Received Date : 13-Dec-2014

Revised Date : 13-Mar-2015

Accepted Date : 18-Mar-2015

Article type : Invited Reviews and Syntheses

SNP genotyping and population genomics from expressed sequences –current advances and future possibilities.

Pierre De Wit[1], Melissa H Pespeni[2] & Stephen R Palumbi[3]

[1] University of Gothenburg, Dept. of Biology and Environmental Science, Sven Lovén Centre for Marine Sciences - Tjärnö, Hättebäcksvägen 7, SE-452 96 Strömstad, Sweden.

[2] University of Vermont, Dept. of Biology, Marsh Life Sciences, Rm 326A, 109 Carrigan Drive, Burlington, Vermont 05405, USA.

[3] Stanford University, Dept. of Biology, Hopkins Marine Station, 120 Ocean view Blvd., Pacific Grove, CA 93950, USA.

Keywords: RNA-Seq, transcriptome assembly, SNP discovery, population genomics, genotyping.

Corresponding author: Pierre De Wit, University of Gothenburg, Department of Biology and Environmental Science, Sven Lovén Centre for Marine Sciences - Tjärnö, Hättebäcksvägen 7, SE-452 96 Strömstad, Sweden. Fax: +46 31 786 1333. Email: pierre.de_wit@bioenv.gu.se

Running title: RNA-Seq for population genomics

Abstract

With the rapid increase in production of genetic data from new sequencing technologies, a myriad of new ways to study genomic patterns in non-model organisms are currently possible. Because genome assembly still remains a complicated procedure, and because the functional role of much of the genome is unclear, focusing on SNP genotyping from expressed sequences provides a cost-effective way to reduce complexity while still retaining functionally relevant information. This review summarizes current methods, identifies ways that using expressed sequence data benefits population genomic inference, and explores how current practitioners evaluate and overcome challenges that are commonly encountered. We focus particularly on the additional power of functional analysis provided by expressed sequence data and how these analyses push beyond allele pattern data available from non-function genomic approaches. The massive datasets generated by these approaches create opportunities and problems as well – especially false positives. We discuss methods available to validate results from expressed SNP genotyping assays, new approaches that sidestep use of mRNA, and review follow up experiments that can focus on evolutionary mechanisms acting across the genome.

Introduction

We currently live in what has been dubbed "the golden age of DNA sequencing". New high-throughput sequencing technologies promise to continue to make DNA sequencing cheaper and easier: DNA sequence costs have dropped five orders of magnitude in the last ten years. In combination with increased capacity of computing infrastructures, this has allowed researchers in the fields of molecular ecology and population genetics to upgrade analysis methods in a

myriad of different ways. However, there are numerous pitfalls within these methods that need to be taken into account in order to avoid drawing false conclusions from massive high-throughput sequence datasets. Using large datasets to find and test genes with particular evolutionary patterns is both the promise and the challenge of these new tools.

Particularly, genome/transcriptome assemblies are often incomplete, poorly annotated, and can contain large fractions of chimaeric sequences (Cahais *et al.* 2012). Also, error rates in sequencing machines, while usually low (Ross *et al.* 2013), can still be a nuisance when the output is extremely high (for example, the Illumina HiSeq 2500 currently outputs 1000 Gb in a single run). In addition to these technical issues, there are also biological problems to consider, such as recent gene duplication events, genomic repeat regions and high polymorphism rates, that complicate assembly.

One way of reducing the complexity of genomes in order to facilitate population genomic analyses, especially in non-model systems, is to focus on expressed sequences (Wang *et al.* 2009; Gayral *et al.* 2013). A focus on expressed genes not only reduces the complexity substantially, but also allows for greater accuracy of functional annotation than in reduced-representation genomic DNA libraries. In addition, due to the nature of data from coding genes, there are a number of quality-control steps that are highly useful for trying to distinguish biological patterns from technical artifacts. These advantages allow the basic raw data of SNP analysis to be tested against neutral expectations at a series of levels beyond typical outlier approaches.

Expressed sequences can be targeted in several ways. A direct approach is to create libraries from mRNA transcribed by individuals in a population, and sequence the full transcriptome using RNASeq or other sequencing approaches. A second is to use exome-capture, in which regions of expressed genes are synthesized as oligonucleotides, attached to beads or other

substrate, and used to capture short DNA regions that are homologous to the oligos (Teer & Mullikin, 2010; Stillman & Armstrong, 2015). Such capture libraries have been printed on arrays for analysis of human polymorphisms because the vast majority of human genetic variants with large disease effects are in the 1% of the genome that is coding (Choi *et al.* 2009).

Unfortunately, development of exome capture arrays is expensive for non-model species, and requires substantial processing of each individual DNA sample. An emerging alternative is to use genomic DNA sequencing at low genome coverage (1-2x) and take advantage of sensitive mapping routines and a transcriptome assembly to sift out the expressed sequence regions (e.g. Doyle *et al.* 2014).

In this review, we attempt to summarize current methods for SNP marker development and genotyping using RNA sequencing, although the principles apply to any source of expressed sequence genotype data. Furthermore, we review how these SNPs are currently used within the field of population genomics and molecular ecology. We try to identify some of the major issues that complicate analysis and potential ways to overcome them. We devote particular attention to the power of expressed sequence data and the ways they can be used to evaluate inferences of SNP genotype data and allele frequency variation.

Assembly quality

Sequencing from mRNA samples generates a wealth of short DNA sequence reads from random places in the transcriptome. As a result, one of the key issues of SNP marker development from genomic/transcriptomic data is the quality of the reference assembly against which these reads are compared (see e.g. Grabherr *et al.* 2011; Cahais *et al.* 2012). The ideal transcriptome assembly for population genomic or comparative genomic analyses has one representative, complete sequence for each gene, i.e. isoforms and allelic variants have a single sequence

representative while gene families and recent gene duplications are maintained as separate sequences (unless the specific goal of a project is to study splice variation-related issues). Attaining this goal has a number of bioinformatic and biological challenges. In this section, we discuss these challenges and review approaches for evaluating transcriptome assembly. Using data from several comparative studies we also aggregate a "best practice" pipeline for assembly creation and evaluation in order to optimize a reference transcriptome for SNP marker quality.

Transcriptome challenges and solutions

Biological complexity and technical challenges can result in errors that clutter a transcriptome assembly, reducing the proportion of complete gene sequences. Examples of biological complexity that can challenge the reconstruction of gene sequences include gene duplications, allelic variants, alternative splicing, and stochastic changes in expression ("transcriptional noise") (Huh & Paulson, 2010). Examples of technical and computational inaccuracies include sequencing errors and the fusion of the ends of two transcripts to form a chimaera artifact (see BOX 1). All together, a large proportion of contigs from an unfiltered initial transcriptome assembly may be composed of sequences that are DNA contamination, incomplete gene fragments, chimaeras, splice and allelic variants considered as two separate gene sequences, and recent gene duplications mistaken to be one gene sequence (see Cahais *et al.* (2012): Fig. 6; see BOX 1).

Fortunately, many of these errors can be identified by working with contigs that have predicted open reading frames (ORFs). ORF prediction can be done easily with publically available programs (e.g. StarORF (http://star.mit.edu/orf/), ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) or TransDecoder.pl (Distributed with the Trinity assembly software)) that recognize start and stop codons and nonsense sequence. For

organisms that have recently undergone full genome duplication events, such as for many plants, the program findorf has been developed that can help simultaneously disentangle homologs and predict ORFs (Krasileva *et al.* 2013). ORF prediction goes a long way to excluding DNA contamination, incomplete sequence fragments, sequencing errors that result in frame shifts and false stop codons, and some chimaeras. This approach will tend to de-emphasize the 3' UTRs of mRNA, which do not have open reading frames. As a result, high quality mRNA preparations are needed so that full length coding gene regions can be assembled. Availability and processing high quality mRNAs can be a major roadblock to using expressed gene approaches, though recent development of DNA-based methods of exome capture are relieving this problem.

The downstream effects of misassembled transcripts are creation of many false SNPs when paralogous sequence changes are mistaken for polymorphisms, and the discarding of true SNPs when allelic differences are treated as two separate genes rather than one. Members of a paralogous gene family can be mistakenly collapsed into a single representative contig. These errors occur during assembly due to the blending of reads from similar transcripts into a single sequence. Such errors can sometimes be identified using results from a tBLASTn search, querying translated assembly contigs against a high quality protein database from a closely related species. This reverse annotation process will reveal erroneously collapsed contigs when multiple orthologous proteins from the reference match a single collapsed contig (O'Neil & Emrich, 2013: Fig 4).

From tBLASTn results, one can also calculate another metric of assembly quality, the "collapse factor", which is simply the mean number of reference proteins that match each contig. Rather than 1:1 orthologous matches between reference and new transcriptomes, there may be several paralogous reference protein sequences that match a single contig of erroneously

collapsed paralogs. Barring true differences in paralog numbers between the new and reference genomes, a better assembly will have a collapse factor near 1, while poorer assemblies will have larger collapse factors (O'Neil & Emrich, 2013). A number of publically available scripts have been developed to calculate such transcriptome quality metrics (including within Transrate v0.2.0 (Smith-Unna *et al.* 2014) and in the Galaxy pipeline associated with Cahais *et al.* (2012)).

However, in the absence of a high quality reference resource, these potential effects on downstream analyses suggest that erring on the conservative side of assembly, that is keeping allelic variants as separate "genes" rather than potentially collapsing paralogs, would reduce the number of potential false positives at the expense of increasing potential false negatives. Other assembly errors, such as chimeras, should not affect variant detection for the purposes of downstream population genomic analyses, though they will affect transcriptome accuracy and gene annotation.

How to evaluate transcriptome assemblies

A number of computational approaches have been developed for evaluating the accuracy and completeness of transcriptome assemblies. Variation between assemblies due to differences in assembler algorithms or assembly parameters can be measured quantitatively through measures of contiguity, such as median contig length, the number of contigs, and N50 (see BOX 2). However, the correctness of an assembly does not correlate well with statistics of contiguity (Salzberg *et al.* 2012). Beyond quantitative metrics of contiguity, there are important qualitative measurements that require comparisons to a reference transcriptome of a closely related species (< 10% sequence divergence (Vijay *et al.* 2013)) or to curated databases such as SwissProt (www.uniprot.org) or core conserved genes in eukaryotic genomes (eukaryotic orthologous groups, COGs (Tatusov *et al.* 2003; Parra *et al.* 2007)). A commonly used strategy to

estimate the quality and completeness of an assembly is based on BLAST hits to public databases such as Uniprot (www.uniprot.org). Though, as expected, this approach can be limited for non-model organisms that are not well represented in such databases (Feldmeyer *et al.* 2011). There are also several publically available packages that incorporate both quantitative and qualitative measures of transcriptome quality (Transrate [http://hibberdlab.com/transrate/], mRNAmarkup [https://github.com/BrendelGroup/mRNAmarkup], or the Galaxy pipeline from Cahais *et al.* [http://kimura.univ-montp2.fr/PopPhyl/resources/datasets/popphyl-galaxy.tar.gz]. In BOX 2, we list different evaluation methods for assembly quality.

"Best practice" transcriptome assembly guidelines suggested by current literature

There are a number of choices that can affect assembly quality and therefore SNP marker development: how much to sequence, what tissues or developmental stages to sequence from, what type of sequencing platform to use, how to process sequence data before assembly, which assembler to use and how to optimize parameters, and finally how to process and evaluate assemblies. Fortunately, a number of comparative studies have been performed to generate recommendations. Here we summarize these "rules of thumb" that have been generated to date.

A recent study by Francis *et al.* (2013) addressed the question of optimal sequencing depth to maximize coverage for *de novo* transcriptome assembly in non-model organisms. Using regular increments of read counts from sequence data from animal taxa across six different phyla, they identified 30-60M reads as the range beyond which the discovery of new genes diminishes and the fraction of sequencing errors in highly expressed genes accumulates. They also used sequence data from mouse heart tissue to be able to compare results to a reference genome. For all seven organisms, BLAST comparisons to conserved orthologs showed that the discovery of additional conserved eukaryotic orthologous genes (COGs) diminished beyond 30 million reads

(see Francis *et al.* (2013); Figs. 4-5). Interestingly, the same optimal range of 30-60M reads for transcriptome assembly was identified using sequence data from human cell cultures and mouse tissue in the study introducing the Oases assembler (Schulz *et al.* 2012). Vijay *et al.* (2013) also make a recommendation regarding coverage of 100M reads or 500-800x coverage for optimal assembly considering the effects on downstream gene expression analyses (Vijay *et al.* 2013).

Regarding starting material, in general, Francis *et al.* (2013) found that having multiple tissues or RNA extracted from whole-animals recovered more transcripts and discovered more conserved genes with less sequencing effort. Sequencing from multiple developmental stages has been an important strategy for maximizing exon coverage as well as complete transcript recovery (Vera *et al.* 2008). The logic behind this is that most genes are alternatively spliced (Wang *et al.* 2008), exon skipping is a major type of alternative splicing (Sultan *et al.* 2008), and exon usage varies substantially depending on the tissue or cell type in which a gene is expressed (Sultan *et al.* 2008; Wang *et al.* 2008). As a result, sampling across multiple developmental stages captures variation in isoform expression. Fortunately, for transcriptome assembly, the Trinity assembler clusters putative isoform variants as "comps" or components with different isoform numbers (Grabherr *et al.* 2011). Components as putative isoform variants can be further collapsed based on sequence similarity using the program CD-HIT-EST (Li & Godzik, 2006). There is however a risk here of accidentally collapsing paralogous transcripts while collapsing the intended isoform variants. The erroneous collapse of paralogs should be assayed using reverse annotation with a protein database of a closely related species as described above. Another method allowing for identification of paralogs is construction of phylogenetic trees of gene families using known sequences from closely related species as outgroups (e.g. Remm *et al.* 2001). This method has the advantage of taking into account the rate of sequence evolution within the gene family of interest, rather than relying on universal estimators of sequence similarity.

Another important consideration is the number of individuals from which to sequence for a *de novo* assembly. In general, sequencing from fewer individuals will reduce the probability of SNP or isoform variants of a single gene among individuals resulting in the erroneous separation of contigs. However, considering the above recommendation to sequence from across a range of developmental stages/sexes/tissues/physiological states to capture transcripts expressed at various conditions requires sampling multiple individuals for study organisms that are not clonal. Therefore, there must be a balance to maximize transcriptome coverage while minimizing the potential of variants that may incorporate error into the assembly. This category of errors, however, may be reduced by using pre-assembly read processing algorithms such as khmer (Crusoe *et al.* 2014), may be captured as "comps" in a Trinity assembly (Grabherr *et al.* 2011), or can be detected via alignments to reference databases post-assembly (Cahais *et al.* 2012).

An important consideration in this respect will be the overall goal of the particular study. In many population genomic studies, for example, the ultimate goal is a comparison of transcripts present in all (or a significant fraction of) samples (e.g. Chen *et al.* 2010). As the inclusion of less common transcripts will also increase the amount of sample-specific transcripts, this might in these cases be a wasted effort. If, however, the ultimate goal is to create an as-complete-as-possible resource for a larger community, one would want to take care to include even less common transcripts.

There are a number of high-throughput sequencing platforms and assembly strategies that can be employed. Intuitively, longer, paired reads and the incorporation of long read data from technologies such as PacBio (although PacBio error rates to date are high (up to 20 % on a single pass) so error correction is in most cases necessary prior to inclusion in assembly, for example using Illlumina short-read data) improves assembly quality and completeness, and results in

longer transcripts (Martin & Wang, 2011; Cahais *et al.* 2012; Koren *et al.* 2012). Cahais *et al.* (2012) use 454 and 100 bp single-end Illumina sequence data individually and combined from five diverse non-model taxa to test the effect of these individual versus combined data types on assembly quality. They find that the combined data perform slightly better than Illumina single-end, long read alone, and those performed much better than 454 alone (Vijay *et al.* 2013). Though not tested by Vijay *et al.* (2013), it could be that Illumina paired-end long-read data would outperform the assemblies from combined 454 and Illumina single-end long-read data. Vijay *et al.* (2013) also showed improved performance with mapping assemblies to a closely related sister species (<10% sequence divergence) compared to *de novo* assemblies. Results from genome and transcriptome assembly studies generally support hybrid assembly techniques, combining different sequencing platforms such as Illumina and PacBio (Koren *et al.* 2012; Utturkar *et al.* 2014). Additionally, assemblers tend to perform better if they incorporate information on the sense versus anti-sense orientation of the RNA-seq data; Trinity is one assembler that does resolve and incorporate strand-specific RNA-seq data (Haas *et al.* 2014).

The typical strategies for processing raw sequence data before assemblies involve removing sequencing adapters and low quality reads and trimming low quality regions of reads. However, relatively little consideration has been given to excluding sequencing errors prior to transcriptome assembly from RNAseq data (Macmanes & Eisen, 2013). This is an important consideration because Illumina sequence data generally have error rates of 1:1000 to 1:10000, primarily substitutions, in a non-random distribution, increasing from the 5' to the 3' end (Yang *et al.* 2010; Liu *et al.* 2014). Macmanes & Eisen (2013) implement the error correction program REPTILE (Yang *et al.* 2010) on modeled and empirical data and find that while error correction does not affect assembly contiguity, there is a 10% reduction in errors incorporated into transcripts (Macmanes & Eisen, 2013). This reduction in the number of substitution errors is

particularly important in population genomic studies aimed at SNP marker development. Processing raw sequence data prior to assembly with REPTILE should reduce the identification of many false positive SNPs.

Another approach that results in the reduction of sequencing errors is digital normalization of sequence data to remove high-coverage sequence reads (Brown *et al.* 2012). This k-mer base process, implemented in the program khmer, removes high-coverage reads to a specified level, reducing sampling variation, and thereby removing many of the sequence errors contained in these high-coverage reads (Brown *et al.* 2012). This process reduces the dataset size to 1/10th the original size and therefore reduces assembly time by 90% with negligible affects on the contiguity of assemblies (see Brown *et al.* (2012); Table 5). However, the effect of digital normalization on qualitative measurements of assembly such as percent of conserved orthologs identified has not yet been reported. It is also possible to normalize RNAseq libraries before sequencing, in order to increase the representation of lowly expressed transcripts. Depending on the goal of the project, this could potentially be useful. In cases where a complete assembly or development of markers in lowly expressed sequences is the goal, normalization can be beneficial. However, it is important to remember that the normalized library no longer contains quantitative information about transcript abundance post-normalization, which makes assessment of expression levels or allele-specific expression impossible.

Post-assembly, Cahais *et al.* (2012) show that filtering contigs based on coverage and length improves the accuracy (percent of correct predictions, see Figure 7) and the proportion of full length transcripts and fragments relative to erroneous transcripts (see Figure 6) though at the expense of raw numbers of transcripts (Cahais *et al.* 2012). A recommendation that balances the number of retained contigs and excluded errors is to filter contigs based on an average 4x coverage and a minimum length of 600bp (Cahais *et al.* 2012).

Summarizing current literature, it is possible to identify a "best-practice" pipeline from experimental design, to platform choice, to raw data processing, to assembler choice, to transcriptome assembly processing and evaluation (Fig 1).

RNA sampling and sequencing: To maximize gene coverage of a new transcriptome assembly within the bounds of diminishing returns, the most effective way seems to be to sample RNA from a breadth of developmental stages and tissues, to prepare non-normalized, strand-specific (Borodina *et al.* 2011) RNA-seq libraries, and to pool and sequence these libraries on one quarter to one half of one Illumina HiSeq 2000 or 2500 lane to generate ~30-100 million paired-end, long (100bp) reads.

Data types and normalization: Paired-end long-read sequence data are most effectively used for creating a reference transcriptome assembly. Additional, longer sequence data can be generated with Ion Torrent or PacBio platforms for the transcriptome assembly (after error corrections). There are methods for hybrid assemblies combining data from different platforms (Lunter & Goodson, 2011; Koren *et al.* 2012; Vijay *et al.* 2013; Utturkar *et al.* 2014) or even *de novo* assembly with Ion Torrent data using Trinity and Oases (Amin *et al.* 2014). However, these platforms are not as readily available and assemblers, including Trinity, have been developed and optimized targeting Illumina read data. To reduce the incorporation of sequencing errors and potential false positive SNP identification, current experience recommends processing raw sequence data for quality as well as errors using the program REPTILE (Yang *et al.* 2010) and then digitally normalizing the data using the program khmer (Crusoe *et al.* 2014), which is now packaged with the latest version of Trinity (v. 20140717). These cleaned, paired-end long-read Illumina data can then ideally be assembled using Trinity because of the ability to resolve splice isoforms and gene paralogs (Grabherr *et al.* 2011). This works well with the ultimate goal of a single transcript for each gene for population genomic SNP marker development.

Contig evaluation and pruning: Post-assembly, a current literature review suggests pruning for coverage (4x), read length (600 bp), and predicting open reading frames (ORFs, e.g. TransDecoder.pl, also packaged with the latest version of Trinity) to remove or reduce DNA contamination, non-coding RNA, chimaeras, and gene fragments. These post-assembly processing steps should be adjusted and tested for each new species. With the goal of maximizing the discovery of new genes to the exclusion of erroneous transcripts, each assembly iteration can be evaluated through BLAST comparisons to a closely related species and measuring completeness, as well as identifying and excluding or correcting errors such as chimaeras, collapsed paralogs, and separated isoforms or allelic variants (Cahais *et al.* 2012). Barring a reference transcriptome or genome of a closely related species, assembly completeness can be estimated by searching for conserved eukaryotic orthologous genes (COGs) using the complete NCBI database (Tatusov *et al.* 2003) or the more restricted dataset of 248 single-copy COGs using the CEGMA program (Parra *et al.* 2007). Finally, assemblies can be further evaluated with packages such as Transrate and/or mRNAmarkup (http://hibberdlab.com/transrate/; https://github.com/BrendelGroup/mRNAmarkup).

Marker development and genotyping

Examining the current literature of this field, it is clear that most studies concerned with SNP marker development from transcriptomic data are based on agri/aquaculture efforts or for conservation genetics (e.g. Bai *et al.* 2011; Ashrafi *et al.* 2012; Helyar *et al.* 2012; Gallardo-Escàrate *et al.* 2013; Montes *et al.* 2013; Pootakham *et al.* 2013; Valenzuela-Muñoz *et al.* 2013; Cui *et al.* 2014). However, more and more data are being generated for natural populations with the goal of identifying differences in populations structure or targets of selection acting among populations or individuals (Bay & Palumbi, 2014).

SNP detection

It is relatively easy to acquire a list of candidate polymorphic loci from RNA-seq data by first aligning the short reads to a reference, then using software such as samtools (http://www.htslib.org/) (Li *et al.* 2009) or GATK (https://www.broadinstitute.org/gatk/) (McKenna *et al.* 2010) to search for consistent patterns of sequence variation and filter out dubious variants. Key parameters when filtering include sequencing depth (coverage), depth of reads with the non-reference allele, sequence quality scores, proximity to other SNP / InDel sites and strand bias. Sequencing errors can usually be reduced as a first step by eliminating SNPs with very low frequencies (it is also of course possible to start with a set of genes of interest and search for SNPs in those rather than using a blind shotgun approach, see e.g. Livaja *et al.* (2013)).

More problematic are artifacts caused by alignment errors due to InDels or multiple gene copies that have incorrectly been grouped together as one contig in an assembly. InDels are particularly troublesome in high-diversity species with large population sizes – for example, in many fish and marine invertebrates, indels are so common in introns that traditional sequencing of amplified loci long ago shifted to exons. The problem of misalignment in multiple gene copies can be particularly acute in gene families where the copies are identical in some regions but variable in others. There are some ways to deal with two issues – most common is to filter out any SNP within a certain distance of an InDel, and to filter out SNP clusters (potentially due to multiple-copy genes). This approach, however, is highly conservative. Another approach used by the Broad Institute's "Variant Quality Score Recalibrator" (VQSR) (DePristo *et al.* 2011) is to use sets of known SNPs in order to train Gaussian mixture models in order to recalibrate the quality scores of a list of raw SNP data, which then makes it possible for the researcher to pull out a set of SNPs with a user-defined probability of being true (Van Der Auwera *et al.* 2013). Paralog identification can also be done by comparing heterozygote excess across contigs with different

sequencing depth. This approach can then be used to compare likelihoods of models with or without paralogy to filter out dubious SNPs (Gayral *et al.* 2013). It is also possible to choose a subset of SNPs in contigs suspected to consist of multiple-copy genes, and verify that they segregate, i.e. that they do not always exhibit the same pattern.

One clear advantage of high-coverage paired-end RNA-Seq data is that they can sometimes be used to infer haplotypes. Particularly, the GATK "haplotype caller" can use reads that cover multiple heterozygous positions to phase these SNPs. If linkage is high enough, the approach can produce longer haplotypes that can provide insights into the patterns and tempo of selection (Nielsen *et al.* 2011).

Once a list of filtered SNPs has been obtained, it is often valuable to confirm them with external validation. One way to perform this is by designing primers from the transcriptomic data, and "Sanger" sequencing genomic DNA or other high-throughput SNP genotyping methods, such as mass spectrometry (Renaut *et al.* 2011), SNP assays (Gagnaire *et al.* 2012; Limborg *et al.* 2012), amplicon sequencing (O'Rawe *et al.* 2013), or high-resolution melting (HRM) (Wittwer *et al.* 2003), if the SNP frequency is not too high (amplicon sequences must be short enough to only span one SNP). There are some complications to this, most notably unknown Intron-Exon boundaries – if you design primers that span a long intron, your genomic DNA will not amplify. This can sometimes be avoided by studying Intron-Exon boundaries in published genomes of related species (boundaries are sometimes quite conserved evolutionarily). Because exons are often very short, such amplification products do not produce much sequence data. However, they can be valuable in expanding the population sample of a study and testing gene frequency differences seen in a preliminary study (e.g. Pespeni & Palumbi, 2013).

Transcriptomic genotyping and allele frequency estimation

A common goal of most population genomic studies is to either genotype each individual at variant sites, or alternatively (and more commonly) use pooled population-wide data to directly estimate allele frequencies (e.g. Kofler *et al.* 2012; Martins *et al.* 2014; Schlötterer *et al.* 2014). It is possible to estimate genotypes and allele frequencies from the GATK /samtools output described above, but it has been shown that for low-medium coverage sites this might introduce biases (Kim *et al.* 2011). Thus, alternative approaches have been developed using maximum likelihood approaches to directly estimate genotypes from the sequences, without first calling SNPs (Tsagkogeorga *et al.* 2012; Gayral *et al.* 2013). Similarly, bias can also be introduced when calculating allele frequencies from low-coverage genotype data, for example due to loss of low-frequency alleles which can affect the site-frequency spectrum (Han *et al.* 2014b). Also in these cases, it seems appropriate to directly estimate allele frequencies directly from the sequence data, using alternative statistical approaches (e.g. Nielsen *et al.* 2012; Han *et al.* 2014a).

One key issue in using pooled data is the representativeness of the number of short reads with one or the other nucleotide, compared to the actual number of alleles present in the genomic DNA of the sequenced tissue. Ideally, a heterozygous individual would always have a 50/50 distribution between alleles in the data, and all individuals in a pool would be equally represented in the sequence data, but in reality the data are most often skewed in some way, due to a number of reasons. PCR artifacts from the library preparation protocol (non-random priming or amplification) can also be potential culprits. Biologically, allele-specific expression (ASE) patterns, where one allele is more highly expressed than the other, could also potentially throw off genotype estimates. On an individual basis, this issue is unlikely to have a large effect, as the expression bias would have to be several orders of magnitude to inaccurately call a heterozygote a homozygote. However, when sequencing pools of individuals, even small

differences in expression could potentially throw off allele frequency estimates. There is currently an active debate on the magnitude of this issue (see e.g. Lemay *et al.* (2013) vs Konczal *et al.* (2013)), but until more is known, it is likely prudent to try to estimate ASE in a dataset before proceeding with analyses of pooled data.

There are several ways to estimate the prevalence of ASE in a dataset, most of which rely on supplementary sequencing of genomic DNA (Degner *et al.* 2009; Montgomery *et al.* 2010; Pickrell *et al.* 2010) to get an idea of the expected distribution of the two alleles in a heterozygote, and then compare the RNA-Seq data to that distribution using binomial exact tests on a locus-by-locus level. Alternatively, it is possible to investigate ASE on a transcriptome-wide level using Bayesian modeling (Skelly *et al.* 2011), which also allows for accurate calculations of false discovery rates. Another approach, taken by Pespeni *et al.* (2013b), uses gene expression data for testing for changes in ASE between different treatments, the rationale being that if ASE is strong, allele frequency changes will be accompanied by changes in gene expression of the same loci (Pespeni *et al.* 2013b). Thus, by testing for significant changes in gene expression, it is possible to filter out transcripts potentially under the influence of ASE.

Applications of expressed sequence datasets (Fig 2)

Transcriptomic SNPs have the advantage of providing functional information, allowing statistical tests to be conducted at levels above that of a single contig. For example, tests can be conducted about whether SNPs with high divergence in allele frequencies from population to population cluster into certain gene categories. The same approach can also in principle detect slight balancing selection by asking if SNPs with particularly low divergence in allele frequencies from population to population cluster into certain gene categories (Leffler *et al.* 2013).

Outliers

One of the most common goals of population genomic studies is to identify loci under selection or adaptive loci (Savolainen *et al.* 2013). The main idea is that in two populations under different selective regimes, genomic regions will exhibit a normal distribution of divergence, and by identifying loci significantly outside of the distribution curve (so-called "outliers"), you will arrive at a set of candidates likely to be affected by selection (balancing or disruptive)(Beaumont & Nichols, 1996). The metric most commonly used for this type of analysis is F_{ST} (the proportion of genetic variation that can be explained by differences among populations) (e.g. Pespeni *et al.* 2012; De Wit & Palumbi, 2013). There are several software packages that identify outliers, most of which are based on the FDIST algorithm, which assumes a certain proportion of the loci to be outliers (Antao *et al.* 2008). However, typical datasets for SNP analysis of transcriptomes or whole genome RAD sequencing include 10,000s to 100,000 of variable positions. Analyzing allele frequencies at these positions for signs of natural selection includes the strong possibility that some will appear to be more differentiated than expected strictly by chance and not by selection. In principle, levels of differentiation between populations will always produce a list of highest F_{ST} loci – the challenge has been to generate other ways to test these candidate loci against neutral expectations.

The outlier approach has been to compare the number and distribution of high F_{ST} loci to that expected under neutral theory. However, it has been difficult to generate this expectation accurately. For example, background selection in subdivided populations can reduce diversity in linked regions with a following increase in F_{ST} (Charlesworth *et al.* 1997), and as a result, outlier approaches can fail to eliminate all spuriously differentiated loci (Lotterhos & Whitlock, 2014). It is also possible to compare loci putatively under selection to outliers generated by permuted

populations from the original dataset, indicating what distribution of F_{ST} you would expect to observe from stochastic processes alone (e.g. Pespeni *et al.* 2013, De Wit *et al.* 2014).

A second method of outlier identification is through the Bayesian framework described by Foll & Gaggiotti (2008). Their Bayesian algorithm uses two models, one incorporating selection and one that does not, and estimates their respective posterior probabilities using an MCMC approach. Finally, it uses the posterior odds ratio to acquire p-values for each locus to be under selection, with user-specified false discovery rates. This method, implemented in the "BayeScan" software (http://cmpg.unibe.ch/software/BayeScan/), is much less prone to false positives (Narum & Hess, 2011).

Non-outlier approaches to identifying loci under selection

Other approaches seek to generate associated data that independently tests high F_{ST} loci for other features associated with selection. Such approaches in testing for groups of loci with 1) high levels of amino acid polymorphism; 2) a skewed distribution of minor allele frequencies; 3) enrichment for certain functional roles; 4) an association with individual fitness; 5) an ontogenetic change in gene frequency, or other links between genotype and phenotype. These analyses can provide additional independent tests that a group of loci showing high F_{ST} differentiation are under selection. In general, when parallel datasets can be used to test the prediction that a set of loci –possibly discovered by outlier analysis - is under natural selection, there is a higher likelihood that the outlier analysis has identified some of the selected loci.

For example, Pespeni *et al.* (2013b) identified a set of outlier loci that had higher levels of differentiation than expected among populations exposed to different levels of ocean acidification. However, the data also showed that this group of highly differentiated loci showed high levels of amino acid polymorphism, and were grouped in functional categories including

skeleton formation (Pespeni *et al.* 2013b). These supportive analyses were particularly important in this case because F_{ST} differences might have been misleading for two reasons: the prevalence of false positives (see above), and the possibility that allele-specific expression in different conditions altered apparent allele frequencies among pooled samples. Differentiation in amino acid replacement rates and in enrichment of important functional genetic categories added important corroborative data increasing the likelihood that selection has affected loci in this group. Likewise, De Wit *et al.* (2014) showed outlier SNP differentiation in abalone populations before and after a major natural mortality event, thought to be due to a harmful algal bloom. Enrichment analyses found that outlier loci grouped into specific metabolic functional categories linked to the effects of algal toxins found at high levels in abalone tissue.

Even with additional datasets, genomic tests of tens of thousands of loci only generate a set of candidate loci hypothesized to be under selection, and further work is usually needed to discern which of these loci are true targets of selection. Such extra work might involve careful surveys of polymorphic loci through targeted sequencing to discern patterns of haplotype variation, or other high resolution patterns of allelic variation over space and time. For example, Pespeni *et al*. (2013b) found that the same functional classes of genes that responded to experimental acidification showed correlations with local pH conditions across six populations in the wild (Pespeni *et al.* 2013a) and putative adaptive loci identified as F_{ST} outliers (Pespeni *et al.* 2010) showed correlations with local temperature conditions in finer scale sampling in the wild (Pespeni & Palumbi, 2013). Further work on the physiological basis of selection, the biochemical ramifications of allelic variation, or the gene expression variation associated with allele differences can help to track the mechanisms by which selection acts (Le Corre & Kramer, 2012).

In this point of view, these analyses first detect that there is a footprint of selection in the data, that the footprint is associated with a particular set of loci, and that the footprints lead in a

particular physiological, biochemical or genetic direction. The initial genome-level datasets should be viewed as a beginning of this process.

Evolutionary transcriptomics

In many cases, selection might not act strongly on single genes, but rather have subtle effects on many loci with similar functions, for example through regulatory or metabolic networks (Fraser *et al.* 2004; 2010). In these cases, it might not be possible to pick up individual loci as outliers, especially at the stringent levels of significance required when 10,000s of individual loci are examined. However, by testing whether loci with high F_{ST} are non-randomly clustered into distinct metabolic or functional categories, it is possible to infer the action of selection even in the absence of individually significant loci (e.g. Pespeni *et al.* 2013b; De Wit *et al.* 2014).

Typically, these non-random associations can be elucidated by overrepresentation analyses (ORA), which compare the proportion of functions in a dataset of interest (e.g. an outlier set) to the transcriptome-wide distribution of gene functions, while correcting for multiple tests (see e.g. Zheng & Wang, 2008). Another, perhaps more powerful enrichment analysis approach, focuses on comparing the transcriptome-wide traits of members in a functional class versus the rest of the transcriptome for amino acid polymorphism, F_{ST} levels, etc.. This approach compares traits in a small number of groups and can easily be simulated in permutation tests to gain statistical support. It is still unclear exactly how much of functionally important genetic variation is located in genic regions compared to regulatory regions (see e.g. Jones *et al.* (2012)), but especially for non-model systems, the genic regions will provide an initial view of the functional targets of a putative selective regime. In this respect, it could also be fruitful to focus on tissues/life stages that are *a priori* determined likely to be enriched for functions of interest,

such as gonadal tissue if reproductive barriers are of interest (Andres *et al.* 2013) or neural tissue for studying behavioral sexual dimorphism (Catalan *et al.* 2012) because RNA from these tissues will be enhanced for expression of these genes.

Another powerful feature of transcriptomic data is the potential to examine changes in gene expression levels among individuals or populations. There has long been a realization that gene expression differences play a strong role in species differentiation and in population adaptation (López-Maury *et al.* 2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Several studies between closely related species indicate that there is a genetic basis for differences in transcript levels (Fraser *et al.* 2004), which could lead to adaptive divergence in the wild (Jeukens *et al.* 2010; Leder *et al.* 2014).

The combination of gene expression measurements (based on read counts) and SNP detection (based on comparing read sequences) from the same individuals and the same RNASeq dataset provides a new look at the functional role of SNPs in gene expression. Quantitative estimates of gene expression can be associated with changes in nucleotide sequence (eGWAS) (Harper *et al.* 2012). Ironically, most SNPs controlling gene expression occur outside the coding regions of genes, and so finding relationship between a SNP genotype and expression levels can signal an indirect link between the SNP and whatever is controlling gene expression (Rose *et al.* submitted). The same method can furthermore be used to study regulatory network changes by analyzing co-expression patterns and associating with nucleotide changes and phenotypic traits (Szeto *et al.* 2014). The promise of this approach is that experiments on natural selection for gene expression differences can now be monitored in ways that require much less effort than in the past.

Finally, cross-species comparisons of of transcriptomic data has recently shown promise for conservation genetics of endangered animals (Loire *et al.* 2013), and also for gaining an enhanced understanding of the fundamental principles of population genomics (Romiguier *et al.*

2014), allowing us to potentially predict the responses of natural populations to future environmental perturbations.

Emerging opportunities

The ultimate data set for population genomics is a comparison of full genome sequences of individuals within and between populations. Such comparisons have been made for humans, yeast, Drosophila and a few other model systems, and are rapidly becoming economical for non-model species. In some senses, the attraction of reduced representation genome data sets (e.g. transcriptomes or RAD) is that they provide a way of increasing sample number in a study while maintaining practical DNA sequencing costs. However, as DNA sequencing costs continue to drop, and as analytical tools continue to become more powerful, there will probably be a move away from reduced representation genome data sets and a move towards full-genome population genetics for species in the wild. Even today such data sets are feasible: a DNA sequence run of 200 million reads at 200 bp each provides 40 Gb of data, or about 40 genome equivalents for a 1 gigabase genome in a single lane. This is not enough data to construct a full genome for all individuals, but it is enough to produce allele frequency data at a good portion of the full genome for a mixture of individuals in the lane.

However, even with these remarkable data in hand, a focus on expressed sequences remains extremely valuable. In this case, mapping genomic reads to a transcriptome can produce the sequences of many regions of the coding genome, allowing many of the analyses suggested above. This approach leaves behind the gene expression data made available by RNASeq but has the advantage of not requiring mRNA as a starting material.

Summary

With all the issues associated with genome assembly, focusing on the transcriptome provides a cost-effective way to reduce complexity while still retaining a large fraction of functionally relevant information. SNP genotyping from transcriptomic (RNA-Seq) data is a field currently growing rapidly, and while many studies to date focus on marker development with no further population genomic analyses, the field is evolving rapidly. Using expressed sequence data, there is potential to study not only patterns of SNP markers but also associations of phenotypes to alternative splice events or gene expression changes, and to start understanding the genetic background causing these patterns. There are some issues remaining to be studied in more detail, especially the effects of allele-specific expression on pooled RNA-Seq data. However, these issues are quite likely to be addressed within the near future, and new statistical frameworks will undoubtedly continue to extend the usefulness of RNA-Seq data for the foreseeable future.

References

Amin S, Prentis PJ, Gilding EK, Pavasovic A (2014) Assembly and annotation of a non-model gastropod (*Nerita melanotragus*) transcriptome: a comparison of *De novo* assemblers. *BMC Research Notes* **7**, 488.

Andres JA, Larson EL, Bogdanowicz SM, Harrison RG (2013) Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. *Genetics* **193**, 501-+.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: A workbench to detect molecular adaptation based on a F(st)-outlier method. *BMC bioinformatics* **9**.

Ashrafi H, Hill T, Stoffel K*, et al.* (2012) De novo assembly of the pepper transcriptome (*Capsicum annuum*): a benchmark for in silico discovery of SNPs, SSRs and candidate genes. *BMC Genomics* **13**, 571.

Bai X, Rivera-Vega L, Mamidala P*, et al.* (2011) Transcriptomic signatures of ash (*Fraxinus spp.*) phloem. *PLoS ONE* **6**, e16368.

Barshis DJ, Ladner JT, Oliver TA, Seneca FO, Traylor-Knowles N, Palumbi, SR (2013) Genomic basis for coral resilience to climate change. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 1387-1392.

Bay RA, Palumbi SR (2014) Multilocus adaptation associated with heat resistance in reef-building corals. *Current Biology*. Available online: doi: 10.1016/j.cub.2014.10.044.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B-Biological Sciences* **263**, 1619-1626.

Borodina T, Adjaye J, Sultan M (2011) A strand-specific library preparation protocol for RNA sequencing. In: *Methods in Enzymology, Vol 500: Methods in Systems Biology* (eds. Jameson D, Verma M, Westerhoff HV), pp. 79-98.

Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH (2012) A reference-free algorithm for computational normalization of shotgun sequencing data. *ArXiv* **arXiv:1203**, 1-18.

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

Catalan A, Hutter S, Parsch J (2012) Population and sex differences in *Drosophila melanogaster* brain gene expression. *BMC Genomics* **13**, 1-12.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research* **70**, 155-174.

Chen S, Yang P, Jiang F*, et al.* (2010) De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS ONE* **5**.

Choi M, Scholl UI, Ji W, *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 19096-19101.

Crusoe MR, Edvenson G, Fish J*, et al.* (2014) The khmer software package : enabling efficient sequence analysis.

Cui J, Wang H, Liu S*, et al.* (2014) SNP Discovery from transcriptome of the swimbladder of *Takifugu rubripes*. *PLoS ONE* **9**, e92502.

De Wit P, Palumbi SR (2013) Transcriptome-wide polymorphisms of red abalone (*Haliotis rufescens*) reveal patterns of gene flow and local adaptation. *Molecular ecology* **22**, 2884-2897.

De Wit P, Rogers-Bennett L, Kudela RM, Palumbi SR (2014) Forensic genomics as a novel tool for identifying the causes of mass mortality events. *Nature communications* **5**, 3652.

Degner JF, Marioni JC, Pai AA*, et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207-3212.

DePristo Ma, 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.

Doyle, SR, Griffith, IS, Murphy, NP, Strugnell JM (2014) Low-coverage MiSeq next generation sequencing reveals the mitochondrial genome of the Eastern Rock Lobster, *Sagmariasus verreauxi. Mitochondrial DNA.* Available online, doi: 10.3109/19401736.2013.855921.

Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M (2011) Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC *Genomics* **12**, 317.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180**, 977-993.

Fos M; Domínguez MA, Latorre A, Moya A (1990) Mitochondrial DNA evolution in experimental populations of *Drosophila subobscura*. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 4198-4201.

Francis WR, Christianson LM, Kiko R*, et al.* (2013) A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics* **14**, 167.

Fraser HB, Hirsh AE, Wall DP, Eisen MB (2004) Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9033-9038.

Fraser HB, Moses AM, Schadt EE (2010) Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proceedings of the National Academy of Sciences* **107**, 2977-2982.

Gagnaire P-A, Normandeau E, Cote C, Hansen MM, Bernatchez L (2012) The genetic consequences of spatially varying selection in the panmictic american eel (*Anguilla rostrata*). *Genetics* **190**, 725-U703.

Gallardo-Escàrate C, Núñez-Acuña G, Valenzuela-Muñoz V (2013) SNP discovery in the marine gastropod *Concholepas concholepas* by high-throughput transcriptome sequencing. *Conservation Genetics Resources* **5**, 1053-1054.

Gayral P, Melo-Ferreira J, Glémin S*, et al.* (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genetics* **9**, e1003457.

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-U130.

Haas BJ, Papanicolaou A, Yassour M*, et al.* (2014) De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols* **8**, 1-43.

Han B, Kang EY, Raychaudhuri S, de Bakker PIW, Eskin E (2014a) Fast pairwise IBD association testing in genome-wide association studies. *Bioinformatics* **30**, 206-213.

Han E, Sinsheimer JS, Novembre J (2014b) Characterizing bias in population genetic inferences from low-coverage sequencing data. *Molecular Biology and Evolution* **31**, 723-735.

Harper AL, Trick M, Higgins J*, et al.* (2012) Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nature biotechnology* **30**, 798-802.

Helyar SJ, Limborg MT, Bekkevold D*, et al.* (2012) SNP discovery using next generation transcriptomic sequencing in Atlantic herring (*Clupea harengus*). *PLoS ONE* **7**, e42089.

Huh D, Paulson J (2010) Non-genetic heterogeneity from random partitioning at cell division. *Nature Genetics* **43**, 95-100.

Jaramillo-Correa JP, Beaulieu J, Bousquet J (2001) Contrasting evolutionary forces driving population structure at expressed sequence tag polymorphisms, allozymes and quantitative traits in white spruce. *Molecular Ecology* **10**, 2729-2740.

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, Lohmueller KE, Albrechtsen A*, et al.* (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC bioinformatics* **12**.

Kofler R, Betancourt AJ, Schloetterer C (2012) Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in Drosophila melanogaster. *PLoS Genetics* **8**.

Konczal M, Koteja P, Stuglik MT, Radwan J, Babik W (2013) Accuracy of allele frequency estimation using pooled RNA-Seq. *Molecular Ecology Resources* **14**, 381-392.

Koren S, Schatz MC, Walenz BP*, et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology* **30**, 693-700.

Krasileva KV, Buffalo V, Bailey P*, et al.* (2013) Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome biology* **14**.

Leder EH, McCairns RJS, Leinonen T*, et al.* (2014) The evolution and adaptive potential of transcriptional variation in sticklebacks—signatures of selection and widespread heritability. *Molecular Biology and Evolution*.

Leffler EM, Gao Z, Pfeifer S*, et al.* (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578-1582.

Lemay Ma, Donnelly DJ, Russello Ma (2013) Transcriptome-wide comparison of sequence variation in divergent ecotypes of kokanee salmon. *BMC Genomics* **14**, 308.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular ecology* **21**, 1548-1566.

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

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.

Limborg MT, Helyar SJ, de Bruyn M*, et al.* (2012) Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). *Molecular ecology* **21**, 3686-3703.

Liu J, McCleland M, Stawiski EW*, et al.* (2014) Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. *Nature communications* **5**, 3830.

Livaja M, Wang Y, Wieckhorst S*, et al.* (2013) BSTA: a targeted approach combines bulked segregant analysis with next- generation sequencing and de novo transcriptome assembly for SNP discovery in sunflower. *BMC Genomics* **14**, 628.

Loire E, Chiari Y, Bernard A*, et al.* (2013) Population genomics of the endangered giant Galapagos tortoise. *Genome biology* **14**.

López-Maury L, Marguerat S, Bähler J (2008) Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics* **9**, 583-593.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of F-ST outlier tests. *Molecular ecology* **23**, 2178-2192.

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

Macmanes MD, Eisen MB (2013) Improving transcriptome assembly through error correction of high-throughput sequence reads. *PeerJ* **1**, e113.

Manel S, Joost S, Epperson BK et al. (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology* **19**, 3760-3772.

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

Martins NE, Faria VG, Nolte V, *et al.* (2014) Host adaptation to viruses relies on few genes with different cross-resistance properties. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 5938-5943.

McKenna A, Hanna M, Banks E, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303.

Montes I, Conklin D, Albaina A, *et al.* (2013) SNP discovery in European anchovy (*Engraulis encrasicolus*, L) by high-throughput transcriptome and genome sequencing. *PLoS ONE* **8**, e70051.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, *et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-U151.

Narum SR, Hess JE (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources* **11**, 184-194.

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**.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**, 443-451.

O'Neil ST, Emrich SJ (2013) Assessing De Novo transcriptome assembly metrics for consistency and utility. *BMC Genomics* **14**.

O'Rawe J, Jiang T, Sun G, *et al.* (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine* **5**, 28.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067.

Pespeni MH, Chan F, Menge BA, Palumbi SR (2013a) Signs of adaptation to local pH conditions across an environmental mosaic in the California Current ecosystem. *Integrative and Comparative Biology* **53**, 857-870.

Pespeni MH, Garfield DA, Manier MK, Palumbi SR (2012) Genome-wide polymorphisms show unexpected targets of natural selection. *Proceedings of the Royal Society B-Biological Sciences* **279**, 1412-1420.

Pespeni MH, Oliver TA, Manier MK, Palumbi SR (2010) Restriction Site Tiling Analysis: accurate discovery and quantitative genotyping of genome-wide polymorphisms using nucleotide arrays. *Genome biology* **11**.

Pespeni MH, Palumbi SR (2013) Signals of selection in outlier loci in a widely dispersing species across an environmental mosaic. *Molecular ecology* **22**, 3580-3597.

Pespeni MH, Sanford E, Gaylord B*, et al.* (2013b) Evolutionary change during experimental ocean acidification. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 6937-6942.

Pickrell JK, Marioni JC, Pai AA*, et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772.

Pootakham W, Uthaipaisanwong P, Sangsrakru D*, et al.* (2013) Development and characterization of single-nucleotide polymorphism markers from 454 transcriptome sequences in oil palm (*Elaeis guineensis*). *Plant Breeding* **132**, 711-717.

Remm M, Storm CEV, Sonnhammer ELL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology* **314**, 1041-1052.

Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L (2011) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (Coregonus spp.). *Molecular ecology* **20**, 545-559.

Romiguier J, Gayral P, Ballenghien M*, et al.* (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**, 261-U243.

Ross MG, Russ C, Costello M*, et al.* (2013) Characterizing and measuring bias in sequence data. *Genome biology* **14**, R51.

Salzberg SL, Phillippy AM, Zimin A*, et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research* **22**, 557-567.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics* **14**, 807-820.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics* **15**, 749-763.

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

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.

Smith-Unna RD, Boursnell C, Kelly S, Hibberd JM (2014) Transrate

Stillman JH, Armstrong E (2015) Genomics are transforming our understanding of responses to climate change. *BioScience*. Available online: doi: 10.1093/biosci/biu219.

Sultan M, Schulz MH, Richard H*, et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956-960.

Szeto CYY, Lin CH, Choi SC*, et al.* (2014) Integrated mRNA and microRNA transcriptome sequencing characterizes sequence variants and mRNA-microRNA regulatory network in nasopharyngeal carcinoma model systems. *FEBS Open Bio* **4**, 128-140.

Tatusov RL, Fedorova ND, Jackson JD*, et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**, 41.

Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics*. Available online, doi: 10.1093/hmg/ddq333.

Tsagkogeorga G, Cahais V, Galtier N (2012) The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. *Genome Biology and Evolution* **4**, 852-861.

Utturkar SM, Klingeman DM, Land ML*, et al.* (2014) Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics* **30**, 2709-2716.

Valenzuela-Muñoz V, Araya-Garay JM, Gallardo-Escàrate C (2013) SNP discovery and high resolution melting analysis from massive transcriptome sequencing in the California red abalone *Haliotis rufescens*. *Marine Genomics* **10**, 11-16.

Van Der Auwera GA, Carneiro MO, Hartl C*, et al.* (2013) From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **11.10**, 1-33.

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, Kuenstner 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.

Wang ET, Sandberg R, Luo S*, et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57-63.

West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, St. Clair DA (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* **175**, 1441-1450.

Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ (2003) High-resolution genotyping by amplicon melting analysis using LCGreen. **860**, 853-860.

Yang X, Dorman KS, Aluru S (2010) Reptile: representative tiling for short read error correction. *Bioinformatics* **26**, 2526-2533.

Zheng Q, Wang XJ (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research* **36**, W358-W363.

Zou F, Chai HS, Younkin CS *et al.* (2012) Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genetics* **8**, e1002707.

Figures and Boxes

Figure 1. Flowchart of a "best-practice" pipeline for transcriptome assembly and evaluation, as suggested by a current literature review.

Figure 2. Examples (not exhaustive) of post-assembly evolutionary applications of transcriptomic datasets discussed in the text. References cited in the figure: [1] Catalan *et al.* 2012; Barshis *et al.* 2013, [2] West *et al*. 2007; Harper *et al.* 2012; Zou *et al.* 2012, [3] Jaramillo-Correa *et al.* 2001; De Wit *et al.* 2014, [4] De Wit *et al.* 2013; Pespeni *et al.* 2013, [5] Manel *et al.* 2010, [6] Fos *et al.* 1990, [7] Szeto *et al.* 2014, [8] Kim *et al.* 2011, [9] Jones *et al.* 2012; Loire *et al.* 2013; Romiguier *et al.* 2014.

BOX 1. Types of errors in transcriptome assemblies

BOX 2. Metrics for evaluating transcriptome assembly quality

BOX 1. Types of errors in transcriptome assemblies

Chimaera – Erroneous fusion of the ends of two separate transcripts; can be detected with blast results when two or more non-overlapping regions of one transcript matches different reference transcripts

Allele – Genetic variant, sequence differences at the same position in a chromosome. Allelic variants can be erroneously assembled as separate transcripts.

Isoform – Transcript variant in the expression of a gene often due to alternative splicing of exons. Alternative spliced isoforms can be erroneously assembled as separate transcripts.

Paralog – Paralogs, separate transcripts related by an historical gene duplication event, can erroneously be assembled as one transcript; can be detected by multiple protein sequences from

a reference assembly matching one contig; can be revealed through collapse factor calculation (see Box 2).

Fragment – A partial, incomplete transcript sequence. Fragments can dominate assemblies due to degradation of the 3' ends of mRNA prior to library preparation or poor assembly.

rRNA – ribosomal RNA, highly abundant RNA that can contaminate sequencing libraries, reads and assemblies.

BOX 2. Metrics for evaluating transcriptome quality

N50 - the length of the contig such that 50% of the sequences in the assembly are longer than the central N50 contig; this metric gives greater weight to longer contigs compared to mean and median contig length.

Recovery or Completeness – If a reference transcriptome is available, recovery or completeness can be calculated as the proportion of bases recovered from the reference transcriptome in the new assembly.

Accuracy – If a reference transcriptome is available, accuracy can be calculated as the proportion of bases correctly matched to orthologous genes between the reference and the new assembly.

Collapse factor – If a transcriptome sequence from a closely related species is available, the collapse factor can be calculated to compare the mean number of reference orthologs that match each contig to evaluate different assemblies. Numbers greater than one suggest the erroneous collapse of paralogous transcripts from a gene family into a single contig.

Ortholog – Genes from two different species that share a common ancestral gene, separated by the event of speciation. Function is normally conserved. Databases of conserved eukaryotic orthologous genes (COGs) can be used to evaluate the completeness of a transcriptome assembly.