REVIEW : VARIOUS TECHNIQUES OF RNA SEQUENCING

Sukhdeep Kaur[1], Ishdeep Singla[2]
[1]Research Scholar,[2]Assistant Professor
Deptartment of Computer Science and Engineering
Mohali, India.

Abstract: ***The ability to quantitatively survey the global behaviour of transcriptomes has been a key milestone in the field of systems biology, enabled by the advent of DNA microarrays. While this approach has literally transformed our vision and approach to cellular physiology, microarray technology has always been limited by the requirement to decide, a priori, what regions of the genome to examine. While very high density tiling arrays have reduced this limitation for simpler organisms, it remains an obstacle for larger, more complex, eukaryotic genomes. The recent development of "next-generation" massively parallel sequencing (MPS) technologies by companies such as Roche (454 GS FLX), Illumina (Genome Analyser II), and ABI (AB SOLiD) has completely transformed the way in which quantitative transcriptomics can be done. These new technologies have reduced both the cost-per-reaction and time required by orders of magnitude, making the use of sequencing a cost-effective option for many experimental approaches.***

Keywords: RNA, transcriptomes, transfrag, genome, RNA-seq, and base pairs.

I. INTRODUCTION

Deep transcriptome sequencing (RNA-Seq) with next generation sequencing (NGS) technologies is providing unprecedented opportunities for researchers to probe the transcriptomes of many species [1-5]. An important goal in these studies is to assess the extent of alternative splicing (AS), a process that increases transcriptome and proteome diversity, and plays a key role in regulating gene expression and protein function [6,7]. Although it is inexpensive and easy to obtain whole transcriptome data using RNA-Seq, one limitation has been the lack of versatile methods to analyze these data. Consequently, there is an increasing demand for methods that can use the short reads produced in these studies to predict patterns of AS. The sequences produced by NGS methods have characteristics that complicate the task of identifying the mRNA transcripts represented in a sample. A sequencing read may consist of fewer than 40 nucleotides, making it difficult to identify a unique origin within a reference sequence. In addition, NGS base-call error rates tend to increase with read length, raising the chance of a mismatch when aligning a read to a reference sequence [8]. These ambiguities are exacerbated by the presence of paralogous genes that can give rise to reads that align well in multiple locations. Much of the work on analyzing NGS reads has focused on aligning reads within exonic regions, and many methods exist for the problem of aligning reads without gaps-for example, MAQ [9], PASS [10] and BowTie [11]. Reads that span splice junctions introduce additional challenges. A splice junction may occur anywhere within a short read, so the read may have just a few bases on one side of a junction. Such a short sequence may align in multiple locations, making it difficult to identify its true origin. One can use heuristics to restrict the number of candidate locations: for example, by establishing limits for permissible intron lengths, or by focusing on locations that are bounded by canonical GT-AG or GCAG splice-site dimers. Several spliced alignment algorithms exist that use these and other approaches to identify unique alignments for spliced reads [12-17]. The first studies that used RNA-Seq data to predict AS focused on exon-skipping events, the most prevalent form of AS in mammals (see, for example, [1,18-21]). To identify splice junctions recapitulated in short read data, these studies used exon sequences flanking annotated splice sites to produce a database of splice junction sequences. Using novel combinations of known acceptor and donor sites, researchers can create a database that consists of both known and putative splice junction sequences. RNA-Seq reads that align to these putative sequences then provide evidence for novel splicing events.

II. SAMPLE PREPARATION

Given the variety of current technical approaches (many of which may be obsolete before this article is published), a precise step-by-step protocol would not be particularly practical for a methodology paper. Instead, this article will focus on the key elements of the procedure which are common to all technologies, and discuss the factors which should be considered when planning such experiments.

A. Amount Requirements

Because the RNA-seq approach is entirely based on the general principles of DNA sequencing, the methodology should be applicable to any organism, subject to the availability of a sufficient amount of RNA. It is worth noting that while published information on the performance of these technologies in high/low GC content genomes is scarce, anecdotally, they do not appear to show any significant bias across a fairly wide spread of GC content (30-70%),suggesting that RNA from most organisms would be suitable.

B. RNA Removal

One of the principal technical hurdles to overcome with RNA-seq is the fact that the vast majority of RNA (>90%) present in cells consists of ribosomal RNA (rRNA). As such, the bulk of the total RNA is not informative as to the true diversity of the transcriptome present in the remaining RNA. In order to avoid wasting effort in re-sequencing the same ribosomal RNA millions of times, several techniques exist to focus the sequencing effort on the non-ribosomal portion.

C. Priming

Following enrichment, the resulting mRNA must be primed for the reverse transcription reaction using either random primers oroligo dT primers. The advantage of using oligo dT (with or without prior mRNA enrichment) is that the majority of cDNA produced should be polyadenylated mRNA, hence more of the sequence obtained should be informative (non-ribosomal). The significant disadvantage of the use of oligo dT primers is that the reverse enzyme will fall off of the template at a characteristic rate, resulting in a bias towards the 30 end of transcripts. For long mRNAs, this bias can be pronounced, resulting in an under representation (or worse, absence) of the extreme 50 end of the transcript in the data.

D. Maintaining Strand Specific Information

An additional consideration in the process of creating the double-stranded cDNA for sequencing is to maintain strand specific information for the RNA. The importance of this consideration will obviously vary depending on the organism being studied, but in more complex genomes (such as mouse and human) where there is clear evidence for wide spread antisense transcription strand specific information should be considered a clear requisite for comprehensive RNA-seq studies.

III. METHODS

A. 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, paper uses 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], Splice Map [18], and Map Splice [19]. Paper chose TopHat (version 1.1.1) (http://tophat.cbcb.umd.edu/) in its 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
```

Figure 1: Mapped Result in SAM Format

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.)[28].

B. Strategy for Discovering Novel Transcripts

RNA-Seq reads are random samples from transcribed genomic regions. To identify the regions that are transcribed, method 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. In this paper, article 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. This method 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 [29] 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.

IV. COMPARISON BETWEEN PASTA AND OTHER TOOLS ON SIMULATION BASED

As a first test of the performance of PASTA, we compared its ability to detect known splice junctions against TopHat. We generated four simulated datasets of 50nt single-ended RNA-Seq reads from mouse transcripts appearing in ENSEMBL gene annotations, corresponding to average depths of coverage ranging from 1 to 8 reads per nucleotide, and we introduced random sequencing errors at a frequency of 1/1000 basepairs and Single Nucleotide Polymorphism (SNP) at a frequency of 5/1000 basepairs. The results show that PASTA consistently exhibits a lower false negative rate than TopHat, especially at low coverage level. Sensitivity is consistently higher than TopHat (on average, 20% to 40% higher), especially for transcripts expressed at a low level. PASTA is therefore well-suited for identifying "rare" splicing events, reducing the risk of missing splicing events critical for AS analysis[1]. This indicates that the use of PASTA may lead to a reduction in sequencing costs, for example by multiplexing more samples in the same run, since it is able to produce reliable results even at low sequencing depths.

V. TRANSCRIPT RECONSTRUCTION

Defining a precise map of all transcripts and isoforms that are expressed in a particular sample requires the assembly of these reads or read alignments into transcription units. Collectively, we refer to this process as transcriptome reconstruction. Transcriptome reconstruction is a difficult computational task for three main reasons. Several methods exist to reconstruct the transcriptome, and they fall into two main classes: 'genome-guided' and 'genomeindependent' (Fig. 2). Genome-guided methods rely on a reference genome to first map all the reads to the genome and then assemble overlapping reads into transcripts.

Genome-Guided Reconstruction

Existing genome-guided methods can be classified in two main categories: "exon identification" and "genome-guided assembly" approaches. Exon identification methods such as G.mor.se were developed early when reads were short (36 bases) and few aligned to exon-exon junctions. Genome-guided assembly methods such as Cufflinks and Scripture28 have been developed. These methods use spliced reads directly to reconstruct the transcriptome. Scripture initially transforms the genome into a graph topology, which represents all possible connections of bases in the transcriptome either when they occur consecutively or when they are connected by a spliced read[30].

Genome-Independent Reconstruction

Rather than mapping reads to a reference sequence first, genome-independent transcriptome reconstruction algorithms such as transAbyss use the reads to directly build consensus transcripts. Consensus transcripts can then be mapped to a genome or aligned to a gene or protein database for annotation purposes. The central challenge for genome-independent approaches is to partition reads into disjoint components, which represent all isoforms of a gene. A commonly used strategy is to first build a de Bruijn graph, which models overlapping subsequences, termed 'k-mers' (k consecutive nucleotides), rather than reads. This reduces the complexity associated with handling millions of reads to a fixed number of possible *k*-mers.

Criterion	Expression arrays	Tiling arrays	RNA-Seq
Resolution of data	N/A	Dependent on genome size but ⩾35 bp for human/mouse	1 bp, at sufficient sequencing depth
Cost per sample (excluding equipment)	Low	Low–high, depending on arrays needed to cover genome	High
Linear dynamic range of expression values	<4 orders of magnitude	<2 orders of magnitude	Limited only by sequencing depth and biological expression levels
Sensitivity (Signal:Noise)	Moderate	Low	High
Discovery of novel transcribed regions	No	Yes	Yes
Monitor splice site usage	No	Limited	Yes
Identification alternative promoters/ UTRs	No	Yes	Yes
Detection of antisense transcripts	Not standard	Not standard	Requires strand specific preparation
Detection of SNPs, mutations, allelic differences	Limited	Limited	Yes
Size of raw data files per experiment	0.01–0.05 Gb	0.1–1 Gb	1–15 Tb
Downstream Bioinformatic requirements	Low	High	Very high

Figure 2: Comparison of current methods for surveying transcriptome.

VI. CONCLUSION

As sequencing technologies mature, existing computational tools will need to evolve to meet new requirements, and new tools will emerge to enable new applications. In this paper we have reviewed some methods of rna-sequencing and their methods. We have also compare the methods for surveying transcriptones and also compare the PASTA tools with the other ones.

References

[1] Mortazavi A, Williams B, McCue K, Schaeffer L, and Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, pages 621–628, 2008.

[2] Wang Z, Gerstein M, and Snyder M. RNA-Seq : A revolutionary tool for transcriptomics. *Nat RevGenet*, pages 57–63, 2009.

[3] Filichkin S, Priest H, Givan S, Shen R, Bryant D, Fox S, Wong W, and Mockler T. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res*, page 45, 2010.

[4] Harr B, and Turner L. Genome: Wide analysis of alternative splicing evolution among Mus subspecies. *Mol Ecol*, pages 228–239, 2010.

[5] Ramani A, Calarco J, Pan Q, Mavandadi S, Wang Y, Nelson A, Lee L, Morris Q, Blencowe B, Zhen M, and Fraser A. Genome-wide analysis of alternative splicing in Caenorhabditis elegans. *Genome Res*, page 342, 2011.

[6] Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj T, and Soreq H. Function of alternative splicing. *Gene*, pages 1–20, 2005.

[7] Hallegger M, Llorian M, and Smith CWJ. Alternative splicing: Global insights. *FEBS J*, pages 856–866, 2010.

[8] Shendure J, and Ji H. Next-generation DNA sequencing. *Nat Biotechnol*, pages 1135–1145, 2008.

[9] Li H, Ruan J, and Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, page 1851, 2008.

[10] Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, Manavski S, Vitulo N, and Valle G. PASS : A program to align short sequences. *Bioinformatics*, page 967, 2009.

[11] Langmead B, Trapnell C, Pop M, and Salzberg S. Ultrafast and memory-efficientalignment of short DNA sequences to the human genome. *Genome Biol*, page 25, 2009.

[12] Homer N, Merriman B, and Nelson S F. BFAST : An alignment tool for large scale genome resequencing. *PLoS One*, page 7767, 2009.

[13] Langmead B, Trapnel C, Pop M, and Salzberg S L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, page 25, 2009.

[14] Li H, Ruan J, and Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, pages 1851–1858, 2008.

[15] Trapnell C, Pachter L, and Salzberg S L. TopHat : Discovering splice junctions with RNA-Seq. Bioinformatics. pages 1105–1111, 2009.

[16] Au K F, Jiang H, Lin L, Xing Y, and Wong W H. Detection of splice junctions from paired-end RNA-Seq data by SpliceMap. Nucleic Acids Research. pages 4570–4578, 2010.

[17] 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, and Liu J. MapSplice : Accurate mapping of RNA-Seq reads for splice junction discovery. *Nucleic Acids Research*, page 178, 2010.

[18] Trapnell C, and Salzberg S L. How to map billions of short reads onto genomes. *Nature Biotechnology*, pages 455–457, 2009.

[19] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/MAP format and SAMtools. *Bioinformatics*, pages 2078–2079, 2009.

[20] Pruitt K D, Tatusova T, and Maglott D R. NCBI reference sequence (RefSeq) : A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, pages 501–504, 2005.

[21] 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, and Clamp M. The Ensemble genome database project. *Nucleic Acids Research*, pages 38–41, 2002.

[22] 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, and Guigo R. GENCODE : Producing a reference annotation for ENCODE. *Genome Biology*, 2006.

[23] Wang L K, Feng Z X, Wang X, Wang X W, and Zhang X G. DEGseq : An R package for identifying differentially expressed genes from RNA-Seq data. *Bioinformatics*, pages 136–138, 2010.

[24] Kent W J, Sugnet C W, Furey T S, Roskin K M, Pringle T H, Zahler A M, and Haussler D. The human genome browser at UCSC. *Genome Research*, pages 996–1006, 2002.

[25] Robinson J T, Thorvaldsd'ottir H, Winckler W, Guttman M, Lander E S, Getz G, and Mesirov J P. Integrative genomics viewer. *Nature Biotechnology,*, pages 24–26, 2011.

[26] Chao YE, Linxi LIU, Xi WANG, and Xuegong ZHANG. Observations on potential novel transcripts from RNA-Seq data.

[27] Cock P J, Fields C J, Goto N, Heier M L, and Rice P M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, pages 1767–1771, 2010.

[28] Manuel Garber, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq.

[29] Brian T., Wilhelma B., and Josette-Renee Landry. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing.

