Front. Electr. Electron. Eng. China 2011, 6(2): 275–282
DOI 10.1007/s11460-011-0148-9

RESEARCH ARTICLE

Chao YE, Linxi LIU, Xi WANG, Xuegong ZHANG

Observations on potential novel transcripts from RNA-Seq data

Abstract With the rapid development of next generation deep sequencing technologies, sequencing cDNA reverse-transcribed from RNA molecules (RNA-Seq) has become a key approach in studying gene expression and transcriptomes. Because RNA-Seq does not rely on annotation of known genes, it provides the opportunity of discovering transcripts that have not been annotated in current databases. Studying the distribution of RNA-Seq signals and a systematic view on the potential new transcripts revealed from the signals is an important step toward the understanding of transcriptomes.

Keywords RNA-Seq, novel transcripts, next generation sequencing, bioinformatics

1 Introduction

Gene expression or the transcription from DNA to RNA is the primary step in transferring information encoded in genes to corresponding functional units. Expression levels of difference genes, defined as the relative abundance of RNA transcripts, are of a wide range. In some recent researches, people found that there were a proportion of transcripts which did not come from known gene regions [1,2]. Besides quantifying expression levels of known genes, it is also crucial to study those transcription events without known gene annotations and investigate their behaviors.

Several approaches have been developed since the 1990s to measure the abundance of many genes simultaneously. The well-known DNA microarray technology is a typical example of such technologies [3]. However, microarrays need to know gene structure and sequences in the design of their probes, which limits their application just on known genes. Benefiting from the rapid development of next generation sequencing (NGS) technologies [4,5], the newly emerging RNA-Sequencing (or RNA-Seq) is becoming a major technology to quantify transcript abundance levels at high resolution without prior gene annotation [6]. Typical NGS technologies include Roche/454, Illumina/Solexa and AB/SOLiD.

The typical protocol of RNA-Seq experiments is as follows. RNA molecules of interest are first extracted from total RNA in cells. Then they are sheared into RNA fragments. RNA fragments are reverse transcribed to cDNA by random priming. After cDNA amplification and size selection, cDNA libraries are prepared for sequencing experiments on NGS platforms. The resulting data are short reads of DNA sequences randomly sampled from the amplified fragments. They are usually stored as short reads in FASTQ format for further data processing and analysis [7].

Many RNA-Seq experiments and data have already been published since 2008. Applications of RNA-Seq include the study of gene expression at mRNA levels [8,9], the study of microRNAs if the RNA fragments are filtered to only include the short ones [10], the study of alternative splicing [11,12], and inference of isoform expressions for alternatively spliced genes [13], etc. Among them, identifying novel transcripts that have not been annotated in existing databases is an important task. It can reveal previously unknown protein-coding or non-coding transcripts present at certain cells, which might imply important biological functions. For this purpose, we developed an RNA-Seq data processing protocol and conducted a series of experiments on two published RNA-Seq datasets and identified two sets of putative novel transcripts. The experiments highlighted observations that are useful for studying potential novel transcripts from RNA-Seq data.

Received March 23, 2011; accepted April 12, 2011

Chao YE, Linxi LIU, Xi WANG, Xuegong ZHANG (✉)
Key Laboratory of Bioinformatics and Bioinformatics Division, Ministry of Education, Tsinghua National Laboratory for Information Science and Technology/Department of Automation, Tsinghua University, Beijing 100084, China
E-mail: zhangxg@tsinghua.edu.cn

2 Methods

2.1 RNA-Seq reads mapping

The first step after obtaining the RNA-Seq reads data is to map the short reads back to the reference genome. For RNA-Seq data on human samples, we use the human genome data from the UCSC website (http://genome.ucsc.edu) as the reference genome. For DNA reads, several mapping algorithms have been developed to map them back to the genome, and several software packages have been published, such as BFAST [14], Bowtie [15], and MAQ [16]. For RNA-Seq data, reads come from part of genome rather than the whole. Some post-transcriptional processing of RNAs like splicing in eukaryotes introduces RNA sequences that are not from any single location of the genome, but rather from junctions of distant parts. These features make the task of RNA-Seq reads mapping different and more challenging than DNA reads mapping. There are several software packages available for RNA-Seq reads mapping, such as TopHat [17], SpliceMap [18], and MapSplice [19]. We chose TopHat (version 1.1.1) (http://tophat.cbcb.umd.edu/) in our protocol. It uses the same core algorithm of Bowtie that is one of the fastest algorithms for aligning short reads and is also memory-efficient. It indexes the human genome with a Burrows-Wheeler index [20] to make the algorithm efficient [15]. It can also identify junction reads that come from spliced exons. It is one of the most widely used tools for RNA-Seq mapping that does not rely on given annotations. In the mapping, a certain number of mismatched nucleotides are allowed as there may be errors in the sequencing data and also there may be polymorphisms in RNA sequences. In our experiments reported in this paper, we allowed for up to 2 mismatches in each alignment.

After mapping RNA-Seq reads back to the reference genome, we get the following information for each read: whether it can be mapped to any particular location on the reference genome, which chromosome it is mapped to and the mapping coordinate on the chromosome, whether it is uniquely mapped or has multiple mapping locations, how many mismatch nucleotides are found in the mapping, etc. All these results would be written in one of the standard formats including SAM, BAM [21], BED or GTF. The information about file formats can be found at http://genome.ucsc.edu/FAQ/FAQformat.html. Figure 1 gives an example of the SAM format we used in our experiments.

```
1 TUPAC:1:52:1193:365#0
2 137
3 chr1
4 4868
5 0
6 50M
7 *
8 0
9 0
10 GACATCAAGTGCCCACCTTGGCTCGTGGCTCTCACTGCAACGGGAAAGCC
11 :@CACCC>ACABCC?BC>C>BCCBB9>CCCCB@>CBCBBBB9>BBBCBCB
12 NM:i:0  NH:i:6  CC:Z:chr15      CP:i:100333634
13
14 TUPAC:2:44:1064:1266#0
15 137
16 chr1
17 4869
18 0
19 33M757N17M
20 *
21 0
22 0
23 ACATCAAGTGCCCACCTTGGCTCGTGGCTCTCTCTTGCTCCTGCTCCTTC
24 ####=314,/;=;99-++32B2?:987A;;=;,<*96-7@07A>86A<AA
25 NM:i:1  XS:A:-  NH:i:7  CC:Z:chr15      CP:i:100332875
26
```

Fig. 1 Mapped result in SAM format. (The fields are split by tabs. Here we list these fields in row. Row 1 to row 12 is a read record and row 14 to row 25 is another one. These fields are: read name, bitwise flag contained the information about the read and its mapping result, chromosome, 1-based leftmost position in plus strand, mapping quality, extended CIGAR string (here are two popular types: 50 M means read length is 50 bp, 33M757N17M means it is a junction read combined with a 33 bp and a 17 bp fragment, distance between their mapped position is 757 bp), mate reference sequence ("=" means mate one were mapped in the same chromosome, "*" means there is not read mated with it), 1-based leftmost mate position in plus strand, distance between two mate reads (0 indicated that there was not mated read), sequence, quality of sequencing of each nucleotide (ASCII-33 format), and optional fields.)

2.2 Strategy for discovering novel transcripts

RNA-Seq reads are random samples from transcribed genomic regions. To identify the regions that are transcribed, we merge mapped reads into longer transcription fragments if reads are overlapping or the spacing of two neighboring reads are less than a given threshold. We call the genomic region formed in this way as "transfrag". The next step is to identify which of these transfrags are from known genes. We get the genome annotation from the UCSC website. There are three major types of annotations for the human genome: RefSeq, Ensemble, and Gencode. RefSeq is a database constructed by the National Center for Biotechnology Information (NCBI) [22]. It provides a complete collection of validated human genes. Ensemble annotation came from the Ensemble project [23] and Gencode annotation was built by ENCODE (the encyclopedia of DNA elements) [24]. These two annotations also include some predicted genes. We extracted the 5' location and 3' location of each annotated gene with all three types of annotations, and formed the table of genomic regions corresponding to known genes. All transfrags are checked with these regions, and those that overlap with any gene region are regarded as transcripts from known genes. Those transfrags that do not overlap with any of the known gene regions are regarded potential novel transcripts. Some of

them contain very few reads and may be due to sequencing noise or transcription noise. We can set a threshold to exclude those transfrags. We can also use some extra criteria to select transfrags according to their length and distance from known genes as putative novel transcripts for further investigations.

Following the random sampling model, the number of reads that are from a known gene region or a transfrag is affected by the expression level (abundance of the RNA transcript), the length of the region, and the sequencing depth (or the total number of reads obtained on the whole sample). We adopted the RPKM method to estimate the expression level of a gene or of a potential novel transcript. RPKM represents reads per kilo bases per million reads [7] and is the most widely used estimation for gene expression. After calculating the expression of transfrags, if the study involves two or more samples, we can detect differentially expressed transfrags with DEGseq, a software tool we developed earlier [25].

After finding some potential novel transcripts with high expression or differential expression between compared samples, we can use the UCSC Genome Browser [26] or visualization tools like integrative genomics viewer (IGV, http://www.broadinstitute.org/igv) [27] to further investigate the details of the read distribution.

2.3 Data

We chose two published RNA-Seq datasets in this experiment. They are datasets GM12878 and K562 from the ENCODE project. Dataset GM12878 was from lymphoblastoid samples. It contains 23416325 reads. Dataset K562 was from leukemia samples and it has 20871303 reads. Both datasets were obtained with the Illumina/Solexa platform. Read length in both datasets was 75 bp paired-end. We did not use the paired information in this experiment. For better quality, we only used the first 50 bp in each read in our experiments as the sequencing errors increase when reads are longer [4]. The reference human genome we used is hg18 (build 36, March 2006).

3 Results

Figure 2 gives the proportion of mapped and unmapped reads in the two datasets. We can see that although we allowed for two mismatches when we map the 50 bp reads to the reference genome, there are still almost 1/3 of the total reads that cannot be mapped. We can also observe that when the read length is short, junction reads only compose a small proportion of all the reads. Therefore, we did not use junction reads information in our discovery for possible novel transcripts.

On dataset GM12878, 40.1%, 13.9%, and 50.7% of

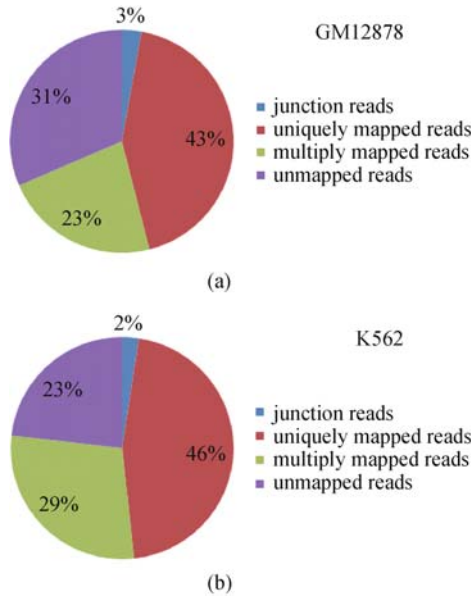


Fig. 2 Percentage of junction reads, uniquely mapped reads, multiply mapped reads, and unmapped reads in datasets GM12878 and K562. (a) GM12878; (b) K562

the mapped reads fall into intergenic regions according to the gene annotation of RefSeq, Ensemble, and Gencode, respectively. We used the threshold of 50 bp when merging mapped reads into transfrags. With this setting, we obtained a total of 782397 transfrags. Among them, 21.4%, 21.6%, and 38.6% fall into intergenic regions according to the RefSeq, Ensemble, and Gencode annotations, respectively. Taking the intersection of these transfrags, we got 119808 transfrags (15.3%) as potential novel transcripts.

The same procedures were applied on dataset K562. The total number of transfrags was 843260. There were 38.3%, 15.7%, and 50.7% mapped reads and 29.8%, 29.4%, and 43.4% transfrags falling into intergenic regions according to the RefSeq, Ensemble, and Gencode annotations, respectively. Also, we got 191426 transfrags (22.7%) as potential novel transcripts.

We studied the length distribution of the obtained potential novel transcripts, and compared them with transfrags located in known gene regions based on the RefSeq annotation. Figure 3 shows the logarithmic histogram of transfrag lengths of known genes and potential novel transcripts of the two datasets. We can see that the general styles of transfrag length distributions are the same between known genes and potential novel transcripts, as well as between the two datasets. However, the tails are heavier in the distribution of transfrags corresponding to known genes, which indicates that there are more longer transfrags in the known genes.

We are more interested in those potential novel transcripts that are relatively far from known genes and contain sufficient reads in each transfrag. They are more possibly from previously un-annotated coding or noncoding transcripts. Figure 4 illustrates the numbers of

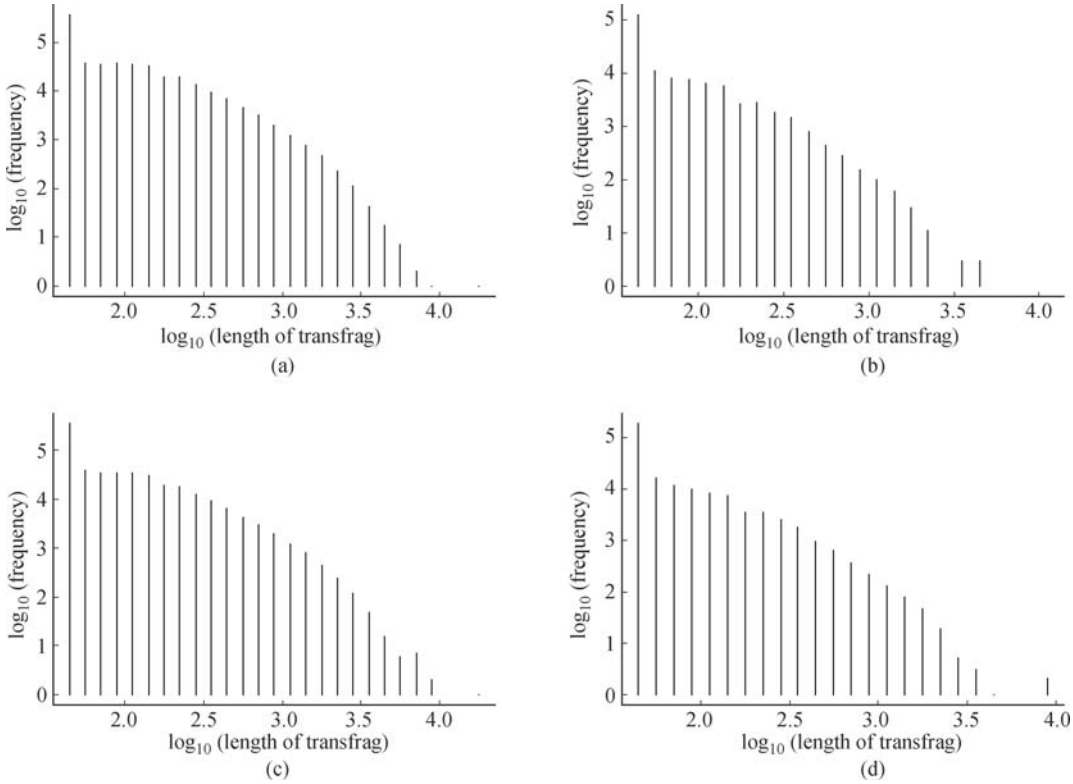


Fig. 3 Logarithmic histogram of length of transfrags located in gene regions and intergenic regions in two datasets. (a) Transfrags in gene region (GM12878); (b) transfrags in intergenic region (GM12878); (c) transfrags in gene region (K562); (d) transfrags in intergenic region (K562)

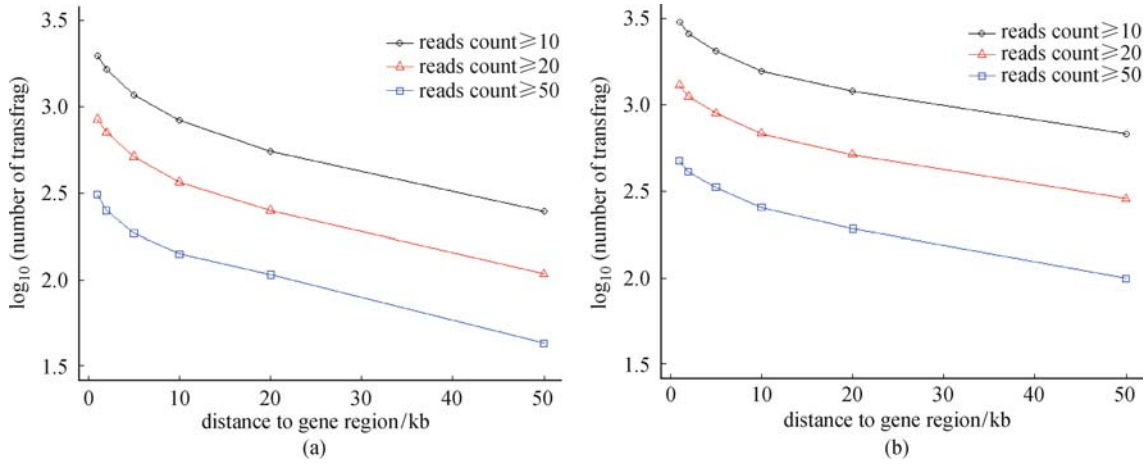


Fig. 4 Logarithm number of transfrags located in intergenic region in dataset GM12878 and K562. (a) GM12878; (b) K562

potential novel transcripts selected by controlling the distance from nearest known genes and the number of reads in each transfrag. We can see that, for example, when we require that the distance from nearest known genes is as far as 10 k and the reads count in each transfrag is no less than 20, there are still 364 and 681 putative novel transcripts in the two datasets, respectively. Among them, 288 pairs of the putative novel transcripts are discovered in both datasets.

We calculated the RPKMs of the transfrags of these putative novel transcripts, and also calculated the RPKMs for the transfrags located in known genes according to the RefSeq annotation. Figure 5 shows the logarithmic histograms of the RPKMs. From the histogram, we can observe that the proportion of relatively high RPKM in putative novel transcripts is less than the known gene, indicating that putative novel transcripts tend to have low level of expression.

We then applied DEGseq to search for putative novel transcripts that are differentially expressed in the two datasets we studied. It utilized a random sampling model based on MA-plot to estimate the P-value and exploit the method about multiple hypothesis testing adjustment introduced by Benjamini and Hochberg [28] to

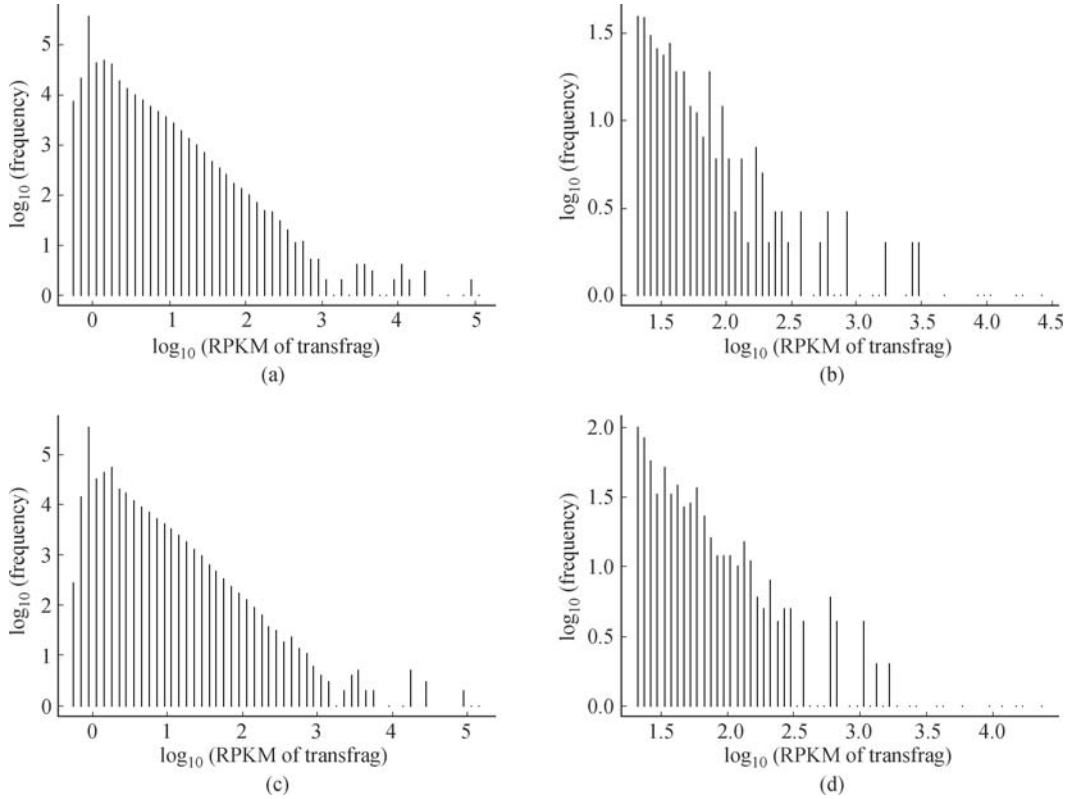


Fig. 5 Histogram of RPKM of transfrags located in gene regions and intergenic regions in two datasets. (a) Transfrags in gene region (GM12878); (b) transfrags in intergenic region (GM12878); (c) transfrags in gene region (K562); (d) transfrags in intergenic region (K562)

derive the Q-value from P-value. We set the Q-value level of 0.01. For the 364 putative novel transcripts obtained in GM12878, we found 85 transfrags differentially expressed with K562. For the 681 putative novel transcripts obtained in K562, 408 of them were detected as differentially expressed with GM12878. These observations show that a noticeable proportion of the detected putative novel transcripts have tissue-specificity in their expression, which indicates that they must have tissue-specific biological functions.

Figure 6 shows the reads distribution in an example region where putative novel transcripts were detected on both datasets. We also put the junction reads in the figure. We can observe coding gene-like structures in the read distribution. Another example is shown in Fig. 7, which shows the reads distribution of a region where putative novel transcripts were detected in only one of the datasets but not in the other. We can also see many reads in this region on dataset GM12878 and gene-like structures. These examples show that the detected putative novel transcripts may very possibly be new coding or noncoding genes that have not been previously reported.

4 Conclusion and discussion

In this paper, we presented a protocol for detecting putative novel transcripts from RNA-Seq data, and described our experiment observations on potential novel transcripts in two RNA-Seq datasets. The experiments show that there are a noticeable number of reads that may come from previously un-annotated transcripts. After applying a more stringent criterion to eliminate possible noises, we still get hundreds of putative novel transcripts. They are far from any known genes on the genome, have significant expression levels, and some are differentially expressed between different tissues. They may be novel protein-coding or non-coding genes that have not been reported, and may have important biological functions.

Identifying putative novel transcripts from RNA-Seq data is only a first step for the discovery of new genes or non-coding functional transcripts. The current protocol is still quite experimental and there are many factors to be refined. For example, noise model in RNA-Seq is necessary to filter possible noises in the sequencing experiment or due to imperfect sample preparation. The genomic sequence features as well as epigenomic features of the potential novel transcript regions also need to be investigated to search for hints of transcription. Comparative genomics analysis may be needed to study the conservation of the potential novel transcript regions to better identify putative novel transcripts. The study on possible motifs in the genomic sequences of and flanking the novel transcripts as well as the possible correlation of expression of novel transcripts with that

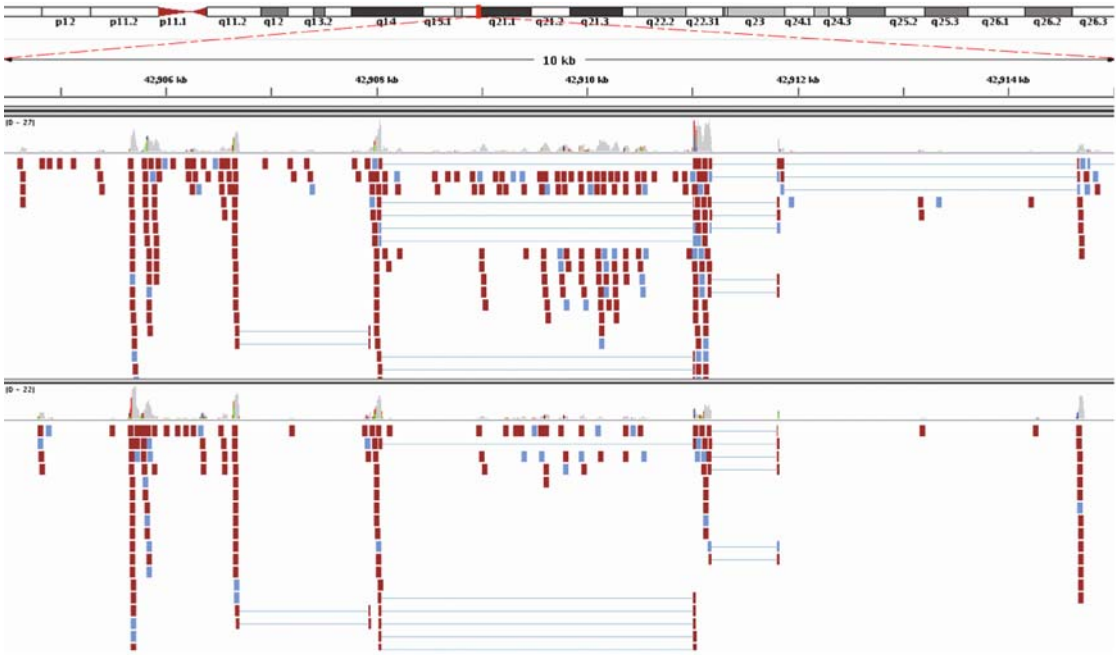


Fig. 6 An example region (on chromosome 15: 42904486–42915097) where putative novel transcripts are found in both datasets (above: GM12878; below: K562)

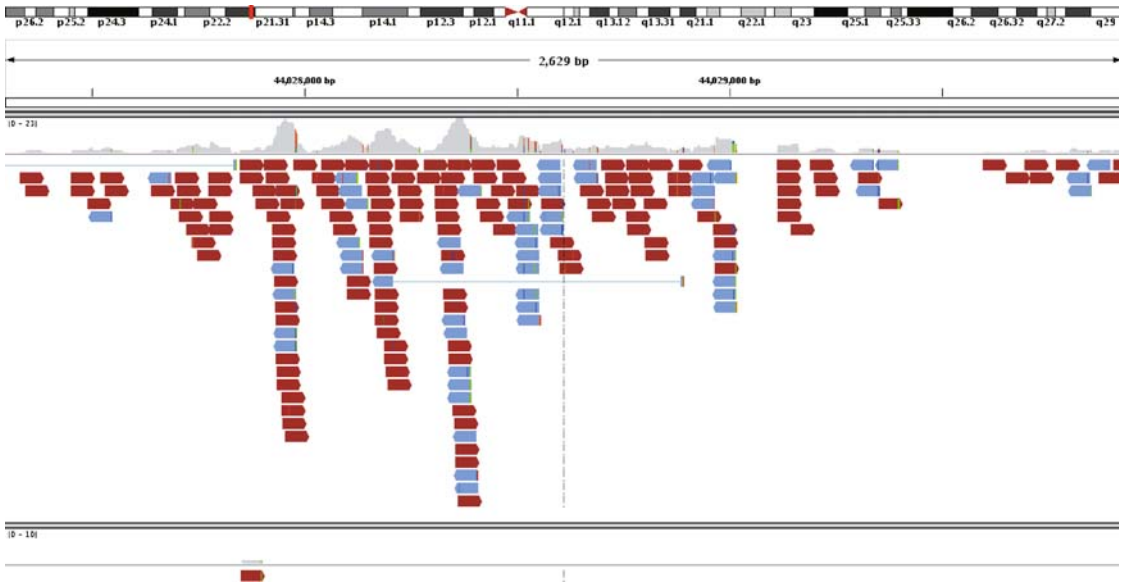


Fig. 7 An example region (on chromosome 3: 44027053–44029689) where putative novel transcript is found in dataset GM12878 (above) but not in K562 (below)

of known genes will also be important to understand the transcription mechanism of the novel transcripts. And finally biological experiments will be needed to validate the expression of the detected putative novel transcripts and to understand their biological functions.

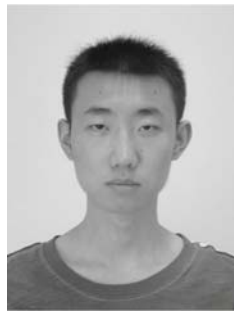
Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 61021063 and 60928007).

References

1. Mercer T R, Dinger M E, Mattick J S. Long non-coding RNAs: insights into functions. Nature Reviews Genetics, 2009, 10(3): 155–159
2. van Bakel H, Hughes T R. Establishing legitimacy and function in the new transcriptome. Briefings in Functional Genomics & Proteomics, 2009, 8(6): 424–436
3. Schena M, Shalon D, Davis R W, Brown P O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995, 270(5235): 467–470
4. Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology, 2008, 26(10): 1135–1145
5. Metzker M L. Sequencing technologies — the next generation. Nature Reviews Genetics, 2010, 11(1): 31–46
6. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 2009,

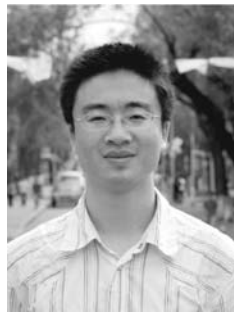
10(1): 57–63

7. Cock P J, Fields C J, Goto N, Heier M L, Rice P M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 2010, 38(6): 1767–1771
8. Mortazavi A, Williams B A, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 2008, 5(7): 621–628
9. Marioni J C, Mason C E, Mane S M, Stephens M, Gilad Y. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Research, 2008, 18(9): 1509–1517
10. Friedlaender M R, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. Discovering microRNAs from deep sequencing data using miRDeep. Nature Biotechnology, 2008, 26(4): 407–415
11. Pan Q, Shai O, Lee L J, Frey B J, Blencowe B J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics, 2008, 40(12): 1413–1415
12. Wang E T, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore S F, Schroth G P, Burge C B. Alternative isoform regulation in human tissue transcriptomes. Nature, 2008, 456(7221): 470–476
13. Jiang H, Wong W H. Statistical inferences for Isoform expression in RNA-Seq. Bioinformatics, 2009, 25(8): 1026–1032
14. Homer N, Merriman B, Nelson S F. BFAST: an alignment tool for large scale genome resequencing. PLoS One, 2009, 4(11): e7767
15. Langmead B, Trapnel C, Pop M, Salzberg S L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 2009, 10(3): R25
16. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, 2008, 18(11): 1851–1858
17. Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25(9): 1105–1111
18. Au K F, Jiang H, Lin L, Xing Y, Wong W H. Detection of splice junctions from paired-end RNA-Seq data by SpliceMap. Nucleic Acids Research, 2010, 38(14): 4570–4578
19. Wang K, Singh D, Zeng Z, Coleman S J, Huang Y, Savich G L, He X, Mieczkowski P, Grimm S A, Perou C M, MacLeod J N, Chiang D Y, Prins J F, Liu J. MapSplice: accurate mapping of RNA-Seq reads for splice junction discovery. Nucleic Acids Research, 2010, 38(18): e178
20. Trapnell C, Salzberg S L. How to map billions of short reads onto genomes. Nature Biotechnology, 2009, 27(5): 455–457
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/MAP format and SAMtools. Bioinformatics, 2009, 25(16): 2078–2079
22. Pruitt K D, Tatusova T, Maglott D R. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 2005, 33(suppl 1): D501–D504
23. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M. The Ensemble genome database project. Nucleic Acids Research, 2002, 30(1): 38–41
24. Harrow J, Denoeud F, Frankish A, Reymond A, Chen C K, Chrast J, Lagarde J, Gilbert J G R, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis S E, Guigo R. GENCODE: producing a reference annotation for ENCODE. Genome Biology, 2006, 7(Suppl 1): S4.1–S4.9
25. Wang L K, Feng Z X, Wang X, Wang X W, Zhang X G. DEGseq: an R package for identifying differentially expressed genes from RNA-Seq data. Bioinformatics, 2010, 26 (1): 136–138
26. Kent W J, Sugnet C W, Furey T S, Roskin K M, Pringle T H, Zahler A M, Haussler D. The human genome browser at UCSC. Genome Research, 2002, 12(6): 996–1006
27. Robinson J T, Thorvaldsdóttir H, Winckler W, Guttman M, Lander E S, Getz G, Mesirov J P. Integrative genomics viewer. Nature Biotechnology, 2011, 29(1): 24–26
28. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 1995, 57(1): 289–300



Chao YE received his B.E. degree of Measuring and Control Technology and Instrumentations in 2009 from Beijing University of Post and Telecommunications, Beijing, China. He is now a Ph.D candidate at MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University. His current research interests include gene regulation and RNA-seq data processing and analysis.

Linxi LIU received her B.S. degree of Mathematics in 2010 from Tsinghua University, Beijing, China, and now is a Ph.D student at Stanford University. She finished her undergraduate thesis at the Bioinformatics Division of TNLIST.



Xi WANG received his B.E. degree in Automation in 2005 from Harbin Institute of Technology, Harbin, China. He is now a Ph.D candidate in Bioinformatics at MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of

Automation, Tsinghua University. His research interests include machine learning, data mining for bioinformatics, DNA sequence analysis, and ChIP-seq/RNA-seq data processing and analysis.

Xuegong ZHANG received his B.E. degree and Ph.D degree in Tsinghua University. He is now a Professor at the Department of Automation, Tsinghua University, and the Director of the Bioinformatics Division, TNLIST. His research interest is in pattern recognition and bioinformatics.