MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) **14**, 381–392 doi: 10.1111/1755-0998.12186

Accuracy of allele frequency estimation using pooled RNA-Seq

M. KONCZAL,* P. KOTEJA,* M. T. STUGLIK,* J. RADWAN† and W. BABIK*
*Institute of Environmental Sciences, Jagiellonian University, Gronostajowa 7, 30-387 Kraków, Poland, †Faculty of Biology, Institute of Environmental Biology, Adam Mickiewicz University, Umultowska 89, 61-614 Poznań, Poland

Abstract

For nonmodel organisms, genome-wide information that describes functionally relevant variation may be obtained by RNA-Seq following *de novo* transcriptome assembly. While sequencing has become relatively inexpensive, the preparation of a large number of sequencing libraries remains prohibitively expensive for population genetic analyses of nonmodel species. Pooling samples may be then an attractive alternative. To test whether pooled RNA-Seq accurately predicts true allele frequencies, we analysed the liver transcriptomes of 10 bank voles. Each sample was sequenced both as an individually barcoded library and as a part of a pool. Equal amounts of total RNA from each vole were pooled prior to mRNA selection and library construction. Reads were mapped onto the *de novo* assembled reference transcriptome. High-quality genotypes for individual voles, determined for 23 682 SNPs, provided information on 'true' allele frequencies; allele frequencies estimated from the pool were then compared with these values. 'True' frequencies and those estimated from the pool were highly correlated. Mean relative estimation error was 21% and did not depend on expression level. However, we also observed a minor effect of interindividual variation in gene expression and allele-specific gene expression influencing allele frequency estimation accuracy. Moreover, we observed strong negative relationship between minor allele frequency and relative estimation error. Our results indicate that pooled RNA-Seq exhibits accuracy comparable with pooled genome resequencing, but variation in expression level between individuals should be assessed and accounted for. This should help in taking account the difference in accuracy between conservatively expressed transcripts and these which are variable in expression level.

Keywords: accuracy estimation, bank vole, pool, RNA-Seq, transcriptome

Received 21 June 2013; accepted 6 October 2013

Introduction

Next-generation sequencing (NGS) technologies have resulted in enormous progress not only in the field of medicine but also in the fields of ecology and evolutionary biology. Comparative studies of natural variation at the molecular level have yielded important insights into the evolutionary history of populations, as well as the genomics of adaptation and speciation (Gilad *et al.* 2009; Rice *et al.* 2011; Radwan & Babik 2012). For example, NGS technologies have recently been instrumental in enabling findings as impressive and varied as evidence of interbreeding between modern humans and Neanderthals (Reich *et al.* 2010), the discovery that adaptive evolution results from standing genetic variation in the stickleback (Jones *et al.* 2012) and the identification of epistasis as one of the most important factors in evolution (Breen *et al.* 2012). Nonmodel, ecologically well-characterized organisms are being studied at a scale and with a precision unimaginable a few years ago (Ekblom & Galindo 2011). Unfortunately, high-quality reference genomes are still lacking for many organisms that are commonly used in evolutionary and ecological studies, mainly because the *de novo* assembly of complex genomes that include a large number of repetitive sequences remains a challenging task (Brenchley *et al.* 2012). In such cases, genome-wide information that describes functionally relevant variation may be obtained through RNA sequencing (RNA-Seq) that utilizes *de novo* reference transcriptome assembly. This approach has been broadly used in ecological genomics (Vera *et al.* 2008; Babik *et al.* 2010; Jeukens *et al.* 2010; Wolf *et al.* 2010; Salem *et al.* 2012).

RNA-Seq is an approach in which RNA molecules are selected, reverse-transcribed and then sequenced using an NGS platform (Mortazavi *et al.* 2008). Genome complexity and redundancy are reduced because only

Correspondence: M. Konczal, Fax: +48 12 664 69 12;
E-mail: mateusz.konczal@uj.edu.pl

transcribed sequences are used, which enable the *de novo* assembly of entire transcripts, even when a relatively modest amount of sequence data are available (Martin & Wang 2011). It is important to note that RNA-Seq does not reduce genomic complexity randomly, but rather produces reads from regions in which a large proportion of functionally relevant variation is expected to be located (Jones *et al.* 2012). Such variation may be assessed and compared with known variation in genes in other organisms without requiring any pre-existing genomic information. Gene expression, alternative splicing patterns and the association of both with phenotypic traits may be also studied using RNA-Seq (Lu *et al.* 2010; Barbosa-Morais *et al.* 2012).

RNA-Seq is usually less costly than genome resequencing. However, if transcripts with low levels of expression are to be assembled, greater sequencing depth may be required, which increases the overall cost. Furthermore, the cost of preparing a large number of RNA-Seq libraries, for example from many individuals, is still prohibitively high. An attractive possible solution to this problem is sample pooling (i.e. a pooled RNA-Seq). However, meaningful inferences from pooled RNA data require that allele frequencies estimated from pooled samples adequately reflect true allele frequencies. In case of RNAseq, uncertainty about population allele frequency arises not only because of sampling finite number of individuals, but also from additional stochasticity introduced due to differences in expression level among genes or even among alleles of the same gene. It may bias allele frequency estimates drastically, and to our knowledge, the extent to which these RNAseq-specific issues bias allele frequency estimates has not been explored.

Pooling strategies using DNA samples (Pool-Seq) have been comprehensively tested (Sham *et al.* 2002; Futschik & Schlötterer 2010; Kim *et al.* 2010; Gompert & Buerkle 2011; Li 2011; Zhu *et al.* 2012), and they share some of the difficulties of the pooled RNA-Seq approach. In both Pool-Seq and pooled RNA-Seq approaches, the error associated with allele frequency estimates is inversely proportional to 'true' allele frequency. Several computational approaches that have been proposed to find rare variants in DNA pools and estimate their frequencies (Druley *et al.* 2009; Bansal 2010) could possibly be applied in pooled RNA-Seq analyses as well. Furthermore, variability introduced by technical errors (inaccurate pipetting, sequencing errors, etc.) is expected to be similar for RNA and DNA samples. However, three sources of error specific to pooled RNA-Seq have not been previously studied: (i) variation in expression levels among individuals, (ii) variation in expression levels among loci and (iii) allele-specific gene expression (Fig. 1).

Substantial differences in gene expression levels commonly occur among individuals of the same sex or developmental stage and are attributable to differences in genetic background and environment. For example, Whitehead and Crawford (2006) showed that 64% of genes are differentially expressed among individuals of the teleost fish genus *Fundulus*. Other studies argue that gene expression varies extensively both within and among populations (Sandberg *et al.* 2000; Morloy *et al.* 2004; Oleksiak *et al.* 2005; Lynch & Wagner 2008;

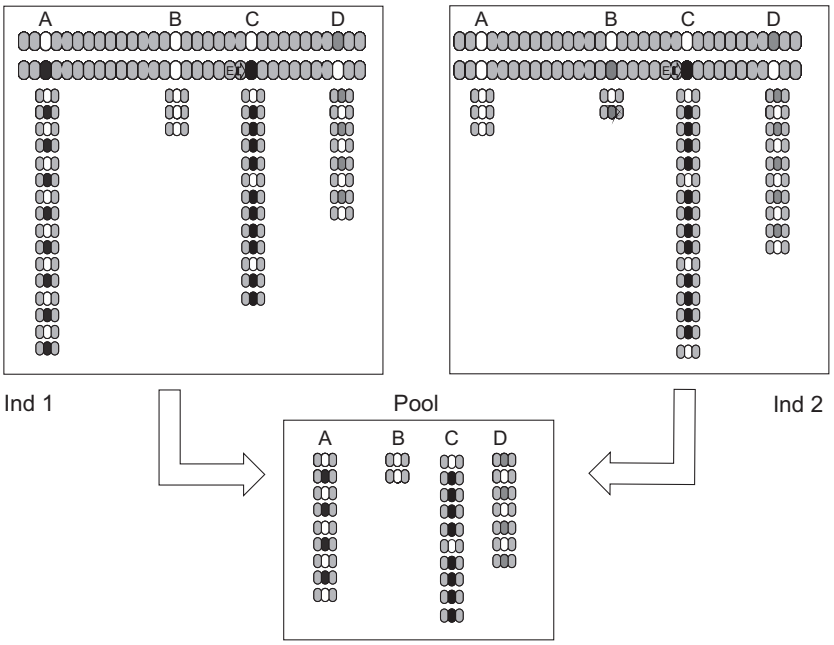

Fig. 1 Transcriptome-specific sources of error in allele frequency estimates obtained from a pooled sample. Interindividual variation in gene expression (A), interlocus variation in gene expression (B) and allele-specific gene expression (C) are compared with a locus for which the allele frequency estimate is not biased (D).

Barbosa-Morais *et al.* 2012). In the pooled RNA-Seq approach, interindividual variation in expression level may bias estimates of allele frequency because different individuals contribute unequal numbers of reads. If individuals differ greatly in their expression of a given gene, allele frequency estimates will be biased towards individuals with higher expression levels.

Interlocus variation in expression levels produces enormous differences in sequencing coverage, which may cause differences in the accuracy of allele frequency estimates for different loci. In non-normalized RNA-Seq analyses, gene expression levels may differ by six orders of magnitude (Mortazavi *et al.* 2008). The estimated allele frequencies for genes expressed at low levels will be less accurate than those obtained for genes covered by millions of reads. This problem is known to occur in transcriptomic studies, but it has not been studied in the context of pooling.

The third major issue is allele-specific gene expression (Serre *et al.* 2008; Ge *et al.* 2009). Cis-acting regulation or epigenetic silencing may cause differential expression of a diploid individual's two alleles. Although allele-specific gene expression is a widespread phenomenon that affects the expression of 20% of genes, allele expression ratios higher than 70:30 are rather rare (Serre *et al.* 2008). As a result, heterozygotes can in the vast majority of cases be successfully identified, given sufficient sequencing depth (Skelly *et al.* 2011). However, using pooling techniques, we expect frequency estimates to be distorted for over- and underexpressed alleles.

Although both potentially attractive and inexpensive, the utility of pooled RNA-Seq may be limited by the above issues, and thus, the accuracy of the allele frequency estimates obtained from pooled data should be characterized empirically. Building on results of Pool-Seq studies, we explore here additional, RNA-Seq specific, aspects of allele frequency estimation. Our general aim is to determine how various aspects of expression level variation influence allele frequency estimation.

To examine the accuracy of allele frequency estimates obtained with a pooled RNA-Seq approach, we used bank vole (*Myodes glareolus*) liver transcriptomes. This rodent species is an important organism in evolutionary, ecological and behavioural studies (Sadowska *et al.* 2005; Radwan *et al.* 2008; Boratyński & Koteja 2009; Mokkonen *et al.* 2011; Tschirren *et al.* 2012). The bank vole genome is not available, and a high-quality reference genome is unlikely to become available in the near future. The bank vole thus serves as a good example of a nonmodel organism for which obtaining genome-wide data is an important but nontrivial task. RNA samples from the livers of 10 voles were sequenced to generate both individually barcoded libraries and a pooled sample. Allele frequencies were estimated from the pool and then compared with the 'true' frequencies obtained from the individual libraries.

Materials and methods

Sample collection

Liver samples were obtained from ten bank voles (*Myodes glareolus*) from a single control line (unselected) of an artificial selection experiment (generation 13), designed to study correlated evolution of behavioural and physiological traits (Sadowska *et al.* 2008). The laboratory colony was created using voles captured in the Niepołomice Forest near Kraków (Poland) in 2000. Details related to colony protocols have been provided elsewhere (Sadowska *et al.* 2008). The experimental protocols were approved by the I Local Ethical Committee in Kraków (decision number 99/2006).

Five male and five female voles, each 75–80 days old, were euthanized using an overdose of isoflurane (Aerrane®). The animals were dissected immediately, and a small part (ca. 0.01 g) of the left liver lobe was placed in RNA*later*®. The samples were stored overnight at 4 °C and then frozen at −20 °C.

Total RNA was extracted using RNAzol® (Molecular Research Center) in accordance with the manufacturer's instructions. Residual DNA was removed with a DNA-*free* Kit (Ambion®). RNA concentration and quality were measured using Nanodrop and Agilent 2100 Bioanalyzer, respectively. All samples had an RNA integrity number (RIN) > 7.0, which indicated quality sufficient for poly-A selection and cDNA library preparation.

The pooled sample was prepared using an equal amount of total RNA from each individual. RNA concentration and quality in the pool were assessed as described earlier.

In the final step, the eleven RNA samples (ten individual and one pooled) were used in poly-A selection and the preparation of barcoded cDNA libraries by the Georgia Genomics Facility. Paired-end 2 × 100 bp sequencing was performed in one lane of an Illumina HiSeq 2000, a process that produced a similar number of reads from all the individually tagged samples together and from the pool.

Reference transcriptome reconstruction

After trimming low-quality reads using DynamicTrim (Cox *et al.* 2010), the Trinity assembler (version 2012-06-08) was employed to reconstruct the bank vole liver transcriptome *de novo* (Grabherr *et al.* 2011). For computational reasons, only reads from the pool were used in the assembly. We then processed the Trinity output by merging transcripts that were probably derived from the

same genomic location and subsequently produced transcriptome-based gene models (M. Stuglik, W. Babik & J. Radwan, unpublished data). In brief, in the first step of this process, we aggregated Trinity transcripts with overlapping ends using CAP3 (Huang & Madan 1999). The cut-offs for overlap length and per cent identity of the overlap were 40 bases and 99%, respectively. In the next step, we discarded contigs that were entirely contained within other sequences using CD-HIT (Li & Godzik 2006) (settings: identity 0.95 and word size 8). Finally, MegaBLAST was employed to merge all sequences that shared at least 70% of the length of the shortest sequence and had a minimum identity value of 0.96. Contigs were clustered, aligned and merged to form a single consensus sequence. The 'reference transcriptome' that results from this procedure should contain sequences from all exons from all genes that are expressed in at least one transcript and should thus correspond to an assembly of transcriptome-based gene models.

Mapping, SNP calling and allele frequency estimation

Because mapping algorithms take into account quality scores, we used nontrimmed reads when mapping and SNP calling. Reads that mapped onto multiple locations in the reference transcriptome were discarded.

Reads were mapped onto the reconstructed reference transcriptome using Bowtie 2 (2.0.0-beta6) and employing a very sensitive alignment approach (Langmead & Salzberg 2012). The resulting bam file was post-processed using SAMtools (Li *et al.* 2009).

SNP calling was performed separately for the 10 individually tagged samples and for the pool using mpileup in SAMtools. For SNP calling in individual samples, the default settings were applied; for SNP calling in the pool, a flat prior for the allele frequency spectrum was used.

Low-quality SNPs were filtered out of the VCF file that contained information on the individual genotypes. We excluded SNPs with individual genotypes that were based on less than five reads, and sites at which more than two variants were present. We then retained only SNPs that were reliably genotyped for all 10 individuals (Phred scores of at least 30 for SNP quality and individual genotype quality). Moreover, we discarded all contigs that contained one or more SNPs that would have led us to classify 9 or 10 of the individuals as heterozygotes. As the probability of obtaining such a sample by chance, even assuming equal allele frequencies for both variants, is only 0.01, reads that mapped onto such contigs were most probably derived from highly similar paralogues. Such stringent filtering practices allowed us to classify the genotypes at these polymorphic sites as high-quality SNPs with known 'true' allele frequencies in the sample. In the next step, we assessed how accurately these 'true' values were reflected by the pool.

Accuracy estimates

For each high-quality SNP position, the number of non-reference bases was calculated (N_O). The expected number of nonreference bases (N_E) was the 'true' allele frequency estimated from individually tagged samples multiplied by the coverage at the SNP position. The accuracy of the allele frequency estimates was quantified as the relative estimation error, which was defined as the absolute value of $(N_E\text{-}N_O)/N_E$.

To quantify the effect of allele-specific expression level (ASE) on relative estimation error, we first selected contigs showing evidence of ASE using the following procedure. For each contig, one SNP with the highest number of heterozygotes (max 8 for the reasons explained earlier) was selected. Then, for each heterozygous individual, the hypothesis of equal expression of both alleles was tested (chi-squared test), using the number of reads derived from each allele. SNPs with at least 80% of heterozygotes showing $P < 0.001$ were considered as indicators of contigs exhibiting ASE. Mean relative estimation error was compared between ASE genes and SNPs randomly selected from the data (sampling the same number of SNPs from each MAF class as in genes with ASE), and the significance of the difference between these two groups was tested using randomization test.

To assess the effect of inaccuracy in allele frequency estimation on the results of a typical population genetic analysis, we simulated a Wright–Fisher population ($N_e = 10\,000$, $u = 10^{-9}$), in which the expected distribution of allele frequencies is given by equation $\phi(i) = 4N_e u/i$ (where $0 < i < 2N$; i is the number of copies of the derived allele) (Charlesworth & Charlesworth 2010). We estimated F_{ST} under two scenarios: (i) differentiation was only due to sampling error (allele frequencies in the sample were known precisely) and (ii) differentiation was due to errors resulting from both sampling a limited number of individuals and form estimation of allele frequency from pool. In each of 10 000 simulations, we sampled one SNP from the expected distribution of allele frequencies and simulated two samples of 10 individuals each (sampling from binomial distribution with P set to population allele frequency) and calculated F_{ST} according to the formula $(H_T\text{-}H_S)/H_T$ (Hartl & Clark 2006). Next, we simulated second scenario by adding estimation error caused by pooling. We replaced the sample allele frequencies by frequencies randomly drawn from our empirical results obtained from pool for the given 'true' allele frequency. We calculated F_{ST} and

compared F_{ST} distributions between two scenarios using the Wilcoxon signed-rank test.

Accuracy of gene expression estimation

To estimate gene expression, we used RSEM package (Li & Dewey 2011). We performed TMM normalization (Robinson & Oshlack 2010) to account for differences in the mass of the RNA-Seq samples and thus provide a scaling parameter for each sample. This parameter was then used to calculate the fragments per kilobase of transcript per million fragments mapped (FPKM). FPKM was calculated for each transcriptome-based gene model in each sample. Accuracy was estimated for each contig with a mean FPKM value higher than one. Relative estimation error was calculated in the same way as for allele frequency. The mean expression level calculated from 10 individuals was used as the expected value, and observed values were the FPKM values calculated using the pool.

Results

Reference transcriptome assembly

A total of 194.1 million read pairs (2 × 100 bp) were obtained; the average per individual was 8.0 (SD 0.41) million pairs, and 114.1 million pairs were re-covered from the pool. Trimming resulted in the removal of 9.6% of the bases. Trimmed reads from the pool were used to assemble the liver transcriptome *de novo*; 181 698 contigs (contig length max: 16 742 bp; mean: 1111.8 bp; median: 429 bp; N50: 2662 bp) totalling 202.0 megabases were generated.

Transcriptome assemblers attempt to reconstruct the sequences of all the transcripts present in the sample, which results in considerable redundancy in the assembled transcriptome – a large fraction of exons will be represented many times, reflecting their presence in multiple alternatively spliced transcripts. While such redundancy reflects biological reality, it is undesirable if one wants to construct transcriptome-based gene models in order to detect polymorphism. We therefore further processed the results generated by Trinity using a custom pipeline that aims to produce transcriptome-based gene models, or a 'reference transcriptome'. The reference transcriptome comprised 146 758 contigs (contig length max, mean, median and N50, respectively: 16 742 bp, 702.7 bp, 353 bp and 1225 bp) and had a total length of 103.1 Mb (Table 1). These contigs represented protein and nonprotein coding sequences expressed in the liver.

Mapping and SNP calling

Reads that mapped uniquely onto the reference transcriptome (83.8% of raw reads) were used to identify polymorphic sites. SNPs were identified separately in individually tagged samples and the pool. SNP calling performed on the 10 individually barcoded libraries yielded 264 310 putative SNPs and 40 277 short indels that had scaled Phred quality scores of greater than 10. The same analysis on the pooled sample (which differed only in the use of the flat prior for the allele frequency distribution) yielded 246 122 putative SNPs and 40 621 short indels.

High-quality SNPs were further analysed. We found 95 contigs that were extremely heterozygous at one or more sites (probably representing pairs of paralogues), and all SNPs from these sequences were discarded. In total, we identified 23 682 high-quality polymorphisms within 4128 contigs. Only 6336 (26.8%) of high-quality SNPs within 2380 (57.7%) contigs were called from the pool. Not surprisingly, polymorphisms with rare variants (Fig. 2) and a low proportion of alternative variant reads were under-represented among the SNPs called from the pool. We found that 7% of SNPs with minor

Table 1 Overview of the assembly of a hepatic transcriptome for bank voles (*Myodes glareolus*). The transcriptome was assembled using Trinity and filtered using CAP3, CD-HIT and MegaBLAST. Statistics for the final transcriptome-based gene models are provided in the last column

	TRINITY	CAP3	CD-HIT	MEGABLAST
Min contig length	201	201	201	201
Max contig length	16 742	16 742	16 742	16 742
Mean contig length	1111.8	1061.6	1030.6	702.7
SD contig length	1529.4	1487.4	1440.9	977.3
Median contig length	429	411	406	353
N50	2662	2587	2504	1225
N contigs	181 698	173 496	171 077	146 758
N contigs > 1 kb	51 988	46 524	44 551	23 512
N contigs in N50	22 252	20 648	20 371	19 101
N bases in contigs	202 007 816	184 191 537	176 303 671	103 123 071
N bases in contigs > 1 kb	151 525 798	135 260 359	127 596 475	56 436 078

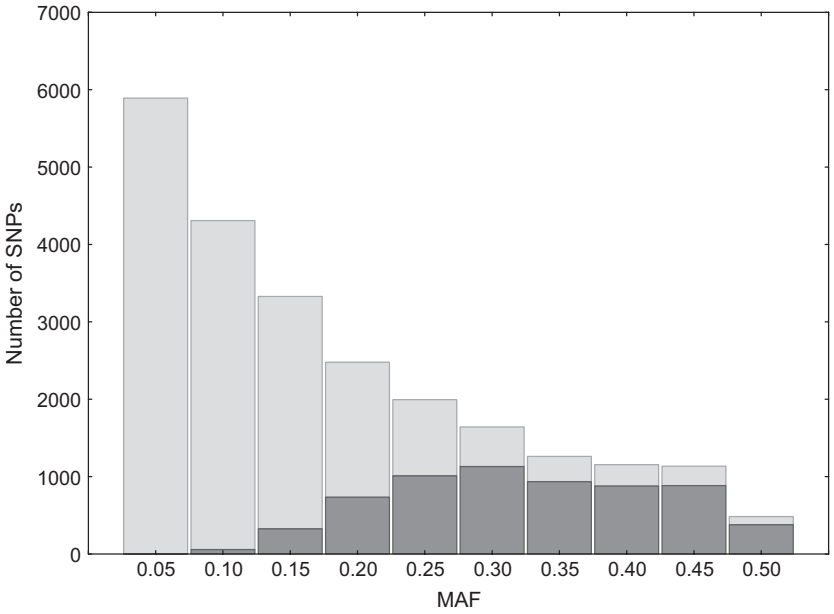

Fig. 2 Polymorphic sites discovered in the pool. The number of identified (dark) and unidentified high-quality SNPs in the pooled sample.

allele frequency (MAF) values of less than 0.25 and 74% of SNPs with MAF values greater than 0.25 were called from the pool.

Accuracy of allele frequency estimation

The observed and expected number of reads were strongly correlated ($R^2 = 0.96$; $P < 10^{-14}$; Fig 3). Mean relative estimation error was 0.21 ± 0.001 SE (median 0.16), and it was negatively correlated with the minor allele frequency (Fig. 4). Relative estimation error was relatively high for SNPs for which MAF equalled 0.05 (mean = 0.33, median = 0.25), but it decreased significantly when MAF was greater than 0.25 (mean = 0.12, median = 0.09) (Fig. 5). However, the absolute differences in frequencies did not decrease with increasing MAF (Fig. 6). We found a very weak negative correlation between the sequencing depth for a given SNP ('SNP expression level') and the relative estimation error ($R^2 = 0.002$; $P < 10^{-11}$; Figs 5 and 7).

Relative estimation error correlated significantly with the coefficient of variation in gene expression level among individuals ($R^2 = 0.04$, $P < 0.10^{-15}$; Fig. 8). We identified 43 contigs (containing 283 SNPs) with

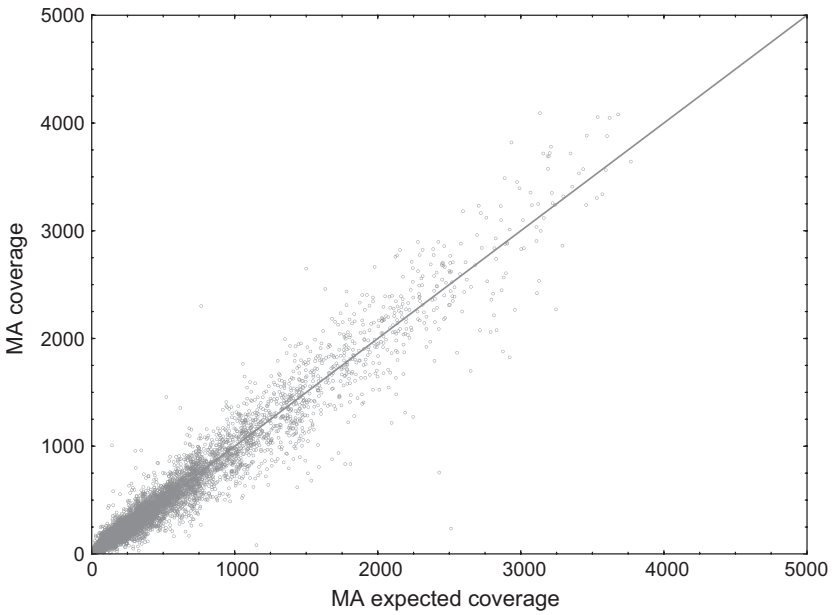

Fig. 3 Relationship between the observed and expected frequencies of minor alleles in the pool. The observed and expected numbers of bases for minor alleles in the pool are represented for 23 682 high-quality SNPs. SNPs were originally identified during individual genotyping, and the expected numbers of minor allele bases were calculated based on allele frequency and coverage

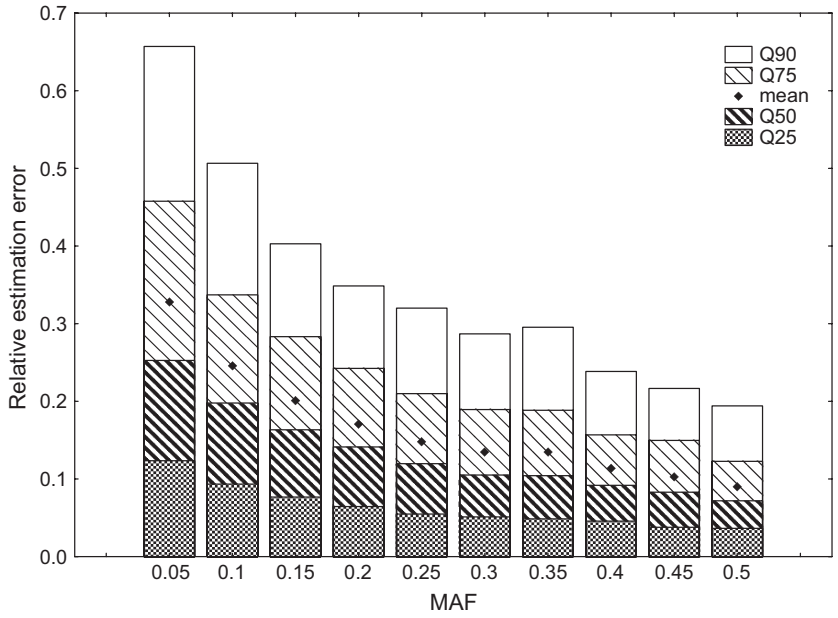

Fig. 4 Relationship between MAF and allele frequency relative estimation error values for the pooled sample. Columns represent the 25% (Q25), 50% (Q50), 75% (Q75) and 90% (Q90) percentiles for all the relative estimation error values associated with the minor allele frequency classes.

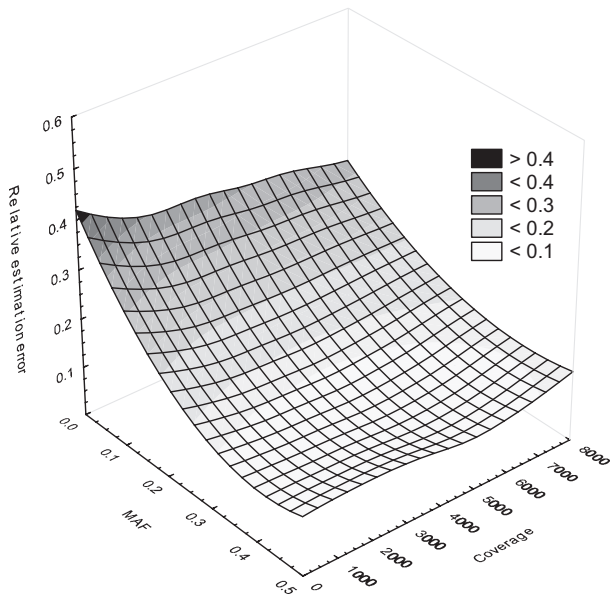

Fig. 5 Allele frequency relative estimation errors for different sequencing coverage and MAF values. The surface contours were obtained using the distance-weighted least squares method for all 23 682 high-quality SNP positions. Relative estimation error was calculated using the expected and observed number of reads of minor frequency alleles in the pooled sample.

signatures of ASE. These genes have higher relative estimation error than randomly sampled genes (mean = 0.32, $P < 0.0001$, randomization test).

F_{ST} values were generally overestimated in the pool simulation ($mean_{ind} = 0.026$, $mean_{pool} = 0.033$; Wilcoxon test: $P < 10^{-15}$). Also, we observed some extreme outliers for pools (0.3% observations higher than twice maximum F_{STind}), which suggests that in some cases, F_{ST} may be strongly overestimated due to inaccuracy in estimation of allele frequencies introduced by pooling.

Accuracy of gene expression estimation

In total, 17 861 contigs were analysed to quantify the accuracy of gene expression estimates. Mean relative estimation error was 0.14 ± 0.001 SE (median 0.12). We found a significant but very weak negative correlation between mean expression level and relative estimation error ($R^2 = 0.0004$; $P = 0.01$). The means of expression levels calculated from the individual samples were highly correlated with those calculated from the pooled sample ($R^2 = 0.998$; $P < 10^{-5}$).

Discussion

We used a nonmodel organism to quantitatively assess the accuracy of allele frequency estimates obtained from pooled RNA samples. Liver RNA samples of ten bank voles were sequenced both separately and as a pool. When we compared the allele frequencies estimated from the pool with the 'true' allele frequencies obtained from the individual samples, we found that the estimates from the pool were generally accurate.

We used only one pooled sample as variability introduced by technical issues should be similar for DNA and RNA pools, and the effect of such variability has been thoroughly explored for DNA pools (Barratt *et al.* 2002; Zhu *et al.* 2012). However, RNA pools differ from DNA

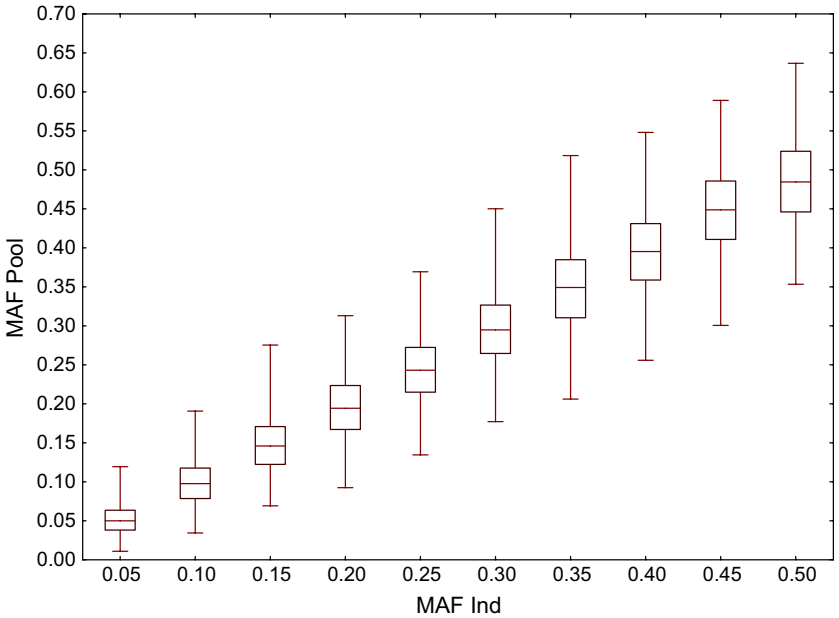

Fig. 6 Allele frequencies estimates from the pool sample within minor allele frequency classes. Boxes indicate 50% of observations, whiskers – 98% of all estimates. Horizontal lines represent medians.

pools in that the biological variation in the RNA pool is due to inherent differences in expression levels among genes and individuals. As a result, it is more important to examine the accuracy of frequency estimates for SNPs called from multiple genes found in a sample of individuals than to examine that of a few genes across a number of pools.

SNP calling in the de novo assembled transcriptome

For nonmodel organisms, transcriptome assembly is the first, crucial step of RNA-based SNP identification (Singhal 2013). This step is challenging, as divergent alleles may be identified as separate transcripts, sequences of similar paralogues may be lumped together and chimeric transcripts may arise as artefacts of the assembly process. The effectiveness of transcriptome reconstruction has been discussed in several other studies (Bao *et al.* 2011; Earl *et al.* 2011; Martin & Wang 2011; De Wit *et al.* 2012; Singhal 2013), in which different sequencing strategies and assemblers were compared. We should note that, for a comprehensive test of this problem *in silico*, a high-quality reference genome is needed (Vijay *et al.* 2013). If no reference genome is available, we have to accept an unknown rate of false positives and subsequently

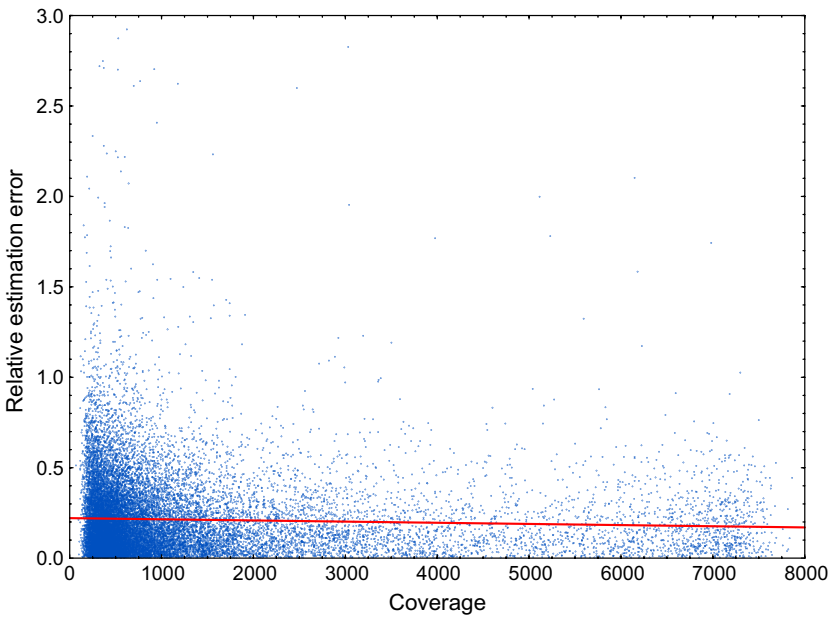

Fig. 7 Relationship between sequencing coverage and the accuracy of the allele frequency estimates. The correlation plot includes all high-quality SNPs. The regression line is given by equation $y = 0.22 - 6 \times 10^{-6}x$.

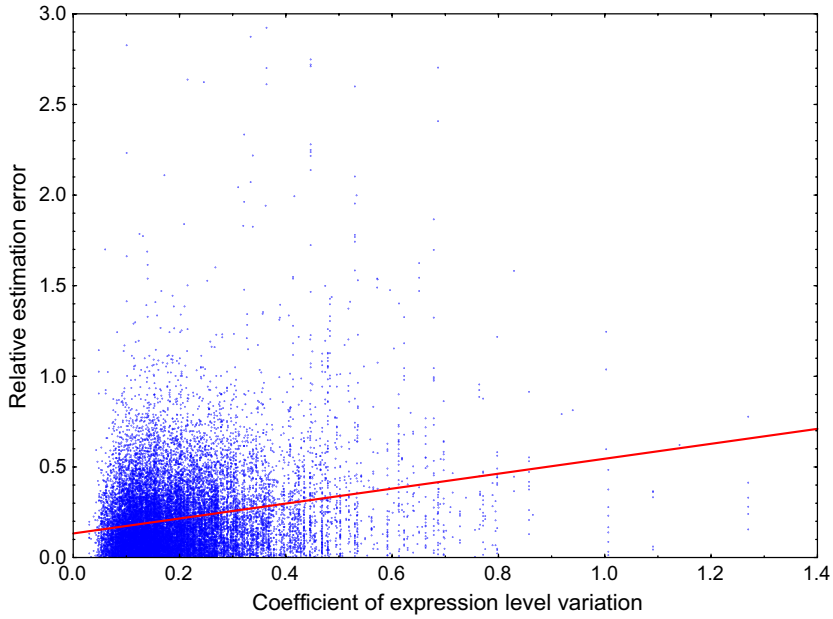

Fig. 8 Relationship between coefficient of expression level variation and the accuracy of the allele frequency estimates. The correlation plot includes all high-quality SNPs. The regression line is given by equation $y = 0.13 + 0.41x$.

test candidate SNPs in future analyses (Singhal 2013). While we recognize that there are problems related to *de novo* transcriptome assembly, we wish to emphasize that the results of our study appear to be robust to the many possible artefacts of transcriptome assembly.

First, we focused on high-quality, high-coverage SNPs that were derived from genes that were at least moderately expressed and had well-assembled transcripts. Second, by discarding SNPs called from contigs that exhibited excessive heterozygosity, we probably filtered out similar paralogues that were represented by a single transcriptome-based gene model (TGM). Heterozygosity at a biallelic locus is not expected to exceed 0.5, and, even then, we are unlikely to observe 9 or 10 heterozygotes of 10 individuals ($P = 0.01$); therefore, sites with high heterozygosity probably indicate that the contigs represent more than one region in the genome. By removing them from our analyses, we reduced the number of falsely positive SNPs caused by merging paralogues during assembly and reference transcriptome reconstruction. Third, in chimeric transcripts, individual SNPs were most probably properly called; consequently, the presence of such chimeras, which may constitute a noticeable fraction of transcripts (Edgar *et al.* 2011), should not systematically bias our results. Taken together, these filtering steps considerably reduced the number of putative SNPs, but the numbers of retained SNPs and genes were still large. This data set provided information on the accuracy of allele frequency estimates for high-quality SNPs varying in sequence coverage and minor allele frequency.

SNP identification in the sample pool

Our results suggest that SNP calling from the pool remains challenging for rarer alleles. However, this problem is common to pooling approaches and has been widely discussed in genome resequencing studies focusing on improving the discovery of SNPs with rare variants (Bansal *et al.* 2010). Several programs dedicated to SNP calling from pools, such as PoPoolation2 (Kofler *et al.* 2011), vipR (Altmann *et al.* 2011) or Varscan (Koboldt *et al.* 2009), are available, but they usually require at least two pooled samples. In most experimental and case–control studies, at least two pools are compared (Sham *et al.* 2002), and thus, the identification of polymorphic positions and the estimation of allele frequency may be considered somewhat separate tasks. If true sample allele frequencies can be accurately estimated from pools, then existing software used to identify SNPs in DNA pools could potentially be successfully applied to RNA-Seq surveys as well (Thumma *et al.* 2012). Moreover, in experimental and case–control studies, the aim is to identify SNPs whose allele frequencies differ between groups. At such sites, alternative variants should occur at least an intermediate frequency in one group and thus be easily detected with available software. For example, in our study, 74% of SNPs for which MAF values were greater than 0.25 were called from the pool. SNP discovery is therefore not a limiting factor in the identification of candidate sites from pooled samples because our results support the ability of a pooled approach to identify most of the relevant genetic variation.

Accuracy of allele frequency estimation from the pooled sample

Many population genetic analyses require estimates of allele frequencies for comparing different natural populations, experimental treatments or phenotypic classes. Sampling a finite number of individuals from population always introduce stochasticity to these estimates, which was studied elsewhere (Futschik & Schlötterer 2010; Buerkle & Gompert 2013). Obviously, as more individuals are sequenced from a population, allele frequencies are estimated more precisely and bias is eliminated. In some cases, however, we are not able to sample as many individuals as required (small groups/populations, laboratory colonies of vertebrates, etc.). Estimates of allele frequencies obtained from small samples have wider confidence intervals and are biased, which should be taken into consideration (Gompert & Buerkle 2011). In this study, we estimated the magnitude of additional uncertainty in estimates of allele frequencies introduced by variation in expression level in pooled RNA sample.

We found that estimates of allele frequency obtained from the RNA pool were acceptable for many purposes. The strong correlation between the observed and expected number of nonreference bases demonstrates the utility of pooled RNA samples in wide range of population genetic analyses. The correlation we found was only slightly weaker than that found in a study in which pooled DNA was used (Sham *et al.* 2002; Ramos *et al.* 2012). Moreover, almost no bias was present for SNPs with lower expression levels, and estimates of expression level obtained from the pool were accurate even for genes exhibiting moderate expression. However, it is important to note that we focused on genes that were at least moderately expressed by all individuals, and thus, extrapolating our results to genes expressed at very low levels would not be justified.

On the other hand, in our analysis, some gene and SNP categories demonstrate elevated estimation error. We found a negative correlation between MAF and relative estimation error, a result that has been observed for DNA pools as well (Guo *et al.* 2013). Along with SNP discovery, the low accuracy of allele frequency estimates for rare alleles remains a challenge in analyses of both DNA and RNA pools.

We found evidence that between-individual variation in expression increases estimation error only slightly but significantly: ca. 4% variation in relative estimation error can be explained by variation in expression level between individuals. Allele-specific expression also significantly influences estimates of allele frequency, but ASE seems to occur only in a minor fraction of genes (ca. 1% in our data set according to the applied criteria). These results suggest that inaccuracy in allele frequency estimation may be higher for some classes of genes, and, ideally, such genes should be identified and excluded or analysed separately. Finally, our simple simulations indicate that variation introduced by pooling systematically increases estimates of population differentiation which may result in some false positives in outliers' analyses.

Using RNA pooling has some additional limitations, namely that a well-assembled reference transcriptome is needed. When using a pooling approach, we do not have access to individual genotypes and thus have no possibility of removing sites with excessive heterozygosity. Therefore, it is worthwhile to invest time and resources in obtaining a high-quality reference transcriptome and sequencing several individually barcoded samples to test and remove the sequences of similar paralogues. These individuals can be used to explore variation in expression level between individuals, and for assessment of ASE. If such resources are available, one can control additional sources of variation in estimating allele frequency and then pooled RNA-Seq is a reliable technique to study nonmodel organisms at the genome- and population-wide scale.

Obviously, pooled approach is not applicable to analyses, which require individual genotypes (e.g. estimating admixture coefficient or estimating linkage disequilibrium among loci). Therefore, clear arguments need to be made for using this approach for molecular ecology studies. Cost effectiveness of large-scale studies is the most obvious such case. Although sequencing itself has become relatively inexpensive, library preparation remains expensive, especially when many samples are processed. With two experimental treatments, 4 replicates within treatment and only ten individuals sampled per treatment, at least 80 libraries need to be prepared. The cost of library preparation for such a modest experiment would be $4800 (NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina®) or even up to $32 000 (Illumina TruSeq Kit (Stranded Total RNA LT)). This can be reduced ten times if samples within replicates are not barcoded. For studying many populations of nonmodel species pool RNA-seq may reduce laboratory costs drastically. If studied organisms and/or organs are very small and pooling is necessary to obtain enough material for library preparation – pooled RNA-seq is the only viable solution.

Our study tested the accuracy of allele frequency estimates obtained from RNA pools sequenced using Illumina technology. We demonstrated that pooled RNA-Seq approach is a reliable, and cost-effective strategy for obtaining genome-wide information about potentially functionally relevant variation, provided that high-quality transcriptome assembly and stringent SNP-calling and filtering criteria based on sequencing of subset of individuals are used. The lack of such filtering can

result in higher inaccuracy for some categories of transcripts, which may in turn result in a higher rate of false positives in some downstream analyses. When aforementioned prerequisites are fulfilled, the accuracy obtained is very similar to that obtained for DNA pools.

Acknowledgements

This manuscript was improved by comments from C.A. Buerkle and three anonymous reviewers. The work was supported by a Polish Ministry of Science Grant N N303 816740 to PK, and Jagiellonian University Grants DSC 855 and DS 757.

References

Altmann A, Weber P, Quast C *et al.* (2011) vipR: variant identification in pooled DNA using R. *Bioinformatics*, **27**, i77–i84.

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

Bansal V (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, **26**, i318–i324.

Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, **11**, 773–785.

Bao S, Jiang R, Kwan W *et al.* (2011) Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics*, **56**, 406–414.

Barbosa-Morais NL, Irimia M, Pan Q *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**, 1587–1593.

Barratt BJ, Payne F, Rance HE *et al.* (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Annals of Human Genetics*, **66**, 393–405.

Boratyński Z, Koteja P (2009) The association between body mass, metabolic rates and survival of bank voles. *Functional Ecology*, **23**, 330–339.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA (2012) Epistasis as the primary factor in molecular evolution. *Nature*, **490**, 535–538.

Brenchley R, Spannagl M, Pfeifer M *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Charlesworth B, Charlesworth D (2010) *Elements of Evolutionary Genetics*. Roberts & Company Publishers. Greenwood Village, Colorado.

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11**, 485.

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

Druley TE, Vallania FLM, Wegner DJ *et al.* (2009) Quantification of rare allelic variants from pooled genomic DNA. *Nature Methods*, **6**, 263–265.

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

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

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

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Ge B, Pokholok DK, Kwan T *et al.* (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nature Genetics*, **41**, 1216–1222.

Gilad Y, Pritchard JK, Thornton K (2009) Characterizing natural variation using next-generation sequencing technologies. *Trends in Genetics*, **25**, 463–471.

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

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Guo Y, Samuels DC, Li J *et al.* (2013) Evaluation of allele frequency estimation using pooled sequencing data simulation. *The Scientific World Journal*, **2013**, 895496.

Hartl D, Clark A (2006) *Principles of Population Genetics*. 4th Edn, Sinauer Associates, Inc., Sunderland, Massachusetts.

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

Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (Coregonus clupeaformis spp., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular Ecology*, **19**, 5389–5403.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Kim SY, Li Y, Guo Y *et al.* (2010) Design of association studies with pooled or un-pooled next-generation sequencing data. *Genetic Epidemiology*, **34**, 479–491.

Koboldt DC, Chen K, Wylie T *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.

Kofler R, Pandey RV, Schlötterer C (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**, 3435–3436.

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

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 B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lu T, Lu G, Fan D *et al.* (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Research*, **20**, 1238–1249.

Lynch VJ, Wagner GP (2008) Resurrecting the role of transcription factor change in developmental evolution. *Evolution*, **62**, 2131–2154.

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

Mokkonen M, Kokko H, Koskela E *et al.* (2011) Negative frequency-dependent selection of sexually antagonistic alleles in Myodes glareolus. *Science*, **334**, 972–974.

Morloy M, Molony CM, Weber TM *et al.* (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621–628.

Oleksiak MF, Roach JL, Crawford DL (2005) Natural variation in cardiac metabolism and gene expression in Fundulus heteroclitus. *Nature Genetics*, **37**, 67–72.

Radwan J, Babik W (2012) The genomics of adaptation. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 5024–5028.

Radwan J, Tkacz A, Kloch A (2008) MHC and preferences for male odour in the bank vole. *Ethology*, **114**, 827–833.

Ramos E, Levinson BT, Chasnoff S *et al.* (2012) Population-based rare variant detection via pooled exome or custom hybridization capture with or without individual indexing. *BMC Genomics*, **13**, 683.

Reich D, Green RE, Kircher M *et al.* (2010) Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature*, **468**, 1053–1060.

Rice AM, Rudh A, Ellegren H, Qvarnström A (2011) A guide to the genomics of ecological speciation in natural animal populations. *Ecology Letters*, **14**, 9–18.

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, r25.

Sadowska ET, Labocha MK, Baliga K *et al.* (2005) Genetic correlations between basal and maximum metabolic rates in a wild rodent: consequences for evolution of endothermy. *Evolution*, **59**, 672–681.

Sadowska ET, Baliga-Klimczyk K, Chrzascik KM, Koteja P (2008) Laboratory model of adaptive radiation: a selection experiment in the bank vole. *Physiological and Biochemical Zoology*, **81**, 627–640.

Salem M, Vallejo RL, Leeds TD *et al.* (2012) RNA-seq identifies SNP markers for growth traits in rainbow trout. *PLoS ONE*, **7**, e36264.

Sandberg R, Yasuda R, Pankratz DG *et al.* (2000) Regional and strain-specific gene expression mapping in the adult mouse brain. *Proceedings of the National Academy of Sciences, USA*, **97**, 11038–11043.

Serre D, Gurd S, Ge B *et al.* (2008) Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic Cis-acting mechanisms regulating gene expression. *PLoS Genetics*, **4**, 2.

Sham P, Bader JS, Craig I, O'Donovan M, Owen M (2002) DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, **3**, 862–871.

Singhal S (2013) De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. *Molecular Ecology Resources*, **13**, 403–416.

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

Thumma BR, Sharma N, Southerton SG (2012) Transcriptome sequencing of Eucalyptus camaldulensis seedlings subjected to water stress reveals functional single nucleotide polymorphisms and genes under selection. *BMC Genomics*, **13**, 364.

Tschirren B, Andersson M, Scherman K, Westerdahl H, Råberg L (2012) Contrasting patterns of diversity and population differentiation at the innate immunity gene toll-like receptor 2 (tlr2) in two sympatric rodent species. *Evolution*, **66**, 720–731.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Vijay N, Poelstra JW, Künstner A, Wolf JBW (2013) 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.

Whitehead A, Crawford DL (2006) Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Science, USA*, **103**, 5425–5430.

Wolf JBW, Bayer T, Haubold B *et al.* (2010) Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Molecular Ecology*, **19**, 162–175.

Zhu Y, Bergland AO, González J, Petrov DA (2012) Empirical validation of pooled whole genome population re-sequencing in Drosophila melanogaster. *PLoS ONE*, **7**, e41901.

M.K., W.B., P.K. and J.R. conceived and designed the experiment; MK performed the experiment; M.K. and W.B. analysed the data; M.K., P.K., W.B. and M.S. contributed reagents/materials/analysis tools; M.K., W.B., P.K., J.R. and M.S. wrote the paper.

Data Accessibility

Raw sequences: NCBI BioProject PRJNA222572; reference transcriptome, variant calling files (VCF), files containing numbers of reads and statistics, custom Python scripts: Dryad Digital Repository entry. doi: 10.5061/dryad.bh23t.