.* (2011) The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, **477**, 587–591.

Altschul S, Madden T, Schaffer A *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Andrews S (2012) FastQC. Available from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Baker M (2012) *de novo* genome assembly: what every biologist should know. *Nature Methods*, **9**, 333–337.

Bao S, Jiang R, Kwan W, Wang B, Ma X, Song Y (2011) Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics*, **56**, 406–414.

Begun D, Holloway A, Stevens K *et al.* (2011) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, 310.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good J (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 403.

Catchen J, Amores A, Hohenlohe P, Cresko W, Postlethwait J (2011) Stacks: building and genotyping loci *de novo* from short read sequences. *G3 Genes Genomes Genetics*, **1**, 171–182.

Chen F, Mackey A, Stoeckert C Jr, Roos D (2005) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, **34**, 363–368.

Chen F, Mackey A, Vermunt J, Roos D (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, **2**, 383.

Comeault A, Sommers M, Schwander T *et al.* (2012) *De novo* characterization of the *Timema cristinae* transcriptome facilitates marker discovery and inference of genetic divergence. *Molecular Ecology Resources*, **12**, 549–561.

Conesa A, Gotz S, Garcia-Gomez J, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **15**, 3674–3676.

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

Drummond D, Wilke C (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **25**, 341–352.

Earl D, Bradnam K, John JS *et al.* (2011) Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Research*, **21**, 2224–2241.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 756–760.

Flicek P, Amode M, Barrell D *et al.* (2012) Ensembl 2012. *Nucleic Acids Research*, **40**, 84–90.

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

Gompert Z, Buerkle C (2011) A hierarchical Bayesian model for next-generation population genomics. *Genetics*, **187**, 903–917.

Gouzy J, Careere S, Schiex T (2009) FrameDP: sensitive peptide detection on noisy matured sequences. *Bioinformatics*, **25**, 670–671.

Grabherr M, Haas B, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **15**, 644–652.

Gutenkunst R, Hernandez R, Williamson S, Bustamante C (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, 1000695.

Hird S, Brumfield R, Carstens B (2011) PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a 'provisional reference genome'. *Molecular Ecology Resources*, **11**, 743–748.

Hohenlohe P, Bassham S, Etter P, Stiffler N, Johnson E, Cresko W (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, 1000862.

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Research*, **9**, 868–77.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, **44**, 226–232.

Kalyana-Sundaram S, Kumar-Sinha C, Shankar S *et al.* (2012) Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell*, **149**, 1622–1634.

Keller I, Wagner C, Greuter L *et al.* (2012) Population genomic signatures of divergent adaptation, geneflow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular Ecology*, doi: 10.1111/mec.12083. in press.

Kent W (2002) BLAT–the BLAST-like alignment tool. *Genome Research*, **12**, 656–64.

Kleinman C, Majewski J (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, **335**, 1302.

Koboldt D, Chena K, Wylie T *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–5.

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

Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, 25.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li W, Godzik A (2006) CD-Hit: a fast program for clustering and coomparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–59.

Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Li H, Handsaker B, Wysoker A *et al.*, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Lin W, Piskol R, Tan M, Li J (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, **335**, 1302.

Lohse M, Bolger A, Nagel A *et al.* (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**, 622–7.

Luca F, Hudson R, Witonsky D, Rienzo AD (2011) A reduced representation approach to population genetic analyses and applications to human evolution. *Genome Research*, **21**, 1087–1098.

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

Magoč T, Salzberg S (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–63.

Martin J, Wang Z (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics*, **12**, 671–682.

Minoche A, Dohm J, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, **12**, 112.

Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, **24**, 319–324.

Nielsen R (2004) Population genetic analysis of ascertained SNP data. *Human Genomics*, **1**, 218–224.

Nielsen R, Paul J, Albrechtsen A, Song Y (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Review Genetics*, **12**, 443–51.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, **7**, 37558.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–67.

R Development Core Team (2011). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org. ISBN 3-900051-07-0.

Ramsköld D, Wang E, Burge C, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Computational Biology*, **5**, 1000598.

Schatz M, Delcher A, Salzberg S (2010) Assembly of large genomes using second-generation sequencing. *Genome Research*, **20**, 1165–73.

Schulz M, Zerbino D, Vingron M, Birney E (2012) *Oases*: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.

Simpson J, Wong K, Jackman S, Schein J, Jones S, Birol I (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Singhal S, Moritz C (2012) Strong selection against hybrids maintains a narrow contact zone between morphologically cryptic lineages in a rainforest lizard. *Evolution*, **66**, 1474–89.

Skelly D, Johansson M, Madeoy J, Wakefield J, Akey J (2011) A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*, **21**, 1728–1737.

Skinner A, Hugall A, Hutchinson A (2011) Lygosomine phylogeny and the origins of Australian scincid lizards. *Journal of Biogeography*, **38**, 1044–1058.

Slater G, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Smith S, Wilson N, Goetz F *et al.* (2011) Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, **480**, 364–367.

SOAP (2012). Soapdenovo-trans. Available from http://soap.genomics.org.cn/SOAPdenovo-Trans.html.

Surget-Groba Y, Montoya-Burgos J (2010) Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Research*, **20**, 1432–40.

Suzek B, Huang H, McGarvey P, Mazumder R, Wu C (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–8.

Vijay N, Poelstra J, Kunstner 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*, **22**, 620–634.

Vinson JP, Jaffe D, O'Neill K *et al.* (2005) Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Research*, **15**, 1127–1135.

Wickham H (2009) *ggplot2: elegant graphics for data analysis*. Springer, New York. Available from http://had.co.nz/ggplot2/book.

Wiedmann R, Smith T, Nonneman D (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics*, **9**, 81.

de Wit P, Pespeni M, Ladner J *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.

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yang S, Tu Z, Cheung F *et al.* (2011) Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. *BMC Genomics*, **12**, 199.

Yi X, Liang Y, Huerta-Sanchez E *et al.* (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, **329**, 75–78.

Zerbino D, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using *de Bruijn* graphs. *Genome Research*, **18**, 821–829.

S.S. designed study, collected data, analyzed data, and wrote the manuscript.

Data Accessibility

Data are available at the following locations:

1 Specimens and tissues used in this study are accessioned at the Museum of Vertebrate Zoology, UC-Berkeley, catalog numbers 269023-269105

2 Original Illumina reads are available on SRA, entry SRA062739 under BioProject PRJNA183544

3 Final assemblies are available on DRYAD, entry doi:10.5061/dryad.7c99f

4 Scripts used are available on DRYAD, entry doi:10.5061/dryad.7c99f and at https://sites.google.com/site/mvzseq/original-scripts-and-pipelines/pipelines.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Pipeline used in this work, annotated to show (1) different approaches tested [pink], (2) the approach used for the final analysis [blue], and (3) scripts used, as named in the DataDryad package [green].

Fig. S2 A. Phylogeny of the lineages studied in this work. Boxes indicate contacts studied; the top percentage reflects the mitochondrial divergence between lineages and the bottom is nuclear. B. A map of the Australian Wet Tropics, with all identified contact zones represented by black lines. Contacts of interest in this study are labelled.

Fig. S3 Quality scores in Phred along a read; top graph shows quality prior to cleaning and filtering, bottom shows quality after cleaning.

Fig. S4 Identified mismatches between reads from a randomly selected individual and the reference sequence, A. expressed in raw numbers and B. as a density distribution.

Fig. S5 Correlation between contig length and coverage for a randomly selected final assembly.

Fig. S6: Correlation between contig length and polymorphism for a randomly-selected final assembly.

Fig. S7: Gene ontology for annotated contigs for a randomly selected lineage, with respect to cellular component, biological process and molecular function.

Fig. S8: Identifying unannotated contigs from a randomly selected assembly, as identified from a BLAST search to the NCBI 'nr' nucleotide database.

Fig. S9: Correlation in coverage between homologous, annotated contigs for a randomly selected lineage pair.

Fig. S10: Summary of SNPs found in a randomly selected lineage pair, annotated with respect to SNP and coding type.

Fig. S11: Top row shows correlation in sequence divergence and bottom row shows correlation in inferred dN dS ratios for homologs for a randomly selected lineage pair for three methods of homolog discovery: annotation, in which contigs which share the same annotation are inferred to be homologous, BLAST, in which reciprocal best-hit BLAST is used to identify homologs, and SNP methods, in which variant information is used to reconstruct one homolog with respect to another.

Table S1: Individuals included in this study and their associated locality data; individuals are accessioned at the Museum of Vertebrate Zoology at University of California, Berkeley.

Table S2: Quality control filtering and their rates for raw data, summarized across seven lineages.

Table S3: Number of contigs annotated according to different reference databases for a randomly selected assembly.

Table S4: Prevalence of chimerism, or percentage of contigs that appeared to consist of multiple genes misassembled together, and stop codons, or percentage of contigs that had nonsense mutations, in assemblies, summarized across seven lineages both before and after the data were run in the annotation pipeline.

Table S5: Number of annotated contigs which have given coverage for each individual; shown for one randomly selected lineage pair.