RNADAP—RNA-Seq Data Annotation Pipeline

Zunming Liu[1, 2†], Jingfa Xiao[1†], Jiayan Wu[1,*], Jun Yu[1, *]
[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics,
Chinese Academy of Sciences, Beijing, China
[2] Graduate University of Chinese Academy of Sciences, Beijing, China
[†] The first two authors contributed equally to this work.
[*]Correspondence to: Jiayan Wu E-mail: wujy@big.ac.cn, Jun Yu E-mail:junyu@big.ac.cn

***Abstract*--RNA-Seq has become one of the most important new approaches for gene expression analysis as well as transcriptome analysis. The issue of how to analysis RNA-Seq data is one of the biggest challenges for current transcriptomics research. In this study, we develop an RNA-Seq data annotation pipeline named RNADAP, which is an efficient transcriptomes analysis tool to evaluate gene expression quantization in isoform level and compatible for reads data from different platforms. RNADAP is a typical Java application so the pipeline could be carried out on Windows as well as Linux. The installation process is convenient and user can grasp it very easily with a friendly user interface.**

RNADAP is a free, open-source software and written in Java. All source code, instructions, testing data and additional scripts are available at http://rnadap.sourceforge.net/.

1. INTRODUCTIONS

With the development of high-throughput next-generation DNA sequencing (NGS) technologies, RNA sequencing (RNA-Seq) has become one of the most important new approaches for both gene expression analysis and transcriptomes analysis [1]. The main commercial available NGS platforms are Illumina/Solexa Genomic Analyzer, Roche/454 Life Sciences GS FLX and ABI/Agencourt SOLiD System [2]. The work flow for these sequencers includes: fragment library preparation, immobilization of fragment, clonal amplification, parallel sequence by synthesis, and sequence read assembly [3]. Due to the sequence throughput and cost, the most universal used platforms in transcirptome and mircoRNA sequencing are Illumina/Solexa and ABI/SOLiD. Roche/454 is usually selected in de novo sequencing because it produces longer reads length.

NGS technology is a large-scale and high throughput technology, which consumes less money and time [1]. On the other side of the coin, short reads length causes more bioinformatics barriers [3]. When we obtained RNA-Seq data from NGS platforms, basic analysis follows two steps: map RNA-Seq reads to reference genome, which is called mapping; count reads number and density correspond to each known exon, splice event or new candidate gene, which is called expression quantization. Challenges for analyzing RNA-Seq data include reads mapping uncertainty, sequencing non-uniformity, novel isoform (alternatively spliced transcript [4]) expression level estimating, efficient storage and sequence alignment [5].

More challenges occur in the second step of RNA-Seq data analysis. There are several annotation tools published, such as ERANGE [6], RSEM [5] and rSeq [7]. Most of the tools work well, however, those pipelines still have some limitations. First, some software, such as ERANGE [6], count reads in gene but not isoform level, which gives more accurate estimation [8] and less uncertainty, especially detects splice events and carries out alternative splicing investigation [9]. Second, several tools cost larger storage or more time, and some others cost much more time in installation and configuration. Last but not the least, some pipelines are not compatible to sequence reads from all platforms or the results from other mapping software.

Here we develop a new software-RNADAP (RNA-Seq Data Annotation Pipeline). RNADAP counts read number in isoform level, works with high speed and less internal storage usage. Most important, our pipeline can be compatible with results from different mapping software and deal with reads data from all NGS platforms. RNADAP is a free, open-source software and written in Java. All source code, instructions, testing data and additional scripts are available at http://rnadap.sourceforge.net/.

2. RESULTS & DISCUSSION

NGS platforms generally produce tens of millions of reads data, how to process them quickly with less internal memory usage is a big challenge for us. In our pipeline, mapped reads data will be preprocessed and presorted, which help us perfectly weigh time cost and space consuming as well as fulfill accuracy requirement. To be compatible with all kinds of mapping results from different software and platforms, we predefine the standard input data format and develop a data convert tool to transfer various mapping results to our standard format.

When reads are mapped to the gene are shared by more than one isoform, we should find a way to estimate gene expression in isoform level. Isoform-based estimation has less uncertainty than gene level estimation. In our pipeline, we use Poisson model for isoform quantification, which was introduced by Jiang and Wang [7]. All the parameters of different isoform expression with a gene, which stand for isoform-specific RPKM [6], are estimated by a maximum likelihood approach [7]: initial parameters defined by least square method, official parameters solved by hill climbing algorithm and the final parameters modified by essential important sampling from the posterior distribution.

Finally, we design man-machine interface with C/S structure (Figure 1), so we can handle the pipeline easily and friendly. The software is written in Java which can be used on Windows or Linux. Compared with other RNA-Seq analysis

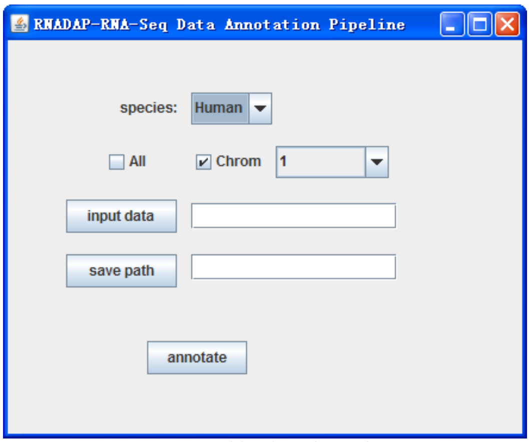
978-1-4577-1666-9/11/$26.00 ©2011 IEEE

pipeline, RNADAP keeps users away from Linux bugs and commands.

Table 1: Validate results comparison between RNADAP and rSeq

Species		Mouse				Human	
Tissue		Testis	Brain	Liver	Stem	Testis	Brain
Platform		SOLiD	Solexa	Solexa	SOLiD	Solexa	Solexa
Case_Id		1	2	3	4	5	6
RNADAP	FPR(%)	5.94	5.35	7.25	7.73	18.62	10.58
	FNR(%)	9.02	6.77	10.47	16.99	5.36	6.5
rSeq	FPR(%)	1.42	3.28	2.35	2.01	3.56	2.58
	FNR(%)	23.16	31.5	22.01	34.26	14.38	15.25

*The false positive rate (FPR) is the proportion of isoforms has no expression in fact but the pipeline detect the isoform have expression in total isoforms. The false negative rate (FNR) is the proportion of isoforms has expression in fact but the pipeline detect the isoform have no expression in total isoforms.


RNADAP-RNA-Seq Data Annotation Pipeline

species: Human

☐ All ☑ Chrom 1

Figure 1 man-machine interface of RNADAP

We compare RNADAP's results with rSeq's for six cases (four from mouse and others from human). Tissue type of these cases includes testis, brain, liver and stem. The real RNA-seq datasets are come from Marioni et al [6], SRA database of NCBI (SRA008403) and our laboratory [10]. All these datasets produced by Solexa or SOLiD. We perform simulations [5] to validate these two methods. False positive rate (FPR) and false negative rate (FNR) in six validated results are summarized in Table 1.

As to FNR, rSeq's are always higher than RNADAP's. For example in case 1 (mouse, testis, SOLiD), FNR value with RNADAP is 9.02%, almost one third of FNR value with rSeq, which is 23.16%. On the other hand, rSeq does better than our method in FPR. Still take case 1 for an instance, rSeq's FPR value is 1.42% but RNADAP's FPR value is 5.94%. Although both methods are based on Poisson model, the concrete realization is in different ways. Thus the results of these two pipelines have their own characteristics, just as coin's two sides. From what we discussed above, we can conclude that the solution about annotation in our pipeline is correct and useable.

To access the performance, we conduct various comparisons among RNADAP, rSeq and RSEM. The test dataset is simulated by RSEM and the reference reads dataset is publicly available from SRA database of NCBI (SRA008403). The comparison results are shown in Table 2.

Compared with other software, RNADAP can be compatible with mapping results from different software and platforms, especially SAM format, which is widely considered as the de facto standard for storing and transferring short read alignment results [11]. For example, RNADAP can deal with reads data from Corona, MAQ, Eland and SAM formats, but the other two pipelines shown in Table 2 can only dispose their own mapping results. In addition, our pipeline is developed in Java language, so we can run it on Windows or Linux; however, the others can only run on Linux system. RNADAP has a friendly user interface, which makes it easier to grasp.

Table 2: Software performance comparison among RNADAP, rSeq and RSEM

Compare item	RNADAP	rSeq	RSEM
Time cost	21 mins*	22 mins	303 mins
Memory cost	10 GB*	10 GB	2.5 GB
Format compatible			
CORONA	YES	NO	NO
MAQ	YES	NO	NO
ELAND	YES	YES	NO
SAM	YES	NO	NO
Operating systems	Windows/Linux	Linux	Linux
Use interface	Graphical	Command	Command

*In order to be fair in comparison, we run the mapping process in rSeq, and use it as the input data for RNADAP. The cost time and memory listed here have been added the mapping process part running with rSeq. As for RNADAP's annotation part alone, time cost is 4 mins and memory usage less than 0.5 GB

3. CONCLUSIONS

We develop a tool for RNA-Seq data annotation named RNADAP. This pipeline can evaluate gene expression in isoform level and balance time consuming and internal storage cost as well. Most important, our pipeline can be compatible with mapping results from different software and all kinds of NGS platforms. RNADAP is written in Java, so the portability is very well.

ACKNOWLEDGMENT

This study was supported by grant from the National Basic Research Program (973 Program) (No. 2010CB126604), grant from the Special Foundation Work Program (No. 2009FY120100), grant from National Key

Technology R&D Program (No. 2008BA164B02), the Ministry of Science and Technology of the People's Republic of China; grant from the National Science Foundation of China (No. 31071163).

REFERENCES

[1] Wang Z, Gerstein M, Snyder M, RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63

[2] Marguerat S, Bähler J, RNA-seq: from technology to biology. Cell Mol Life Sci. 2010 Feb;67(4):569-79. Epub 2009 Oct 27

[3] Fatih Ozsolak & Patrice M. Milos, RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics 12, 87-98 (February 2011) | doi:10.1038/nrg2934

[4] Sika Zheng and Liang Chen, A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. Nucleic Acids Research, 2009, Vol. 37, No. 10

[5] Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson and Colin N. Dewey1, RNA-Seq gene expression estimation with read mapping.Uncertainty. December 18, 2009, Vol. 26 no. 4 2010, pages 493–500 doi:10.1093/bioinformatics/ btp692

[6] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer1 & Barbara Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. 30 May 2008, DOI:10.1038/ NMETH.1226

[7] Hui Jiang and Wing Hung Wong, Statistical Inferences for Isoform Expression in RNA-Seq. Bioinformatics, February 25, 2009

[8] [8] Brian E Howard, Steffen Heber, Towards reliable isoform quantification using RNA-SEQ data. IEEE International Conference on Bioinformatics and Biomedicine 2009 Washington, DC, USA. 1-4 November 2009

[9] Wang X, Wu Z, Zhang X, Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. J Bioinform Comput Biol. 2010 Dec;8 Suppl 1:177-92

[10] [10] Cui P, Lin Q, Xin C, Han L, An L, Wang Y, Hu Z, Ding F, Zhang L, Hu S, Hang H, Yu J. Hydroxyurea-induced global transcriptional suppression in mouse ES cells. Carcinogenesis. 2010 Sep;31(9): 1661-8. Epub 2010 May 31.

[11] 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 Aug 15;25(16):2078-9. Epub 2009 Jun.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE