Journal of Biosciences and Medicines, 2013, 1, 1-5 **JBM**
doi:10.4236/jbm.2013.11001 Published Online February 2013 (http://www.SciRP.org/journal/jbm/)

Isoform Inference From RNA-Seq Samples Based on Gene Structures on Chromosomes

Yan Ji, Jia Wei

R&D Information, AstraZeneca, Shanghai, China
Email: yan.ji@astrazeneca.com, jenny.wei@astrazeneca.com

Received 2013.

ABSTRACT

The emerging RNA-Seq technology makes it possible to infer splicing variants from millions of short sequence reads. Here we present a method to identify isoforms by their specific signatures on chromosomes including both exons and junctions. By applying this method to a RNA-Seq dataset of gastric cancer, we showed that our method is more accurate and sensitive than other isoform inference tools such as RSEM and Cufflinks. By constructing a network from gene list identified by our method but missed by other tools, we found that some cancer-related genes enriched in network modules have significant implications for cancer drug discovery.

Keywords: RNA-Seq; isoform inference; cancer genes; Cufflinks; RSEM

1. INTRODUCTION

Next–generation sequencing (NGS) platforms have been widely available recently [1]. A massively parallel sequencing technology termed RNA-Seq has made it possible to sequence cDNA derived from cellular RNA [2]. Recently, many studies have applied RNA-Seq to various biological and medical studies. Quantification of alternative slicing in tissues [3] and new transcript identification [4] have all benefited from this new technology. To fully enable RNA-Seq technology to solve biological problems, powerful computational tools are required. Since identifying disease driven differentially expressed isoforms is extremely important for drug research and development, we focus on inferring isoforms from RNA-Seq datasets.

In this work, we present an accurate and sensitive method to identify isoforms by their specific signatures. Expression values of an exon or junction can be calculated by counting the number of reads of a RNA-Seq dataset falling within an exon or spanning a junction if both the genome sequence and gene structure annotation are available. For each isoform, there must be single exon (junction) or a combination of exons (junctions) which uniquely belong to the isoform. We term such exons (junctions) as specific exons (junctions) of an isoform. Comparing with other isoform inference tools such as RSEM and Cufflinks, our method has two significant advantages [5, 6]. First, for all the isoforms in the gene structure annotations, our method can determine whether or not they are expressed in a RNA-Seq dataset. In the result section, we show that our method has the capability to identify isoforms which are missed by both RSEM and Cufflinks. Second, because some exons or junctions of an isoform may be overlapped by other isoforms, the estimation of the expression value of an isoform is always affected by expression levels of other isoforms. Therefore, instead of using expression values of an isoform, using expression values of specific exons or junctions of an isoform calculated in our method are better choice for finding differently expressed isoforms by comparing samples from patient and healthy donors.

2. METHODOLOGIES AND RESULTS

2.1. The Description of Our method for Inferring Isoforms from RNA-Seq Datasets

First, RPKM expression values of both exons and junctions are calculated. In this step, two files are needed including a BAM file of the mapping result against reference genome and an annotation file in GFF format containing the structure information of genes on chromosomes describing the relationship between exons and isoforms. The RPKM expression value of an exon is calculated by the read counts mapped to a chromosome region of the exon normalized by the total read counts in a sample and the length of the exon. The RPKM expression value of a junction is calculated in the same way except normalization by an extension length spanning the junction position. Second, a set of exons or junctions specific for an isoform are used to determine whether an isoform exists in a sample or not. The rationale to determine the expression status of an isoform in a sample

is described as following. If more than sixty percent of exons in an isoform are not expressed or with very small expression values, the expression status of the isoform is assigned a term “no expression”. Then, if either the set of specific exons or the set of specific junctions of an isoform are all expressed in a sample, the expression status of the isoform is assigned a term “existence”. Then, if one or more of specific exons or junctions are not expressed and they are located in the 5’ end of the isoform, the expression status of the isoform is assigned a term “uncertainty”. It is because that in the process of RNA- Seq sample preparation, a small fraction of sequences at the 5’ end of an isoform may be degraded or lost due to various reasons. Otherwise the expression status of the isoform is assigned a term “absence”.

2.2. Datasets

We apply this method to a Gastric cancer sample obtained from gastric cancer patient. mRNA was fragmented and plus- and minus- strand cDNA were synthesized for illumina pair-end sequencing. A 300-bp fragment size was selected by gel excision and the sample was sequenced twice to avoid technical variance. There are totally 30,121,416 and 17,510,256 read pairs for each replica. We use TopHat with reference genome hg19 and UCSC gene structure annotation to align the two RNA-Seq datasets [7].

3. Results

In the UCSC gene structure annotation, there are 22920 genes. Our method is designed to determine expression status of isoforms of all the genes having at least two isoforms. That is, 18967 isoforms of 7209 genes can be processed by our method. Because we only processed genes mapped by RNA-Seq reads, our results on the two RNA-Seq datasets, as shown in **Figure 1**, determined expression status of 16151 isoforms of 5891 genes.

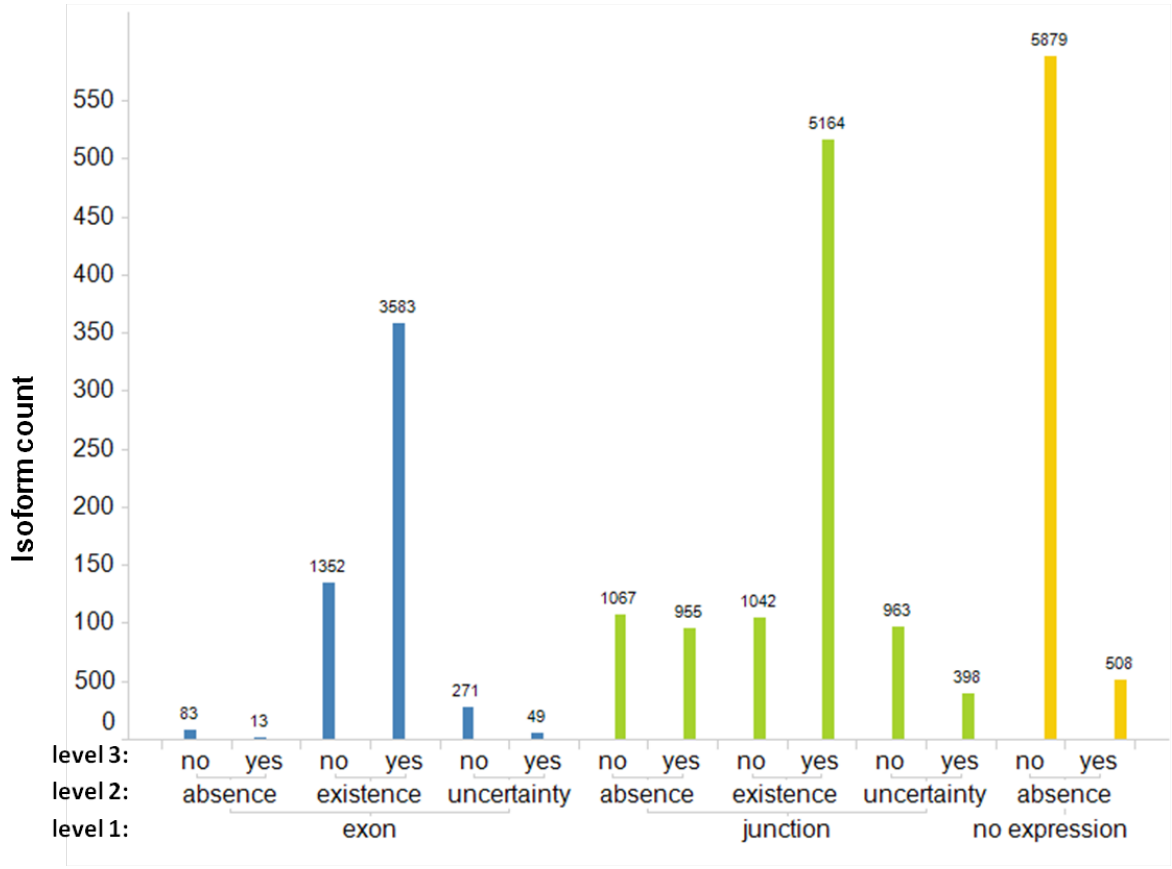

Figure 1. Comparisons of expression status of isoform between our method and RSEM. Level 1 is the evidence used to infer isoform by our method: exon, junction and no expression. Level 2 is the type of isoform expression status inferred by our method: existence, uncertainty and absence. Level 3 is the expression level of an isoform inferred from their expression values calculated by RSEM: isoforms marked with ‘yes’ are taken as expressed isoforms if their expression values are bigger than 0.3; isoforms marked with ‘no’ are taken to be absent in a RNA-seq dataset if their expression values are no bigger than 0.3.

Copyright © 2013 SciRes.

3.1. Comparisons with Cufflinks and RSEM

The method based on specific exons or junctions has its advantages over current popular isoform quantification tools, such as RSEM and Cufflinks. In **Figure 1**, we showed comparisons of expression status of isoforms inferred by our method and the two other tools. First, the two isoform expression quantification tools quantified isoforms of some genes marked as 'no expression' by our method with big expression values. ANO7 and TBXAS1 are two examples. As shown in **Figure 2**, most of exons of the two genes are not expressed. We guess that the two tools quantify expressions of isoforms by total read counts even if only a small fraction of exons such as UTR exons were expressed. So it is necessary for our method to check expression profiles of exons before isoform quantification. Second, some isoforms are marked as "existence" by our method, but they were quantified by the two tools with very small expression values. For example, as shown in **Figure 3**, while RSEM and cufflinks quantified the isoform NM_001135685 with 3.93E-05 and 0, there were 10 or 15 reads spanning its specific junctions with RPKM expression values 1.6 or 2.4. For NM_000548, its unique specific exon has expression value 15.03, while RSEM and cufflinks quantified it as 5.05E-23 and 0. Third, some isoforms are marked as "absence" by our method, but they were quantified by the two tools with big expression values. For example, as shown in **Figure 4**, while RSEM and Cufflinks quantified the isoform NM_001135730 with big expression values 5.66 and 4.66, there are no reads spanning its specific junction.

3.2. Biological Significance of Identified Isoforms

Our method marked 1860 isoforms with 'existence' while they were determined to be absent according to their expression values quantified by RSEM. Expression values of these isoforms are smaller than 0.3 (expression values of 62% of these isoforms were zero). Actually, by comparing expression values of isoforms and those of their specific exons or junctions, we found that expression values of most of these isoforms were under-estimated. There are 468 cancer-related isoforms out of 1860 isoforms. By using Analyze network algorithm with default settings in GeneGo (http://www.genego.com/), biological networks were constructed from cancer-related isoforms. In **Table 1**, some networks with annotated GO Processes were shown. In the first network, NM_003376 is the isoform b of VEGFA (vascular endothelial growth factor A) which plays a major role in vascular cell migration and proliferation. It mediates the vascular permeability, mitogenesis and angiogenesis. Inhibition of VEGF-A

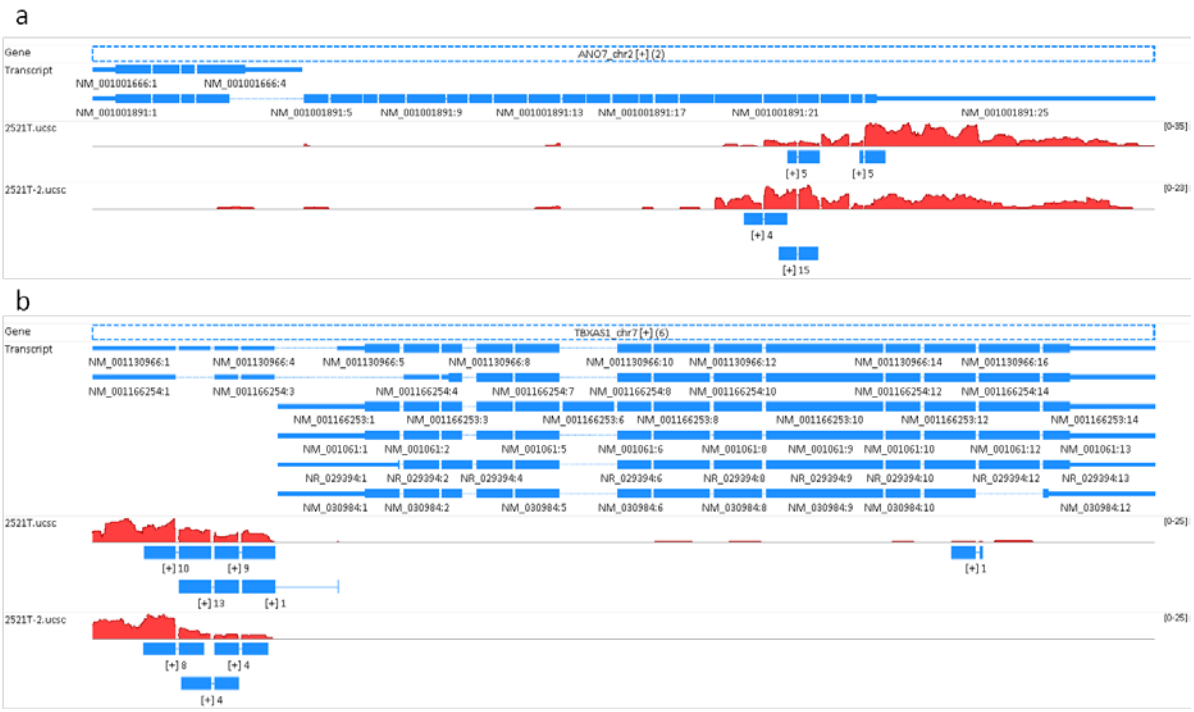

Figure 2. Visualization of the coverage of genes mapped by reads of two RNA-Seq datasets. Isoforms of the two genes (ANO7 and TBXAS1) are marked with 'no expression' by our method while they are quantified with big expression values by RSEM and Cufflinks.

Copyright © 2013 SciRes.

Figure 3. Visualization of the coverage of genes mapped by reads of two RNA-Seq datasets. As shown in 'a', NM_001135685 has two specific junctions highlighted by arrows. As shown in 'b', NM_000548 has a specific exon. The two isoforms are marked with 'existence' by our method while they are quantified with very small or zero expression values by RSEM and Cufflinks.

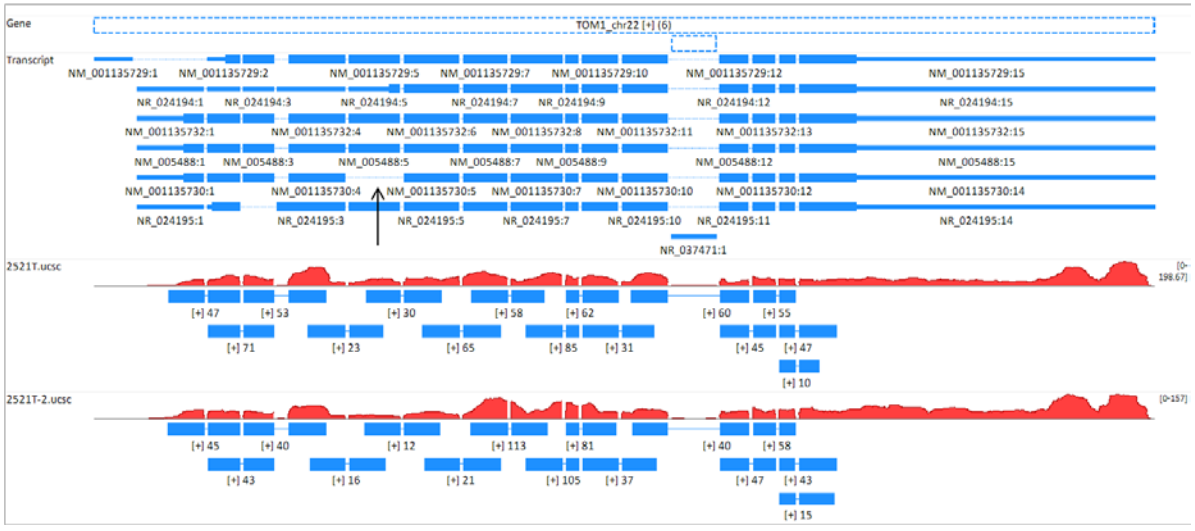

Figure 4. Visualization of the coverage of genes mapped by reads of two RNA-Seq datasets. NM_001135730 has a specific junction highlighted by an arrow. It is marked with 'absence' by our method while it is quantified with big expression values by both RSEM and Cufflinks.

promoter may be a useful approach for treating diseases in which aberrant angiogensis is present such as inflammatory diseases, cancer and diabetic retinopathy among others [8]. In the second network, NM_021872 is the smallest isoform of CDC25B (M-phase inducer phosphatase 2) which plays a role in endogenous tyrosine

Copyright © 2013 SciRes.

Table 1. Networks constructed from cancer related isoforms

Index	Key Network Objects	GO Processes
1	ACACA, eIF4G1, VEGF-A, MNK1, ABCC1	integrin-mediated signaling pathway, cell-matrix adhesion, protein transport within lipid bilayer, calcium-independent cell-matrix adhesion, cell migration involved in sprouting angiogenesis
2	CDC25B, VAV-2, alpha-1/beta-1 integrin, ZO-2	positive regulation of cell proliferation
3	Dynamin-2, TFIID, MALT1, PDK (PDPK1), Alpha-synuclein	protein phosphorylation, cell-cell signaling, synaptic transmission, phosphorylation, signal transduction
4	MAP3K3, FANCA, KLF11 (TIEG2), AP-1, WASP	adenylate cyclase-modulating G-protein coupled receptor signaling pathway

phosphatase activity and G2/M phase transition of cell cycle. Blocking CDC25B expression leads to inhibition of cell proliferation, migration and invasion. CDC25B may be a potential marker and therapeutic target for hepatocellular carcinoma and pancreatic cancer [9]. In the third network, NM_002613 is the longer isoform of PDPK1 (3-phosphoinositide-dependent protein kinase 1) which is involved in the regulation of cell proliferation, cell survival and signal transduction. PDK1-expressing MCF-7 cells showed up-regulation of genes of the Wnt signaling pathway and down-regulation of putative tumor suppressor genes. In the fourth network, NM_203351 is the longest isoform of MAP3K3 (mitogen-activated protein kinase kinase kinase 3) which is involved in regulating interleukin 1 receptor (IL-1R), toll-like receptor 4 (TLR4), SAPK and ERK signaling pathways. MEKK3 may be a therapeutic target in controlling the apoptosis resistance of some cancers [10].

Therefore, accurately inferring isoforms from RNA-Seq datasets of tumor samples and identifying tumor driven isoforms have important implications in the field of cancer drug research and development.

4. CONCLUSIONS

We have developed a method for inferring isoforms from RNA-Seq samples. Its advantages over current isoform inference tools have been illustrated in previous sections. Its limitations are that complete sequences and gene structure annotation of transcriptome of targeted species have to be available. With advances of sequencing technologies and genome biology, our method can be applied to more species. Furthermore, our method can detect isoform switching if multiple disease and normal samples are available. If expression status of isoforms are different between disease and normal samples, isoform switching can be identified. If expression status of isoforms are both marked with ‘existence’, expression of specific exons or junctions between samples can be used to determine whether isoforms are differentially expressed.

5. ACKNOWLEDGEMENTS

This work was supported in part by a project from AstraZeneca.

REFERENCES

[1] Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1), 31-46.

[2] Mortazavi, A. Williams, B. A. McCue, K. Schaeffer, L. Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods, 2008. 5(7): p. 621-8.

[3] Wang, E.T., et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), 470-6.

[4] Guttman, M., et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*, 28(5): p. 503-10.

[5] Li, B. and C.N. Dewey (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12: p. 323.

[6] Trapnell, C., et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5): p. 511-5.

[7] Trapnell, C., L. Pachter, and S.L. Salzberg (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9): p. 1105-11.

[8] Snowden, A.W., et al. (2003) Repression of vascular endothelial growth factor A in glioblastoma cells using engineered zinc finger transcription factors. *Cancer Res*, 63(24): p. 8968-76.

[9] Ngan, E.S., et al. (2003) Overexpression of Cdc25B, an androgen receptor coactivator, in prostate cancer. *Oncogene*, 22(5): p. 734-9.

[10] Samanta, A.K., et al.(2004) Overexpression of MEKK3 confers resistance to apoptosis through activation of NFkappaB. *J Biol Chem*, 279(9): p. 7576-83.

Copyright © 2013 SciRes.