SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Signaling and regulatory biology
• REVIEW •

February 2017 Vol.60 No.2: 116–125
doi: 10.1007/s11427-015-0349-4

Characterizing and annotating the genome using RNA-seq data

Geng Chen[1], Tieliu Shi[2*] & Leming Shi[1,3,4,5**]

[1]*Center for Pharmacogenomics, School of Pharmacy and School of Life Sciences, Fudan University, Shanghai 201203, China;*
[2]*The Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China;*
[3]*State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China;*
[4]*Fudan-Zhangjiang Center for Clinical Genomics, Shanghai 201203, China;*
[5]*Zhangjiang Center for Translational Medicine, Shanghai 201203, China*

Received November 18, 2015; accepted December 16, 2015; published online June 12, 2016

Bioinformatics methods for various RNA-seq data analyses are in fast evolution with the improvement of sequencing technologies. However, many challenges still exist in how to efficiently process the RNA-seq data to obtain accurate and comprehensive results. Here we reviewed the strategies for improving diverse transcriptomic studies and the annotation of genetic variants based on RNA-seq data. Mapping RNA-seq reads to the genome and transcriptome represent two distinct methods for quantifying the expression of genes/transcripts. Besides the known genes annotated in current databases, many novel genes/transcripts (especially those long noncoding RNAs) still can be identified on the reference genome using RNA-seq. Moreover, owing to the incompleteness of current reference genomes, some novel genes are missing from them. Genome-guided and *de novo* transcriptome reconstruction are two effective and complementary strategies for identifying those novel genes/transcripts on or beyond the reference genome. In addition, integrating the genes of distinct databases to conduct transcriptomics and genetics studies can improve the results of corresponding analyses.

RNA-seq, genome-guided transcriptome reconstruction, *de novo* assembly, long noncoding RNA, genetic variants

Citation: Chen, G., Shi, T., and Shi, L. (2017). Characterizing and annotating the genome using RNA-seq data. Sci China Life Sci 60, 116–125. doi: 10.1007/s11427-015-0349-4

INTRODUCTION

RNA-seq technologies have been greatly improved in terms of sequencing time, cost, throughput and accuracy. These improvements have tremendously facilitated the transcriptomic studies for diverse species (Chen et al., 2011b; Oshlack et al., 2010; Ozsolak and Milos, 2011; Wang et al., 2009). However, in order to accomplish corresponding research goals, the matter of how to efficiently and fully explore the RNA-seq data using appropriate approaches faces many challenges. Notably, the annotation of genes on the reference genome of many organisms is still far from complete and many novel genes/transcripts (including both protein-coding and noncoding) remain to be identified. On the other hand, those reference genomes constructed for diverse species may be incomplete and some genomic sequences that contain genes are missing (Chen et al., 2011a, 2013b; Li et al., 2010). These problems raise the urgent need for identifying and characterizing those novel genes/transcripts on or out of the reference genome. Many studies, using RNA-seq, have identified a number of novel protein-coding and noncoding genes/transcripts of interested organisms (Cabili et al., 2011; Chettoor et al., 2014; Guttman et al.,

*Corresponding author (email: tlshi@bio.ecnu.edu.cn)
**Corresponding author (email: lemingshi@fudan.edu.cn)

2010; Pauli et al., 2012; Roberts et al., 2011). Because long noncoding RNAs (lncRNAs) were less explored previously, they are a significant portion of those identified novel genes/transcripts.

Transcriptomic and genetic studies are generally conducted based on annotated genes, thus the completeness of gene annotation could substantially influence related analyses (Chen et al., 2013a, 2015). RefSeq database was the most popular one used in previously published research. This is because of the high-confidence of its annotated genes. Nevertheless, RefSeq is very conservative and annotated a limited number of genes (Pruitt et al., 2014). The ENCODE project has greatly improved the human gene annotation and provided a more comprehensive gene set in GENCODE database (Consortium, 2012; Harrow et al., 2012). Moreover, GENCODE (corresponding to Ensembl (Cunningham et al., 2015)) annotated many lncRNAs. The transcripts annotated in GENCODE/Ensembl databases can be divided into multiple distinct categories, whereas the majority of genes annotated in RefSeq database are protein-coding. Besides, UCSC (Rosenbloom et al., 2015) and AceView (Thierry-Mieg and Thierry-Mieg, 2006) databases also provide the genes for distinct organisms. Each database used their own pipeline and criteria to annotate different types of genes. Thus, the genes annotated in those databases can vary in quantity and quality, but each database may contain specific genes that were not annotated in other databases (Chen et al., 2013a). Knowing how to integrate those databases for comprehensively carrying out gene expression analysis using RNA-seq could be very helpful. Furthermore, RNA-seq has the potential to capture all the genes/transcripts expressed in cells, which could provide the possibility of detecting novel genes/transcripts unannotated in present databases. Although several reviews have discussed how to process the RNA-seq data properly in order to characterize the gene expression profile, they were mainly focused on the common analyses of known genes (Chen et al., 2011b; Garber et al., 2011; Oshlack et al., 2010; Pepke et al., 2009). A knowledge of how to characterize and annotate those novel genes/transcripts on or beyond the reference genome using RNA-seq is currently lacking.

In this review, we first discussed genome-based and transcriptome-based approaches for characterizing the expression profile of known genes. Then, we summarized how to use genome-guided and *de novo* transcriptome reconstruction methods to identify the novel genes/transcripts unannotated on the reference genome. We also provided different approaches to characterize the novel genes/transcripts that were missing from the reference genome. Furthermore, we described current strategies for identifying and annotating lncRNAs. In the end, we further discussed how to combine different gene databases and RNA-seq data to improve the annotation of genetic variants.

EXPLORING THE EXPRESSION PROFILES OF KNOWN GENES

Integrating different gene databases

To comprehensively characterize the gene expression profiles, it is important to make the gene set used in the study as complete as possible. There are different gene databases that can be used for conducting related RNA-seq data analysis, such as RefSeq (Pruitt et al., 2014), Ensembl (Cunningham et al., 2015), UCSC (Rosenbloom et al., 2015) and AceView (Thierry-Mieg and Thierry-Mieg, 2006). However, previously, we showed that thousands of Ensembl genes were not annotated in AceView database and vice versa, and an integrated gene set of these two databases significantly improved diverse transcriptomic analyses (Chen et al., 2013a). The specific genes in each database could result from the differences of resources and methodologies used in their gene annotation. Therefore, incorporating the genes in different databases could both, increase the completeness of transcriptome and benefit associated studies.

To fully explore the expression profiles of all known/annotated genes, combining the Ensembl genes with those annotated in other databases for the purpose of generating a more complete gene set would be a good choice. Our previous study indicated that those human genes from one database, but unannotated in another, are mainly in the intergenic and intronic regions (Chen et al., 2013a). However, two genes from two distinct databases may partially overlap each other in sequence, and it is hard to determine whether they are the same ones or not due to the incomplete annotation of each database. To avoid counting duplicated genes, one can only add those genes that are in the intergenic and intronic regions of Ensembl gene annotation. Although a portion of genes in those databases might be predicted based on certain pipelines, RNA-seq data can be used to validate the authenticity of those predicted genes. Until now, a large number of RNA-seq data sets regarding various tissues and cell lines were published and can be accessed from distinct public databases, such as GEO (Barrett et al., 2013), ArrayExpress (Kolesnikov et al., 2015) and SRA (Kodama et al., 2012). Those RNA-seq data sets are valuable resources for examining the expression profiles of genes.

Strategies for quantifying the expression of genes/transcripts

After choosing an appropriate gene set, gene quantification can be carried out using related quantification tools (Figure 1). Two different methods could be used to quantify the expression of known genes: (i) one is based on the genome and gene annotation file (genome-based), such as Cufflinks

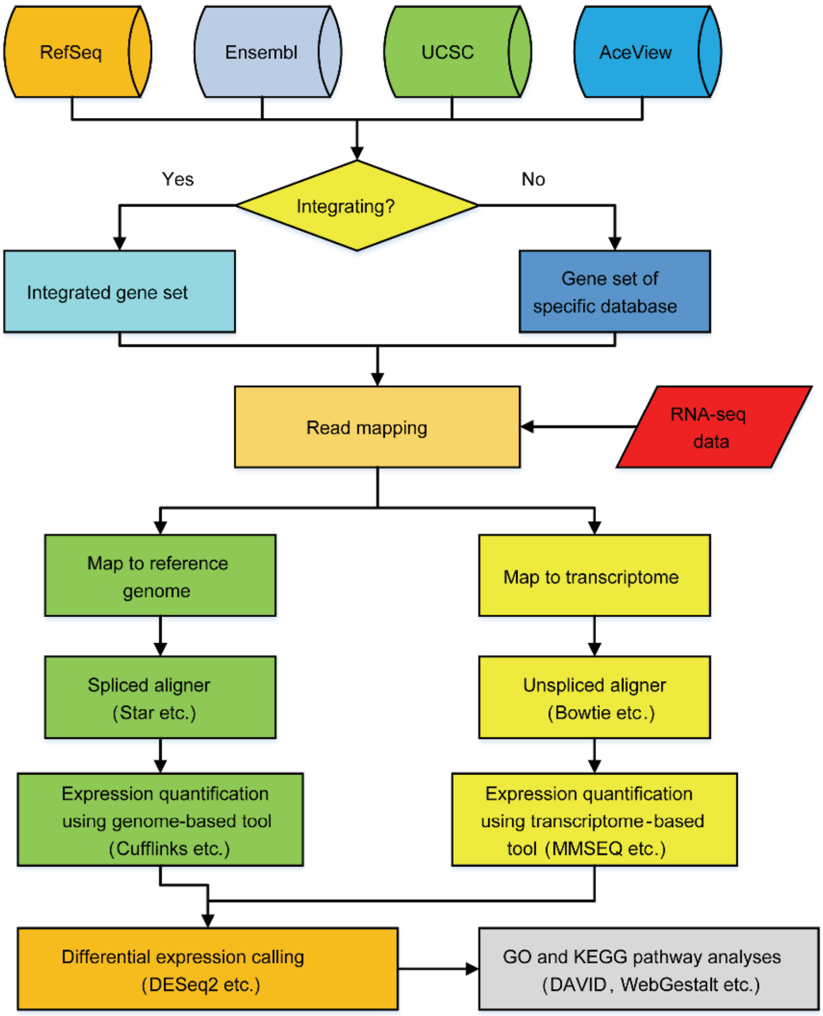
RefSeq
Ensembl
UCSC
AceView
Yes
Integrating?
No
Integrated gene set
Gene set of specific database
Read mapping
RNA-seq data
Map to reference genome
Map to transcriptome
Spliced aligner (Star etc.)
Unspliced aligner (Bowtie etc.)
Expression quantification using genome-based tool (Cufflinks etc.)
Expression quantification using transcriptome-based tool (MMSEQ etc.)
Differential expression calling (DESeq2 etc.)
GO and KEGG pathway analyses (DAVID, WebGestalt etc.)

Figure 1 Strategies for quantifying the expression of known genes. The gene set used for quantifying expression directly decides how many genes can be profiled. One can use the genes of specific database or integrate the genes of distinct databases. RNA-seq reads can be mapped to the reference genome or transcriptome to quantify the gene/transcript expression.

(Trapnell et al., 2010) and Scripture (Guttman et al., 2010); and (ii) another is based on the transcriptome sequences (transcriptome-based), for example, MMSEQ (Turro et al., 2011) and rSeq (Jiang and Wong, 2009). Genome-based approaches need to map the RNA-seq reads to the reference genome first, and then quantify the expression of those annotated genes according to the annotation file and read mapping information. In this case, spliced alignment tools are required for the mapping step in order to identify exon-exon splice junctions on the genome. Many spliced alignment programs (such as HISAT (Kim et al., 2015), Star (Dobin et al., 2013) and TopHat2 (Kim et al., 2013)) have been developed for aligning RNA-seq reads to the genome. However, the performances of these spliced aligners may vary greatly and the corresponding comparison for them can be found in a relevant review (Engstrom et al., 2013) and the recent paper of HISAT (Kim et al., 2015). In contrast, transcriptome-based methods employ unspliced aligners (such as BWA (Li and Durbin, 2009)) to map RNA-seq reads directly to the transcriptome sequences and do not require considering splice junctions between exons. Different unspliced aligners may also have distinct mapping performances and some reviews have compared them in detail (Fonseca et al., 2012; Li and Homer, 2010). In general, genome-based methods take more time to quantify gene/isoform expression compared with transcriptome-based approaches. However, genome-based strategies can be used to identify novel genes/transcripts. Such cannot be done by transcriptome-based approaches.

CHARACTERIZING UNANNOTATED GENES/TRANSCRIPTS ON THE REFERENCE GENOME

Genome-guided transcriptome reconstruction

Gene annotation of the reference genome for diverse organisms is undergoing continuous improvement and many novel genes/transcripts could be annotated using RNA-seq. To identify the novel genes/transcripts on the reference genome, two distinct approaches of genome-guided and *de novo* transcriptome (see next paragraph) assembly can be applied (Figure 2). For genome-guided transcriptome reconstruction, RNA-seq reads are first mapped to the reference genome using the aforementioned spliced aligners. Then *ab initio* transcriptome assembly can be carried out based on the mapping results in sam/bam format by employing corresponding tools, such as Cufflinks (Trapnell et al., 2010) and Scripture (Guttman et al., 2010). After transcriptome reconstruction, those novel genes/transcripts can be identified by removing the known genes/transcripts annotated in relevant databases based on certain filtering criteria.

De novo transcriptome reconstruction

De novo transcriptome assembly provides an alternative way for detecting novel genes on the reference genome (Figure 2). Firstly, RNA-seq reads are assembled using a *de novo* transcriptome assembler (such as Bridger (Chang et al., 2015), Trinity (Grabherr et al., 2011), Oases (Schulz et al., 2012), Trans-ABySS (Robertson et al., 2010) and etc.). Next, the assembled transcripts/contigs are mapped to the reference genome by employing an aligner for aligning long sequences (for example, Blat (Kent, 2002)). After mapping the assembled transcripts/contigs to the reference genome, one can compare the assembly with known genes to designate those not annotated in current databases as novel. However, assembly errors tend to increase for those assembled transcripts/contigs with short length and low coverage. In order to minimize the false positives, it is better to remove those assembled transcripts/contigs shorter than a certain length (such as <100 bp) and lower than a certain coverage.

Comparison of genome-guided and *de novo* transcriptome reconstruction

Genome-guided and *de novo* transcriptome reconstruction strategies have their own strengths and weaknesses (Martin and Wang, 2011). Choosing which approach to explore the novel genes/transcripts is based on the particular research goal and the properties of the data at hand. Generally, genome-guided approaches require lower sequencing depth as compared to *de novo* assembly methods (Garber et al., 2011). Because of the high accuracy and flexibility of genome-guided assembly, this is the main strategy for characterizing novel genes/transcripts. However, a major limitation of this approach is that it can only be applied to species for which a reference genome is available. Moreover, assembly quality of genome-guided transcriptome reconstruction approaches heavily depends on the read

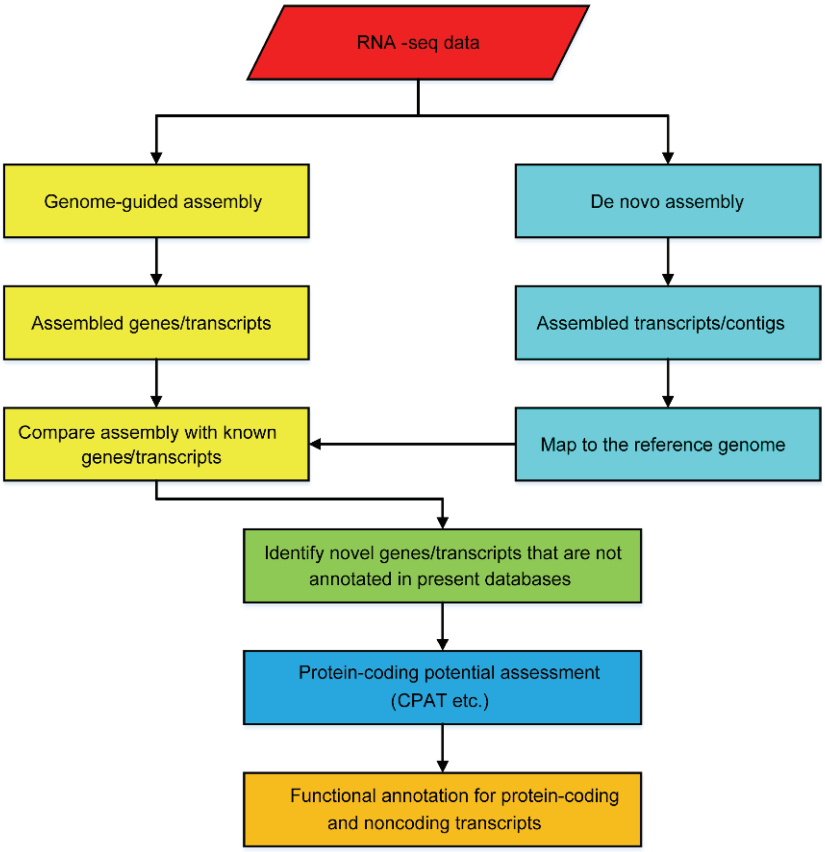
RNA -seq data
Genome-guided assembly
De novo assembly
Assembled genes/transcripts
Assembled transcripts/contigs
Compare assembly with known genes/transcripts
Map to the reference genome
Identify novel genes/transcripts that are not annotated in present databases
Protein-coding potential assessment (CPAT etc.)
Functional annotation for protein-coding and noncoding transcripts

Figure 2 Approaches for identifying novel genes/transcripts on the reference genome. Both genome-guided and *de novo* transcriptome reconstruction are the two distinct methods that can be used to identify novel genes/transcripts.

mapping quality. Splice junctions, single nucleotide polymorphisms (SNPs), indels (insertions and deletions) and other variants in the reads cause difficulties for correctly allocating reads to the right loci of the genome. Alternatively, one can use *de novo* transcriptome reconstruction methods to assess the expression profiles of known and novel genes/transcripts. Such an approach does not require reference genome. The two major advantages of *de novo* transcriptome assembly are: (i) it is applicable for any organisms whether a reference genome is available or not; and (ii) it can be used to detect genes that are missing from the reference genome. However, a *de novo* transcriptome assembly could not successfully reconstruct genes with long repetitive and highly similar sequences (like homologous genes). Moreover, high sequencing depth and large memory are needed for accurately and smoothly reconstructing the transcriptome. It is important to pay much attention to the quality of novel genes/transcripts identified from genome-guided or *de novo* transcriptome reconstruction methods. This is because those novel genes/transcripts reconstructed from RNA-seq data might not be in full-length. Rather, they may be just gene fragments. Stringent criteria and other validation methods are required to evaluate the quality of novel genes/transcripts. Therefore, combining genome-guided and *de novo* transcriptome reconstruction could achieve better assembly results and help to reliably identify more novel genes/transcripts.

IDENTIFYING NOVEL GENES ABSENT FROM THE REFERENCE GENOME

Factors responsible for the incompleteness of reference genomes

Technological limitations and complexity of genomes cause present reference genomes of various species, including that of humans, to contain assembly errors and missing genomic sequences. Several studies have revealed that due to its incompleteness, some functional genes were missing from the human reference genome (Chen et al., 2011a, 2013b; Li et al., 2010). Three major factors could be responsible for the incompleteness of constructed reference genomes. The first one is that genomes (especially for mammals) usually contain a large number of repetitive sequences, a situation which is a big challenge for *de novo* genome assembly (Gongora-Castillo and Buell, 2013). The second one is that different individuals may have their own specific genomic sequences. A significant number of human genomic sequences specific to Asians and Africans have been identified previously (Li et al., 2010). The last, but not the least, the limitation of sequencing technologies and assembly algorithms that can cause assembly errors and lead to missing certain genomic sequences. Two different approaches can be applied to identify the missing genes of the reference genome (Chen et al., 2013b) (Figure 3).

Strategy based on genome-wide comparison coupled with genome-guided transcriptome reconstruction

The first strategy is to obtain the gene sequences specific to other assembled genomes that are absent from the reference genome for the same organism. Genome-wide comparison between reference genome and other non-reference genomes (such as LAST (Kielbasa et al., 2011) tool) of the same organism is required to get the genomic sequences of non-reference genomes that are missing from the reference genome. Genome-guided transcriptome reconstruction can be conducted using those genomes as references to assess whether those specific genomic sequences harbor genes.

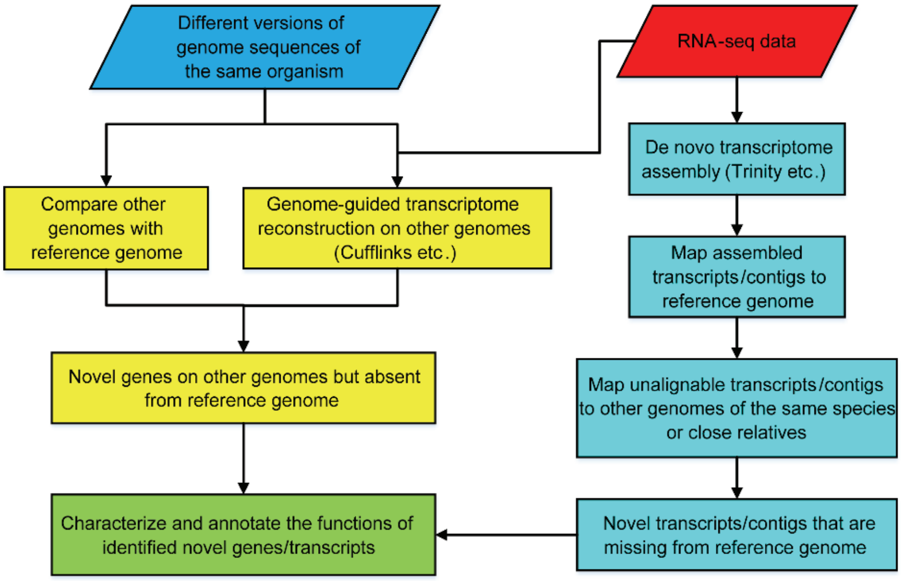
Different versions of genome sequences of the same organism
RNA-seq data
De novo transcriptome assembly (Trinity etc.)
Compare other genomes with reference genome
Genome-guided transcriptome reconstruction on other genomes (Cufflinks etc.)
Map assembled transcripts/contigs to reference genome
Novel genes on other genomes but absent from reference genome
Map unalignable transcripts/contigs to other genomes of the same species or close relatives
Characterize and annotate the functions of identified novel genes/transcripts
Novel transcripts/contigs that are missing from reference genome

Figure 3 Methods for identifying the novel genes beyond the reference genome. In order to identify the genes missing from the reference genome, those genes that can be annotated on the reference genome should be removed from the assembled transcriptome.

Strategy based on *de novo* transcriptome assembly

Another approach is based on *de novo* transcriptome assembly (Figure 3). The first step is to reconstruct the transcriptome using a *de novo* assembler as mentioned before. Next, one can map the assembled transcripts/contigs to the reference genome and remove those alignable sequences. The last step is to discriminate those bona fide missing gene sequences from those unalignable transcripts/contigs to the reference. If other assembled genomes of the same organism are available, one can align those reference-unalignable transcripts/contigs to those genomes to identify the alignable sequences. Those alignable transcripts/contigs could be the truly missing ones. If no other genomes of the same organism are available, mapping those reference-unalignable transcripts/contigs to the genomes of close relatives is an alternative possibility. For instance, one can map the reference-unalignable transcripts/contigs of human to chimpanzee, macaque, mouse and rat to identify the bona fide missing human genes (Chen et al., 2013b).

CHARACTERIZING AND ANNOTATING LONG NONCODING RNAS USING RNA-SEQ DATA

Identifying long noncoding RNAs

Long noncoding RNAs (lncRNAs) have important regulatory functions and they have become a hot research field in recent years (Cabili et al., 2011; Chen et al., 2011c; Derrien et al., 2012; Fan and Zhang, 2015; Lee and Kikyo, 2012; Pauli et al., 2012). In the past decades, researchers mainly focused on exploring the functions of protein-coding genes, but neglected an important category of genes called lncRNAs. Until now, only a small portion of lncRNAs have clear functions (Quek et al., 2015), while the functions of most lncRNAs are still unclear. RNA-seq provides unprecedented opportunities for exploring those novel lncRNAs. Some studies identified a number of novel lncRNAs based on the RNA-seq data in different organisms including humans (Cabili et al., 2011), mouse (Guttman et al., 2010) and zebrafish (Pauli et al., 2012). A popular way to identify lncRNAs is to conduct transcriptome reconstruction using genome-guided methods (such as Cufflinks (Trapnell et al., 2010)) first (Figure 4). Then those novel transcripts with high confidence can be obtained using a series of criteria (see the pipelines used in related research (Cabili et al., 2011; Guttman et al., 2010; Pauli et al., 2012)) and their protein-coding potential can be further assessed. Several programs have been developed for efficiently assessing the protein-coding capacity of transcripts, such as CPAT (Wang et al., 2013), CPC (Kong et al., 2007), lncRNA-MFDL (Fan and Zhang, 2015) and CONC (Liu et al., 2006). After protein-coding potential assessment, one can discriminate the noncoding RNAs from protein-coding ones based on their scores of protein-coding capacity. However, it is hard to correctly determine the protein-coding potential of some transcripts, especially for those bifunctional RNAs that have both protein-coding and noncoding functional properties (Ruiz-Orera et al., 2014). Tandem mass spectrometry is a powerful technology to identify proteins encoded by the corresponding transcripts (Nesvizhskii, 2007). Therefore, a complementary approach is to use mass spectrometry data to determine the protein-coding capacity of transcripts directly.

Functional annotation of long noncoding RNAs

After obtaining the novel lncRNAs from RNA-seq data, the next step is to annotate their functions (Figure 4). For protein-coding genes, one can characterize their functions by carrying out GO (gene ontology) term and KEGG pathway analyses directly. Functional annotation of lncRNAs is not so direct because the knowledge about lncRNAs is very limited so far. However, lncRNAs may be involved in the same pathways with those protein-coding genes in the same co-expressed module (Liao et al., 2011; Zhao et al., 2014). Thus one can annotate the functions of lncRNAs through construction of the co-expression network using WGCNA (Langfelder and Horvath, 2008) or other tools. Then the functions of lncRNAs can be inferred by assessing the functions of their co-expressed protein-coding genes. On the other hand, one can also predict the interactions between lncRNAs and proteins or miRNAs based on the CLIP-Seq data (Konig et al., 2011) or the known/predicted interactions in corresponding databases. For example, starBase catalogued many interactions of potein-RNA and miRNA-ncRNA (Li et al., 2014), which can be used to explore the potential regulatory functions of lncRNAs.

INTEGRATING DISTINCT DATABASES TO ANNOTATE GENETIC VARIANTS

With the development of genome- and exome-sequencing technologies, a large number of genetic variants have been identified in diverse diseases (Nielsen et al., 2011). Moreover, thousands of SNPs associated with various traits/diseases have been characterized in genome-wide association studies (Welter et al., 2014). However, most of these genetic variations are located in the intergenic and intronic regions, and only a small portion (~5%) is in the RefSeq exonic regions. How to interpret the functions of those intergenic and intronic variations represents a big challenge. In fact, the incomplete gene annotation is an important factor that hinders the functional annotation of those noncoding genetic variants. In our previous study, we showed that integrating the genes annotated in different databases could generate a more comprehensive gene set, which can locate

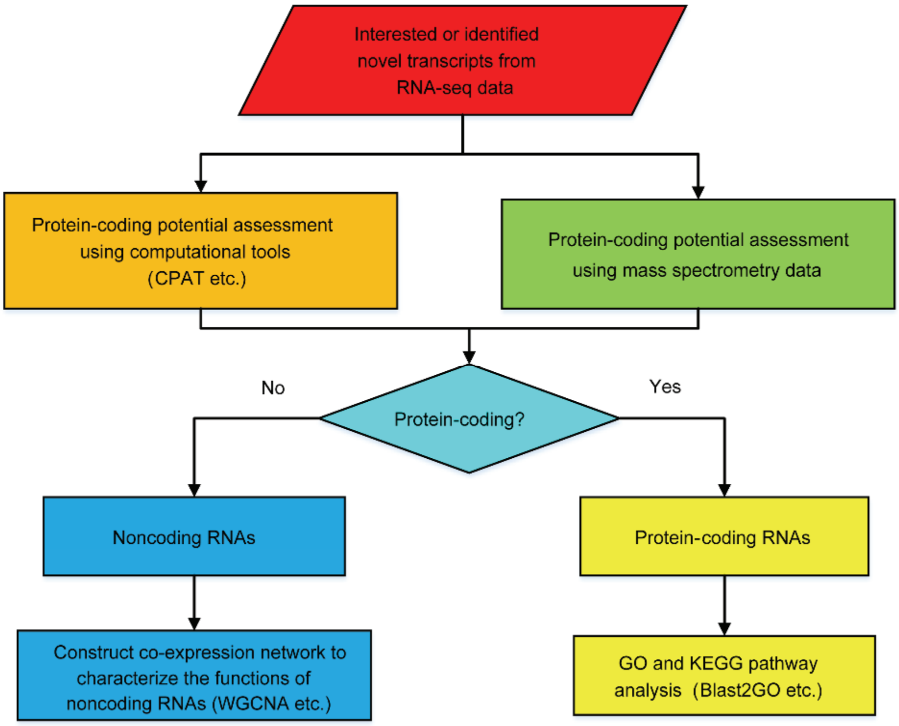
Interested or identified novel transcripts from RNA-seq data
Protein-coding potential assessment using computational tools (CPAT etc.)
Protein-coding potential assessment using mass spectrometry data
Protein-coding?
No
Yes
Noncoding RNAs
Protein-coding RNAs
Construct co-expression network to characterize the functions of noncoding RNAs (WGCNA etc.)
GO and KEGG pathway analysis (Blast2GO etc.)

Figure 4 Identifying and annotating novel long noncoding RNAs. Novel transcripts identified from genome-guided or *de novo* transcriptome reconstruction are used to assess their protein-coding potential. Long noncoding RNAs are those that are longer than 200 nt, but lack of protein-coding capacity.

more variants correctly into corresponding genic regions (Chen et al., 2015).

To fully explore the functions of genetic variations, it is crucial to accurately determine the relationship between variants and genes. Therefore, the first important step is to make the gene set used for annotation as complete as possible. At present, RefSeq, Ensembl, UCSC, AceView and other databases annotated the genes/transcripts for different organisms. To assign more variants into associated genes, one can annotate their genetic variants using distinct databases separately and then combine the annotation results to interpret the functions of those variants (Figure 5). After allocating those genetic variants to related genes, one may wonder how many of those genes harboring genetic variants are bona fide. Two complementary methods can be used to further determine the validity of genes: (i) if a variant was located to the same gene in two or more different databases, this variant was likely mapped correctly; and (ii) using related RNA-seq data to check whether those genes associated with variants are expressed. If the variants are in the genic regions, only those variants located in expressed genes are probably functional. Furthermore, although integrating the genes of distinct databases can obtain a more complete gene set, many genes/transcripts still have not been annotated in any databases. If necessary, one can further identify the novel genes on the reference genome using the approaches we mentioned above to locate more variants to relevant genic regions.

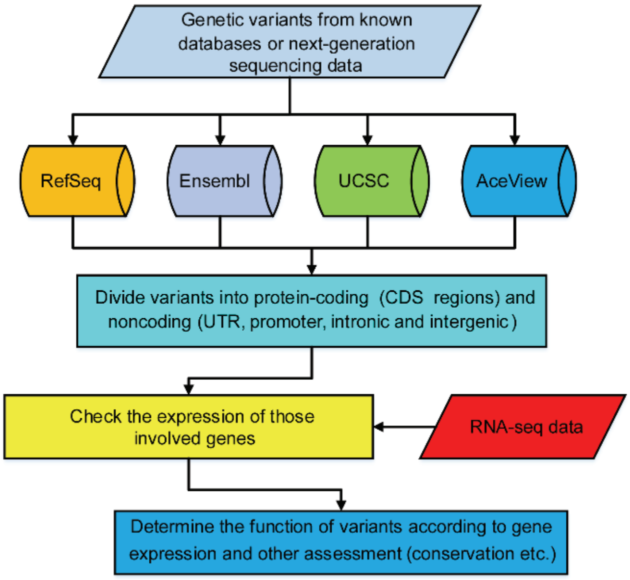
Genetic variants from known databases or next-generation sequencing data
RefSeq
Ensembl
UCSC
AceView
Divide variants into protein-coding (CDS regions) and noncoding (UTR, promoter, intronic and intergenic)
Check the expression of those involved genes
RNA-seq data
Determine the function of variants according to gene expression and other assessment (conservation etc.)

Figure 5 Integrating different gene databases to comprehensively annotate genetic variants. RNA-seq data can be used to examine whether those involved genes are expressed.

DISCUSSION

Here we reviewed how to use RNA-seq data to fully explore the gene expression of the whole genome as well as to annotate genetic variants. In the past ten years, great progress has been achieved in sequencing cells in bulk. Recently, single-cell sequencing technologies have become increasingly popular for characterizing the genome or transcriptome of any organism at the single cell level (Wu et al., 2014). However, both bulk and single-cell RNA-seq encounter the same challenge of how to use RNA-seq data to comprehensively investigate the transcriptional profile of all genes. Using a general pipeline, one can first map the RNA-seq data to the reference genome and then quantify the expression of known genes from a certain database like Ensembl or RefSeq. However, the genes annotated in all current databases are incomplete and many novel genes remain to be uncovered. Furthermore, the reference genome is incomplete as well, and a portion of genomic sequences harboring genes is still missing. Integrating the genes in different databases is an efficient way to obtain a more comprehensive gene set for corresponding research. Besides those genes annotated in databases, one can also use RNA-seq to detect novel genes on or beyond the reference genome using genome-guided and *de novo* transcriptome reconstruction approaches.

RNA-seq data from various tissues and cell lines are still in exponential growth. These abundant data are valuable resources for exploring the expression patterns of known and novel genes. Identifying lncRNAs and annotating their functions will continue to be a hot field of study in the future. On the other hand, the development of mass spectrometry technology can tremendously facilitate the identification of proteins generated by corresponding transcripts. Computational assessment in conjunction with mass spectrometry data can greatly increase the accuracy of identifying lncRNAs and bifunctional RNAs. Furthermore, the innovation of sequencing technologies and bioinformatics methods, and the refinement of reference genomes will continuously benefit the exploration of novel genes/transcripts. If the genes are more accurate and comprehensive, the functional assessment of genetic variants on genes will also be more precise.

In summary, RNA-seq technologies are powerful to characterize the gene expression and unravel the complexity of transcriptome. Continual improvements regarding both genes and corresponding reference genomes are required to make them more complete and accurate. Combining the genes in different databases can give rise to a more comprehensive gene set for improving diverse transcriptomic and genetic analyses. One can identify novel protein-coding and noncoding genes using RNA-seq data by employing genome-guided and/or *de novo* transcriptome assembly. These two different transcriptome reconstruction approaches are complementary and have their respective advantages and disadvantages. The completeness of gene set is crucial for correctly determining the association between the genetic variants and genes. Collectively, different analytical strategies of RNA-seq data coupled with distinct gene databases could dramatically improve the various analyses of transcriptomics and genetics.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the National High Technology Research and Development Program of China (2015AA020104), the China Human Proteome Project (2014DFB30010), the National Science Foundation of China (31471239, to Leming Shi), and the 111 Project (B13016).*

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41, D991–D995.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25, 1915–1927.

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., Cramer, C.L., and Huang, X. (2015). Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data. Genome Biol 16, 30.

Chen, G., Li, R., Shi, L., Qi, J., Hu, P., Luo, J., Liu, M., and Shi, T. (2011a). Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. BMC Genomics 12, 590.

Chen, G., Wang, C., Shi, L., Qu, X., Chen, J., Yang, J., Shi, C., Chen, L., Zhou, P., Ning, B., Tong, W., and Shi, T. (2013a). Incorporating the human gene annotations in different databases significantly improved transcriptomic and genetic analyses. RNA 19, 479–489.

Chen, G., Wang, C., Shi, L., Tong, W., Qu, X., Chen, J., Yang, J., Shi, C., Chen, L., Zhou, P., Lu, B., and Shi, T. (2013b). Comprehensively identifying and characterizing the missing gene sequences in human reference genome with integrated analytic approaches. Hum Genet 132, 899–911.

Chen, G., Wang, C., and Shi, T. (2011b). Overview of available methods for diverse RNA-Seq data analyses. Sci China Life Sci 54, 1121–1128.

Chen, G., Yin, K., Shi, L., Fang, Y., Qi, Y., Li, P., Luo, J., He, B., Liu, M., and Shi, T. (2011c). Comparative analysis of human protein-coding and noncoding RNAs between brain and 10 mixed cell lines by RNA-Seq. PLoS One 6, e28318.

Chen, G., Yu, D., Chen, J., Cao, R., Yang, J., Wang, H., Ji, X., Ning, B., and Shi, T. (2015). Re-annotation of presumed noncoding disease/trait-associated genetic variants by integrative analyses. Sci Rep 5, 9453.

Chettoor, A.M., Givan, S.A., Cole, R.A., Coker, C.T., Unger-Wallace, E., Vejlupkova, Z., Vollbrecht, E., Fowler, J.E., and Evans, M.M. (2014). Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biol 15, 414.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kahari, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., and Flicek, P. (2015). Ensembl 2015. Nucleic Acids Res 43, D662–669.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J., and Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22, 1775–1789.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Engstrom, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Ratsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigo, R., Bertone, P., and Consortium, R. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 10, 1185–1191.

Fan, X.N., and Zhang, S.W. (2015). lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. Mol Biosyst 11, 892–897.

Fonseca, N.A., Rung, J., Brazma, A., and Marioni, J.C. (2012). Tools for mapping high-throughput sequencing data. Bioinformatics 28, 3169–3177.

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8, 469–477.

Gongora-Castillo, E., and Buell, C.R. (2013). Bioinformatics challenges in *de novo* transcriptome assembly using short read sequences in the absence of a reference genome sequence. Nat Prod Rep 30, 490–500.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644–652.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S., and Regev, A. (2010). *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature Biotechnol 28, 503–510.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R., and Hubbard, T.J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22, 1760–1774.

Jiang, H., and Wong, W.H. (2009). Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25, 1026–1032.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res 12, 656–664.

Kielbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. Genome Res 21, 487–493.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12, 357–360.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.

Kodama, Y., Shumway, M., Leinonen, R., and International Nucleotide Sequence Database, C. (2012). The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40, D54–D56.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., Megy, K., Pilicheva, E., Rustici, G., Tikhonov, A., Parkinson, H., Petryszak, R., Sarkans, U., and Brazma, A. (2015). ArrayExpress update—simplifying data submissions. Nucleic Acids Res 43, D1113–D1116.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res, W345–W349.

Konig, J., Zarnack, K., Luscombe, N.M., and Ule, J. (2011). Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet 13, 77–83.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559.

Lee, C., and Kikyo, N. (2012). Strategies to identify long noncoding RNAs involved in gene regulation. Cell Biosci 2, 37.

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

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Briefings Bioinform 11, 473–483.

Li, J.H., Liu, S., Zhou, H., Qu, L.H., and Yang, J.H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 42, D92–97.

Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., Zhou, G., Zhu, X., Wu, H., Qin, J., Jin, X., Li, D., Cao, H., Hu, X., Blanche, H., Cann, H., Zhang, X., Li, S., Bolund, L., Kristiansen, K., Yang, H., Wang, J., and Wang, J. (2010). Building the sequence map of the human pan-genome. Nat Biotechnol 28, 57–63.

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., Zhao, G., Luo, H., Bu, D., Zhao, H., Skogerbo, G., Wu, Z., and Zhao, Y. (2011). Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res 39, 3864–3878.

Liu, J., Gough, J., and Rost, B. (2006). Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet 2, e29.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nat Rev Genet 12, 671–682.

Nesvizhskii, A.I. (2007). Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol Biol 367, 87–119.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12, 443-451.

Oshlack, A., Robinson, M.D., and Young, M.D. (2010). From RNA-seq reads to differential expression results. Genome Biol 11, 220.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12, 87–98.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22, 577–591.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nat Methods 6, S22–32.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., Murphy, M.R., O'Leary, N.A., Pujar, S., Rajput, B., Rangwala, S.H., Riddick, L.D., Shkeda, A., Sun, H., Tamez, P., Tully, R.E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D.R., Murphy, T.D., and Ostell, J.M. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42, D756–D763.

Quek, X.C., Thomson, D.W., Maag, J.L., Bartonicek, N., Signal, B., Clark, M.B., Gloss, B.S., and Dinger, M.E. (2015). lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res 43, D168–D173.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27, 2325–2329.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y.S., Newsome, R., Chan, S.K., She, R., Varhol, R., Kamoh, B., Prabhu, A.L., Tam, A., Zhao, Y., Moore, R.A., Hirst, M., Marra, M.A., Jones, S.J., Hoodless, P.A., and Birol, I. (2010). *De novo* assembly and analysis of RNA-seq data. Nat Methods 7, 909–912.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hickey, G., Hinrichs, A.S., Hubley, R., Karolchik, D., Learned, K., Lee, B.T., Li, C.H., Miga, K.H., Nguyen, N., Paten, B., Raney, B.J., Smit, A.F., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., and Kent, W.J. (2015). The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 43, D670–681.

Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a source of new peptides. eLife 3, e03523.

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

Thierry-Mieg, D., and Thierry-Mieg, J. (2006). AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 7 Suppl 1, S12 11–14.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515.

Turro, E., Su, S.Y., Goncalves, A., Coin, L.J., Richardson, S., and Lewin, A. (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol 12, R13.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res 41, e74.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nat Rev Genet 42, D1001–D1006.

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., and Quake, S.R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods 11, 41–46.

Zhao, Y., Luo, H., Chen, X., Xiao, Y., and Chen, R. (2014). Computational methods to predict long noncoding RNA functions based on co-expression network. Methods Mol Biol 1182, 209–218.