Summer 8-4-2011

Computational Pipeline for Human Transcriptome Quantification Using RNA-seq Data

Guorong Xu
University of New Orleans, gxu2@uno.edu

Computational Pipeline for Human Transcriptome Quantification using RNA-seq data

A Thesis

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Master of Science
in
Computer Science
Bioinformatics

by

Guorong Xu

August, 2011

Acknowledgments

First and foremost, I want to thank my advisor Dr. Dongxiao Zhu. His endless support and wisdom helped me to finish this thesis. His enthusiasm for Bioinformatics was contagious—and I definitely caught it. His depth of knowledge and very precise academic guidance brought me to develop a computational pipeline for characterizing human transcriptome using RNA-seq.

I would like to thank Dr. Erik Flemington. The major experiment data and biology knowledge for this study were provided by his lab. I really appreciate his consistent support.

I wish to thank our bioinformatics group members: Nan Deng, Lipi Rani Acharya, Thair Judeh, Tin Chi Nguyen, Kristen Marie Johnson. Each individual provided insights that guided and challenged my thinking, substantially improving the finished product.

Lastly, I would like to thank my family members, especially my wife Yan Gao for supporting and encouraging me to pursue this degree. Without my wife's encouragement, I would not have finished the degree.

Table of Contents

List of Figures ... vi
List of Tables ... vii
Abstract ... viii
Chapter 1. Background and Introduction ... 1
1.1 Introduction to transcriptome ... 1
1.2 Gene expression ... 1
1.3 Microarray technology ... 2
1.4 Next-generation sequencing ... 2
1.4.1 RNA-seq technology ... 3
1.4.2 Short reads alignment ... 4
1.4.3 Junction mapping ... 5
1.4.4 Data Format – SAM/BAM ... 6
1.4.5 Quantification of gene expression using RNA-seq ... 8
1.5 Motivation ... 9
1.6 Overview ... 9
Chapter 2. Transcriptome and Targetome Analysis in MIR155 Expressing Cells using RNA-seq ... 11
2.1 Introduction ... 11
2.2 Results ... 14
2.2.1 Model system ... 14
2.2.2 Read mapping ... 15
2.2.2 Expression analysis ... 16
2.2.3 Relative expression and distribution of genes with 3’ UTR MIR155 seed sequences ... 17
2.2.4 NGS *versus* microarray studies ... 19
2.2.5 3’ UTR reporter analysis ... 21
2.2.5 Validation of RNA-seq analysis using 3’UTR luciferase reporter assay ... 23
2.2.6 Global analysis of terminal exon relative expression ... 25
2.3 Discussion ... 27
2.4 Materials and methods ... 32
2.4.1 Biological experiments ... 32
2.4.2 Sequencing and base calling ... 32
2.4.3 Read mapping to genome and across splice sites ... 33
2.4.4 Gene expression analysis using RNA-seq data ... 34
2.4.5 Analysis of Agilent MIR155 arrays ... 34
2.4.6 Statistical treatment of sequencing and microarray data ... 35
2.4.7 Cross-species gene mapping ... 36
2.4.8 Cross-platform comparison of targetome prediction using bitmap ... 36
2.4.9 Statistical comparison of a pair of populations ... 36
2.4.10 3’ UTR-luciferase reporter analysis ... 37
2.4.11 Accession numbers ... 37
2.5 Supplemental Figures ... 38
Chapter 3. RNA-seq Analysis of EBV Transcriptome ... 40

3.1 Background.. 40
3.2 Analysis of EBV gene expression in Mutu I and Akata cells............................ 41
3.3 Splicing evidence in Mutu I and Akata.. 45
3.4 Conclusions.. 47
Chapter 4. SAMMate GUI software system.. 49
4.1 Introduction.. 49
4.2 Key features 50
4.2.1 Key feature: calculating genomic feature abundance scores 50
4.2.2 Key feature: generating signal map for peak detection 53
4.2.3 Key feature: generating wiggle files for visualization..................................... 53
4.2.4 Key feature: generating alignment report 55
4.3 Conclusion 55
Chapter 5. Conclusion and Future Works.. 56
5.1 Conclusion 56
5.2 Future works 57
References.. 59
Vita.. 66

List of Figures

Figure 1.1. Combination of exon reads with junction reads to accurately calculate gene expression RPKM scores 5
Figure 1.2. Example of single-end reads in SAM format. 7
Table 1. 1 Overview of 11 mandatory fields in alignment line 7
Table 1. 2 Overview of optional fields in alignment line 8
Figure 2.1 The biological system 15
Figure 2.2. Genes containing MIR155 seed sequences are enriched in down-regulated fractions. 19
Figure 2.3. Cross-platform comparison of transcriptome quantification using bitmap. ... 21
Figure 2.4. Validation of RNA-seq analysis using 3' UTR assay. 22
Figure 2.5. Isoform-dependent regulation of TAF5L by MIR155. 28
Figure 2.6. Mitigated response of endogenous genes to MIR155 is mediated in part by alternate transcript structure 29
Supplemental Figure 2.1. Lack of read evidence for MIR155 seed containing isoform of PCDH9. 38
Supplemental Figure 2.2. Read evidence for the last exon of MAP3K10 shows a sharp drop in mapped reads upstream from the MIR155 seed site. 39
Supplemental Figure 2.3. Top biological functions implicated by Ingenuity Pathways Assessment (IPA) analysis of regulated genes 39
Figure 3.1. Visualization of RNA-seq coverage across the EBV genome 42
Figure 3.2. RPKM values for EBV genes in Mutu I (A) and Akata (B) cells. 45
Figure 3.4. Visualization of junction evidence for EBNA1 47
Figure 4.1. Key features of SAMMate 50
Figure 4.2. Combination of exon reads with junction reads to accurately calculate gene expression RPKM scores 52
Figure 4.3. Visualization of gene structure sariation 54

List of Tables

Table 1. 1 Overview of 11 mandatory fields in alignment line. 7
Table 1. 2 Overview of optional fields in alignment line. 8

Abstract

The main theme of this thesis research is concerned with developing a computational pipeline for processing Next-generation RNA sequencing (RNA-seq) data. RNA-seq experiments generate tens of millions of short reads for each DNA/RNA sample. The alignment of a large volume of short reads to a reference genome is a key step in NGS data analysis. Although storing alignment information in the Sequence Alignment/Map (SAM) or Binary SAM (BAM) format is now standard, biomedical researchers still have difficulty accessing useful information. In order to assist biomedical researchers to conveniently access essential information from NGS data files in SAM/BAM format, we have developed a Graphical User Interface (GUI) software tool named SAMMate to pipeline human transcriptome quantification. SAMMate allows researchers to easily process NGS data files in SAM/BAM format and is compatible with both single-end and paired-end sequencing technologies. It also allows researchers to accurately calculate gene expression abundance scores.

Keywords

Transcriptome

Gene Expression

Next-Generation Sequencing

RNA-seq Pipeline

SAMMate

SAM/BAM Format

Single-end/Paired-end

Chapter 1. Background and Introduction

1.1 Introduction to transcriptome

The transcriptome is the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA produced in one or a population of cells, or it can be referred to as the total of transcripts (or called isoform) or the specific subset of transcripts in a living cell. Unlike the genome that nearly does not change in a living cell except for mutation cases, the transcriptome varies according to different external environmental conditions, such as specialized tissues or cell lines. Most of the transcripts are processed by splicing to remove introns and generate a mature transcript or messenger RNA (mRNA) that only contains exons. Transcriptome is highly diverse, dynamic, complex and overlapping. Importantly, the range of transcriptome is enhanced by alternative splicing. Alternative splicing is a fundamental molecular process of multiple transcripts from a single gene due to variations in the splicing reaction of pre-mRNA. An exon can be either included or excluded from the mature transcripts. Thus, different splicing variants are generated from the same gene.

1.2 Gene expression

Gene expression is the synthesis process of a functional gene product by using the genetic information from a gene. These functional gene products include proteins and functional RNAs, the latter are in non-protein coding genes such as ribosomal RNA (rRNA) genes or transfer RNA (tRNA) genes. There are two major stages in gene expression. The first stage is transcription. In this stage, a single RNA molecule (a primary transcript) with basically the same sequence as the gene is produced. Most human genes consist of exons and introns, but only the exons carry information required for protein synthesis. Therefore, most mature primary transcripts or mRNAs only contain exons by splicing to remove intron regions. The second stage is translation. In this stage, mRNA along with transfer RNA (tRNA) and ribosomes work together to produce proteins. Other stages of gene expression include RNA splicing and post-translational modification of a protein, and any step of gene expression may be modulated. In genetics, gene expression fundamentally interprets the genetic code stored in DNA.

1.3 Microarray technology

Microarray technology is referred to as a multiplex lab-on-a-chip technology that is widely used in molecular biology. A small solid glass slide or silicon thin-film cell attaches a large amount of different nucleic acid probes to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. The relative abundance of nucleic acid sequences in the target can be usually determined by detection and quantification of the probe-target hybridization. Microarray technology was once as the experiment of choice for transcriptome analysis. Applications of microarray technology widely involve gene discovery, disease diagnosis, drug discovery, toxicological research and so on. In contrast to digital counts of transcript abundance produced by Next-Generation Sequencing (NGS), the fluorescent dye–based microarrays generate analogous signals from image intensity. Although the use of microarrays remains active in a number of research areas, the promising NGS is becoming the method of choice due to the intrinsic experimental limitations of microarrays.

1.4 Next-generation sequencing

For many years, the microarray-based analysis of transcriptomes plays critical roles in interrogating a large portion of genes expressed in a cell. Nevertheless, microarray technologies have several intrinsic limitations, such as signal saturation, biasness of probe design and non-specific hybridization. High-throughput sequencing technologies have overcome many limitations of microarray technologies. NGS technologies sample the mRNA with fewer biases and generate tens of millions of short fragments from a library of nucleotide sequences. Currently, a range of genetic analyses, including whole genome resequencing, gene expression analysis and small ribonucleic acid (RNA) analysis, were supported by NGS platforms. For example, using the Illumina (http://www.illumina.com/) Genome Analyzer platform, recent applications include sequencing mammalian transcriptomes [Mortazavi et al. 2008], ABI Solid Sequencing to profile stem cell transcriptomes [Cloonan et al. 2008] or Life Science's 454 Sequencing to discover SNPs in maize [Barbazuk et al. 2007]. Even though technical differences or applications exist in each platform, the information gathered from each share similar principle. Compared with microarray technology, RNA-seq experiments also provide much higher resolution measurements of expression at comparable costs [Marioni et al. 2008].

The following sections will introduce a series of relevant methods and analyses of whole-transcriptome sequencing data (RNA-seq).

1.4.1 RNA-seq technology

RNA-seq, also called "Whole Transcriptome Shotgun Sequencing" [Ryan et al. 2008] ("WTSS") and dubbed "a revolutionary tool for transcriptomics" [Wang et al. 2009], refers to the use of high-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content. RNA-seq quickly becomes invaluable in the study of diseases like cancer [Maher et al. 2009]. For many years, the standard method for determining the sequence of transcribed genes has been to capture and sequence messenger RNA using expressed sequence tags (ESTs) [Adams et al. 1993] or full-length complementary DNA (cDNA) sequences using conventional Sanger sequencing technology. However, RNA-seq has a number of advantages over the conventional EST sequencing. RNA-seq samples the mRNA with fewer biases and generates tens of millions of short reads per experiment, and these data can be used in measurement of the level of gene expression.

As an emerging RNA-seq technology, we are facing the challenge of complex alignment problem. In an RNA-seq experiment, the computing power to track all the possible alignments is nontrivial when aligning tens of millions of short reads to the reference genome. Besides, millions of reads unable to accurately align to the references genome when the reads originating from exon-exon junctions. Aligning reads originating from exon-exon junctions to references genome is also a hard nut to crack for researchers. Thanks to the deep coverage and base level resolution provided by next-generation sequencing instruments, RNA-seq provides researchers with efficient ways to interrogate transcriptome [Maher CA et al. 2009].

1.4.2 Short reads alignment

In a typical RNA-seq experiment, tens of millions of short reads are generated from a library of nucleotide sequences. We need to map these short reads of mRNA to identify regions of similarity on a reference genome. Since these short reads are sequenced from exonic and junction regions, we need to pay attention to short reads aligned to those regions. Due to the short read length, aligning a large volume of short reads to a long reference genome poses a great challenge to analysis of RNA-seq data. For measuring gene expression, we often have to align short reads to original positions on a reference genome using alignment tools. There are several tools Maq [Li H et al. 2008] , SOAP [Li R et al. 2008], RMAP [Smith et al. 2008], Bowtie [Ben et al. 2009] and Novoalign (http://www.novocraft.com) available for aligning genomic reads to a reference genome.

1.4.3 Junction mapping

Regarding alignment of short reads, a special attention is needed when processing the alignment of the short reads on both sides of the exon-exon junctions. Although most of the short reads can be mapped onto exon regions, a large set of short reads originating from exon-exon junctions still cannot be aligned against to reference genome. Thus, working with the short reads originating from exon-exon junctions in cDNA (around 10%) is a unique challenge for researchers. These short reads fail to map to the reference genome since the exons are separated by introns (Figure 1.1). Millions of unmapped short reads originating from exon-exon junctions, denoted as Initially Unmapped Reads (IUM's), need to be accounted for when measuring gene expression. To address the IUM problem, ERANGE [Mortazavi et al. 2008], Tophat [Trapnell et al. 2009] and rSeq [Jiang et al. 2008] are among the recently developed approaches to map IUM's originating from exon-exon junctions back to individual genes. ERANGE uses a union of known and novel junctions while Tophat *de novo* assembles IUM's using a module in Maq [Li H et al. 2008].

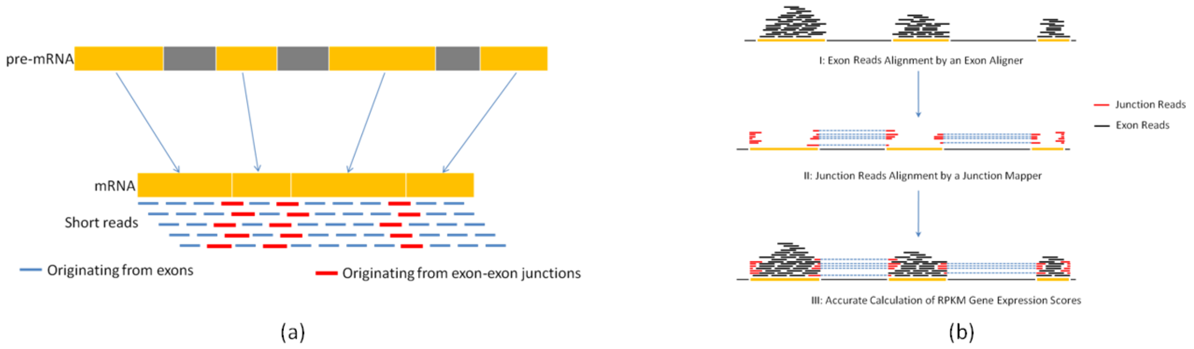
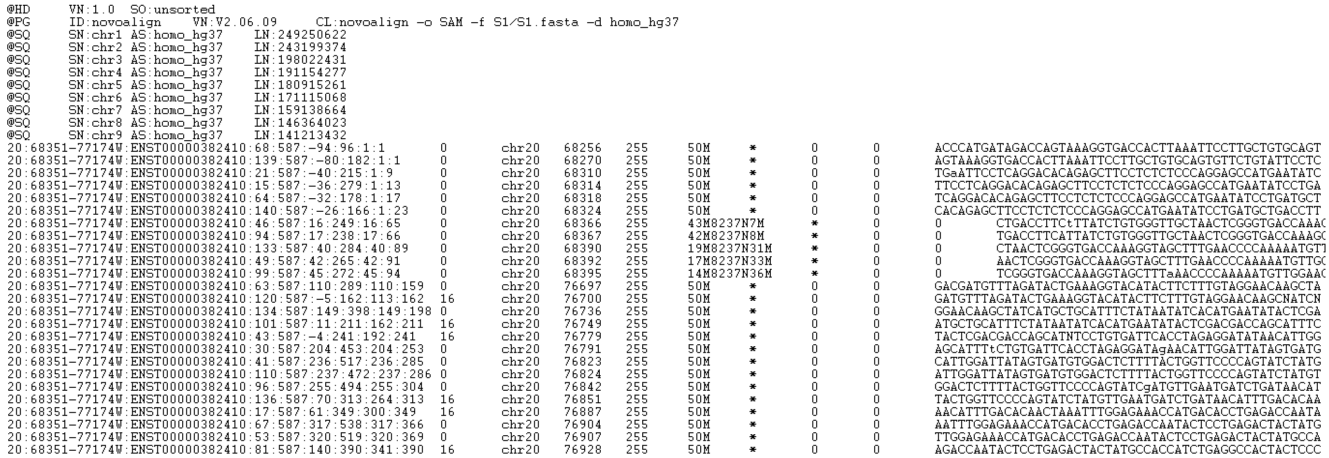

pre-mRNA
mRNA
Short reads
Originating from exons
Originating from exon-exon junctions
(a)
I: Exon Reads Alignment by an Exon Aligner
Junction Reads
Exon Reads
II: Junction Reads Alignment by a Junction Mapper
III: Accurate Calculation of RPKM Gene Expression Scores
(b)

Figure 1.1. Combination of exon reads with junction reads to accurately calculate gene expression RPKM scores. (a) A unique challenge for researchers working with RNA-seq data. The junction reads (red) fail to map back to the reference genome because exons are separated by introns. (b) A demonstration of the ideas of combing exon reads (black) and junction reads (red) to calculate gene expression RPKM scores. [Xu et al, 2011]

1.4.4 Data Format – SAM/BAM

Sequence Alignment/Map (SAM) format is defined as a generic nucleotide alignment format which describes the alignment of query sequences or sequencing reads to a reference sequence or assembly [Li et al. 2009]. Many of these tools output the alignment results in the SAM and Binary SAM (BAM) formats [Li et al. 2009], which are widely considered the *de facto* standards for storing and transferring short read alignment results. Although SAM format is easy to understand and straight-forward, SAM is still a bit slow to parse. Therefore, Binary SAM (BAM) format is introduced for intensive data storing and parsing. SAM/BAM file can store both sing-end and paired-end reads. A mapped read pair is stored in two (or more if multiple hits are stored) separate alignment records [Li et al. 2009].

SAM is a TAB-delimited text format. It generally consists of a header section and an alignment section. The header section is optional and it must be placed before the alignments section if present. Header lines start with '@' and include format version, sorting order of alignments, related reference sequence dictionary and read group. In the alignments section, each alignment line has 11 mandatory fields for essential alignment information. For example, mapping position, and variable number of optional fields for flexible or aligner specific information (Figure 1.2).



```
@HD     VN:1.0  SO:unsorted
@PG     ID:novoalign    VN:V2.06.09     CL:novoalign -o SAM -f S1/S1.fasta -d homo_hg37
@SQ     SN:chr1 AS:homo_hg37    LN:249250622
@SQ     SN:chr2 AS:homo_hg37    LN:243199374
@SQ     SN:chr3 AS:homo_hg37    LN:198022431
@SQ     SN:chr4 AS:homo_hg37    LN:191154277
@SQ     SN:chr5 AS:homo_hg37    LN:180915261
@SQ     SN:chr6 AS:homo_hg37    LN:171115068
@SQ     SN:chr7 AS:homo_hg37    LN:159138664
@SQ     SN:chr8 AS:homo_hg37    LN:146364023
@SQ     SN:chr9 AS:homo_hg37    LN:141213432
20:68351-77174W:ENST00000382410:68:587:-94:96:1:1       0       chr20   68256   255     50M     *       0       0       ACCCATGATAGACCAGTAAAGGTGACCACTTAAATTCCTTGCTGTGCAGT
20:68351-77174W:ENST00000382410:139:587:-80:182:1:1     0       chr20   68270   255     50M     *       0       0       AGTAAAGGTGACCACTTAAATTCCTTGCTGTGCAGTGTTCTGTATTCCTC
20:68351-77174W:ENST00000382410:21:587:-40:215:1:9      0       chr20   68310   255     50M     *       0       0       TGaATTCCTCAGGACACAGAGCTTCCTCTCTCCCAGGAGCCATGAATATC
20:68351-77174W:ENST00000382410:15:587:-36:279:1:13     0       chr20   68314   255     50M     *       0       0       TTCCTCAGGACACAGAGCTTCCTCTCTCCCAGGAGCCATGAATATCCTGA
20:68351-77174W:ENST00000382410:64:587:-32:178:1:17     0       chr20   68318   255     50M     *       0       0       TCAGGACACAGAGCTTCCTCTCTCCCAGGAGCCATGAATATCCTGATGCT
20:68351-77174W:ENST00000382410:140:587:-26:166:1:23    0       chr20   68324   255     50M     *       0       0       CACAGAGCTTCCTCTCTCCCAGGAGCCATGAATATCCTGATGCTGACCTT
20:68351-77174W:ENST00000382410:46:587:16:249:16:65     0       chr20   68366   255     43M8237N7M      *       0       0       CTGACCTTCtTTATCTGTGGGTTGCTAACTCGGGTGACCAAAC
20:68351-77174W:ENST00000382410:94:587:17:238:17:66     0       chr20   68367   255     42M8237N8M      *       0       0       TGACCTTCATTATCTGTGGGTTGCTAACTCGGGTGACCAAAGC
20:68351-77174W:ENST00000382410:133:587:40:284:40:89    0       chr20   68390   255     19M8237N31M     *       0       0       CTAACTCGGGTGACCAAAGGTAGCTTTGAACCCCAAAAATGTT
20:68351-77174W:ENST00000382410:49:587:42:265:42:91     0       chr20   68392   255     17M8237N33M     *       0       0       AACTCGGGTGACCAAAGGTAGCTTTGAACCCCAAAAATGTTGC
20:68351-77174W:ENST00000382410:99:587:45:272:45:94     0       chr20   68395   255     14M8237N36M     *       0       0       TCGGGTGACCAAAGGTAGCTTTaAACCCCAAAAATGTTGGAAC
20:68351-77174W:ENST00000382410:63:587:110:289:110:159  0       chr20   76697   255     50M     *       0       0       GACGATGTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTA
20:68351-77174W:ENST00000382410:120:587:-5:162:113:162  16      chr20   76700   255     50M     *       0       0       GATGTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCNATCN
20:68351-77174W:ENST00000382410:134:587:149:398:149:198 0       chr20   76736   255     50M     *       0       0       GGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATATACTCGA
20:68351-77174W:ENST00000382410:101:587:11:211:162:211  16      chr20   76749   255     50M     *       0       0       ATGCTGCATTTCTATAATATCACATGAATATACTCGACGACCAGCATTTC
20:68351-77174W:ENST00000382410:43:587:-4:241:192:241   16      chr20   76779   255     50M     *       0       0       TACTCGACGACCAGCATNTCCTGTGATTCACCTAGAGGATATAACATTGG
20:68351-77174W:ENST00000382410:30:587:204:453:204:253  0       chr20   76791   255     50M     *       0       0       AGCATTTtCTGTGATTCACCTAGAGGATAgAACATTGGATTATAGTGATG
20:68351-77174W:ENST00000382410:41:587:236:517:236:285  0       chr20   76823   255     50M     *       0       0       CATTGGATTATAGTGATGTGGACTCTTTTACTGGTTCCCCAGTATCTATG
20:68351-77174W:ENST00000382410:110:587:237:472:237:286 0       chr20   76824   255     50M     *       0       0       ATTGGATTATAGTGATGTGGACTCTTTTACTGGTTCCCCAGTATCTATGT
20:68351-77174W:ENST00000382410:96:587:255:494:255:304  0       chr20   76842   255     50M     *       0       0       GGACTCTTTTACTGGTTCCCCAGTATCgATGTTGAATGATCTGATAACAT
20:68351-77174W:ENST00000382410:136:587:70:313:264:313  16      chr20   76851   255     50M     *       0       0       TACTGGTTCCCCAGTATCTATGTTGAATGATCTGATAACATTTGACACAA
20:68351-77174W:ENST00000382410:17:587:61:349:300:349   16      chr20   76887   255     50M     *       0       0       AACATTTGACACAACTAAATTTGGAGAAACCATGACACCTGAGACCAATA
20:68351-77174W:ENST00000382410:67:587:317:538:317:366  0       chr20   76904   255     50M     *       0       0       AATTTGGAGAAACCATGACACCTGAGACCAATACTCCTGAGACTACTATG
20:68351-77174W:ENST00000382410:53:587:320:519:320:369  0       chr20   76907   255     50M     *       0       0       TTGGAGAAACCATGACACCTGAGACCAATACTCCTGAGACTACTATGCCA
20:68351-77174W:ENST00000382410:81:587:140:390:341:390  16      chr20   76928   255     50M     *       0       0       AGACCAATACTCCTGAGACTACTATGCCACCATCTGAGGCCACTACTCCC
```

Figure 1.2. Example of single-end reads in SAM format.

In each alignment line, these mandatory fields are required to present in the same order, but their values can be '0' or '*' (depending on the field) if the corresponding information is not available. The following table gives an overview of the mandatory fields in the SAM format:

Col	Filed	Type	Regxp/Range	Brief description
1	QNAME	String	[!-?A-~] {1,255}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0, 2^{29}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0, 2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next fragment
8	PNEXT	Int	[0, 2^{29}-1]	Position of the mate/next fragment
9	TLEN	Int	[-2^{29}+1, 2^{29}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	fragment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Table 1. 1 Overview of 11 mandatory fields in alignment line.
[http://samtools.sourceforge.net/SAM1.pdf.]

In each alignment line, all optional fields follow the **TAG:TYPE:VALUE** format where **TAG** is a two-character string that matches /[A-Za-z][A-Za-z0-9]/. Each **TAG** can only appear once in one alignment line. A **TAG** containing lowercase letters are reserved for end users. In an optional field, **TYPE** is a single case sensitive letter which defines the format of **VALUE**:

Type	Regexp matching VALUE	Description
A	[!-~]	Printable character
i	[-+]?[0-9]+	Singed 32-bit integer
f	[-+]?[0-9]*\.?[0-9]+([eE][-+]?[0-9]+)?	Single-precision floating number
Z	[!-~]+	Printable string, including space
H	[0-9A-F]+	Byte array in the Hex format1
B	[cCsSiIf](,[-+]?[0-9]*\.?[0-9]+([eE][-+]?[0-9]+)?)+	Integer or numeric array

Table 1. 2 Overview of optional fields in alignment line.
[http://samtools.sourceforge.net/SAM1.pdf.]

1.4.5 Quantification of gene expression using RNA-seq

Quantifying gene expression in cells via measurement of mRNA levels arouses researchers’ interest all the time. In RNA-seq experiment, for each gene, ERANGE reports the number of mapped **R**eads **P**er **K**ilobase of exon per **M**illion mapped reads (RPKM), a measure of transcription activity [Trapnell et al. 2009]. For paired-end short reads, we measure the transcript-level relative abundance in **F**ragments **P**er **K**ilobase of exon model per **M**illion mapped fragments (FPKM).

$$\mathrm{RPKM/FPKM} = 10^9 \times \frac{C}{L \times N}$$

C is the total number of mapped short reads or fragments, *L* is the length of exons and *N* is the total number of short reads in one lane of one experiment. When scaled to range [0, 1000], this value stands for the normalized depth of coverage for each gene. Even though it has been shown that there is no strong correlation between the abundance of mRNA and the related proteins [Greenbaum et al. 2003], measurement of mRNA levels is still very useful in determining how cells differ between a healthy state and a diseased state and other research problems.

1.5 Motivation

Intrinsic technical limitations to microarray technology constrain its ability to fully quantify gene expression. Fortunately, high-throughput multiplexed next-generation sequencing provides a digital readout of absolute transcript levels and imparts a higher level of accuracy and dynamic range than microarray platforms. In a typical experiment, tens of millions of short reads are sampled from RNA with fewer biases for accurately measuring gene expression. However, a significant challenge in analyzing gene expression is the short read alignment to a reference genome. Although storing alignment information in the Sequence Alignment/Map (SAM) or Binary SAM (BAM) format is a generic alignment format, biomedical researchers are still facing a technical obstacle in accessing the useful information stored SAM/BAM files. In this thesis, we present a GUI computational pipeline for researchers to efficiently process and extract information from NGS data for quantifying human transcriptome. We show our GUI software can accurately calculate the gene expression abundance scores for genomic intervals using short reads originating from both exons and exon-exon junctions for RNA-seq data.

1.6 Overview

This thesis is organized into 5 chapters. In Chapter 1 we introduce background, motivation and overview. In Chapter 2 we describe transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq. In Chapter 3 we present RNA-seq Analysis of EBV Transcriptome. In Chapter 4 we discuss a GUI tool for processing short read alignments in SAM/BAM format named SAMMate. In Chapter 5 we draw a conclusion of the thesis and future efforts.

The thesis is largely based on the following list of relevant publications and software,

- Transcriptome and targetome analysis using RNA-seq

Xu, G, Fewell, C, Taylor, C, Deng, N, Hedges, D, Wang, X, Zhang, K, Lacey, M, Zhang, H, Yin, Q, Cameron, J, Zheng, L, *Zhu, D and *Flemington, EK: Transcriptome and targetome analysis in mir155 expressing cells using rna-seq. *RNA (New York, N.Y.), 16(8):1610–1622, August 2010.*

#Lin, Z, #**Xu, G**, Deng, N, Taylor, C, *Zhu, D and *Flemington, EK: Analysis of EBV transcriptome using RNA-seq. *J. Virology, doi:10.1128/JVI.01521-10.*

- SAMMate: a GUI tool for processing short read alignments in SAM/BAM format

***Xu,G**, Deng, N, Zhao, Z, Zhang, K, Judeh, T, Flemington, EK and *Zhu, D: SAMMate: A GUI tool for processing short read alignment information in SAM/BAM format. Source Code for Biology and Medicine, 6:2.*

- Software

SAMMate with Graphic User Interface (GUI) [available from, http://SAMMate.sourceforge.net]

Chapter 2. Transcriptome and Targetome Analysis in MIR155 Expressing Cells using RNA-seq

2.1 Introduction

MicroRNAs play critical roles in controlling biological processes through their ability to post-transcriptionally regulate gene expression. A testament to their importance in normal organismal biology is that dysregulation of microRNA function through genetic or epigenetic alterations is at the root of an array of disparate diseases including cancer [Calinet al. 2004; Iorio et al. 2005; Lu et al. 2005; Volinia et al. 2006; Zhang et al. 2006]. The gene encoding microRNA-155(MIR155) was classified as an oncogene [Clurman and Hayward 1989; Tam et al. 1997; Costinean et al. 2006] many years before it was identified as a microRNA [Eiset al. 2005] and is now among the most highly implicated microRNAs in cancer. Despite its link to hematologic [Eiset al. 2005; Kluiver et al. 2005] and other cancers [Volinia et al. 2006], there is currently little information regarding direct targets or pathways through which MIR155 signals to promote the tumor phenotype.

The phenotypic consequences of microRNAs are facilitated through the combination of not only direct, but also indirect, influences on gene expression. Nevertheless, the identification of direct target mRNAs is a topic of intense interest because it provides insights into the entry point through which microRNAs regulates a respective pathway. The recognition of targets through a predominantly Watson–Crick base-pairing mechanism has capacitated informatics based prediction approaches that have lent considerable support to global target identification efforts [Li et al. 2010b]. Despite the applicability of this approach, there are less tangible flanking sequence criteria that also play a role in specifying targeting [Hammell 2010]. This limits the veracity of informatics-based target prediction and necessitates the concomitant application of experimental methods to search for and/or validate microRNA targets. High-throughput sequencing of RNAs isolated by cross-linking immunoprecipitation (HITS-CLIP) is a recently developed method that enables identification of direct target sequences through the sequencing of RNAs from immunoprecipitated cross-linked Argonaute–miRNA–mRNA complexes [Chi et al. 2009]. The use of SILAC (stable isotope labeling with amino acids in cell culture) and state-of-the-art mass spectroscopy approaches has allowed the interrogation of up to 5000 members of a proteome for changes in protein output in response to micro-RNA expression [Selbach et al. 2008]. Whereas HITS-CLIP can directly identify targeting sequences, proteomics approaches identify an inferred targetome. On the other hand, SILAC-based proteomics approaches also provide expression information for directly and indirectly affected genes that can give additional insights (relative to HITS-CLIP) into the biological outcomes of a microRNA's function. In spite of this, proteomics methods likely miss up to half of all expressed proteins and may be biased against the less abundant proteins that often play critical regulatory roles in cell signaling.

Microarray-based analysis of transcriptomes has the potential to interrogate a much larger portion of all genes expressed in a cell, and this approach has been used to globally characterize microRNA-mediated transcriptome changes and inferred targetomes (for example, see Grimson et al. 2007). Nevertheless, microarray technologies have several intrinsic characteristics that limit their utility in fully exploiting RNA changes to assess microRNA-induced transcriptome changes and microRNA targetomes. High-throughput sequencing of RNAs has overcome many of these limitations, and we reasoned that next-generation sequencing (NGS) may provide an improvement over microarray technologies. The higher level of accuracy of NGS [Marioni et al.

2008; Mortazavi et al. 2008] should make it more suitable for assessing the relatively moderate influences that microRNAs have on their target mRNAs. Its broader dynamic range allows the analysis of both high- and low-abundance transcripts and should therefore facilitate the analysis of genes spanning a wide spectrum of expression levels. Unlike the most commonly used microarray platforms, which only interrogate the ends of terminal exons to derive expression information, NGS gathers expression information throughout the entire locus of all expressed genes. This feature of NGS is important in light of recent studies showing the occurrence of widespread upstream transcription termination shifts during immune cell development and in cancer cells, for example, Sandberg et al. (2008) and Mayr and Bartel (2009). For genes simultaneously expressing shortened and long transcripts, expression changes obtained by microarray studies reflect only changes in the subset of extended isoforms that may be more responsive to microRNA targeting but do not accurately mirror overall changes in expression of the locus. Finally, NGS allows the user to simultaneously assess isoform structure, which is critically important for many regulatory processes including microRNA targeting. This information can be used to elucidate targeting failures in a particular system. We also anticipate that as the potential targetomes of microRNAs become well characterized and as NGS data become broadly accumulated for different cell systems, actual targetomes can be predicted for each cell system based on transcript structure. This should allow for informed prediction of biological outputs of microRNAs in different cell systems based on publicly available information alone.

2.2 Results

2.2.1 Model system

Epstein-Barr virus (EBV) infected B-cells expressing the full repertoire of latency genes (type III latency) manifest high levels of MIR155, whereas EBV-infected B-cells exhibiting a limited viral gene expression program (type I latency) do not [Jiang et al. 2006; Kluiver et al. 2006]. To investigate the utility of next generation RNA sequencing (RNA-seq) in assessing a microRNA targetome, we infected the type I latency B-cell line, Mutu I, in duplicate with an MIR155 expressing retrovirus (or an empty vector control etrovirus) to achieve high MIR155 expression in a low expression background. Infections were judged to be highly efficient based on the low level of cell death after selection and the high percentage of GFP-positive cells (>60%) 2 d after infection. We therefore considered the infected cell populations to be highly polyclonal. Expression of MIR155 was found to be ~100, 000-fold higher in MIR155 transduced Mutu I cells than in control transduced Mutu I cells or in two other type I latency cell lines (Figure 2.1A, Akata and Rael). Additionally, expression in MIR155 transduced Mutu I cells was comparable to endogenous MIR155 expression in the type III latency cell lines, JY, X50-7, and IB4 (Figure 2.1A). The level of MIR155 in retrovirally transduced Mutu I cells was found to be sufficient to exert suppression of the previously identified MIR155 target, BACH1 [Figure 2.1B; Gottwein et al. 2007; Skalsky et al. 2007; Yin et al. 2008]. Furthermore, ectopic MIR155 exerted a phenotypic influence on Mutu I cells since MIR155 transduced cells formed more colonies in soft agar than their control transduced counterparts (Figure 2.1C), a result that is consistent with the known oncogenic properties of MIR155. We therefore considered this system to be suitable for carrying out MIR155 transcriptome and targetome analysis using RNA-seq.

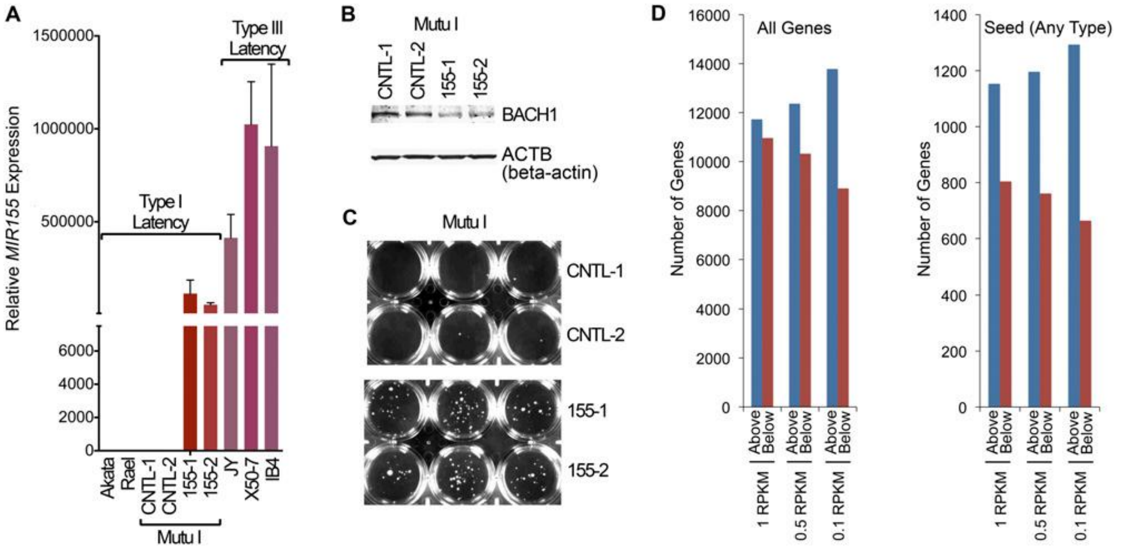


Figure 2.1 The biological system. (A) Mature MIR155 was analyzed by real-time RT-PCR in the EBV type I latency cell lines—Akata, Rael, and Mutu I—and the type III latency cell lines—JY, X50-7, and IB4. CNTL-1 and CNTL-2 and 155-1 and 155-2 refer to duplicate biological replicate infections with control and MIR155 expressing retroviruses. Expression values are reported relative to the average expression levels in Akata and Rael. (B) BACH1 and ACTB (b-actin) Western blots were performed using protein extracts isolated 14 d after retroviral infections. (C) Newly established control and MIR155 expressing Mutu I cells were plated in soft agar and cultured for 2 wk prior to photo-documentation. (D) The total number of genes and the number of genes containing any MIR155 seed type that are expressed above and below the indicated RPKM cutoffs in control Mutu I cells were counted and graphed.

2.2.2 Read mapping

Control and MIR155 transduced Mutu I RNAs were analyzed by NGS using an Illumina Genome Analyzer II. Reads per kilobase of exon model per million mapped reads (RPKM) [Mortazavi et al. 2008] was used to score gene expression abundance. A unique challenge for working with RNA-seq data is short reads originating from exon–exon junctions in cDNAs (～10%) that fail to map back to the reference genome because in this context, exons are separated by introns. The millions of unmapped short reads originating from exon–exon junctions, denoted as initially unmapped reads (IUMs), needed to be accounted for when calculating RPKM scores [Trapnell et al. 2009]. While most aligners (named ‘‘exon aligner’’ hereinafter) map only short reads originating from exons (ignoring IUMs), ERANGE [Mortazavi et al. 2008], TopHat [Trapnell et al. 2009], and rSeq [Jiang and Wong 2008] (named ‘‘junction mapper’’ hereinafter) are among recently developed approaches to assign IUMs originating from exon–exon junctions back to individual genes.

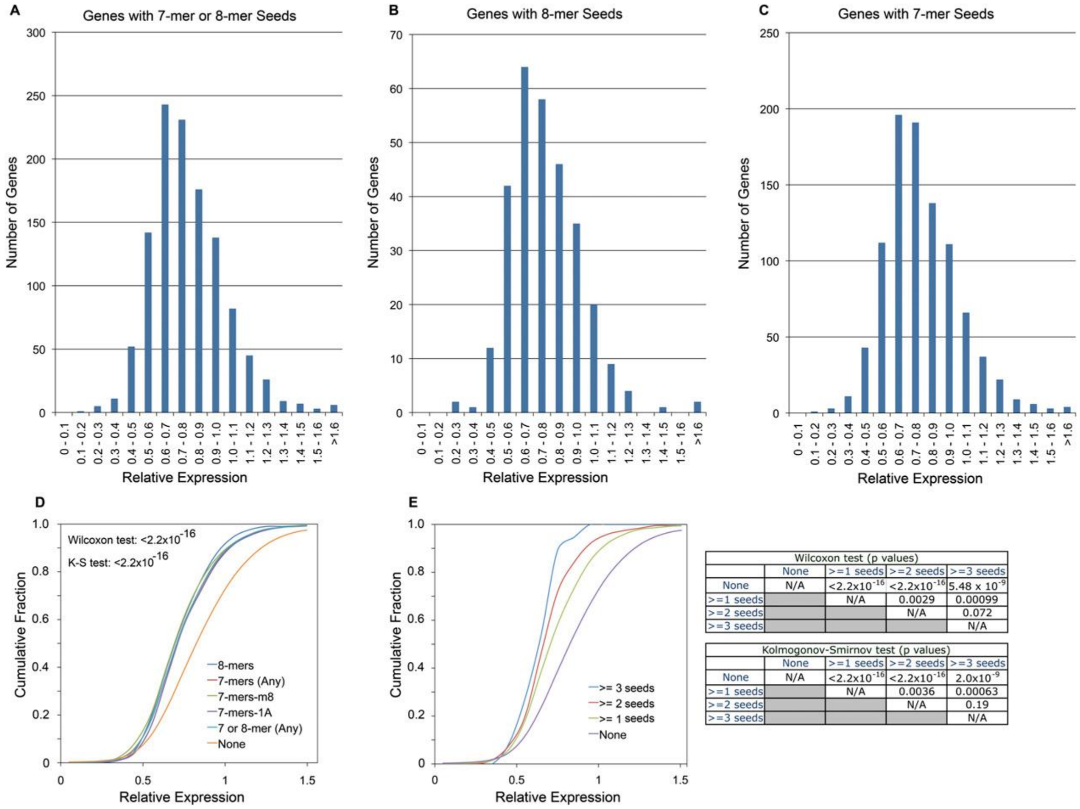
The exon aligners and junction mappers differ vastly in algorithms and accuracy. For RNA-seq data, it is necessary to choose and combine the most accurate exon aligner and junction mapper to estimate the gene expression RPKM scores. In our analysis, alignments were first carried out using the exon aligner Novoalign (http://www.novocraft.com), with reads aligning to more than one locus excluded from further analysis. IUMs were secondly de novo aligned to the assembled IUMs using the junction mapper TopHat [Trapnell et al. 2009]. IUMs mapping to exon junctions using TopHat were then combined with exon mapped reads from the Novoalign output, and RPKM calculations were carried out using University of California Santa Cruz (UCSC)–annotated gene loci [Rhead et al.2010].

2.2.2 Expression analysis

Microarray technologies can readily be used to generate information regarding the relative expression of genes between samples. Due in part to cross-hybridization and sensitivity issues as well as the analog nature of microarray platforms, however, the determination of absolute transcript levels is challenging. In contrast, NGS provides a digital readout of the number of reads mapping to each gene, and Li et al. (2010a) have shown that when the mean expressed transcript length is 1 kb, 1 RPKM corresponds to roughly one transcript per cell in mouse. It is reasonable to assume that transcript levels falling below this threshold may be of limited functional significance to the overall cell population. At the very fundamental level of RPKM analysis, NGS allows the user to tentatively absolve this group of genes from playing a direct global role in altering cell signaling/phenotype in a particular system. Such genes can be set aside and perhaps considered at a later point in the context of paracrine or subpopulation effects. In control Mutu I cells, approximately half of all annotated genes were found to be expressed below 1 RPKM (Figure 2.1D). Approximately 800 genes bearing MIR155 3' UTR seed sequences were found to be expressed below 1 RPKM. Even at a 0.1-RPKM cutoff, more than 600 seed containing genes were found to fall below this threshold and are therefore likely to have limited functional significance in these cells irrespective of whether they are true MIR155 targets. Notably, a higher percentage of MIR155 seed containing genes are expressed in Mutu I B-cells compared to the percentage of all genes expressed, possibly reflecting the critical role of MIR155 in immune cell development [O'Connell et al. 2007; Rodriguez et al. 2007]. Alternatively, this enrichment may simply reflect a general bias toward targeting a group of genes that are universally expressed.

2.2.3 Relative expression and distribution of genes with 3' UTR MIR155 seed sequences

We next assessed whether genes containing MIR155 3' UTR seed sequences are preferentially distributed among down-regulated genes. Genes expressed below 0.5 RPKMs were excluded from this analysis since these genes are likely of arguable functional significance and technical reproducibility begins to wane in this RPKM range [Mortazavi et al.2008]. In line with expectations, we observed preferential distribution of genes containing 7-mer and 8-mer MIR155 seed sequences in the down-regulated fractions (Figure 2.2A–C). Statistical analysis demonstrated preferential distribution (P-value $< 2.2 \times e^{-16}$) in terms of mean and variance (for details, refer to the Materials and Methods section). Cumulative distribution plots showed enrichment of all classes of 7-mers and 8-mers in the down-regulated fractions relative to genes with no 7-mer or 8-mer seeds (P-value < $2.2 \times\ e^{-16}$) (Figure 2.2D). Notably, even among genes without MIR155 seed sequences, there is a greater number of down-regulated compared to induced genes, indicating a more global redistribution of gene expression. This is not inconsistent with previous microarray studies showing that MIR155 down-regulates more genes than it induces [Gottwein et al. 2007; Skalsky et al. 2007]. Because RPKM calculations normalize to all mappable reads, these results presumably mean that MIR155 decreases the expression of genes within the lower expression class through a mechanism other than 3' UTR 7-mer or 8-mer basepairing. Irrespective of this issue, the enrichment for genes with 7-mer or 8-mer seeds in the down-regulated fractions relative to genes without 7-mer or 8-mer seeds is consistent with expectations based on previous microarray studies. As further evidence that RNA-seq analysis performed within expectations, we observed a down regulation bias for genes containing 3' UTRs with greater numbers of 7-mer or 8-mer seeds (Figure 2.2E).



Wilcoxon test (p values)				
	None	>=1 seeds	>=2 seeds	>=3 seeds
None	N/A	$<2.2\times10^{-16}$	$<2.2\times10^{-16}$	5.48×10^{-9}
>=1 seeds		N/A	0.0029	0.00099
>=2 seeds			N/A	0.072
>=3 seeds				N/A

Kolmogonov-Smirnov test (p values)				
	None	>=1 seeds	>=2 seeds	>=3 seeds
None	N/A	$<2.2\times10^{-16}$	$<2.2\times10^{-16}$	2.0×10^{-9}
>=1 seeds		N/A	0.0036	0.00063
>=2 seeds			N/A	0.19
>=3 seeds				N/A

Figure 2.2. Genes containing MIR155 seed sequences are enriched in down-regulated fractions. Relative expression (expression in MIR155 transduced Mutu I cells divided by the expression in control transduced Mutu I cells) histograms for genes containing 7-mer or 8-mer (A), 8-mer (B), or 7-mer (C) seeds in their 3’ UTR. (D) Cumulative distribution of genes with different seed classes. P-values for all seed classes as determined by one-sided Wilcoxon test (Wil.) and one-sided Kologorov–Smirnov test (K-S) were $<2.2 \times 10^{16}$. (E) Cumulative distribution of genes with different numbers of seeds. P-values for all seed number comparisons were determined using a one-sided Wilcoxon test and a onesided Kologorov–Smirnov test and are shown in the matrix to the *right*.

2.2.4 NGS *versus* microarray studies

Having demonstrated a preferential distribution of genes with MIR155 seeds in down-regulated fractions, we next sought to assess the performance of next generation sequencing relative to microarray analysis. We performed differential expression analysis on our RNA-seq data set and on data sets from two previously published microarray studies in which MIR155 expression vectors were introduced into either a mouse macrophage cell line [O’Connell et al.

2008] or human 293 cells [Skalsky et al. 2007]. Using only gene identifiers common to all four platforms and using a false discovery rate (FDR) of zero, 2165 genes were determined to be down-regulated by NGS, whereas 38 and 58 down-regulated genes were identified in our analysis of the two published microarray data sets (Figure 2.3). NGS identified 102 down-regulated genes (FDR = 0) with 3' UTR 8-mer seeds, while seven and two down-regulated genes with 8-mer seeds were identified in the two microarray data sets.

To more stringently assess the relative robustness of NGS for transcriptome and targetome studies, we generated additional control and MIR155 retrovirally transduced Mutu I cell lines and subjected four control and four MIR155 expressing cell lines to Agilent microarray analysis with dye swaps for each comparison. This resulted in better concordance relative to previous microarray studies, but the number of downregulated genes and the number of down-regulated genes with 8-mer seed sequences were threefold and 2.6-fold lower than that observed using the NGS data set (Figure 2.3).

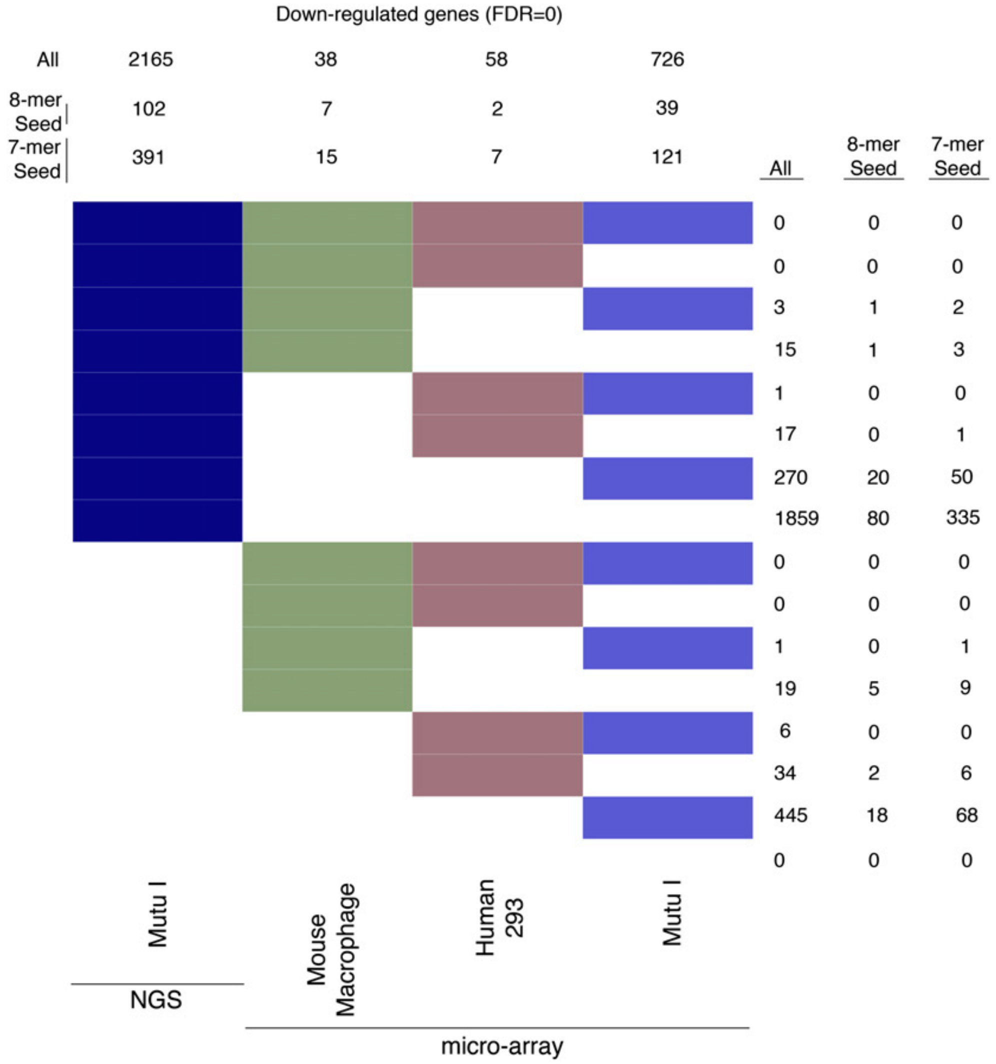


Figure 2.3. Cross-platform comparison of transcriptome quantification using bitmap. Downregulated genes were identified at a false discovery rate (FDR) = 0 for NGS and each microarray platform. Each gene was determined to be significantly down-regulated (at FDR = 0) or not in each of the four platforms; down-regulated genes were assigned to one of the 2^4 = 16 possible clusters, represented by color/white patterns and corresponding to 16 rows in the bitmap. Numbers at the *top* refer to the total number of down-regulated genes for the indicated platform (summation of the number of genes represented by all colored patterns in column). Numbers to the *right* refer to the number of genes common to platforms with colored patterns in each respective row.

2.2.5 3’ UTR reporter analysis

Luciferase reporter plasmids bearing ectopic 3’ UTRs can be used to assess microRNA targeting through the respective 3’ UTR. To further analyze the inferred targetome derived from

NGS, a 3’ UTR reporter data set was generated using 170 3’ UTRs containing MIR155 7-mer or 8-mer seeds and nine 3’ UTRs with no MIR155 seeds (Figure 2.4A). The relative expression of reporters lacking MIR155 seeds in cells cotransfected with a MIR155 expression vector versus a control expression vector fell in the range of 0.8 to 1.04. The relative expression of genes with MIR155 seed sequences spanned a range from 0.13 to 1.2 (Figure 2.4A; Supplemental Data 1), allowing us to analyze a spectrum of MIR155 target regulatory classes.

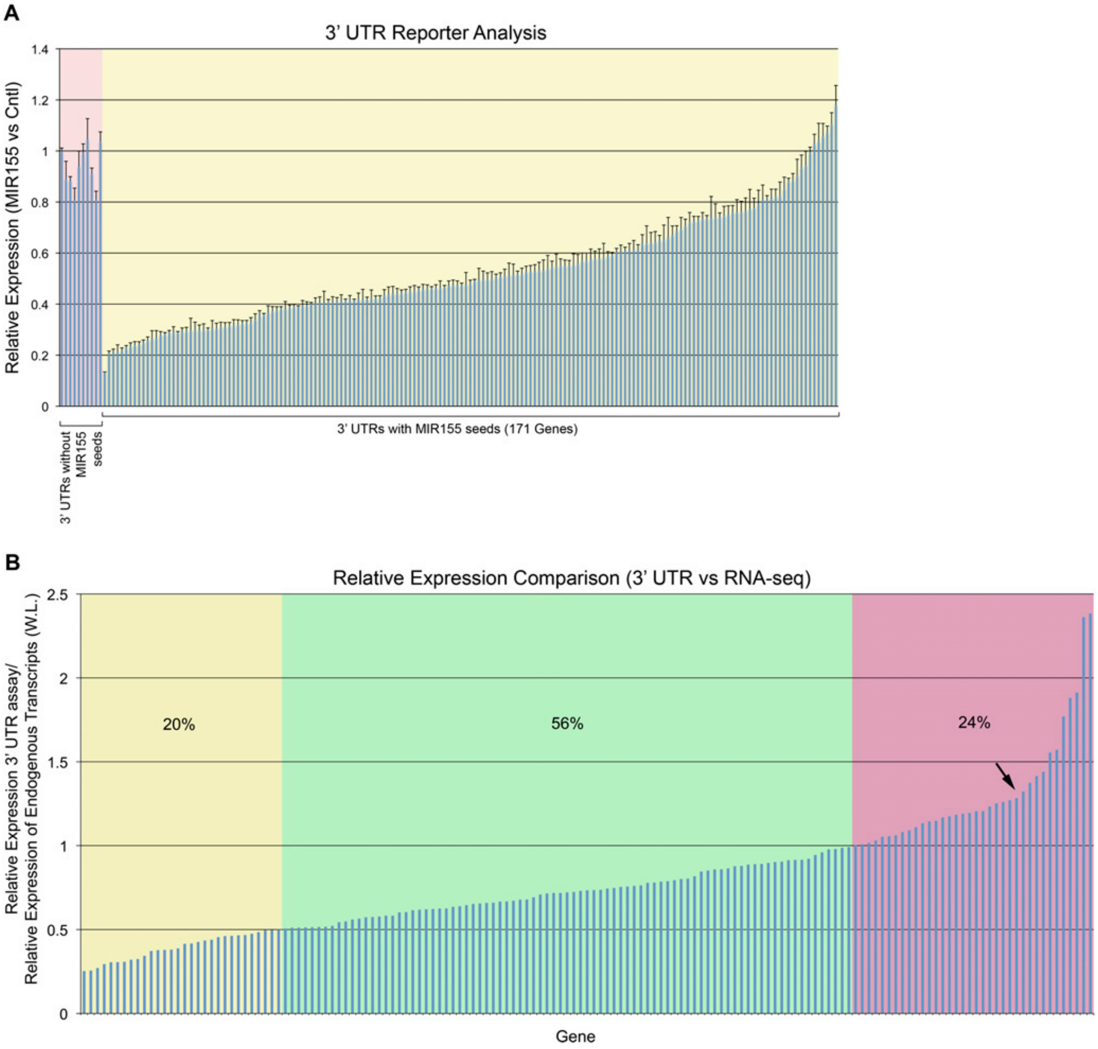


Figure 2.4. Validation of RNA-seq analysis using 3’ UTR assay. (A) Distribution of 3’ UTR suppression by MIR155 in reporter assays. Detailed information including gene names for each data point is presented in Supplemental Data 1. Genomic coordinates for each 3’ UTR in reporter constructs are listed in Supplemental Data 4. Values are the expression of reporters in cells cotransfected with the MIR155 expression vector versus cells cotransfected with the control

expression vector divided by the expression of the control reporter cotransfected with the MIR155 expression vector versus cells cotransfected with the control expression vector. Results are based on biological triplicate transfections with control and MIR155 expression vectors. Error bars are standard error of change. (B) Comparison of relative expression observed in 3' UTR assays with relative expression of endogenous transcripts at the whole locus level. The *y*-axis values are the relative expression observed in 3' UTR assays divided by the relative expression of the endogenous transcripts (at the whole locus [W.L.] level). Genes with yellow shading are genes considered to lack full response at the endogenous transcript level. Genes with green shading fall within the expected difference range. Genes with pink shading represent candidate genes whose transcripts are regulated through additional mechanisms in Mutu I cells. The arrow indicates an inflection point in the curve.

2.2.5 Validation of RNA-seq analysis using 3'UTR luciferase reporter assay

A total of 150 of the 170 genes tested in the 3' UTR reporter study were expressed at 0.5 or more RPKMs in Mutu I cells, allowing us to do comparisons with these 150 genes. Whereas changes observed by RNA-seq only reflect influences at the transcript level, changes observed in 3' UTR assays reflect the combined influence of transcript level changes plus the influence of a microRNA on translation, the latter of which is expected to vary depending on the target. In line with expectations, the majority (76%) of genes tested showed greater than or equal suppression in 3' UTR reporter assays relative to the RNA-seq analysis (Figure 2.4B; Supplemental Data 2). Of the remaining 24% of genes that displayed greater suppression as observed by RNA-seq versus 3' UTR analysis (values >1 in the *y*-axis of Figure 2.4B), at least some of these genes are likely to be modulated endogenously through additional mechanisms such as transcriptional regulation (especially those to the right of the apparent inflection point; see arrow in Figure 2.4B).

Although the relative contribution of translational inhibition varies from target to target, it appears that decreases in protein levels are rarely more than twice the decrease in RNA levels

[Selbach et al. 2008]. In our case, there are 29 genes (19%) with more than twofold greater inhibition in 3’ UTR assays compared to the inhibition observed at the endogenous transcript level (Figure 2.4B; Supplemental Data 2). We operationally treated these genes as outliers that show a disproportionately lesser degree of regulation at the level of RNA-seq.

To investigate these poor performers, we first directed our attention to some of the most extreme examples, those with little or no regulation observed by RNA-seq but good regulation in our 3’ UTR analysis. Three examples—PCDH9, MAP3K10, and TAF5L (Supplemental Data 3)—proved to be illustrative of scenarios accounting for some of these targeting failures/inefficiencies. First, suppression of the PCDH9 3’ UTR in the reporter assay was found to be substantial (relative expression of 0.26), whereas there was no change in RNA-seq (relative expression of 1.0). Visualization of the read pile-ups on the USCS Genome Browser [Rhead et al. 2010) showed no evidence of reads mapping to exons 3, 4, or 5, the latter of which contains the MIR155 seed sequence (Supplemental Figure 2.1). In contrast, ample read evidence supports transcription through exons 1 and 2, and splicing evidence supports an exon 1 and exon 2 junction. In this example, PCDH9 is likely to be a true target of MIR155 as evidenced by the 3’ UTR reporter analysis. The lack of any change in PCDH9 transcript levels in Mutu I cells is likely an accurate reflection of the dominant utilization of an isoform that is recalcitrant to MIR155 targeting.

Several recent studies have demonstrated 3’ UTR shortening in activated lymphocytes and in tumors [Sandberg et al. 2008; Mayr and Bartel 2009]. The lack of a change in MAP3K10 transcript levels as determined by RNA-seq analysis in the face of a relative expression of 0.47 observed by 3’ UTR analysis likely reflects such a scenario. In this case, read evidence falls off upstream from the MIR155 seed sequence, a localization with proximity to the end of a shorter 3’

UTR isoform entry identified from the GenBank database (Z48615) (Supplemental Figure 2.2). MAP3K10 is a representative example of a lack of responsiveness resulting from the dominant usage of an upstream poly-adenylation site.

Expression of MIR155 in Mutu I cells resulted in no observable change in overall TAF5L RNA levels despite the presence of a 3’ UTR with five MIR155 sites and a relative expression of 0.3 observed in our 3’ UTR analysis. TAF5L has two annotated isoforms with two distinct 3’ UTRs (Figure 2.5A, B). In contrast to the 3’ UTR tested in our 3’ UTR analysis, the other 3’ UTRs contains no MIR155 seed sequences. Although both isoforms are expressed in Mutu I cells, RPKM analysis of exon 5 (no MIR155 seeds) and the unique region of exon 4 (5 MIR155 seeds) shows that the relative abundance of the isoforms containing seeds is ~20- fold lower than the isoforms lacking seeds (Figure 2.5C). This dominant expression of the non-seed-containing isoform can explain the lack of observable changes in overall TAF5L transcripts in MIR155 expressing Mutu I cells. To determine whether the seed-containing isoform is regulated by MIR155 in Mutu I cells, we calculated the differential expression of 3’ UTR sequences that are unique to this transcript. Whereas the relative expression of exon 4 (common to both isoforms) and exon 5 are ~1, the relative expression of the unique region is 0.52 (Figure 2.5D), which is in line with the relative expression observed for this 3’ UTR in the reporter assay (Figure 2.5E). Therefore, while the dominance of the nonseed containing TAF5L isoform results in no change at the whole locus level, the seed containing isoform is specifically regulated by MIR155 in concordance with the 3’ UTR results.

2.2.6 Global analysis of terminal exon relative expression

We used these examples of targeting failures as a basis to design a general approach to assess the degree to which differential isoform usage contributes to mitigated suppression in

Mutu I cells. We reasoned that the relative expression of terminal exons would better reflect the degree of suppression observed in 3’ UTR assays in cases in which multiple transcript isoforms account for mitigated suppression in our cells (as observed by RNA-seq analysis of the whole transcript locus). First, for genes using more than one terminal exon, the terminal exons bearing MIR155 seed sequences should, on average, show a higher degree of suppression than the whole locus. Second, for genes bearing only one terminal exon but using more than one poly(A) signal, there should be a higher relative read representation of longer MIR155 seed containing transcripts within terminal exon coordinates compared to the whole gene locus.

For this analysis, RPKM and corresponding relative expression calculations were carried out for all terminal exon loci.We first applied this terminal exon differential expression data to the group of genes (excluding PCDH9, MAP3K10, and TAF5L) showing a disproportionately greater suppression in the 3’ UTR assays relative to RNA-seq (3’ UTR relative expression/RNA-seq [transcript] relative expression <0.5). Genes with terminal exon RPKM levels of <0.5 were excluded from this analysis. Of the 23 testable genes, better concordance with 3’ UTR data was achieved with the terminal exon analysis in a little more than half of the genes (Figure 2.6A). This indicates that, in these cases, the inability of MIR155 to suppress expression of these genes to its full potential is likely due to the expression of alternate nonsuppressible isoforms.

At the global level, cumulative distribution analysis of terminal exons with or without MIR155 seed sequences shows a relative enrichment of seed containing genes in the down-regulated fractions (P-value $< 2.2 \times 10^{16}$) (Figure 2.6B), in line with the whole locus analysis shown in Figure 2.2. However, greater inhibition is observed at the terminal exon level than the whole locus level (Figure 2.6C), indicating that isoform usage likely dictates the degree of suppression for a broad spectrum of MIR155 targets. Together, these data support the contention that for some targets, a lack of miRNA repression responses at the endogenous transcript level can be attributed to altered transcript structure.

2.3 Discussion

There are now a number of tangible methods to globally assess the influence of microRNAs on cell signaling, which range from an assessment of RNA binding sites (e.g., HITSCLIP), to an assessment of changes in RNA (e.g., microarray and NGS) and protein output (e.g., SILAC). Each of these methods has its own niche in determining the overall cascade of events leading to a microRNA's influence on cell signaling pathways. A unique advantage of NGS is its capacity to simultaneously provide accurate transcript level information while, at the same time, providing unprecedented clarity regarding transcript structure at a relatively low cost. Our data support the contention that mRNA isoform utilization is a critical determinant in specifying microRNA targeting. It also advocates transcriptome and miRNA targetome characterization at the isoform level as opposed to the typically used gene level analysis. In our experiments, transcript structure information obtained by NGS provided clarity with respect to at least half of all targeting failures and/or inefficiencies (Figs. 2.5, 2.6; Supplemental Figs. 2.1, 2.2). In addition, NGS provides the user with information on ‘‘failures to detect’’ changes that are attributed to low (or no) target gene expression (in our case, 39% of all genes with seed sequences were expressed below 0.5 RPKMs). NGS makes these transcriptome level targeting inefficiencies and/or failures transparent to the user.

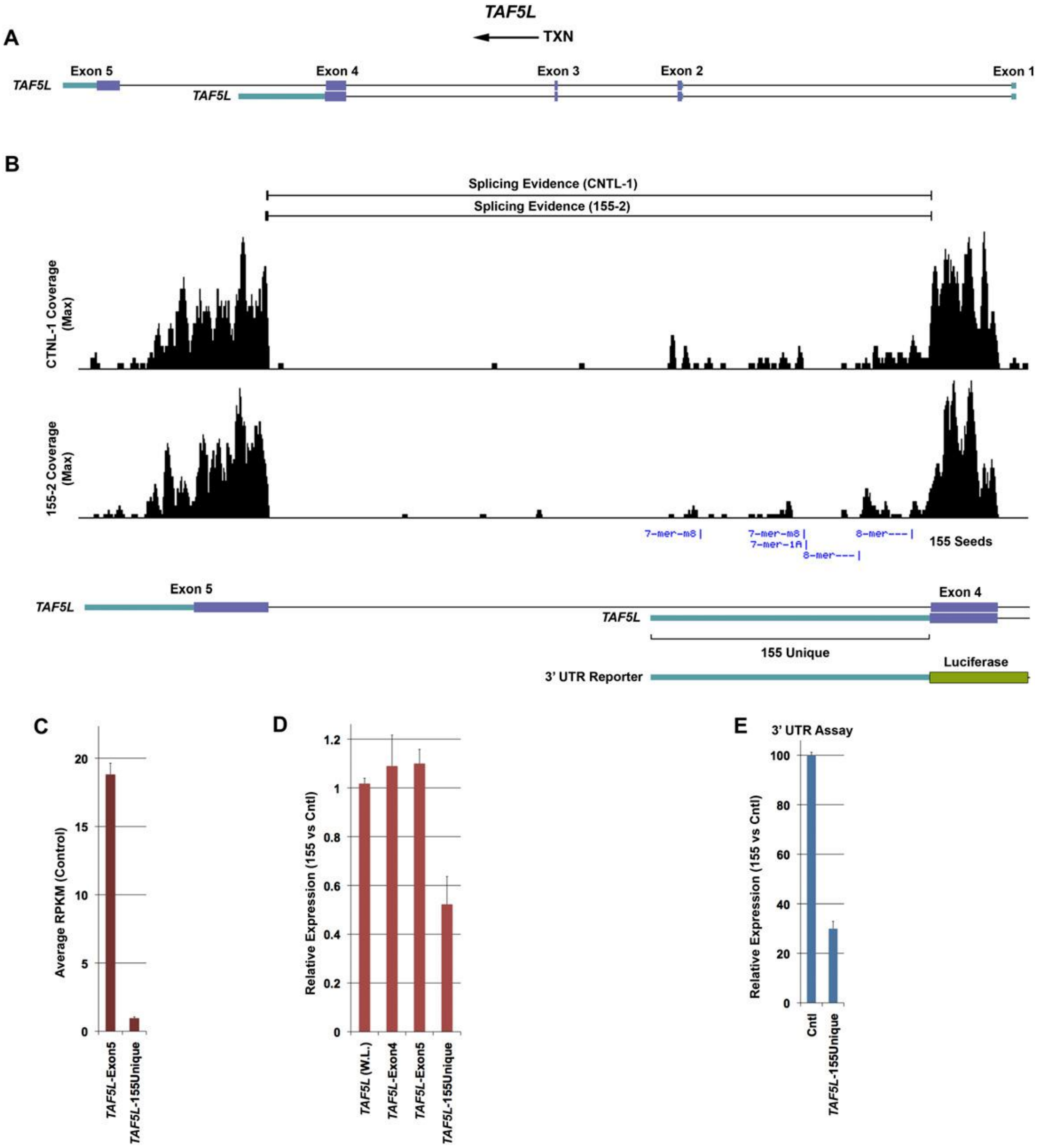


Figure 2.5. Isoform-dependent regulation of TAF5L by MIR155. (A) Gene structures of annotated TAF5L transcripts. (B) Expanded depiction of genomic sequences spanning exons 4 and 5 with accompanying read evidence. Splicing evidence indicates that reads spanning the exon 4/5 junction were found in control (CNTL-1) and MIR155 (155-2)–expressing cells. Pile-ups are a representation of the number of reads identified at each nucleotide position. (C) Quantitative analysis of read coverage (RPKMs) across exon 5 and the unique 3' UTR sequences containing MIR155 seed sequences. RPKM calculations were performed based on the number of reads in control Mutu I cells. (D) Relative expression of TAF5L whole locus (W.L.), TAF5L exons 4 and 5, and unique 3' UTR containing MIR155 seed sequences were calculated based on RPKM calculations performed using the appropriate respective feature annotations. Error bars are standard error of change. (E) 3' UTR assay showing suppression of 3' UTR containing MIR155 seed sequences. These results are extracted from the data shown in Figure 2.4.

The accumulation of publically available RNA-seq data for different cell lines and different model systems willlikely accelerate rapidly in the near future. We also envision that microRNA targetomes will soon become much better elucidated. Using informatics alone, RNA structure information derived from publically available NGS sequencing databases could be merged with a well-characterized targetome data set to allow investigators to make predictions regarding the cell type–restricted targetome. This will facilitate a much better informed prediction of the functional impact of a microRNA in the context of a particular cell system.

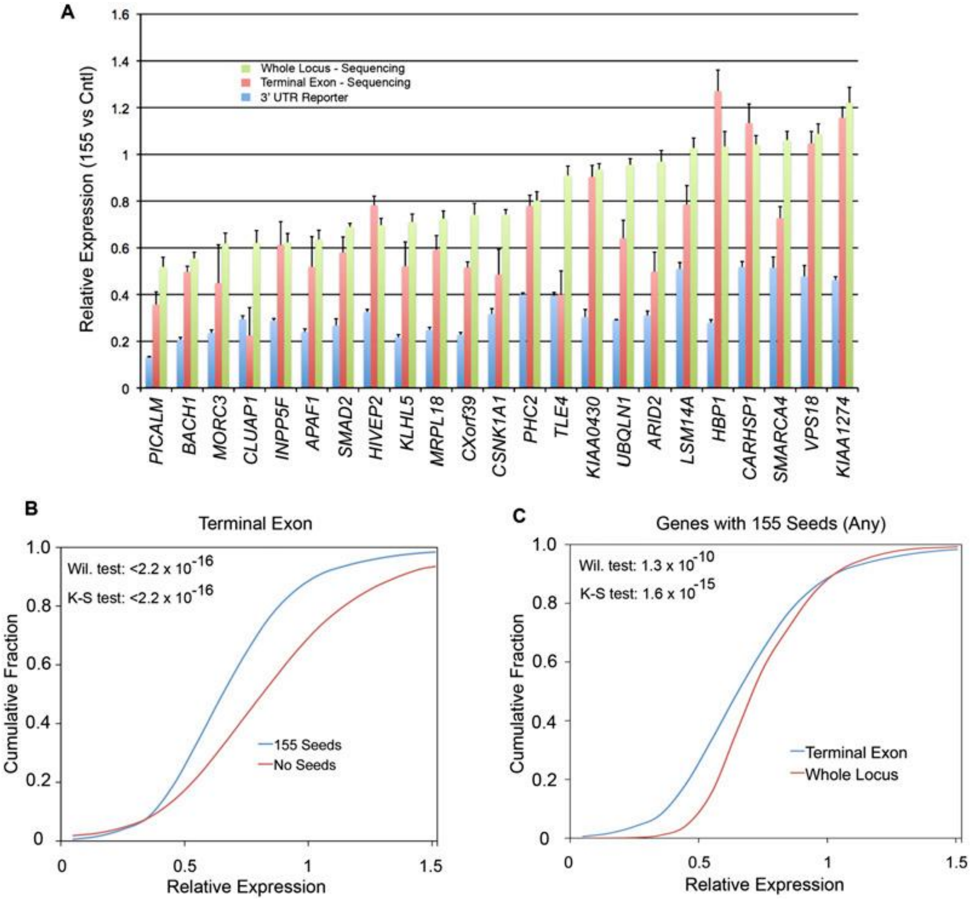


Figure 2.6. Mitigated response of endogenous genes to MIR155 is mediated in part by alternate transcript structure. (A) Comparison of relative expression at the whole locus level, the terminal exon level, and in 3’ UTR assays. The genes depicted are only those showing greater than 0.5 RPKMs in whole locus sequencing and in the terminal exon analysis. Error bars are standard error of change. (B) Cumulative distribution plot of terminal exon analysis shows a highly significant relative expression shift for genes with MIR155 seeds versus genes without seeds. P-values as determined by a one-sided Wilcoxon test (Wil.) and a one-sided Kologorov–Smirnov

test (K-S) were found to be $<2.2 \times 10^{16}$. (C) Terminal exon analysis shows significantly more down-regulation than whole locus analysis implicating transcript structure in mitigating MIR155 responses at the whole locus level. P-values were determined by a one-sided Wil. test and a one-sided K-S test.

Probes used for most nontiling microarrays are typically at or near the 39 end of terminal exons. Such probes exclusively interrogate longer isoforms that are expected on average, to be more susceptible to microRNA targeting. This may account for some of the down-regulated genes uniquely identified in the Agilent microarray platform relative to NGS (Figure 2.3). Despite the intrinsic bias that these microarrays have toward more regulatable isoforms, NGS identified 2.6 times more down-regulated genes with 8-mer seeds than our Agilent microarray data from the same cell system (FDR = 0) (Figure 2.3). NGS is therefore a more robust platform to identify an inferred targetome while providing a better reflection of overall gene expression since read frequency throughout the entire locus (or through the open reading frame [ORF]) can be considered. At the same time, NGS provides the flexibility to assess isoform-specific regulation by considering read frequency through isoformspecific exons or through isoform-specific 3’ UTR sequences only.

Identification of microRNA targets is critical for understanding the initial contact point between a microRNA and an affected pathway and biological function. Nevertheless, the overall biological impact of a microRNA is manifested by the combination of these direct interactions and downstream regulatory processes that are influenced by direct interactions. An illustration of this point was educed by Gene Ontology analysis (Ingenuity Pathway Analysis software; Ingenuity) of MIR155-regulated genes. The most highly implicated biological process predicted from a gene set containing either regulated genes with MIR155 seeds or any regulated gene was the cell cycle (which may contribute to the oncogenic function of MIR155) (Supplemental Figure 2.3). Despite this consistency, there are a far greater number of genes implicated in this process from the latter case (total of 92 genes) than the number of regulated genes with seeds (total of 10 genes). Many of these additional genes are likely regulated through downstream regulatory processes, yet they presumably have a contributing impact on this biological process.

To our knowledge, there is no reason to expect that indirect effects of microRNA targeting would be restricted to nontargets. The addition of an indirect component to direct targets could serve to either reinforce down-regulation, thereby exerting a stronger impact on a pathway, or provide a negative feedback loop to transiently influence a pathway. Both scenarios may explain discordant 3’ UTR reporter/NGS data that cannot be explained by altered transcript structure. KLRA1 and HAL (this study), and SMAD1 and MYO10 (see below) [Yin et al. 2010; this study], are examples of genes showing relative expression of 0.6 or less in 3’ UTR assays but whose relative expression at the endogenous RNA level was found to be ~50% of their 3’ UTR relative expression levels. These genes appear to be true targets of MIR155, but the greater observed regulation at the transcript level suggests a reinforcing component to the regulation of these genes.

Based in part on the RNA-seq data described here, we have recently validated that MIR155 inhibits bone morphogenetic protein signaling by targeting several mediators including SMAD1 (relative expression [RNA-seq] of 0.27, FDR = 0) and SMAD5 (relative expression [RNA-seq] of 0.48, FDR = 0), which are known to activate transcription of MYO10 (relative expression [RNA-seq] of 0.3, FDR = 0) [Yin et al. 2010]. Quantitative PCR analysis of all three of these genes and Western blot analysis of SMAD1 and SMAD5 showed inhibition by MIR155 in Mutu I cells [Yin et al. 2010]. The targeting of the MYO10 regulators, SMAD1 and SMAD5, in addition to the targeting of the MYO10 3’ UTR, illustrates a reinforcing mechanism that leads to greater suppression of MYO10 function. While we do not currently have similar regulatory information on KLRA1 or HAL, they too are candidate genes that may be suppressed through direct and indirect mechanisms to enforce inhibition of cellular processes influenced by these genes.

2.4 Materials and methods

2.4.1 Biological experiments

The sections related to cell culture, plasmid construction, generation of stable MIR 155 expressing cell lines, RNA isolation and real-time RT-PCR analysis of MIR 155 levels can be found in the original paper [Xu et.al 2010]. And the details of western blot analysis and soft agar assay are discussed in section 2.4.15 and 2.4.16 in original paper [Xu et.al 2010].

2.4.2 Sequencing and base calling

Preparation of transcription libraries for sequencing on the Illumina GA2x platform was carried out using the RNA-seq kit (Part no. 1004898 Rev. A) according to the manufacturer's standard protocol. Briefly, purified RNA was fragmented via incubation for 5 min at 94 ° C with the Illumina-supplied fragmentation buffer. The first strand of cDNA was next synthesized by reverse transcription using random oligo primers. Second-strand synthesis was conducted by incubation with RNase H and DNA polymerase I. The resulting double-stranded DNA fragments were subsequently endrepaired, and A-nucleotide overhangs were added by incubation with Taq Klenow lacking exonuclease activity. After the attachment of anchor sequences, fragments were PCR-amplified using Illumina-supplied primers and loaded onto the GA2x flow cell. Image analysis and base calling were conducted with Firecrest and Bustard programs, respectively, and initial sequence alignment for QC purposes was performed with Eland.

2.4.3 Read mapping to genome and across splice sites

Each short read file (sample), in the FASTQ format, was individually aligned against the Human Reference Genome (hg19) following a two-step procedure. The whole set of short reads was initially aligned to annotated exons using an exon aligner, Novoalign (http://www.novocraft.com). We used the following parameter settings to build novoindex and run Novoalign:

1. Novoindex -k 14 -s 1 an index file name (e.g., hg37) reference genome files name (e.g., chr1.fa chr2.fa chr3.fa chr4.fa chr5.fa chr6.fa chr7.fa chr8.fa chr9.fa chr10.fa chr11.fa chr12.fa chr13.fa chr14.fa chr15.fa chr16.fa chr17.fa chr18.fa chr19.fa chr20.fa chr21.fa chr22.fa chrX.fa chrY.fa chrM.fa), where: -k 14 is the k-mer length to be used for the index; and -s 1 is the step size for the index.

2. For searching the full human genome on a 16-GB RAM server, the recommended settings are k=14, s=1 or k =15, s =2.

3. Novoalign -o SAM -f short read data file name (e.g., s_1_sequence.txt) -d file location (e.g., /Volumes/Macintosh HD 2/workspace/hg37) > s_1_novo_output, where -o SAM: the output file is in the SAM format; -f: the following file name is the input FASTQ file; and -d: directory where the output file should locate (note: users need to change it to the actual directory).

4. Other options were set to the default.

The Initially UnMapped reads are likely originated from exon–exon junctions that do not exist in the reference genome. IUMs were aligned to de novo assembled exon–exon junctions using the junction mapper, TopHat [Trapnell et al. 2009]. We used the following parameter settings to run TopHat:

1. TopHat -p 6 -G reference genome file name (e.g., output.gff.txt) bowtieindex file name (e.g., bowtieindex/hg37_bowtie) short read data file name (e.g., s_1_sequence.txt): -p: Uses 6 threads to align reads; and -G: Supply TopHat with a list of gene model annotations in gff format (output.gff.txt).

2. Other options were set to the default.

2.4.4 Gene expression analysis using RNA-seq data

We performed a genome-wide gene expression analysis using gene annotation downloaded from the UCSC Genome Browser. The expression abundance for gene i was quantified using the RPKM measure: $10^9 \times C_i/(N_i \times L_i)$, where i is the gene index, C_i is the sum of short read counts mapped to exons and exon–exon junctions, N_i represents all mapped read counts in the lane, and L_i is the sum of exon lengths [Mortazavi et al. 2008].

RPKM calculation was performed using an in-house-developed graphical user interface (GUI) software, SAMMate, which is freely available from http://sammate.sourceforge.net/. SAMMate takes the inputs of exon alignment files in .SAM format [Li et al. 2009], exon–exon junction alignment file in .BED format (optional), and genome annotation in a variety of formats to export a matrix of RPKM values for annotated genes. In addition, it also calculates RPKM values for each customized genomic interval.

2.4.5 Analysis of Agilent MIR155 arrays

Agilent microarray data for MIR155 transduced and control Mutu I samples were imported into the R software environment, version 2.7 (The R Development Core Team 2008), using the Bioconductor package ‘‘limma’’ [Gentleman et al. 2004; Smyth 2005]. Quality control was performed using limma functions and independently written R scripts. Within-array normalization was performed using eCADS [Dabney and Storey 2007] to correct for gene-specific dye bias. Normalized log-ratios were then analyzed using the one-class framework in SAM [Tusher et al. 2001] to identify probes differentially expressed with an estimated FDR of 0.

2.4.6 Statistical treatment of sequencing and microarray data

We performed differential expression analysis for the NGS data set (this study) and two previously published microarray data sets—mouse macrophage [O’Connell et al. 2008] and Human 293 cell line [Skalsky et al. 2007]. Differentially expressed genes were identified by significance analysis of microarrays (SAM) [Tusher et al. 2001]. SAM is a statistical technique in which significantly differentially expressed genes between control and miRNA-155 transfected cell lines can be identified by assimilating a set of gene specific t-tests. Briefly, SAM computes a nonparametric score for each gene by dividing the between-group difference of (normalized log) gene expression levels and adjusted within-group gene expression variance across the whole genome. The score is then compared with random permutation scores that are computed in the same manner as the original score but based on randomly sampled gene expressions. The per-gene P-value was calculated by the percentage of scores that are larger than the original score in a fixed number of simulations, say, 1000. The per-gene P-values were further adjusted to the false discovery rate (FDR) [Storey and Tibshirani 2003], indicating the percentage of genes identified as being significant by chance alone. Here, we have used SAM’s two class analysis function to call significantly differentially expressed genes with an FDR of 0.

2.4.7 Cross-species gene mapping

Cross-species gene mapping between human and mouse gene orthologs was done using the mapping file downloaded from http://www.informatics.jax.org/mgihome/projects/aboutmgi.shtml.

2.4.8 Cross-platform comparison of targetome prediction using bitmap

To systematically compare the technical capabilities of NGS and microarray platforms in detecting down-regulated genes, we used only gene identifiers that are common to all four platforms. Among these gene identifiers, we identified a down-regulated gene list with FDR = 0 for each platform. Each of the genes was determined to be significantly down-regulated (at FDR = 0) or not in each of the four data sets; all genes were assigned to one of the $2^4 = 16$ possible clusters, which is represented by color/white patterns and corresponding to 16 rows in the bitmap. The bitmap can be viewed as a generalized Venn Diagram to compare more than three groups. For example, the eighth row corresponds to 1859 (total) and 102 (8-mer seed containing) significantly downregulated genes identified at an FDR = 0 in NGS data.

2.4.9 Statistical comparison of a pair of populations

We used a rank-based two-sample one-sided Wilcoxon test to test the equality of two population means (H_0) versus one is greater than another (H_α). The R function wilcox.test() was used to perform this test.

We used Levene's test to test the equality of two population variances (H_0) versus not equal (H_α). The R function levene.test() was used to perform this test. We also used a Kolmogorov–Smirnov test to test the overall equality of two populations (H_0) versus not equal

(H_α). The R function ks.test() was used to perform this test. In all the above tests, small P-values (e.g., < 0.05) will reject the H_0.

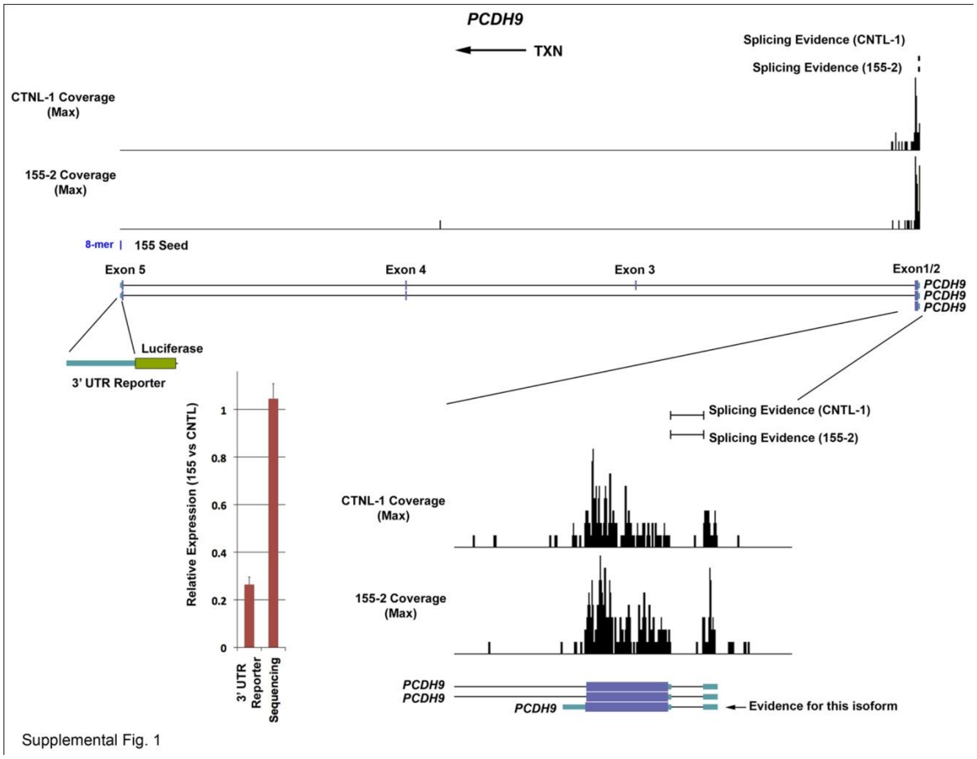
2.4.10 3' UTR-luciferase reporter analysis

For reporter analysis, 3.75 μg of either the control (pMSCV-puro-GFP-miR-CNTL) or MIR155 (pMSCV-puro-GFP-MIR155) expression vector was cotransfected with 0.25 μg of the appropriate pMIR-REPORT-dCMV or pGL4.11 3' UTR reporter plasmid into 1×10^6 Mutu I cells using Lipofectamine 2000 (Invitrogen). Cells were harvested 48 h post-transfection and analyzed using the Promega firefly luciferase assay. The values reported are the expression change of a given 3' UTR relative to change in the control reporter.

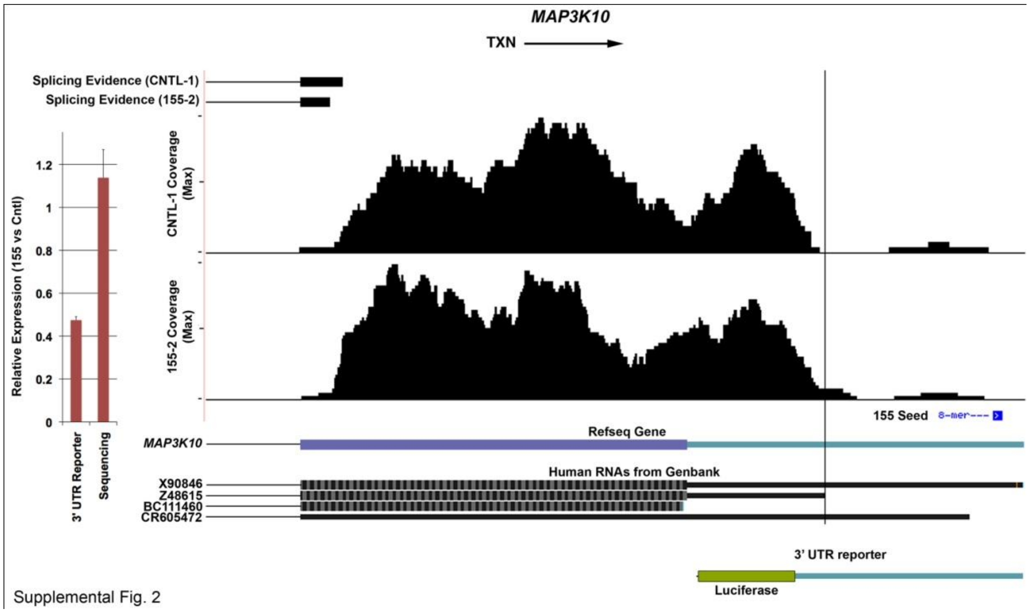
2.4.11 Accession numbers

Sequencing data will be available in the NCBI Short Read Archive, SRA011001. Microarray data submission to NCBI GEO database is in process.

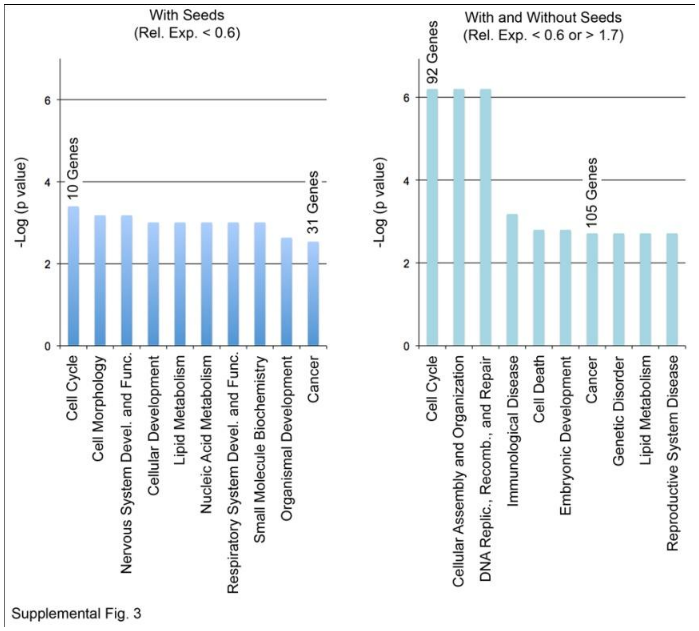
2.5 Supplemental Figures



Supplemental Figure 2.1. Lack of read evidence for MIR155 seed containing isoform of PCDH9. Schematic representation of PCDH9 locus with read and splicing evidence is shown in upper panel. Transcription (TXN) proceeds leftward. The reporter used for the 3' UTR analysis contains the indicated 8-mer seed in the 3' UTR from exon 5. No reads were mapped to exons 3, 4 or 5 but pileups are observed at exons 1 and 2 supporting expression of the third (short) isoform only.



Supplemental Figure 2.2. Read evidence for the last exon of MAP3K10 shows a sharp drop in mapped reads upstream from the MIR155 seed site. The read dropoff corresponds to approximately the end of a transcript supported by Genbank evidence (Z48615).



Supplemental Figure 2.3. Top biological functions implicated by Ingenuity Pathways Assessment (IPA) analysis of regulated genes. Left panel shows the top 10 implicated biological functions identified and their associated p values from list of genes with MIR155 seeds with relative expression values of less than 0.6. The right panel shows the top 10 implicated biological functions identified and their associated p values from list of genes with or without MIR155 seeds with relative expression values of less than 0.6 (log_2 = -0.737) or greater than 1.7 (log_2= 0.737). Y axis shows $-log_{10}$ of p value. The number of genes in the overlapping biological functions, cell cycle and cancer, are indicated.

Chapter 3. RNA-seq Analysis of EBV Transcriptome

3.1 Background

Microarrays have been used to assess the levels of EBV gene expression in experimental and clinical settings [Bernasconi et al. 2006; Li et al. 2006; Yuan et al. 2006; Zhang et al. 2006; Zheng et al. 2008]. Nevertheless, this analysis typically requires the use of custom arrays with user specified probes against each EBV gene of interest. Despite their utility, microarrays have a limited dynamic range, being limited at the low end by the level of background and limited at the high end by signal saturation. Further, the accuracy of microarray data can be a concern because of chip defects, cross hybridization, and the analog nature of the approach. Accordingly, back-to-back comparisons of microarray and RNA-seq data have demonstrated the enhanced performance of RNA-seq in the quantitative assessment of cellular transcripts [Marioni et al. 2008; Mortazavi et al. 2009; Xu et al. 2010]. Using RNA-seq, transcript structure information can also be deduced from a relatively unbiased data set whereas transcript structure information derived from tiling microarrays is dependent on the probe design and is therefore subject to investigator biases.

Due to the perceived potential of RNA-seq in transcriptome analysis, there has been intense interest in the development of informatics approaches to analyze cellular transcriptomes [Costa et al. 2010]. For the most part, these approaches should be directly applicable to the analysis of viral transcriptomes. Nevertheless, the appropriately formatted annotation files for viruses or other ectopic organisms and the incorporation of this annotation information into existing pipelines have been lacking. We have created the necessary annotation files for EBV and merged them with annotation files for the human cellular genome so that EBV specific transcript data can be generated in the context of cellular data. This pipeline allows for the simultaneous analysis of cellular and viral transcriptomes, the digital quantification of EBV transcripts, and the visualization of EBV specific reads and splice junctions on a genome browser (see supplemental file S1 for pipeline details). Using this approach, we have analyzed EBV transcriptomes for the EBV-positive Burkitt's lymphoma type I latency cell lines, Mutu I and Akata.

3.2 Analysis of EBV gene expression in Mutu I and Akata cells

Sequencing data used for the Mutu I analysis were control samples from a previous study where we assessed miR-155 mediated cellular transcriptome changes [Xu et al. 2010] (available from the National Center for Biotechnology Information Sequence Read Archive (SRA011001)). For this study, two separate control RNA preps were generated and single-end 50 base technical replicates were run for each poly(A)+ selected RNA. Akata sequencing data was generated anew from whole cell RNA prepared using a miRNeasy kit (Qiagen) according to the vendor's protocol. Akata sequencing libraries were generated using the Illumina RNA-seq kit (Part #1004898) and run on a GA2x machine for single-end 74 base extensions (deposition to National Center for Biotechnology Information Sequence Read Archive is in progress). Sequences were simultaneously aligned to all human chromosomes plus the EBV genome (AG876 strain [Dolan et al. 2006], GenBank DQ279927) (see supplemental file S1 for general and detailed pipeline information). Reads per kilobase of exon per million mapped reads (RPKM – a measure of relative gene expression) values for all genes were calculated using SAMMate (http://sammate.sourceforge.net).

Figure 3.1A shows the sequence read distribution across the entire EBV genome for Mutu I and Akata RNAs (an expanded/high resolution view can be seen in supplemental file S2).

Ample read evidence is observed across the majority of the EBV genome. Despite this, however, the relatively few intergenic regions that exist within the EBV genome tend to lack reads (for example, see Figure 3.1B), supporting the contention that possible contaminating DNA does not represent a major source of read evidence. The abundance of reads across all latency genes were relatively low and consistent with these cell lines exhibiting type I latency, no reads mapped to the EBNA2 open reading frame (Figure 3.1C and Figure 3.2).

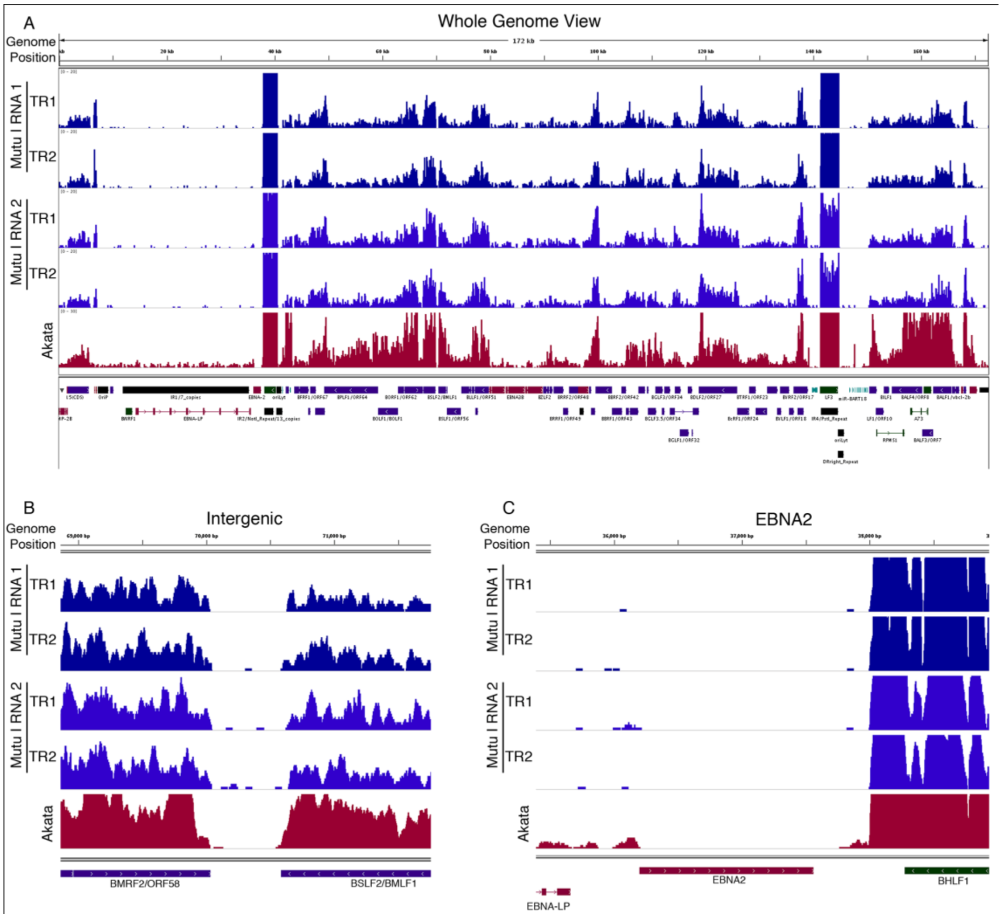
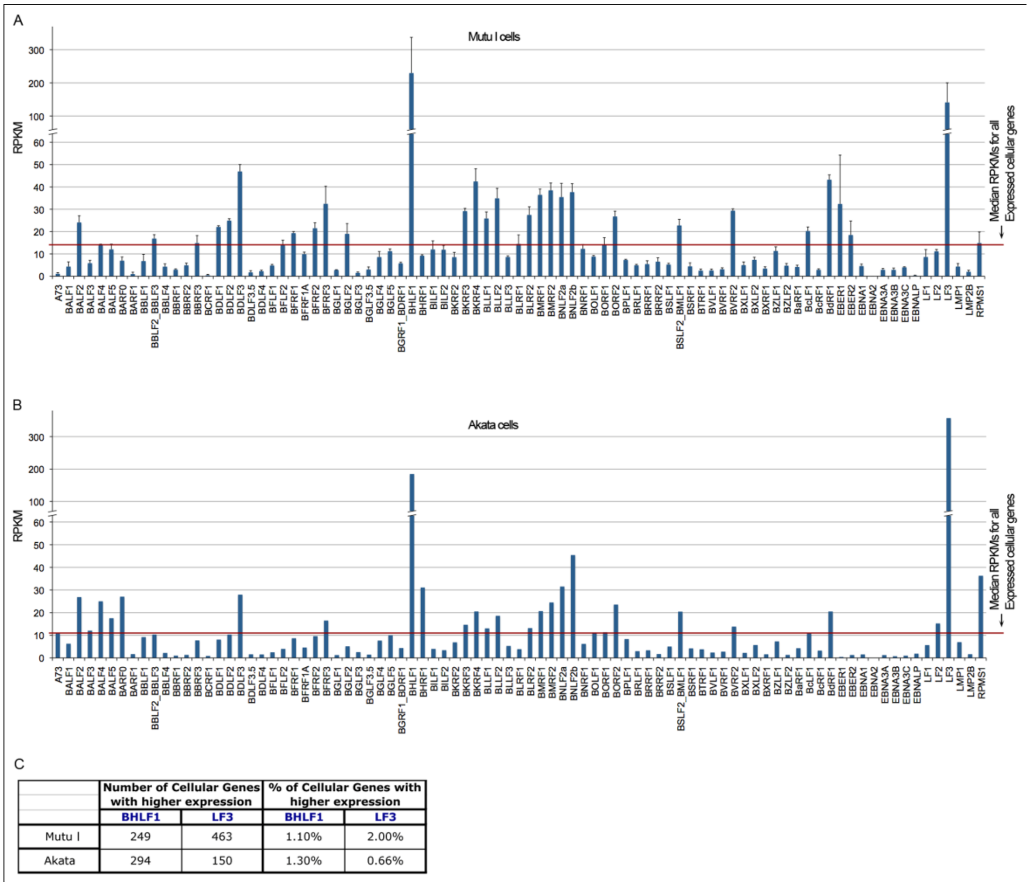


Figure 3.1. Visualization of RNA-seq coverage across the EBV genome. Coverage (Wiggle) files generated from SAMMate and the EBV annotation file were loaded onto the Integrated Genome Viewer (IGV - developed at the Broad Institute (www.broadinstitute.org)). The Y axis shows the number of reads mapping to each location of the genome. Panel A shows the whole genome view, panel B shows a zoomed view of the intergenic region between the BMRF2 and BSLF2 genes, and panel C shows the lack of reads corresponding to the EBNA2 locus. Data range for coverage data was set to 20 (for Mutu I) and 30 (for Akata) meaning that maximal

peaks represent genomic positions where there were at least 20 or 30 reads that crossed that genomic position.

In contrast to the low levels of latency gene expression observed, we were surprised by the robust levels of many of the lytic genes in both Mutu I and Akata cells (Figures. 1 and 2). Many of these lytic genes show expression that is well above the median for all expressed cellular genes (median RPKM = 14.1 (Mutu I) and = 10.9 (Akata) - calculated as the median RPKM of genes with greater than 1 RPKM (1 RPKM typically represents approximately 1 transcript per cell [Li et al. 2009]) (Figure 3.2A and B). Strikingly, the BHLF1 and LF3 transcripts are represented at such high levels that only between 0.66 to 2% of all annotated poly(A)+ cellular genes are expressed at higher levels in Mutu I and Akata cells (Figure 3.2C). The expression values observed here for EBV genes are not due to background since we ran RNA-seq data from the EBV negative cell lines, A549 [Reddy et al. 2009] and MCF7 [Wang et al. 2008] through our pipeline and obtained no alignments to the EBV genome (Figure 3.3). The substantial average expression levels observed here for some lytic genes could arise from either their expression in latency and/or from very high expression in a small proportion of cells that are actively undergoing lytic replication. The latter scenario most likely explains sequences obtained for most of these lytic genes. Nevertheless, it is intriguing to speculate that the former scenario may account for at least some of these genes. For example, BHLF1 and LF3 transcripts have been shown to be derived from multiple promoters, some of which are induced upon reactivation and others of which are constitutive [Gao et al. 1997; Xue et al. 2007]. The high transcript levels that we observed under non-induced conditions suggest that these genes may play a role in the latent phase of the EBV life cycle. Overall, these data illustrate the sensitivity of RNA-seq for assessing transcript levels. Further, the BHLF1 and LF3 examples described here illustrate how the digital nature of RNA-seq allows the user to compare the abundance of transcripts from one gene with the abundance of transcripts of other genes within the transcriptome.



	Number of Cellular Genes with higher expression		% of Cellular Genes with higher expression	
	BHLF1	LF3	BHLF1	LF3
Mutu I	249	463	1.10%	2.00%
Akata	294	150	1.30%	0.66%

Figure 3.2. RPKM values for EBV genes in Mutu I (A) and Akata (B) cells. Mutu I results are the average of two technical replicates (TR) from each of two separate RNA preps. Error bars indicate standard deviation for each gene. C) The number and percentage of genes showing higher RPKM values than BHLF1 and LF3 in Mutu I and Akata cells out of a total of 22,803 annotated cellular and viral genes.

Notably, despite carrying out poly(A)+ RNA selection prior to sequencing, we still detect the expression of non-polyadenylated transcripts such as the EBERs in Mutu I cells (Figure 3.2). However, we note that the error for non-polyadenylated transcripts tend to be high, probably due to differences in the efficacy of poly(A)+ RNA selection between the two biological replicates in

Mutu I cells. Only low levels of EBERs were detected in Akata cells indicating that the poly(A)+ RNA selection was more effective in our newest RNA-seq experiment.

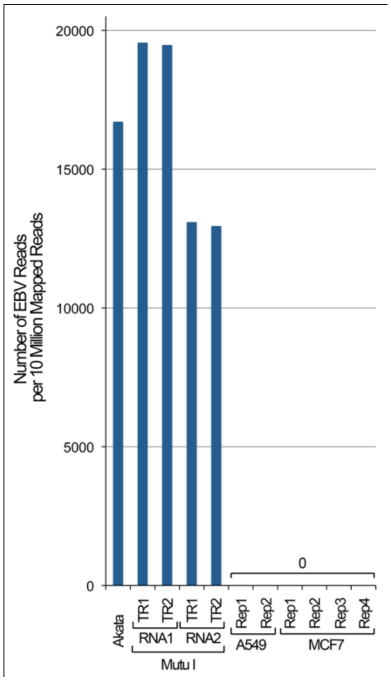


Figure 3.3. Illustration of specificity for RNA-seq in assessing EBV transcriptomes. The total number of reads that mapped to the EBV genome per 10 million mapped reads from the EBV positive cell lines, Akata and Mutu I, and the EBV negative cell lines, A549 and MCF7. No EBV specific reads were identified in either of the EBV negative cell lines. RNA1 and RNA2 refer to biological replicate RNA samples from Mutu I cells. TR1 and TR2 refer to technical sequencing replicates.

3.3 Splicing evidence in Mutu I and Akata

While RNA-seq can provide digital quantification of gene expression, reads that span exon junctions can provide information about gene isoform usage. We used the junction mapper, Tophat [Trapnell et al. 2009], to identify junction mapped reads throughout the EBV genome (see supplemental file S1) for Mutu I and Akata. While no evidence of Cp or Wp derived EBNA1 transcripts was found, evidence for Qp derived EBNA1 splice junctions was observed in both Mutu I and Akata cells (Figure 3.4A). Junction reads were also detected for EBV lytic genes in both Mutu I and Akata cells including junction reads for both BZLF1 (Figure 3.4B) and

BSLF2/BMLF1 (supplemental file S3). Further, evidence for multiple isoform expression (i.e. alternative splicing events) was detected for many genes such as BLLF1/BLLF2 (Figure 3.4C) as well as the complex BamHI A region [Edwards et al. 2008; Reddy et al. 2009] (supplemental file S4). Within the BamHI A region (supplemental file S4), for example, there is evidence for alternative splicing at the A73 gene in both Akata and in Mutu I with JUNC00000180 from Mutu I cells providing evidence of exon skipping (skipping of exons 2 and 3). Within the genomic regions spanning the two BART microRNA clusters, there are very few reads, consistent with these microRNAs being produced from excised introns that are presumably unstable and non-polyadenylated (and therefore not enriched during our poly(A)+ selection procedure). In both Mutu I and Akata, there is evidence for two large introns that span the entire region of both of these clusters of microRNAs (JUNC00000094 and JUNC00000178 in Mutu I and JUNC00000053 and JUNC00000084 in Akata). Consistent with this junction evidence, there are pronounced read spikes in Akata cells immediately upstream from the first junction (centered at position 139,270), between these two junctions (centered at position 147,770), and immediately downstream from the second junction (centered at position 151,115) (supplemental file S4) supporting the idea that a stable, poly(A)+ spliced transcript is generated from this transcription unit. The two introns excised from this transcript can conceivably give rise to all BART microRNAs within these two clusters.

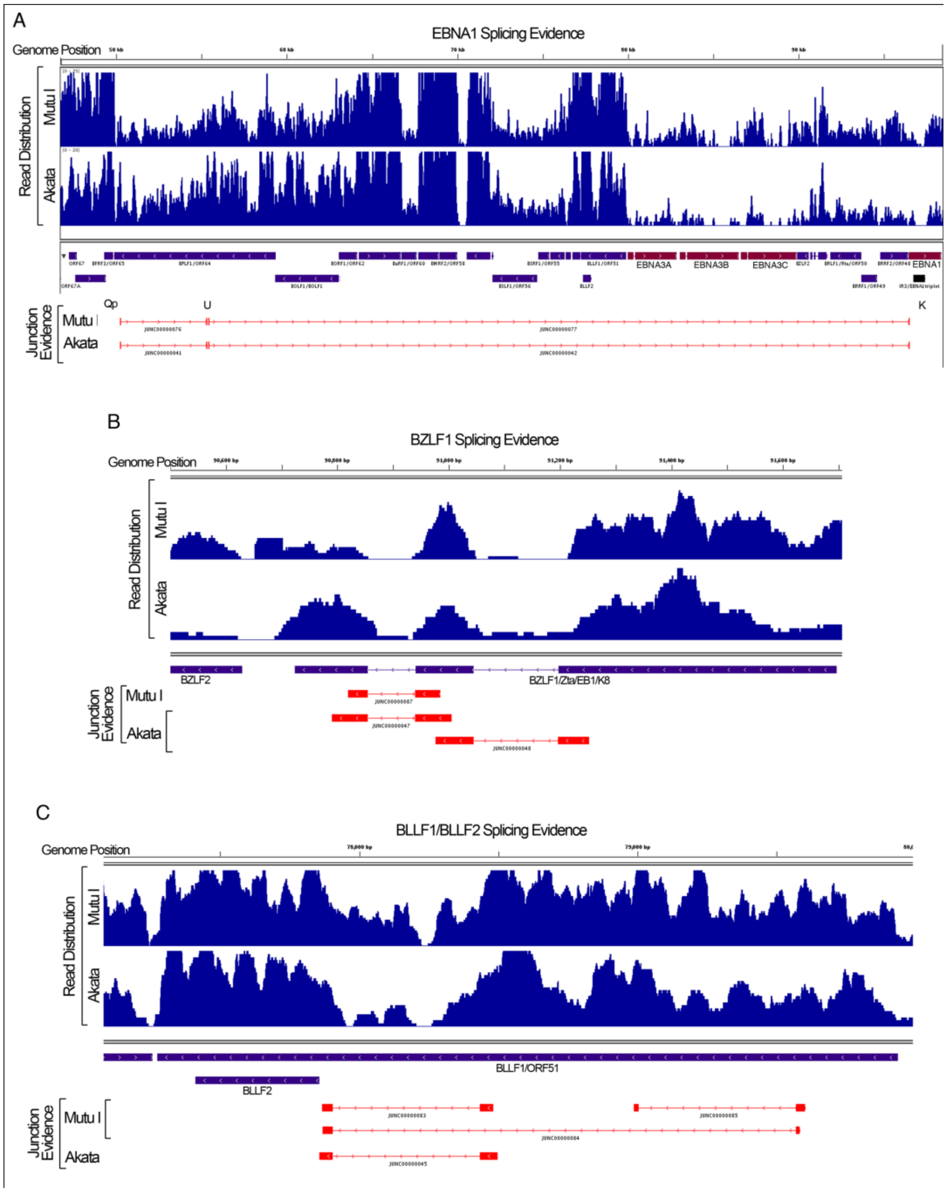


Figure 3.4. Visualization of junction evidence for EBNA1 (A), BZLF1 (B), and BLLF1/BLLF2 (C). Junction (BED) files were generated by the junction mapper, Tophat as outlined in supplemental file S1.

3.4 Conclusions

Our results show robust detection of EBV derived transcripts by RNA-seq using the pipeline outlined here (supplemental file S1). From a quantitative standpoint, several studies have shown this approach to outperform microarrays since it is more accurate [Marioni et al. 2008; Mortazavi et al. 2008; Xu et al. 2010] and since there is an inherently broad dynamic range.

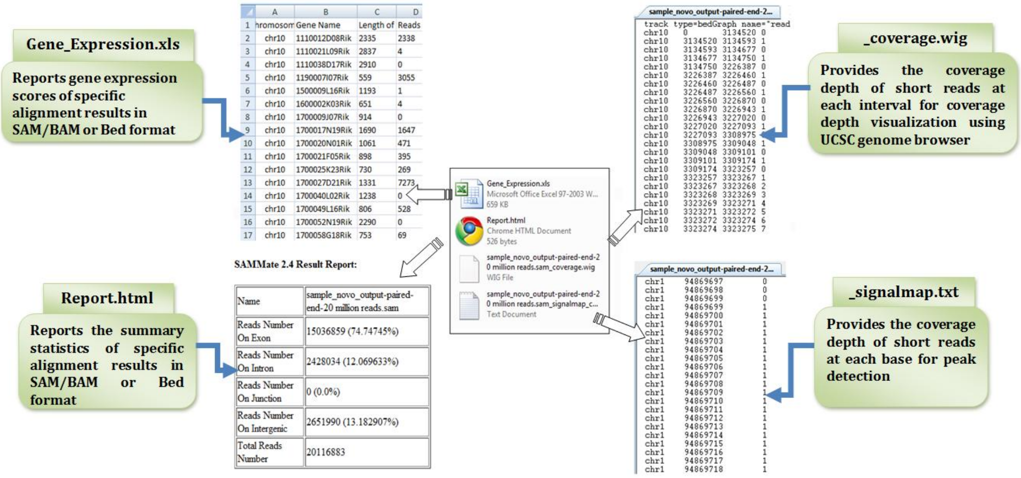
For example, a previous report [Bernasconi et al. 2006] documented the difficulty in attaining confidence in detecting most EBV derived transcripts using microarrays because of low signal-to-noise ratios. Here, we show excellent coverage of the bulk of EBV genes (including lytic genes) in predominantly latently infected cell lines while at the same time detecting no EBV specific reads in two EBV-negative cell lines. The digital nature of RNA-seq allows the user to better compare the relative expression of distinct genes through the calculation of RPKMs. This allowed us to determine that BHLF1 and LF3 are among the most abundant genes expressed even in predominantly latently infected cell populations. Lastly, RNA-seq inherently contains splice junction information that can be readily exploited to garner viral isoform expression patterns.

Our approach can also be readily applied to other viruses by manual conversion of the respective annotation information (generally available at the National Center for Biotechnology Information database) to the appropriate format and its subsequent conjugation to cellular annotation files. This should result in an improvement over microarrays in the analysis of virus-associated transcriptomes not only for EBV but for other viruses.

Chapter 4. SAMMate GUI software system

4.1 Introduction

In chapter 2 and 3, we introduced transcriptome and targetome analysis using RNA-seq. A unique challenge for working with RNA-seq data is to extract useful information from short reads alignments stored in SAM/BAM format. Even though there are some software can process SAM/BAM file, most of them are not very convenient for researchers because they are using in command-line interface. In this section we introduce SAMMate which is a GUI software tool that allows biomedical researchers to easily access essential information stored in SAM/BAM files. A detailed documentation and a quick walkthrough are available at SAMMate's homepage [http://sammate.sourceforge.net]. SAMMate possesses the following key features (Figure 4.1): (1) For RNA-seq alignment SAMMate uses short reads originating from both exons and exon-exon junctions to calculate gene expression scores. SAMMate's versatility allows biomedical researchers to combine the output from an exon alignment program, such as Novoalign [http://www.novocraft.com/], with the output of a splice junction analysis program, such as Tophat [Trapnell et al. 2009]. This intuitive combination results in a more accurate estimation of gene expression abundance scores. (2). Using SAM/BAM files generated from short read alignments, SAMMate implements an efficient and fast algorithm to calculate a base-wise signal map. For large SAM/BAM files with more than 2.0×10^7 records, it only takes approximately one minute on a standard desktop or laptop computer to generate the base-wise signal map. (3). SAMMate also exports a wiggle file for visualization of alignment results on the UCSC genome browser. (4). Lastly, SAMMate exports an alignment statistics report. In addition, SAMMate has nice utilities for manipulating SAM/BAM files that include merging and sorting. We have designed several case studies in the context of miRNA-155 target prediction to demonstrate the key features of SAMMate since these features are essential for solving a wide range of biological problems using NGS data.


Gene_Expression.xls
Reports gene expression scores of specific alignment results in SAM/BAM or Bed format
Report.html
Reports the summary statistics of specific alignment results in SAM/BAM or Bed format
_coverage.wig
Provides the coverage depth of short reads at each interval for coverage depth visualization using UCSC genome browser
_signalmap.txt
Provides the coverage depth of short reads at each base for peak detection

	A	B	C	D
1	hromosom	Gene Name	Length of	Reads
2	chr10	1110012D08Rik	2335	2338
3	chr10	1110021L09Rik	2837	4
4	chr10	1110038D17Rik	2910	0
5	chr10	1190007I07Rik	559	3055
6	chr10	1500009L16Rik	1193	1
7	chr10	1600002K03Rik	651	4
8	chr10	1700009J07Rik	914	0
9	chr10	1700017N19Rik	1690	1647
10	chr10	1700020N01Rik	1061	471
11	chr10	1700021F05Rik	898	395
12	chr10	1700025K23Rik	730	269
13	chr10	1700027D21Rik	1331	7273
14	chr10	1700040L02Rik	1238	0
15	chr10	1700049L16Rik	806	528
16	chr10	1700052N19Rik	2290	0
17	chr10	1700058G18Rik	753	69

SAMMate 2.4 Result Report:

Name	sample_novo_output-paired-end-20 million reads.sam
Reads Number On Exon	15036859 (74.74745%)
Reads Number On Intron	2428034 (12.069633%)
Reads Number On Junction	0 (0.0%)
Reads Number On Intergenic	2651990 (13.182907%)
Total Reads Number	20116883

Gene_Expression.xls
Microsoft Office Excel 97-2003 W...
659 KB

Report.html
Chrome HTML Document
526 bytes

sample_novo_output-paired-end-20 million reads.sam_coverage.wig
WIG File

sample_novo_output-paired-end-20 million reads.sam_signalmap_c...
Text Document

sample_novo_output-paired-end-2...

```
track type=bedGraph name="read
chr10   0       3134520 0
chr10   3134520 3134593 1
chr10   3134593 3134677 0
chr10   3134677 3134750 1
chr10   3134750 3226387 0
chr10   3226387 3226460 1
chr10   3226460 3226487 0
chr10   3226487 3226560 1
chr10   3226560 3226870 0
chr10   3226870 3226943 1
chr10   3226943 3227020 0
chr10   3227020 3227093 1
chr10   3227093 3308975 [illegible]
chr10   3308975 3309048 1
chr10   3309048 3309101 0
chr10   3309101 3309174 1
chr10   3309174 3323257 0
chr10   3323257 3323267 1
chr10   3323267 3323268 2
chr10   3323268 3323269 3
chr10   3323269 3323271 4
chr10   3323271 3323272 5
chr10   3323272 3323274 6
chr10   3323274 3323275 7
```

sample_novo_output-paired-end-2...

```
chr1    94869697        0
chr1    94869698        0
chr1    94869699        0
chr1    94869699        1
chr1    94869700        1
chr1    94869701        1
chr1    94869702        1
chr1    94869703        1
chr1    94869704        1
chr1    94869705        1
chr1    94869706        1
chr1    94869707        1
chr1    94869708        1
chr1    94869709        1
chr1    94869710        1
chr1    94869711        1
chr1    94869712        1
chr1    94869713        1
chr1    94869714        1
chr1    94869715        1
chr1    94869716        1
chr1    94869717        1
chr1    94869718        1
```

Figure 4.1. Key features of SAMMate: A schematic diagram of the four key features of SAMMate. (1) Fast calculation of gene expression RPKM scores combining exon reads and junction reads. (2) Fast calculation of whole-genome base-wise signal map. (3) Generation of wiggle files for short read alignment visualization. (4) Generation of an alignment statistics report.

4.2 Key features

4.2.1 Key feature: calculating genomic feature abundance scores

Ideally, transcriptome characterization and quantification should be done on the isoform level. However, many existing approaches that quantify transcript abundance on the isoform level depend upon stringent assumptions such as a *priori* known isoform structures and suffer from identifiability problems. Moreover, the accuracy of these algorithms for high throughput studies is in doubt. It is also not known how sensitive these algorithms are to error-prone isoform annotation databases. Many existing approaches [Li et al. 2009; Jiang et al. 2009; Zheng et al. 2009] have proved their merit as pilot studies. However, they were only validated using RT-PCR for a limited number of genes with simple and identifiable isoform transcript structures. With the aforementioned in mind, transcript abundance quantification on the gene level remains one of the most demanded outputs from high throughput molecular profiling experiments such as microarray and NGS. The latter platform is much more sensitive in detecting low level gene expression and provides a much broader dynamic range of expression quantification. Taking gene expression profiling using Illumina Genome Analyzer as an example, Read Per Kilobase of exon model per million Mapped reads (RPKM) was used to score gene expression abundance. The values obtained can be interpreted as the number of copies of each transcript in the living cell where the average length of transcripts is 1KB [Mortazavi et al. 2008]. The RPKM scores can range from < 0.01 to > 10,000. There are now unprecedented and unparalleled opportunities to detect novel transcripts with ultra-low or ultra-high abundance.

A unique challenge for researchers working with RNA-seq data are short reads originating from exon-exon junctions in cDNA (around 10%). These short reads fail to map back to the reference genome since the exons are separated by introns (Figure 4.2a). The millions of unmapped short reads originating from exon-exon junctions, denoted as Initially Unmapped Reads (IUM's), need to be accounted for when calculating RPKM scores [Trapnell et al. 2009]. Unfortunately, most alignment tools only map the short reads originating from exons completely ignoring IUMs in the process. Hereinafter, we denote such aligners as "exon aligner". To address the limitations of "exon aligners", ERANGE [Mortazavi et al. 2008], Tophat [Trapnell et al. 2009] and rSeq [Jiang et al. 2008] are among the recently developed approaches to map IUM's originating from exon-exon junctions back to individual genes. ERANGE uses a union of known and novel junctions while Tophat de novo assembles IUM's using a module in Maq [Li et al. 2008]. Hereinafter, we denote an aligner of this type as "junction mapper". Thus, there are now two types of aligners that complement each other.

Performance-wise, aligners vary vastly in accuracy as well as the underlying algorithms used. It is highly desirable for RNA-seq data analysis to allow users the freedom to choose and combine a pair of their favorite exon aligner and junction mapper to estimate gene expression scores. SAMMate fulfills this role by calculating and exporting a gene expression score matrix using a user-defined combination of an exon aligner and a junction mapper (Figure 4.2b). SAMMate then calculates the gene expression RPKM or FPKM score for gene i, R as $R_i = \frac{10^9(C_i^A+C_i^B)}{NL_i}$; where i represents the gene index. C_i^A is the short read counts uniquely mapped to exons using an exon aligner (e.g. Novoalign), and C_i^B is the IUM short read counts uniquely mapped to the exon-exon junctions using a junction mapper (e.g. Tophat). N represents all uniquely mapped read counts in a cell extract sample, and L_i is the summation of the exon lengths. Thus, SAMMate combines short reads mapped to exons (e.g. available in SAM/BAM format) and to exon-exon junctions (e.g. available in BED format) to accurately estimate gene expression scores (Figure 4.2b).

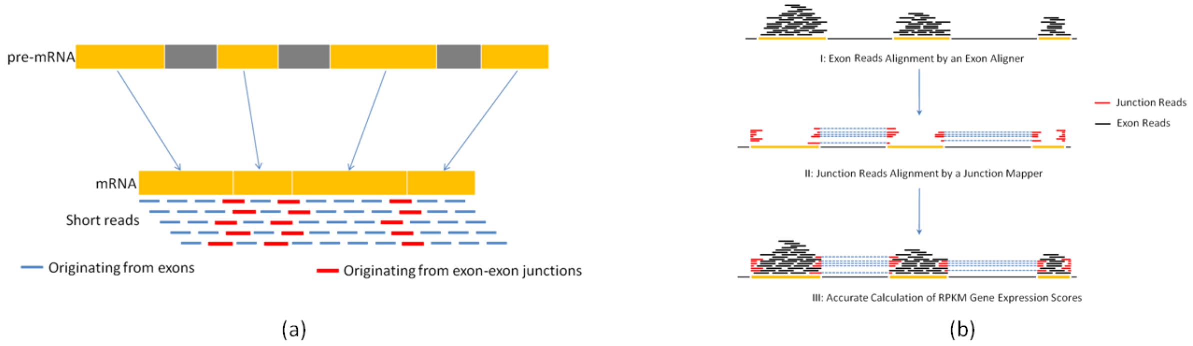

pre-mRNA
mRNA
Short reads
Originating from exons
Originating from exon-exon junctions
(a)
I: Exon Reads Alignment by an Exon Aligner
Junction Reads
Exon Reads
II: Junction Reads Alignment by a Junction Mapper
III: Accurate Calculation of RPKM Gene Expression Scores
(b)

Figure 4.2. Combination of exon reads with junction reads to accurately calculate gene expression RPKM scores (a) A unique challenge for researchers working with RNA-seq data. The junction reads (red) fail to map back to the reference genome because exons are separated by introns. (b) A demonstration of the ideas of combing exon reads (black) and junction reads (red) to calculate gene expression RPKM scores.

SAMMate can also take many pairs of SAM(BAM)/BED files simultaneously, one for each cell sample, to calculate a Microsoft EXCEL compatible gene expression matrix. In this matrix rows correspond to genes or the customized genome coordinate intervals, and columns correspond to different cell samples. It must be noted that SAMMate is more flexible and accurate than other software, such as Tophat [Trapnell et al. 2009], that also export the gene expression scores. We validate our claim using experimental data obtained from 3' UTR assay as a case study shown below. SAMMate's reporting utility for gene expression abundance score is also quite versatile as this utility is not limited to the annotated genes. In fact, SAMMate calculates genomic feature abundance scores for any user-defined genomic intervals. This utility dramatically simplifies the technical barriers for discovering novel genes.

4.2.2 Key feature: generating signal map for peak detection

A signal map is a frequently demanded data format for NGS data analysis. In a signal map file, alignment results are represented in the per-base "pileup" format. In this format the single nucleotide short read coverage depth is calculated whereas the whole genome coverage is provided as a vector of integers with length 3.2×10^9. A signal map is a common input for a number of frequently performed sequential analyses to detect a wide range of genomic features. For ChIP-seq and Methyl-seq data, significant peaks in a signal map may indicate potential transcription factor binding sites and DNA methylation sites, respectively. For DNA-seq data, significant change points in the signal map might indicate a true copy number change, which is often a hallmark of cancer [Chen et al. 2009].

4.2.3 Key feature: generating wiggle files for visualization

Biomedical researchers also need to visualize the alignment results stored in SAM files in order to examine possible gene structure alterations between case and control studies. For example, shortened 3'-UTR's in cancer cells are reflected as an abrupt dropout of the short read coverage. This visualization need is addressed by another key feature of SAMMate. SAMMate can take the alignment results stored in SAM files and export the genome information to wiggle (.wig) files where the wiggle format is compatible with the UCSC genome browser and other browsers used for visualization. This feature will allow biomedical researchers to visually check the alignment quality of selected genes in selected genomic regions. For the miRNA-155 target prediction research, Figure 4.3 presents two typical scenarios: the left and right panels show the alignment results in the pile-up format for gene CXorf39 on Chromsome X and gene LBA1 on Chromsome 3, respectively. Figure 4.3a indicates no overall expression change in the codon regions, but a significant dropout in the 3'-UTR region occurs. On the contrary, Figure 4.3b shows no significant difference in the 3'-UTR region but a significant difference in the codon region instead. These two examples demonstrate SAMMate's ability to generate wiggle files for biomedical researchers allowing them to visually look for possible gene structure alterations. While there are a number of existing alignment visualization software (e.g. [Bao et al. 2009; Arner et al. 2010]), these systems do not allow many annotation tracks in parallel, which is the deterministic feature for knowledge discovery.

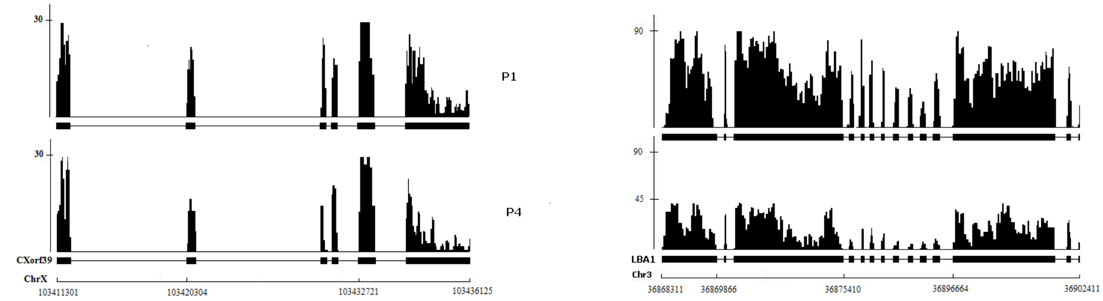


Figure 4.3. Visualization of gene structure sariation. Two typical examples were shown: (1) Gene CXorf39 was called by the Change Point Analysis as a potential miRNA-155 target due to it's abrupt read dropout on the 3'-UTR end. (2) Gene LBA1 was called by the Differential Expression Analysis as a potential miRNA-155 target due to the overall read coverage decrease in codon region.

4.2.4 Key feature: generating alignment report

Short read alignment statistics provide indispensable resources to examine the alignment quality as well as comparing alignment results. SAMMate calculates and exports a number of alignment statistics including the percentage of uniquely mapped short reads and the percentage of short reads mapped to intergenic, exonic and intronic regions.

4.3 Conclusion

We have implemented a GUI software to allow biomedical researchers to parse, process and integrate alignment information stored in SAM/BAM files. With this tool biomedical researchers are able to calculate gene expression abundance using either standard or customized annotations. They are also able to visualize and compare alignment results with great ease. These utilities and their biological impact are adequately demonstrated via the case studies of miRNA target prediction. The biological applications of SAMMate, however, are not limited to miRNA target prediction alone. In fact, SAMMate applies to any biological problem whose solution depends on the gene expression abundance score and base-wise short read coverage signal. SAMMate is also highly modular and extensible providing a programmer friendly interface for ease of updates and the incorporation of contributions from the community. Our tool will greatly facilitate the downstream analysis of genomic sequencing data.

Chapter 5. Conclusion and Future Works

5.1 Conclusion

We provided a GUI software pipeline to analyze transcriptome changes induced by the human microRNA MIR155 using RNA-seq. A comparison with 3’ UTR reporter assay demonstrated general concordance between NGS and corresponding 3’ UTR reporter results. Nonharmonious results were investigated more deeply using transcript structure information assembled from the NGS data. This analysis revealed that transcript structure plays a substantial role in mitigated targeting and in frank targeting failures.

In analysis of EBV transcriptome, our results also showed robust detection of EBV derived transcripts by RNA-seq using the pipeline outlined here. From a quantitative standpoint, several studies have shown this approach to outperform microarrays since it is more accurate [Marioni et al. 2008; Mortazavi et al. 2008; Xu et al. 2010] and since there is an inherently broad dynamic range. The digital nature of RNA-seq allows users to better compare the relative expression of distinct genes through the calculation of RPKMs. This should result in an improvement over microarrays in the analysis of virus-associated transcriptomes not only for EBV but for other viruses. With its high level of accuracy, its broad dynamic range, its utility in assessing transcript structure, and its capacity to accurately interrogate global direct and indirect transcriptome changes, NGS is a useful tool for investigating the biology and mechanisms of action of microRNAs.

For efficiently processing NGS data, our GUI software SAMMate allows biomedical researchers to quickly process SAM/BAM files and is compatible with both single-end and paired-end sequencing technologies. SAMMate also automates some standard procedures in

DNA-seq and RNA-seq data analysis. Using either standard or customized annotation files, SAMMate allows users to accurately calculate the short read coverage of genomic intervals. In particular, for RNA-seq data SAMMate can accurately calculate the gene expression abundance scores for customized genomic intervals using short reads originating from both exons and exon-exon junctions. Furthermore, SAMMate can calculate a whole-genome signal map at base-wise resolution in a short time allowing researchers to solve an array of bioinformatics problems. Finally, SAMMate can export both a wiggle files for alignment visualization in the UCSC genome browser and an alignment statistics report. The biological impact of these features has been already demonstrated via several case studies that predict miRNA targets using short read alignment information files.

With just a few mouse clicks, SAMMate will provide biomedical researchers easy access to important alignment information stored in SAM/BAM files. Our software is constantly updated and will greatly facilitate the downstream analysis of NGS data. Both the source code and the GUI executable are freely available under the GNU General Public License at http://sammate.sourceforge.net.

5.2 Future works

Genome-wide analysis of transcriptomes at the isoform level is our ultimate goal in the future research work. Alternative splicing is one of the key gene expression regulation mechanisms at the transcription level, giving rise to transcriptome diversity. It is estimated that as many as 90% human genes are alternatively spliced in different tissues and conditions, and point mutations in splice sites are responsible for at least 15% of all disease-causing mutations. The problem itself is challenging due to the fact that the observed exonic expression signal can be aggregated from a set of sibling isoforms encoded by the same gene with diverse alternative splicing mechanisms. In essence, the problem finds its root in latent variable models where we infer the latent variables (isoform expression) from the observed variables (exonic expression). However, the development of computational algorithms to deconvolve the gene expression signal emitted from each splicing isoform is not a trivial task.

Therefore, my future research work for transcriptome quantification is to estimate isoform abundance using base-wise RNA-seq data and then integrate the algorithm into SAMMate. We will finally develop a GUI software pipeline for quantifying human transcriptome (whole set of mRNA transcripts) using the NGS data.

References

[1] Arner E, Hayashizaki Y, Daub CO: NGSView: an extensible open source editor for next-generation sequencing data. Bioinformatics (Oxford, England) 2010, 26:125-126, [http://dx.doi.org/10.1093/bioinformatics/btp611].

[2] Bao H, Guo H, Wang J, Zhou R, Lu X, Shi S: MapView: visualization of short reads alignment on a desktop computer. Bioinformatics 2009, 25(12):1554-1555.

[3] Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007). "SNP discovery via 454 transcriptome sequencing". The Plant Journal 51 (5): 910-918.

[4] Cameron JE, Yin Q, Fewell C, Lacey M, McBride J, Wang X, Lin Z, Schaefer BC, Flemington EK. 2008. Epstein-Barr virus latent membrane protein 1 induces cellular MicroRNA miR-146a, a modulator of lymphocyte signaling pathways. J Virol 82: 1946–1958.

[5] Chen J, Wang YP: A Statistical Change Point Model Approach for the Detection of DNA Copy Number Variations in Array CGH Data. IEEE/ACM Trans. Comput. Biol. Bioinformatics 2009, 6(4):529-541.

[6] Chi SW, Zang JB, Mele A, Darnell RB. 2009. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature 460: 479–486.

[7] Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. (2008).

[8] Clurman BE, Hayward WS. 1989. Multiple proto-oncogene activations in avian leukosis virus-induced lymphomas: Evidence for stage-specific events. Mol Cell Biol 9: 2657–2664.

[9] Costinean S, Zanesi N, Pekarsky Y, Tili E, Volinia S, Heerema N, Croce CM. 2006. Pre-B cell proliferation and lymphoblastic leukemia/high-grade lymphoma in Em-miR155 transgenic mice. Proc Natl Acad Sci 103: 7024–7029.

[10] Dabney AR, Storey JD. 2007. Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. Genome Biol 8: R44. doi: 10.1186/gb-2007-8-3-r44.

[11] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 5: R80. doi: 10.1186/gb-2004-5-10-r80.

[12] Gottwein E, Mukherjee N, Sachse C, Frenzel C, Majoros WH, Chi JT, Braich R, Manoharan M, Soutschek J, Ohler U, et al. 2007. A viral microRNA functions as an orthologue of cellular miR-155. Nature 450: 1096–1099.

[13] Greenbaum D, Colangelo C, Williams K, Gerstein M. (2003)."Comparing protein abundance and mRNA expression levels on a genomic scale". Genome Biology 4 (9): 117.

[14] Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. Mol Cell 27: 91–105.

[15] Hammell M. 2010. Computational methods to identify miRNA targets. Semin Cell Dev Biol doi: 10.1016/j.semcdb.2010.01.004.

[16] Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M, et al. 2005. MicroRNA gene expression deregulation in human breast cancer. Cancer Res 65: 7065–7070.

[17] Jiang H, Wong WH: SeqMap : mapping massive amount of oligonucleotides to the genome.Bioinformatics 2008, 24(20):btn429-2396, [http://dx.doi.org/10.1093/bioinformatics/btn429].

[18] Jiang H, Wong WH: Statistical inferences for isoform expression in RNA-seq. Bioinformatics 2009, 25(8):1026-1032, [http://dx.doi.org/10.1093/bioinformatics/btp113].

[19] Jiang J, Lee EJ, Schmittgen TD. 2006. Increased expression of micro-RNA-155 in Epstein-Barr virus transformed lymphoblastoid cell lines. Genes Chromosomes Cancer 45: 103–106.

[20] Kluiver J, Haralambieva E, de Jong D, Blokzijl T, Jacobs S, Kroesen BJ, Poppema S, van den Berg A. 2006. Lack of BIC and microRNA miR-155 expression in primary cases of Burkitt lymphoma. Genes Chromosomes Cancer 45: 147–153.

[21] Kluiver J, Poppema S, de Jong D, Blokzijl T, Harms G, Jacobs S, Kroesen BJ, van den Berg A. 2005. BIC and miR-155 are highly expressed in Hodgkin, primary mediastinal and diffuse large B cell lymphomas. J Pathol 207: 243–249.

[22] Langmead B, Trapnell C, Pop M, Salzberg S: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 2009, 10(3):R25, [http://genomebiology.com/2009/10/3/R25].

[23] Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010a. RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics 26: 493–500.

[24] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis GR, Durbin R: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25(16):2078-2079, [http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics25.html#LiHWFRHMAD09].

[25] Li H, Ruan J, Durbin R: Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008, [http://dx.doi.org/10.1101/gr.078212.108].

[26] Li L, Xu J, Yang D, Tan X, Wang H. 2010b. Computational approaches for microRNA studies: A review. Mamm Genome 21:1–12.

[27] Li R, Li Y, Kristiansen K, Wang J SOAP: short oligonucleotide alignment program. Bioinformatics 2008 , 24:713-714.

[28] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, et al. 2005. MicroRNA expression profiles classify human cancers. Nature 435: 834–838.

[29] Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM (January 2009)."Transcriptome sequencing to detect gene fusions in cancer". Nature458 (7234): 97-101.

[30] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNAseq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509–1517.

[31] Mayr C, Bartel DP. 2009. Widespread shortening of 39UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138: 673–684.

[32] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008)."Mapping and quantifying mammalian transcriptomes by RNA-Seq".Nature Methods 5 (7): 621-628.

[33] O’Connell RM, Rao DS, Chaudhuri AA, Boldin MP, Taganov KD, Nicoll J, Paquette RL, Baltimore D. 2008. Sustained expression of microRNA-155 in hematopoietic stem cells causes a myeloproliferative disorder. J Exp Med 205: 585–594.

[34] O’Connell RM, Taganov KD, Boldin MP, Cheng G, Baltimore D. 2007. MicroRNA-155 is induced during the macrophage inflammatory response. Proc Natl Acad Sci 104: 1604–1609.

[35] Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, et al. 2010.The UCSC Genome Browser database: Update. Nucleic Acids Res 38:D613–619.

[36] Rodriguez A, Vigorito E, Clare S, Warren MV, Couttet P, Soond DR, van Dongen S, Grocock RJ, Das PP, Miska EA, et al. 2007. Requirement of bic/microRNA-155 for normal immune function. Science 316: 608–611.

[37] Ryan D. Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor J. Pugh, Helen McDonald, Richard Varhol, Steven J.M. Jones, and Marco A. Marra. (2008). "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing". BioTechniques 45 (1): 81-94.

[38] Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 39 untranslated regions and fewer microRNA target sites. Science 320: 1643–1647.

[39] Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. 2008. Widespread changes in protein synthesis induced by microRNAs. Nature 455: 58–63.

[40] Skalsky RL, Samols MA, Plaisance KB, Boss IW, Riva A, Lopez MC, Baker HV, Renne R. 2007. Kaposi’s sarcoma-associated herpesvirus encodes an ortholog of miR-155. J Virol 81: 12836–12845.

[41] Smith AD, Xuan Z, Zhang MQ Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 2008 , 9:128.

[42] Smyth GK. 2005. Limma: Linear models for microarray data. In Bioinformatics and computational biology solutions using R and Bioconductor (ed. R Gentleman et al.), pp 397–420. Springer, New York.

[43] Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci 100: 9440–9445.

[44] Tam W, Ben-Yehuda D, Hayward WS. 1997. bic, a novel gene activated by proviral insertions in avian leukosis virus-induced lymphomas, is likely to function through its noncoding RNA. Mol Cell Biol 17: 1490–1502.

[45] The R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://cran.r-project.org/doc/manuals/refman.pdf.

[46] Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq.Bioinformatics 2009, 25(9):1105-1111, [http://dx.doi.org/10.1093/bioinformatics/btp120].

[47] Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 98: 5116–5121.

[48] Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, et al. 2006. A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci 103: 2257–2261.

[49] Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 2009, 10:57-63, [http://dx.doi.org/10.1038/nrg2484].

[50] Xu G, Fewell C, Taylor C, Deng N, Hedges D, Wang X, Zhang K, Lacey M, Zhang H, Yin Q, Cameron J, Zhen L, Zhu D and Flemington EK: Transcriptome and targetome analysis in mir155 expressing cells using rna-seq. *RNA* (New York, N.Y.), 16(8):1610-1622, 2010.

[51] Xu G, Deng N, Zhao Z, Flemington EK, Zhu D. (2011) SAMMate: A GUI tool for processing short read alignment information in SAM/BAM format. Source Code for Biology and Medicine, 6:2.

[52] Yin Q, McBride J, Fewell C, Lacey M, Wang X, Lin Z, Cameron J, Flemington EK. 2008. MicroRNA-155 is an Epstein-Barr virusinduced gene that modulates Epstein-Barr virus-regulated gene expression pathways. J Virol 82: 5295–5306.

[53] Yin Q, Wang X, Fewell C, Cameron J, Zhu H, Baddoo M, Lin Z, Flemington EK. 2010. MiR-155 inhibits bone morphogenetic protein (BMP) signaling and BMP mediated Epstein-Barr virus reactivation. J Virol 84: 6318–6327.

[54] Zhang L, Huang J, Yang N, Greshock J, Megraw MS, Giannakakis A, Liang S, Naylor TL, Barchetti A, Ward MR, et al. 2006. microRNAs exhibit high frequency genomic alterations in human cancer. Proc Natl Acad Sci 103: 9136–9141.

[55] Zheng S, Chen L: A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. Nucl. Acids Res. 2009, 37(10):e75+, [http://dx.doi.org/10.1093/nar/gkp282].

[56] Zheng, Z. B., Y. D. Wu, X. L. Yu, and S. Q. Shang. 2008. DNA microarray technology for simultaneous detection and species identification of seven human herpes viruses. J Med Virol 80:1042-50.

Vita

Guorong Xu was born in Hunan province, China on October 3rd 1977. He received his first Bachelor degree in the department of Mathematics and Computer Science from Jishou University in 2001 and received the second Bachelor degree in the School of Software from Tsinghua University in 2003. And then he came to University of New Orleans to continue his further study in Computer Science area.