Chapter 1
Advanced Datamining Using RNAseq Data

Yan Guo
Vanderbilt University, USA

Margot Bjoring
Vanderbilt University, USA

Shilin Zhao
Vanderbilt University, USA

Leng Han
MD Anderson Cancer Center, USA

ABSTRACT

In recent years, RNA sequencing (RNAseq) technology has experienced a rapid rise in popularity. Often seen as a competitor of and the ultimate successor to microarray technology given its more accurate and quantitative gene expression measurement, RNAseq also offers a wealth of additional information that is often overlooked, and given the massive accumulation of RNAseq data available in public data repositories over the past few years, these data are ripe for discovery. Abundant opportunities exist for researchers to conduct in-depth, non-traditional analyses that take advantage of these secondary uses and for bioinformaticians to develop tools to make these data more accessible. This is discussed in this chapter.

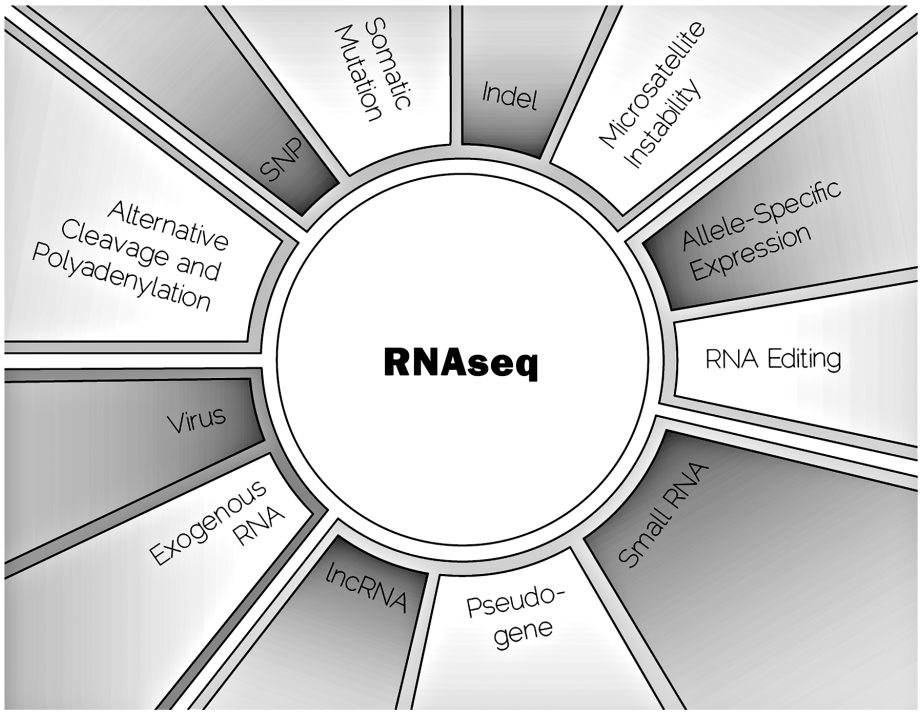
INTRODUCTION

One major drawback of next-generation sequencing (NGS) technology is the imperfect capture technology, which produces a variety of genomic sequences as byproducts in addition to the specific region targeted. Rather than ignore these byproducts, we can turn them into useful information; for example, in DNA sequencing, people have used these byproduct reads to study mitochondria (Guo, Li, Li, Shyr, & Samuels, 2013; Picardi & Pesole, 2012), viruses (Samuels et al., 2013), and non-capture region SNPs (Guo et al., 2012). These and similar ideas can also be applied to RNAseq data (Z. Wang, Gerstein, & Snyder, 2009).

To this point, RNAseq technology has largely been used as a strict replacement for microarrays, its scope limited to gene expression analysis and occasionally the detection of structural variants. While RNAseq is fully capable of filling this role, its extensive inventory of sequencing byproducts make it important and fruitful for researchers to move beyond the past constraints of microarrays and embrace the full potential of RNAseq data. These additional opportunities offered by RNAseq data fall into five major categories (Figure 1):

DOI: 10.4018/978-1-4666-6611-5.ch001

Figure 1. Additional areas of research offered by RNAseq technology beyond gene expression analysis



1. **Mutations:** The mutations identified by RNAseq data can include single nucleotide polymorphisms (SNPs), somatic mutations, insertions and deletions (indels), and microsatellite instability. Because RNAseq by definition contains only the expressed portions of the DNA, the variations contained in these data sets may be enriched for functional relevance.
2. **RNA Editing and Allele-Specific Expression:** RNAseq data can be used to identify both RNA editing, a post-transcriptional mechanism that diversifies the transcriptome by changing nucleotides at the RNA level, and allele-specific expression, or a difference in the expression levels of the two alleles in a gene.
3. **Non-Coding RNA:** There are three types of RNAseq which result for different species of RNA: mRNAseq, miRNAseq, and total RNAseq. Total RNAseq is naturally the most inclusive of RNA species, but even in mRNAseq and miRNAseq, other species of RNA are present, including long non-coding RNA, transfer RNA, and small nucleolar RNA.
4. **Exogenous RNA:** RNAseq data sets will contain some exogenous RNA content such as viruses and exogenous miRNA. Viruses such as the human papillomavirus have known oncogenic effects, and RNAseq can be used to identify exogenous sequences and determine insertion points.
5. **Alternative Cleavage and Polyadenylation:** Alternative cleavage and polyadenylation (APA) is common in mRNA with the great majority of genes having multiple cleavage and polyadenylation sites. While challenging, RNAseq can and has been used to identify APA.

This chapter will discuss in detail these five datamining opportunities including the techniques, tools, challenges, and potential rewards