Transcriptome Analysis in Domesticated Species: Challenges and Strategies

Jessica P. Hekman, Jennifer L. Johnson and Anna V. Kukekova

Department of Animal Sciences, College of ACES, University of Illinois at Urbana-Champaign, Urbana, USA.

Supplementary Issue: Current Developments in Domestic Animal Bioinformatics

ABSTRACT: Domesticated species occupy a special place in the human world due to their economic and cultural value. In the era of genomic research, domesticated species provide unique advantages for investigation of diseases and complex phenotypes. RNA sequencing, or RNA-seq, has recently emerged as a new approach for studying transcriptional activity of the whole genome, changing the focus from individual genes to gene networks. RNA-seq analysis in domesticated species may complement genome-wide association studies of complex traits with economic importance or direct relevance to biomedical research. However, RNA-seq studies are more challenging in domesticated species than in model organisms. These challenges are at least in part associated with the lack of quality genome assemblies for some domesticated species and the absence of genome assemblies for others. In this review, we discuss strategies for analyzing RNA-seq data, focusing particularly on questions and examples relevant to domesticated species.

KEYWORDS: RNA-seq, transcriptomics, NGS, domestication

SUPPLEMENT: Current Developments in Domestic Animal Bioinformatics

CITATION: Hekman et al. Transcriptome Analysis in Domesticated Species: Challenges and Strategies. *Bioinformatics and Biology Insights* 2015:9(S4) 21–31 doi: 10.4137/BBI.S29334.

TYPE: Review

RECEIVED: October 21, 2015. **RESUBMITTED:** December 21, 2015. **ACCEPTED FOR PUBLICATION:** December 26, 2015.

ACADEMIC EDITOR: J. T. Efird, Associate Editor

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 1605 words, excluding any confidential comments to the academic editor.

FUNDING: The project was supported by USDA Federal Hatch Project 538922, AKCCHF grant 02193-A, and FCP grant 84–74. Jessica Hekman is supported by the Jonathan Baldwin Turner Fellowship, College of ACES, University of Illinois at Urbana-Champaign. The authors confirm that the funders had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: hekman2@illinois.edu

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Over thousands of years, humans have selectively bred domesticated animals for different uses and environments, resulting in a wide diversity of morphology and behavior. In fact, for some traits, the variation observed in domesticated animals is much greater than that found in laboratory or wild species. The striking diversity observed among individuals within domesticated species provides advantages for genetic studies of traits with direct relevance to biomedical research as well as traits with economic and cultural value. For example, some traits, such as herding behavior in some varieties of dogs or a comfortable riding gait in a gated horse, facilitate specific animal uses. Other traits, such as milk or wool production, represent an increasingly salient avenue of study as the global demand for food and fiber increases. Some adaptations may even be shared between humans and domesticated animals, such as altered metabolism to facilitate living in extreme environments. A better understanding of the underlying biological mechanisms associated with these traits will help further selection for increased productivity, utility, and health. Indeed, the health of domesticated animal populations is closely tied to that of our own societies and the environments that we share with them.[1] Domesticated animals are threatened by zoonotic diseases that also threaten us[2] and have risk factors for hereditary diseases that often closely mimic our own.[3–5] Many of these diseases represent natural models for the corresponding human conditions, and clinical studies in domesticated animals receiving advanced veterinary care may facilitate the development of innovative treatment strategies also of use in human medicine.[3,6] Progress in understanding human inherited diseases, the genetic architecture of complex phenotypes, and development of treatments against infectious or inherited diseases can, therefore, be significantly advanced through studies of similar traits and conditions in domesticated animals.

The history and population structure of domesticated species make them well suited for genetic studies.[7,8] Selection of domesticated strains for different morphological and other characteristics has led to the formation of "breeds" that are maintained in complete or partial reproductive isolation from each other. This practice of closed breeding generally results in reduced effective population size (ie, the number of individuals in a population who contribute genetically to the next generation) within breeds. This reduction is due to three effects: small founding populations, population bottlenecks, and the popular sire effect.[9,10] As a result of closed breeding and intense selection for specific traits, individuals within a breed commonly share long stretches of homozygosity at genomic loci related to the traits under selection.[7,11,12] Although the haplotype associated with a specific trait can be relatively long within a breed, a comparison of haplotypes across breeds characterized by the same phenotype allows a reduction of the critical interval and thereby facilitates identification of the genes under selection.[7,13,14] For example, small size in dogs is linked to a specific haplotype for *IGF1*, which is shared by a majority of toy dog breeds.[15,16] Closed breeding also leads to accumulation of disease-associated mutations, as has been demonstrated in dogs,[17] cattle,[18–20] and other species.[21–25] The identification of disease-associated genes in domesticated animals as compared to human populations is facilitated by domesticated species population structures and access to samples from many individuals in a pedigree.[11,26,27] For example, limited within-breed genetic diversity in dogs and the elevated occurrence of particular cancers in particular breeds facilitates the study of cancer susceptibility loci in that species.[28,29]

With the sequencing of the dog, cow, and pig genomes,[7,30,31] genome-wide association studies (GWAS) have become a common approach for the identification of genomic regions implicated in traits of interest in domesticated animals. This approach tests the probability of association of genetic markers with a trait; in domesticated animals, GWAS have commonly found a relatively small number of loci for such complex phenotypes as height, skull shape, or coat quality.[4,8,32,33] The fact that variation in these traits is explained by a small number of loci in these species is likely due to intense selective pressure.[34–36] As a result, many GWAS in domesticated species have successfully identified causal genes both for Mendelian traits and for complex traits controlled by loci with large effect size (for a review, see Schoenebeck and Ostrander[4] and Andersson.[8]) However, not all traits are amenable to investigation by this technique. Although the longer stretches of homozygosity common to breeds result in relatively large target regions that provide a strong signal, such large loci may contain dozens of genes and therefore may provide poor resolution for the identification of causative genes lying within them. Additionally, study design may be complicated by a lack of knowledge about the underlying genotype of a trait shared by multiple breeds, which may be identical by descent or different due to distinct causal mutations. Finally, not all traits in domesticated animals are controlled by a small number of genes of large effect; some complex traits, such as behavior, weight, meat quality, milk production, and some diseases, such as hip dysplasia and cancer, are controlled by many genes of small effect.[37–40] As has been demonstrated by GWAS in humans, loci of small effect may prove particularly challenging in the elucidation of molecular mechanisms, as they may require large sample sizes to achieve statistical significance.[41,42]

When the identified regions of interest are large, when many loci of small effect are implicated, or when the function of the discovered genes is unknown, GWAS alone will not be sufficient to elucidate genetic mechanisms associated with the phenotype under investigation. An alternative approach employs the analysis of gene expression differences to pinpoint changes in pathways rather than in specific genes. For example, gene expression studies have proven particularly well suited to investigate genomic changes in neoplastic cells, illuminating the molecular distinctions between different types of breast cancer[43] and contributing, along with copy number variation analysis, to the identification of oncogenes.[44]

Microarray gene expression studies pioneered the use of genome-wide techniques in the hunt for sets of genes or gene networks implicated in complex phenotypes in domesticated species.[45–47] However, microarray technology is limited by its dependence on the use of known probes, requiring a species-specific chip for most accurate results. Cross-species microarray use may result in decreased specificity of hybridization and can therefore be used only for closely related species, preferably <10 million years divergent from each other.[45,48] Moreover, even with a chip designed for the species under study, the dependence of microarray technology on known probes implies that transcripts that do not correspond to known sequences will not be detected, and novel isoforms will not be distinguished from known splice forms.

The advent of next-generation sequencing (NGS) revolutionized gene expression studies by obviating the need for preexisting probes for transcripts. RNA sequencing, or RNA-seq, uses the high-throughput reads produced by NGS to represent the entire transcriptome: in other words, all transcripts produced in a tissue sample including previously uncharacterized transcribed sequences and novel isoforms. RNA-seq is used for a variety of applications, most commonly to discover lists of genes that are differentially expressed between experimental groups, such as samples from different tissues,[28,49] samples from different treatment groups,[50,51] or samples from different populations.[52,53] To identify gene networks associated with inherited diseases or other genetic traits, individuals can be grouped by disease status (affected versus unaffected) or by different haplotypes at the mapped loci.[28,52,54–56] Differential gene expression may complement association studies when used to provide differential expression information about genes in the genomic regions of interest identified by GWAS.[28,29] In addition to gene expression differences, some RNA-seq studies may seek differences in isoform expression[57,58] and allele-specific gene expression.[59,60] Concurrently with the analysis of gene expression, RNA-seq data may be used for calling variants such as single-nucleotide polymorphisms (SNPs) or simple sequence repeats for subsequent use in marker studies.[61,62] This use is well suited to nonmodel species with limited genomic resources and to call variants that are novel for the species or that are enriched in the population under study. Finally, RNA-seq is used to improve genomic annotations through the identification of novel transcripts.[63–65]

RNA-seq has a particular advantage in nonmodel species, specifically those less common domesticated species for which species-specific chips for microarray studies are lacking. However, despite the promise of RNA-seq technology, performing differential gene expression experiments with RNA-seq may be challenging in domesticated species with low-quality genomes or a lack of high-quality reference annotation. In this review, we discuss strategies for performing differential expression analysis in nonmodel species, focusing particularly on the challenges common to studies in domestic species.

RNA-Seq Experimental Design

Number of replicates. One of the first steps in the experimental design of an RNA-seq study is the selection of the optimum number of biological replicates. At least a few replicates are necessary in order to characterize biological variation and separate it from technical variation,[66] and additional replicates provide additional benefits. Specifically, increasing the number of replicates in an RNA-seq study results in increased power available for differential gene expression analysis. When a trade-off must be made between the number of replicates and the depth of sequencing, replicates may be more instrumental than sequencing depth in increasing the power to detect differential expression.[67,68] Additionally, studies employing samples from outbred domesticated animals require a larger number of replicates than those employing a group with reduced genetic diversity, such as inbred mouse strains.[69]

Sample preparation. A variety of commercial RNA extraction kits are available for RNA isolation; though the kits generally extract similar amounts of RNA, they differ in the quality of RNA extracted. Therefore, the choice of kit may affect study results.[70] After RNA extraction, messenger RNA (mRNA) is isolated using either the polyA capture or rRNA depletion protocol. The polyA capture protocol results in a bias to the 3' end of transcripts, while the rRNA depletion protocol results in more variation in depth of coverage throughout the length of transcripts.[71] Decreased depth on the 5' end of transcripts sequenced from libraries built with the polyA capture protocol may result in decreased likelihood of identifying differential exon expression on the 5' end of transcripts, decreased depth of sequencing of long genes compared to short genes, and poor coverage of 5' untranslated regions, particularly important in the use of RNA-seq for improvement of transcriptome annotation. The rRNA depletion protocol depends on known ribosomal RNA sequences, and the probes have not been tested on all species. For example, while the Ribo-Zero Kit (Illumina) is predicted to work on all mammalian species due to probe homology, it has only been tested on human, mouse, rat, and dog, as stated in its manual. Its efficiency on avian genomes such as chicken and turkey is also unknown. After mRNA isolation, cDNA libraries are constructed. Libraries may be nonstrand-specific, or may support strand-specific RNA reads, which allow transcripts to be identified as sense or antisense. Strand-specific reads have been used in transcriptome assembly[72,73] and may facilitate differentiation of reads from adjacent or overlapping genes transcribed from opposite strands.[74] Additionally, Illumina sequencers support either single- or paired-end read sequencing. Paired-end sequencing may be more expensive, but increases the percentage of reads successfully mapped to the genome. Its use is recommended for the detection of distinct isoforms; however, its increase in the mapping of unique reads may be only marginal, so use of paired-end reads is not recommended unless maximizing unique read mapping is critical to the project.[69]

Sequencing strategy. After the construction of the library, transcripts are typically sequenced on an NGS platform; currently, Illumina sequencers are the most common. Sequencing considerations include determining the appropriate read length and the number of lanes (ie, sequencing depth). Illumina HiSeq 2500 sequencers produce reads of 50–150 bp in length; they employ flow cells with eight lanes, and multiple samples may be run on a single lane. To differentiate reads from different samples after sequencing on the same lane, a unique bar code may be attached to each sample during library preparation. The appropriate number of lanes must be determined by taking into account the necessary depth of sequencing; for example, studies that rely on the detection of rare transcripts or polymorphisms will require greater depth.[75] Artifactual variation per lane may contribute technical variation to a study, but this can be avoided by the use of multiplexing, eg, ensuring that each lane contains a balanced number of samples from each treatment group.[66]

Longer reads result in an increased percentage of mapped transcripts and improved handling of splice junctions during alignment.[76] Longer reads may therefore prove particularly useful for projects using species without an existing reference genome sequence that require de novo transcriptome assembly or identification of alternative transcripts; otherwise, a 50–bp read length should be sufficient.[69] In the past, 454 pyrosequencing (454 Life Sciences) has been used to produce RNA-seq reads between 100 and 500 bp in length.[77–86] However, this technology has proven prohibitively expensive and is currently not widely available. Emerging platforms such as PacBio (Pacific Biosciences), which provide longer read lengths, may prove popular in the future, perhaps even providing the ability to sequence entire transcripts in a single read.

RNA-Seq Bioinformatic Workflow

A typical bioinformatic workflow using a reference genome and aimed at the identification of differentially expressed genes is described below and summarized in Table 1. The workflow begins with raw reads, which are aligned to a reference genome. Gene counts are then quantified from the alignment files and used in differential gene expression analysis.

Read Filtering. Post sequencing, several filtering steps are recommended in order to produce a high-quality dataset. Common tools for removal of low-quality sequences as well as barcodes and platform-specific adapters added during library construction are the fastx-toolkit,[87] FLEXBAR,[88] and

Table 1. RNA-seq bioinformatic workflow for calling differentially expressed genes.

STEP	TOOLS	CHALLENGES
1. Remove low-quality reads, barcodes, and adapters	Fastx-toolkit, FLEXBAR, or Trimmomatic	Follow recommended protocol
2. Remove mitochondrial and ribosomal sequences	Bowtie2	Sequences from the same or related species should be used
3. Align to reference genome	TopHat2	Incomplete or nonexistent reference genome
4. Call differentially expressed genes	DESeq2, edgeR, or limma	Incomplete or nonexistent reference genome annotation

Trimmomatic.[89] These tools operate on FASTQ-format files and accept command-line parameters to specify the minimum length or Phred score below which a read should be discarded. Ribosomal and mitochondrial sequences may subsequently be removed, although these types of RNA should have been depleted in large part during library construction.[90] Removal of mitochondrial or ribosomal reads may be accomplished through alignment of all reads to mitochondrial and ribosomal sequences from related species, obtained, for example, from National Center for Biotechnology Information's (NCBI's) Reference Sequence collection (RefSeq) database (http://www.ncbi.nlm.nih.gov/refseq/). Unaligned reads kept for further processing will then be depleted of mitochondrial and ribosomal sequences. Although the exact percentage of removed reads will depend on the details of the chosen pipeline, in our laboratory, following this protocol in three separate tissues sequenced in different runs resulted in removal of 5.7%–13.3% of reads (unpublished data).

Alignment. After filtering, the RNA reads are typically aligned to a reference genome. This mapping process is complicated by the presence of splice junctions in the reads, originating from post processing of mRNA. Two approaches may be used to reduce the number of reads mismapped as a result of splice junctions. First, a splice-junction-aware aligner should always be used for mapping RNA reads, as alignment of this type of reads with a nonsplice-aware aligner such as Bowtie2 or BWA results in a higher percentage of mapping errors.[91] TopHat2[92] and STAR[93] are two splice-aware aligners that are widely used for RNA read mapping; STAR operates at greater speeds than other aligners but has a correspondingly larger memory footprint.[69] These aligners have similar mapping performance with a low median error rate.[91] Second, annotation of splice junction locations in the genome should be provided to the aligner when available. Typically, species annotations are archived at Ensembl (ftp://ftp.ensembl.org/), RefSeq (ftp://ftp.ncbi.nlm.nih.gov), and UCSC (http://genome.ucsc.edu/). Selection of an annotation with a broader gene coverage will result in an increased percentage of reads mapped to genes. In human, the Ensembl annotation provides the broadest gene coverage and, as a result, corresponds to the highest gene mapping rates.[94] Ensembl also provides increased coverage of dog, including intron and untranslated regions (UTR) annotation, which is not available through the other two annotations (personal observation). The appropriate choice of annotation may vary from species to species, but, in general, the annotation with the broadest coverage should be selected to maximize mapping rates.

Alignment of reads to the genome is further complicated by reads that map to complementary sequences at multiple locations in the genome. Ambiguous mapping may be due to conserved domains of paralogous genes, pseudogenes, and repeats.[76] Such reads are particularly problematic in gene differential expression studies, as some gene count quantification tools discard them.[91] A paired-end sequencing approach results in an increased percentage of uniquely mapped reads, though this improvement may be minimal.[69]

Visualization of aligned reads offers the opportunity to evaluate the dataset before continuing. Such evaluation can provide opportunities to better understand problems such as coverage bias, intronic or intergenic reads, or overlapping genes. Two such visualizers are GenomeBrowse (Golden Helix) and IGV (Broad Institute). Although exonic reads make up the preponderance of RNA-seq datasets, introns, untranslated regions, and intergenic regions are often retained, albeit at lower depth. These nonexonic regions may not be artifactual but may be a result of pervasive transcription of the genome.[95] Reads that align to intergenic regions may also represent unannotated exons[96] or long noncoding RNA (lncRNA) transcripts. At least 15,512 lncRNAs have been identified in human[97] and 7,224 in dog.[65]

Differential gene expression analysis. Gene expression must be quantified in reads before differences in expression can be identified. An assessment of quantification tools shows that while results from different tools are often highly correlated, results from a subset of genes may display differences as great as 10-fold. Identification of the quantification tool with the greatest accuracy is difficult, as accurate counts may not be known for comparison. However, in a comparison of different pipelines composed of a variety of quantification tools and aligners, pipelines including the HTSeq-count quantification tool numbered among those with the best performance.[91]

Outlier samples may influence differential expression results, and should be identified and removed prior to differential gene analysis. George et al.[98] describe a leave-one-out approach for the detection of outliers. Alternatively, some differential expression analysis tools, such as DESeq2,[99] perform outlier detection and removal automatically. Another differential expression tool, edgeR,[100] incorporates outlier detection into its estimate of genewise dispersion when the robust = TRUE parameter is specified in the estimateDisp() method.[101]

Differential gene expression analysis tools are confronted with normalization difficulties that are inherent in the analysis of RNA-seq reads, namely, bias due to different depths of sequencing per sample or to gene length.[75] Additionally, these tools must contend with the small replicate numbers that are typical of RNA-seq experiments, often as low as 2–3 replicates.[99] A comparison of 11 different methods for differential expression analysis showed that while some widely used methods have similar accuracy, the sets of differentially expressed genes found by different methods vary significantly.[102] Therefore, analysis of a single dataset by several methods may provide increased sensitivity, by considering genes identified by any tool as differentially expressed, as well as increased specificity, by considering genes identified only by multiple tools as differentially expressed. The tools DESeq2, edgeR, and limma have been found to have superior specificity and sensitivity[67] and are widely used. To ensure that the findings from the RNA-seq analysis are not artifactual, it is recommended that real-time quantitative polymerase chain reaction (RT-qPCR) be used to evaluate a representative set of differentially expressed genes.[50,51,103] The use of RT-qPCR, a well-established method for the evaluation of gene expression, provides technical validation of the RNA-seq procedure and data analysis used for the identification of differentially expressed genes.

Although differential expression has conventionally been performed using the gene or the exon as the base unit, it may alternatively be performed at the level of the nucleotide or region of sequential nucleotides. DER Finder[104] analyzes differential expression by nucleotide, and therefore does not require annotation of gene locations. It does require a sequenced genome, which may be a draft assembly. This tool may, therefore, prove useful in species lacking genome annotation. It may also be used concurrently with the pipeline described above to provide additional information about expression differences at the base, rather than the gene, level.

Analysis of differentially expressed gene lists. Typical differential expression analyses produce lists of hundreds of differentially expressed genes, requiring further analysis to construct a high-level overview of changes between the groups being compared. A commercial package, Ingenuity Pathway Analysis (QIAGEN), provides a graphical user interface to assist in discovery of pathways enriched in differentially expressed genes, generates publication-quality figures, and offers links to peer-reviewed articles about differentially expressed genes and related pathways. For studies with smaller budgets, the Database for Annotation, Visualization, and Integrated Discovery (DAVID)[105] offers a freely available alternative with a web-based interface. DAVID associates differentially expressed genes with other genes that have similar functions. While DAVID does not provide figure generation, visualization may be accomplished by use of the freely available tool Cytoscape.[106,107] This tool, which will run on OS X, Windows, or Linux, provides a graphical interface to allow the user to specify how to visualize a network of genes. Finally, weighted gene coexpression network analysis (WGCNA) may be used to identify clusters of genes with highly correlated coexpression patterns. This tool identifies networks of genes that are perturbed together, thereby suggesting biological pathways affected by the model in question.[108]

SNP detection. In addition to their use in gene expression studies, RNA-seq reads may be used to call SNPs. The Genome Analysis ToolKit (Broad Institute) provides a pipeline for calling SNPs specifically in RNA data. Alternatively, the SAMtools and BCFtools toolkits may be used in conjunction to call SNPs.[109,110] Both pipelines require alignment files (BAM) as input rather than raw reads (FASTQ). The identified SNPs will for the most part be in exonic or UTR; calling of intronic or intergenic SNPs requires genomic, not transcriptomic, data.

Computing resources. Alignment and SNP calling processes may be computationally expensive and are best performed on a high-performance server or cluster rather than a desktop computer. An eight-core cluster with 32 GB of RAM has been recommended as the minimum hardware requirement for a typical alignment process. However, a desktop computer is generally sufficient for calling differentially expressed genes.[111] Additionally, file sizes for reads from an individual sample aligned to a genome may be expected to reach ~10 GB in size, depending on the read depth (unpublished data).

RNA-Seq in Domesticated Species

Challenges of RNA-seq in domesticated species. The RNA-seq analysis pipeline described above has been successfully employed in species with "finished" reference genomes, such as human, mouse, and fruit fly, to identify genes that are differentially expressed between different samples.[51,63,112] These genomes were constructed using Sanger sequencing,[113–115] resulting in high-quality assemblies with low scaffold numbers, high scaffold N50 lengths, and low total assembly gap lengths. These widely used reference genomes are also extensively annotated, making available a large number of transcripts for gene expression studies, with high percentages of curated transcripts. Curated transcripts have been manually reviewed to remove sequence errors and ensure association with the correct genomic locus.[116]

A small number of the genomes of domesticated species were also constructed entirely or in part using Sanger sequencing, including dog,[7] chicken,[117] cow,[30] and pig.[31] The assemblies for these species are of variable quality compared to human and mouse. All four have a larger number of scaffolds, and chicken, pig, and cow have a significantly shorter N50 length (Table 2). Annotation of these widely used domesticated animal genomes is also less complete than in human or mouse, with fewer total transcripts in RefSeq and a much smaller percentage of manually curated transcripts (Table 3). As a result, gene expression pipelines in these species will have the use of many fewer isoforms and may encounter a higher percentage of transcripts with sequence errors. Moreover, annotation of gene and exon locations may be inadequate to

Table 2. Properties of assemblies of human, mouse, and domesticated species.

SPECIES	ASSEMBLY	SEQUENCING TECHNOLOGY	SCAFFOLDS	SCAFFOLD N50 (BP)	TOTAL GAP LENGTH (BP)
Human	GRCh38.p5	Sanger	797	59,364,414	161,368,151
Mouse	GRCm38.p4	Sanger	293	52,589,046	79,356,756
Dog	CanFam3.1	Sanger	3,310	45,876,610	18,261,639
Chicken	Gallus_gallus–4.0	Sanger	16,847	12,877,381	14,074,301
Cow	Btau_4.6.1	Sanger	13,387	2,599,288	176,429,395
Pig	Sscrofa10.2	Sanger and NGS	9,906	576,008	289,397,178
Turkey	Turkey 5.0	NGS	233,806	3,801,642	35,294,427
Yak	BosGru 2.0	NGS	41,192	1,407,960	120,154,638
Ferret	MusPutFur1.0	NGS	7,783	9,335,154	132,851,443

Note: All data were downloaded from ncbi.nlm.nih.gov on December 2, 2015.

identify important genes even in relatively well-annotated genomes.[65] For example, the *POMC* gene is not present in the dog RefSeq database (verified by a search on the UCSC Genome Browser, October 20, 2015). This gene encodes pre-pro-opiomelanocortin, which is cleaved to produce β-endorphin and met-enkephalin, which are endogenous opioid peptides; α-melanocyte-stimulating hormone, which is important in feeding behavior and energy homeostasis; and adrenocorticotropic hormone, which is a component of the hypothalamic–pituitary–adrenal axis. *POMC*'s function is therefore well characterized and this gene may be expected to be of interest in various studies. Incomplete annotation is a significant limitation for an RNA-seq pipeline, as existing differential expression tools rely on accurate annotation.[104]

Although domesticated animal genomes constructed using Sanger sequencing have some limitations, they are of relatively high quality. However, in the face of the plummeting cost of NGS, the number of unfinished "draft" genomes has increased.[118,119] Many genomes of less studied domesticated species were constructed entirely with NGS, such as Illumina or 454, as for example turkey,[120] yak,[121] and ferret.[122] NGS uses libraries with smaller insertions than does Sanger sequencing, usually not longer than 10–20 kb; as a trade-off, it produces shorter scaffolds. These shorter, sometimes misassembled scaffolds result in fragmented genes and a significant number of missing coding exons.[119,123] The assemblies for the reference genomes of turkey, yak, and ferret, for example, have many more scaffolds than do the human and mouse assemblies (Table 2). Annotation of these newer genomes may also lag behind that of more widely studied species, as is evidenced by the lower number of total transcripts and dramatically lower number of curated transcripts in RefSeq for turkey, yak, and ferret (Table 3). Assembly errors in draft genomes such as these have been shown to result in misannotation, particularly by automated annotators; moreover, the completeness of draft genome annotations is difficult to assess.[124]

Overall, domestic animal reference sequences have a wide range of qualities. Some may prove to have assemblies and annotations that are complete enough to support the described pipeline as a sole approach to RNA-seq analysis.

Table 3. Numbers of curated and uncurated transcripts annotated in the RefSeq database for human, mouse, and domesticated species.

SPECIES	SPECIES NAME USED IN REFSEQ SEARCH	TOTAL REFSEQ TRANSCRIPTS	CURATED REFSEQ TRANSCRIPTS	UNCURATED REFSEQ TRANSCRIPTS
Human	*Homo sapiens*	100,068	39,623	60,445
Mouse	*Mus musculus*	78,241	29,900	48,341
Dog	*Canis lupus familiaris*	47,095	1,675	45,420
Chicken	*Gallus gallus*	32,244	6,197	26,047
Cow	*Bos taurus*	70,342	13,329	57,013
Pig	*Sus scrofa*	47,445	4,154	43,291
Turkey	*Meleagris gallopavo*	26,450	93	26,357
Yak	*Bos mutus*	28,868	7	28,861
Ferret	*Mustela putorius furo*	48,113	61	48,052

Note: All data were downloaded by searching "Species name"[porgn] AND refseq[filter] AND biomol_mrna[PROP] (eg, "Canis lupus familiaris"[porgn] AND refseq[filter] AND biomol_mrna[PROP]) at http://www.ncbi.nlm.nih.gov/nuccore/ on December 10, 2015.

Analysis of others using the described pipeline may prove difficult if the associated reference genome and annotation contain significant missing information. Still others may have no reference genome at all. Therefore, an alternate approach may be required for the analysis of RNA-seq reads from less widely studied species.

Alignment of RNA-seq reads to a related reference. One solution to this impasse is the use of a reference genome and annotation from a closely related species. For example, a study of gene expression changes in macrophages of red deer in response to paratuberculosis used the cow genome for alignment of deer RNA-seq reads.[125] This solution may be practical for an increasing number of less common domesticated species as more genomes are assembled and annotated. Critically, the reference used must itself be mature and well annotated for this approach to provide real benefit; moreover, the related genome should be not more than 15% divergent for the best results.[111]

Alignment of reads from one species to the genome of another may prove challenging due to species divergence, even in closely related species. Genomic differences such as SNPs and indels may decrease mapping accuracy when reads are aligned to the genome of a different species, resulting in a decreased depth of coverage due to the loss of reads that cannot be mapped. The default parameters of splice-aware aligners may therefore not be appropriate for use in this situation. For example, without modification of its default parameters, TopHat2 will not accept an alignment of a read to a location if that alignment has more than two mismatches. As a result, three or more differences in a read of 100 or 150 base pairs (2%–3% divergent) will result in a rejected alignment. A difference may be a mismatch due to a SNP, or an edit due to either a SNP or a gap (indel). TopHat2's tolerance for mismatches and edits may be increased using the –read-mismatches and –read-edit-dist parameters. In our laboratory, alignment of *Vulpes vulpes* reads to the *Canis familiaris* 3.1 reference using TopHat2 with default parameters resulted in 69% alignment. Alignment with –read-mismatches = 3 and –read-edit-dist = 3 increased alignment to 79% (unpublished data). Increasing the allowed mismatches and edit distance should be performed with caution, for fear of false positive alignments.[126]

An alignment tool designed to handle divergent genomes, Stampy,[127] is tolerant of up to 15% sequence divergence. Stampy assumes 0.001 substitutions per site, but this default may be modified by the command-line parameter –substitutionrate = . Stampy has been successfully used to map RNA-seq reads from white-throated sparrow, song sparrow, and white-crowned sparrow to the zebra finch genome with the substitution rate set to 0.05.[55]

De Novo Assembly of RNA-Seq Reads

If a high-quality reference genome of a closely related species is not available, an alternative solution is to use the de novo assembly of a reference transcriptome from available RNA-seq reads. This approach was successfully used in our laboratory for the assembly of the brain transcriptome of silver fox (*Vulpes vulpes*) using 454 reads.[85] De novo assembly may be performed in the absence of any genome, or may be guided by a reference genome if one is available. This approach suffers from difficulties in annotating the assembled contigs as well as increased computation requirements.[104] Additionally, the use of a reference genome from a related species that is ≤15% divergent was found to recover more bases than use of a de novo genome from the species under investigation.[111] In practice, however, the choice to use a reference approach over a de novo assembly approach may depend not just on the divergence of the two genomes but also on the quality of the reference genome assembly and completeness of its annotation. When no reference of a closely related species is available, or the available reference is not sufficiently annotated, the de novo approach may be a valid alternative.

Widely used assemblers that will operate independently of a genome include Trinity,[73] Velvet,[128] and Oases.[129] Genome-independent assembly typically requires significant time on a high-end server or cluster. Exact requirements vary depending on the number of reads, but may include hundreds of gigabytes of memory and hundreds of hours of runtime.[130,131] For example, recommendations for use of Trinity include allocation of ~1 GB for every one million reads assembled, and from 256 GB to 1 TB of memory.[132] Laboratories that do not have access to their own high-end server may consider purchasing time on a campus cluster or Amazon Web Services,[133] or applying for time through XSEDE.[134]

The assembly process produces putative transcripts, known as contigs. Ideally, a single contig is equivalent to a single isoform, but in practice a contig may represent an entire isoform (a complete transcript), part of an isoform (an incomplete transcript), or a chimera (a transcript consisting of two transcripts that are biologically independent). Therefore, after assembly, chimeric contigs must be identified and discarded, and remaining contigs must be annotated with the appropriate gene symbol. Some programming, using a scripting language such as Python (The Python Software Foundation, python.org), may be necessary to accomplish this. A set of protocols and scripts exists to aid in the analysis of de novo assemblies,[135] or the following protocol may be observed.

Initial analysis of a de novo assembly should include masking of repetitive sequences, to avoid false positives during chimera identification. RepeatMasker[136] uses a database of known repetitive elements to substitute Ns or Xs for repetitive sequences.

Assembly annotation may be accomplished using BLAST[137] or BLAT.[138] The contigs should be compared to a well-annotated genome that is as closely related as possible. If no closely related genome exists and results using distantly related genomes are insufficient, a protein–protein comparison may be made instead of a nucleotide–nucleotide comparison, thereby eliding synonymous variations. Typically, multiple matches will be found for many genes. This is to be expected, as assemblers may not successfully differentiate between different isoforms, so that multiple contigs may represent different isoforms of the same gene. Additionally, a contig representing an incomplete transcript may match to one set of exons in a gene, while a different contig may match to a different set of exons, together completing the transcript. Chimeric contigs that match multiple genes should be removed from the assembly.

The BLAST or BLAT results may be used to rename individual contigs according to the gene they best match. Tools exist to aid in this process; for example, BioPython[139] provides tools to handle and rename sequences as Python objects. The BLAST or BLAT results may then further be used to identify chimeric contigs that match to multiple genes. Some apparently chimeric contigs may actually match two genes from the same gene family that have very similar sequences, and therefore do not need to be removed. To identify these apparently chimeric but actually legitimate contigs, a BLAST or BLAT search comparing the reference genome to itself may be performed. Genes that match to other genes may be considered pseudo-chimeric, and contigs matching multiple pseudo-chimeric genes need not be removed from the assembly. Additionally, contigs with multiple matches of extremely different lengths (for example, one match 10 times longer than the second match) need not be removed from the assembly, as differential gene expression tools will be able to choose the appropriate (longer) match and discard the other (shorter) match.

Nonexonic sequences may be retained in some contigs, comprising both the UTR, intronic sequence from pre-splicing mRNA, and intergenic sequence, now known to be pervasively transcribed though at lower levels than intra-genic sequence.[95] Intron removal is a challenging proposal for de novo assemblers.[130]

The differential expression tools described in the reference genome RNA-seq pipeline require a genome, not a transcriptome, as their reference. This makes them inappropriate for use with a de novo assembled transcriptome. However, an alternative differential expression tool, Cuffdiff, was designed to work with the genome-guided assembler Cufflinks, and will perform differential expression analysis against assemblies constructed by alternative assemblers.[140]

Conclusions

Species from a marked diversity of taxa have been domesticated, from mammals to fish to birds, including both commonly studied species such as the dog and less commonly studied species such as the yak. Many domesticated species have phenotypes of biomedical or economic value for humans, making them important subjects of research. Compared to wild animals, domesticated animals are well suited for study with RNA-seq, which is a new technology for evaluating transcriptional activity across the entire genome. First, the striking phenotypic diversity of these species provides opportunities for comparison of traits of interest among individuals. Second, reduced genetic diversity within domesticated breeds results in increased statistical power in many studies. Third, the breeding of domesticated species is under human control, so samples may more easily be collected from individuals with specific phenotypes and at specific time points in their development. Finally, domesticated animals commonly receive veterinary care, providing an opportunity for sample collection from individuals with well-characterized disease status or subject to advanced treatment. Domesticated animals remain critical for human well-being, and molecular genetic studies provide insights into the mechanisms involved in the regulation of the complex phenotypes for which these animals have been selected. Using this knowledge, we can not only advance human medicine but also select animals better suited to the changing climate and to human needs.

While many of the most common domesticated species, such as dog, cow, pig, and chicken, have high-quality genomes, other species have lower quality, fragmented NGS genomes, and still others are not yet sequenced. While the genomes of all domesticated species may well be sequenced in the coming decades, newly sequenced genomes can be expected to be subject to the limitations of NGS assemblies. RNA-seq is becoming a standard method for the annotation of NGS genome assemblies, and its use in improving the annotations of the high-quality genome assemblies produced using Sanger sequencing has been demonstrated.

In this review, we have discussed two strategies for the analysis of RNA-seq data in species with lower quality genome assemblies. Use of a mature, well-annotated genome from a closely related species may prove sufficient, especially if stringent requirements are relaxed during alignment to tolerate an increased rate of sequence divergence. If even a closely related genome annotation is lacking, a de novo assembly may be constructed and used as a reference. Using either a reference genome or a de novo transcriptome assembly, differentially expressed genes may be called. This list of genes may be further analyzed to ascertain groups of differentially expressed genes with similar functions or networks of genes that are coexpressed. Therefore, even in the absence of the resources available for RNA-seq analysis of model species, RNA-seq analysis is a powerful tool for use in the investigation of the genomic underpinnings of phenotypes in domesticated species.

Acknowledgments

We are grateful to Dr. Andrew G. Clark and Dr. Xu Wang for the thoughtful discussion on RNA-seq data analysis.

Author Contributions

Wrote the first draft of the manuscript: JPH. Contributed to the writing of the manuscript: JPH, AVK, JLJ. Jointly developed the structure and arguments for the paper: JPH, AVK. Made critical revisions and approved final version:

JPH, AVK, JLJ. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Zinsstag J, Schelling E, Waltner-Toews D, Tanner M. From “one medicine” to “one health” and systemic approaches to health and well-being. *Prev Vet Med*. 2011;101(3–4):148–56.
2. Kahn L. Confronting zoonoses, linking human and veterinary medicine. *Emerg Infect Dis*. 2006;12(4):556–61.
3. Khanna C, Lindblad-Toh K, Vail D, et al. The dog as a cancer model. *Nat Biotechnol*. 2006;24(9):1065–6.
4. Schoenebeck J, Ostrander E. Insights into morphology and disease from the dog genome project. *Annu Rev Cell Dev Biol*. 2014;30(1):535–60.
5. Stefaniuk M, Ropka-Molik K. RNA sequencing as a powerful tool in searching for genes influencing health and performance traits of horses. *J Appl Genet*. 2016.
6. Paoloni M, Khanna C. Translation of new cancer treatments from pet dogs to humans. *Nat Rev Cancer*. 2008;8(2):147–56.
7. Lindblad-Toh K, Wade C, Mikkelsen T, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*. 2005;438(7069):803–19.
8. Andersson L. Molecular consequences of animal breeding. *Curr Opin Genet Dev*. 2013;23(3):295–301.
9. Ostrander E, Kruglyak L. Unleashing the canine genome. *Genome Res*. 2000;10(9):1271–4.
10. Gibbs R, Taylor J, Van Tassell C, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*. 2009;324(5926):528–32.
11. Andersson L, Georges M. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet*. 2004;5(3):202–12.
12. Purfield D, Berry D, McParland S, Bradley D. Runs of homozygosity and population history in cattle. *BMC Genet*. 2012;13(1):70.
13. Van Laere A, Nguyen M, Braunschweig M, et al. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature*. 2003;425(6960):832–6.
14. Setoguchi K, Furuta M, Hirano T, et al. Cross-breed comparisons identified a critical 591-kb region for bovine carcass weight QTL (CW-2) on chromosome 6 and the Ile-442-Met substitution in NCAPG as a positional candidate. *BMC Genet*. 2009;10(1):43.
15. Chase K, Carrier D, Adler F, et al. Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. *Proc Natl Acad Sci U S A*. 2002;99(15):9930–5.
16. Sutter N, Bustamante C, Chase K, et al. A single IGF1 allele is a major determinant of small size in dogs. *Science*. 2007;316(5821):112–5.
17. Patterson D. Companion animal medicine in the age of medical genetics. *J Vet Intern Med*. 2000;14(1):1.
18. Jolly R, Leipold H. Inherited diseases of cattle – a perspective. *N Z Vet J*. 1973;21(7):147–55.
19. Meyers S, McDaneld T, Swist S, et al. A deletion mutation in bovine SLC4A2 is associated with osteopetrosis in Red Angus cattle. *BMC Genomics*. 2010;11(1):337.
20. Venhoranta H, Pausch H, Flisikowski K, et al. In frame exon skipping in UBE3B is associated with developmental disorders and increased mortality in cattle. *BMC Genomics*. 2014;15(1):890.
21. Charlier C, Segers K, Karim L, et al. The callipyge mutation enhances the expression of coregulated imprinted genes in cis without affecting their imprinting status. *Nat Genet*. 2001;27(4):367–9.
22. Beever J, Smit M, Meyers S, et al. A single-base change in the tyrosine kinase II domain of ovine FGFR3 causes hereditary chondrodysplasia in sheep. *Anim Genet*. 2006;37(1):66–71.
23. Rosengren Pielberg G, Golovko A, Sundström E, et al. A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. *Nat Genet*. 2008;40(8):1004–9.
24. Sironen A, Thomsen B, Andersson M, Ahola V, Vilkki J. An intronic insertion in KPL2 results in aberrant splicing and causes the immotile short-tail sperm defect in the pig. *Proc Natl Acad Sci U S A*. 2006;103(13):5006–11.
25. Sironen A, Uimari P, Venhoranta H, Andersson M, Vilkki J. An exonic insertion within Tex14 gene causes spermatogenic arrest in pigs. *BMC Genomics*. 2011;12(1):591.
26. Chase K. Teaching a new dog old tricks: identifying quantitative trait loci using lessons from plants. *J Hered*. 1999;90(1):43–51.
27. Acland GM, Ray K, Mellersh CS, et al. Linkage analysis and comparative mapping of canine progressive rod-cone degeneration (prcd) establishes potential locus homology with retinitis pigmentosa (RP17) in humans. *Proc Natl Acad Sci U S A*. 1998;95(6):3048–53.
28. Tonomura N, Elvers I, Thomas R, et al. Genome-wide association study identifies shared risk loci common to two malignancies in golden retrievers. *PLoS Genet*. 2015;11(2):e1004922.
29. Arendt M, Melin M, Tonomura N, et al. Genome-wide association study of golden retrievers identifies germ-line risk factors predisposing to mast cell tumours. *PLoS Genet*. 2015;11(11):e1005647.
30. Zimin A, Delcher A, Florea L, et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol*. 2009;10(4):R42.
31. Groenen M, Archibald A, Uenishi H, et al. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*. 2012;491(7424):393–398.
32. Cadieu E, Neff M, Quignon P, et al. Coat variation in the domestic dog is governed by variants in three genes. *Science*. 2009;326(5949):150–3.
33. Makvandi-Nejad S, Hoffman G, Allen J, et al. Four loci explain 83% of size variation in the horse. *PLoS One*. 2012;7(7):e39929.
34. Jones P, Chase K, Martin A, Davern P, Ostrander E, Lark K. Single-nucleotide-polymorphism-based association mapping of dog stereotypes. *Genetics*. 2008;179(2):1033–44.
35. Boyko A, Quignon P, Li L, et al. A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol*. 2010;8(8):e1000451.
36. Akey J, Ruhe A, Akey D, et al. Tracking footprints of artificial selection in the dog genome. *Proc Natl Acad Sci U S A*. 2010;107(3):1160–5.
37. Ashwell M, Heyen D, Sonstegard T, et al. Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. *J Dairy Sci*. 2004;87(2):468–75.
38. Zhou Z, Sheng X, Zhang Z, et al. Differential genetic regulation of canine hip dysplasia and osteoarthritis. *PLoS One*. 2010;5(10):e13219.
39. Karlsson E, Sigurdsson S, Ivansson E, et al. Genome-wide analyses implicate 33 loci in heritable dog osteosarcoma, including regulatory variants near CDKN2 A/B. *Genome Biol*. 2013;14(12):R132.
40. Ma J, Yang J, Zhou L, et al. Genome-wide association study of meat quality traits in a White Duroc × Erhualian F2 intercross and Chinese Sutai pigs. *PLoS One*. 2013;8(5):e64047.
41. Ng M, Levinson D, Faraone S, et al. Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry*. 2008;14(8):774–85.
42. Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*. 2010;467(7317):832–8.
43. Reis-Filho J, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet*. 2011;378(9805):1812–23.
44. Huang N, Shah P, Li C. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Brief Bioinform*. 2011;13(3):305–16.
45. Saetre P, Lindberg J, Leonard J, et al. From wild wolf to domestic dog: gene expression changes in the brain. *Mol Brain Res*. 2004;126(2):198–206.
46. Everts R, Band M, Liu Z, et al. A 7872 cDNA microarray and its use in bovine functional genomics. *Vet Immunol Immunopathol*. 2005;105(3–4):235–45.
47. Klopfleisch R, Lenze D, Hummel M, Gruber A. Metastatic canine mammary carcinomas can be identified by a gene expression profile that partly overlaps with human breast cancer profiles. *BMC Cancer*. 2010;10(1):618.
48. Renn S, Aubin-Horth N, Hofmann H. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC Genomics*. 2004;5:42.
49. Roy M, Kim N, Kim K, et al. Analysis of the canine brain transcriptome with an emphasis on the hypothalamus and cerebral cortex. *Mamm Genome*. 2013;24(11–12):484–99.
50. Matulova M, Rajova J, Vlasatikova L, et al. Characterization of chicken spleen transcriptome after infection with *Salmonella enterica* serovar enteritidis. *PLoS One*. 2012;7(10):e48101.
51. Duffy D, Krstic A, Schwarzl T, Higgins D, Kolch W. GSK3 inhibitors regulate MYCN mRNA levels and reduce neuroblastoma cell viability through multiple mechanisms, including p53 and Wnt signaling. *Mol Cancer Ther*. 2013;13(2):454–67.
52. Bottomly D, Walter N, Hunter J, et al. Evaluating gene expression in C57BL/6 J and DBA/2J mouse striatum using RNA-seq and microarrays. *PLoS One*. 2011;6(3):e17820.
53. Voineagu I, Wang X, Johnston P, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*. 2011;474(7351):380–4.
54. Gautier E, Shay T, Miller J, et al. Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nat Immunol*. 2012;13(11):1118–28.
55. Balakrishnan C, Mukai M, Gonser R, et al. Brain transcriptome sequencing and assembly of three songbird model systems for the study of social behavior. *PeerJ*. 2014;2:e396.
56. Rickard A, Petek L, Miller D. Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. *Hum Mol Genet*. 2015;24(20):5901–14.
57. Shalek A, Satija R, Adiconis X, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*. 2013;498(7453):236–40.

58. Cheng A, Shi J, Wong P, et al. Muscleblind-like 1 (Mbnl1) regulates pre-mRNA alternative splicing during terminal erythropoiesis. *Blood*. 2014;124(4):598–610.
59. Gregg C, Zhang J, Butler J, Haig D, Dulac C. Sex-specific parent-of-origin allelic expression in the mouse brain. *Science*. 2010;329(5992):682–5.
60. Wang X, Miller D, Harman R, Antczak D, Clark A. Paternally expressed genes predominate in the placenta. *Proc Natl Acad Sci U S A*. 2013;110(26):10705–10.
61. Chepelev I, Wei G, Tang Q, Zhao K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res*. 2009;37(16):e106–e106.
62. Iorizzo M, Senalik D, Grzebelus D, et al. De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. *BMC Genomics*. 2011;12(1):389.
63. Rogers R, Shao L, Sanjak J, Andolfatto P, Thornton K. Revised annotations, sex-biased expression, and lineage-specific genes in the *Drosophila melanogaster* group. *G3 (Bethesda)*. 2014;4(12):2345–51.
64. Du L, Li W, Fan Z, et al. First insights into the giant panda (*Ailuropoda melanoleuca*) blood transcriptome: a resource for novel gene loci and immunogenetics. *Mol Ecol Resour*. 2015;15(4):1001–13.
65. Hoeppner M, Lundquist A, Pirun M, et al. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One*. 2014;9(3):e91172.
66. Auer P, Doerge R. Statistical design and analysis of RNA sequencing data. *Genetics*. 2010;185(2):405–16.
67. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*. 2013;14(9):R95.
68. Liu Y, Zhou J, White K. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*. 2014;30(3):301–4.
69. Williams A, Thomas S, Wyman S, Holloway A. RNA-seq data: challenges in and recommendations for experimental design and analysis. *Curr Protoc Hum Genet*. 2014;83:11.13.1–11.13.20.
70. Sellin Jeffries M, Kiss A, Smith A, Oris J. A comparison of commercially-available automated and manual extraction kits for the isolation of total RNA from small tissue samples. *BMC Biotechnol*. 2014;14(1):94.
71. Lahens N, Kavakli I, Zhang R, et al. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol*. 2014;15(6):R86.
72. Perkins T, Kingsley R, Fookes M, et al. A strand-specific RNA–seq analysis of the transcriptome of the typhoid bacillus *Salmonella* typhi. *PLoS Genet*. 2009;5(7):e1000569.
73. Grabherr M, Haas B, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–52.
74. Levin J, Yassour M, Adiconis X, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods*. 2010;7(9):709–15.
75. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. *Genome Res*. 2011;21(12):2213–23.
76. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods*. 2009;6(11 s):S22–32.
77. Sugarbaker D, Richards W, Gordon G, et al. Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc Natl Acad Sci U S A*. 2008; 105(9):3521–6.
78. Hahn D, Ragland G, Shoemaker D, Denlinger D. Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. *BMC Genomics*. 2009;10(1):234.
79. Maher C, Kumar-Sinha C, Cao X, et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature*. 2009;458(7234):97–101.
80. Meyer E, Aglyamova G, Wang S, et al. Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*. 2009;10(1):219.
81. Schwarz D, Robertson H, Feder J, et al. Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. *BMC Genomics*. 2009;10(1):633.
82. Zhao Q, Caballero O, Levy S, et al. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc Natl Acad Sci U S A*. 2009;106(6):1886–91.
83. Ferguson L, Lee S, Chamberlain N, et al. Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus. *Mol Ecol*. 2010;19:240–54.
84. Wang H, Zhang H, Wong Y, et al. Rapid transcriptome and proteome profiling of a non-model marine invertebrate, Bugula neritina. *Proteomics*. 2010;10(16):2972–81.
85. Kukekova AV, Johnson JL, Teiling C, et al. Sequence comparison of prefrontal cortical brain transcriptome from a tame and an aggressive silver fox (*Vulpes vulpes*). *BMC Genomics*. 2011;12:482.
86. Ekblom R, Slate J, Horsburgh G, Birkhead T, Burke T. Comparison between normalised and unnormalised 454-sequencing libraries for small-scale RNA-seq studies. *Comp Funct Genomics*. 2012;2012:1–8.
87. Gordon A, Hannon G. Fastx-toolkit. *FASTQ/A Short-Reads Preprocessing Tools (Unpublished)*. 2010. Available at: http://hannonlab. cshl. edu/fastx_toolkit. Accessed December 12, 2015.
88. Dodt M, Roehr J, Ahmed R, Dieterich C. FLEXBAR – flexible barcode and adapter processing for next-generation sequencing platforms. *Biology*. 2012;1(3):895–905.
89. Bolger A, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20.
90. Wilhelm B, Landry J. RNA-Seq – quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*. 2009;48(3):249–57.
91. Fonseca N, Marioni J, Brazma A. RNA-seq gene profiling – a systematic empirical comparison. *PLoS One*. 2014;9(9):e107026.
92. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.
93. Dobin A, Davis C, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2012;29(1):15–21.
94. Zhao S, Zhang B. A comprehensive evaluation of Ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*. 2015;16(1):97.
95. Clark M, Amaral P, Schlesinger F, et al. The reality of pervasive transcription. *PLoS Biol*. 2011;9(7):e1000625.
96. Pickrell J, Marioni J, Pai A, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010;464(7289):768–72.
97. Harrow J, Frankish A, Gonzalez J, et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res*. 2012;22(9):1760–74.
98. George N, Bowyer J, Crabtree N, Chang C. An iterative leave-one-out approach to outlier detection in RNA-seq data. *PLoS One*. 2015;10(6):e0125224.
99. Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
100. Robinson M, McCarthy D, Smyth G. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26(1):139–40.
101. Chen Y, McCarthy D, Robinson M, Smyth G. *edgeR: Differential Expression Analysis of Digital Gene Expression Data: User's Guide*. 2015. Available at: https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf. Accessed October 12, 2015.
102. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. 2013;14(1):91.
103. Pérez-Montarelo D, Madsen O, Alves E, et al. Identification of genes regulating growth and fatness traits in pig through hypothalamic transcriptome analysis. *Physiol Genomics*. 2013;46(6):195–206.
104. Frazee A, Sabunciyan S, Hansen K, Irizarry R, Leek J. Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics*. 2014;15(3):413–26.
105. Huang D, Sherman B, Lempicki R. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2008;4(1):44–57.
106. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–504.
107. Smoot M, Ono K, Ruscheinski J, Wang P, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. 2010;27(3):431–2.
108. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559.
109. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
110. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–93.
111. Vijay N, Poelstra J, Künstner A, Wolf J. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol Ecol*. 2012;22(3):620–34.
112. Liao W, Spolski R, Li P, et al. Opposing actions of IL-2 and IL-21 on Th9 differentiation correlate with their differential regulation of BCL6 expression. *Proc Natl Acad Sci U S A*. 2014;111(9):3508–13.
113. Adams M. The genome sequence of *Drosophila melanogaster*. *Science*. 2000; 287(5461):2185–95.
114. Lander E, Linton L, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
115. Chinwalla A, Cook L, Delehaunty K, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002;420(6915):520–62.
116. Pruitt K, Tatusova T, Maglott D. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 2007;35(Database):D61–5.
117. Hillier L, Miller W, Birney E, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*. 2004;432(7018):695–716.

118. Chain P, Grafham D, Fulton R, et al. Genome project standards in a new era of sequencing. *Science*. 2009;326(5950):236–7.
119. Alkan C, Sajjadian S, Eichler E. Limitations of next-generation genome sequence assembly. *Nat Methods*. 2010;8(1):61–5.
120. Dalloul R, Long J, Zimin A, et al. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol*. 2010;8(9):e1000475.
121. Qiu Q, Zhang G, Ma T, et al. The yak genome and adaptation to life at high altitude. *Nat Genet*. 2012;44(8):946–9.
122. Peng X, Alfoeldi J, Gori K, et al. The draft genome sequence of the ferret (Mustela putorius furo) facilitates study of human respiratory disease. *Nat Biotechnol*. 2014;32(12):1250–55.
123. Ye L, Hillier L, Minx P, et al. A vertebrate case study of the quality of assemblies derived from next-generation sequences. *Genome Biol*. 2011;12(3):R31.
124. Norgren R. Improving genome assemblies and annotations for nonhuman primates. *ILAR J*. 2013;54(2):144–53.
125. Marfell B, O'Brien R, Griffin J. Global gene expression profiling of monocyte-derived macrophages from red deer *(Cervus elaphus)* genotypically resistant or susceptible to *Mycobacterium avium* subspecies *paratuberculosis* infection. *Dev Comp Immunol*. 2013;40(2):210–7.
126. Quinn A, Juneja P, Jiggins F. Estimates of allele-specific expression in *Drosophila* with a single genome sequence and RNA-seq data. *Bioinformatics*. 2014;30(18):2603–10.
127. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Res*. 2010;21(6):936–9.
128. Zerbino D, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821–9.
129. Schulz M, Zerbino D, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*. 2012;28(8):1086–92.
130. Garber M, Grabherr M, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods*. 2011;8(6):469–77.
131. Wolf J. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Mol Ecol Resour*. 2013;13(4):559–72.
132. Haas B, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013;8(8):1494–512.
133. Jackson K, Ramakrishnan L, Muriki K, et al. Performance analysis of high performance computing applications on the Amazon web services cloud. 2010 *IEEE Second International Conference on Cloud Computing Technology and Science*. 2010. Indianapolis.
134. Towns J, Cockerill T, Dahan M, et al. XSEDE: accelerating scientific discovery. *Comput Sci Eng*. 2014;16(5):62–74.
135. Brown C, Scott C, Crusoe M, Sheneman L, Rosenthal J. *Khmer-Protocols* 0.8.4 *Documentation*.2013.Availableat:http://dx.doi.org/10.6084/m9.figshare.878460. Accessed August 26, 2015.
136. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinform. 2009;10(4):1–14.
137. Altschul S. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.
138. Kent W. BLAT – the BLAST-like alignment tool. *Genome Res*. 2002;12(4):656–64.
139. Cock P, Antao T, Chang J, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422–3.
140. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7(3):562–78.