Getting Started with RNA-Seq Data Analysis

Your path to superior transcriptome mapping and gene expression profiling

Introduction

RNA sequencing (RNA-Seq) is a powerful method for discovering, profiling, and quantifying RNA transcripts. Using Illumina's next-generation sequencing technology, RNA-Seq allows scientists to count the number of reads that align to one of thousands of different cDNAs, producing results similar to those of gene expression microarrays.

Although the end results are similar, the informatics challenges of generating cDNA counts and performing differential expression tests are fundamentally different than those of measuring relative intensity and differential expression using microarrays. Sequences generated from an RNA-Seq experiment are usually mapped to libraries of known exons in known transcripts; this step is absent in microarray expression analysis. Unlike hybridization-based array technologies that measure continuous probe intensities, RNA-Seq measures the number of discrete sequence reads. These counting statistics often require different statistical tests to determine differential expression and other types of patterns[1]. For new RNA-Seq users, there are a host of mature software applications that vastly simplify the data analysis process.

Because RNA-Seq does not require species- and transcript-specific probes, the data are not biased by previous assumptions about the nature of the transcriptome. RNA-Seq provides an opportunity to make hypothesis-free experimental designs, and investigate species with poor or missing genomic annotations. Beyond gene expression analysis, RNA-Seq can be used for discovery applications such as identifying alternative splicing events, allele-specific expression, and rare and novel transcripts.

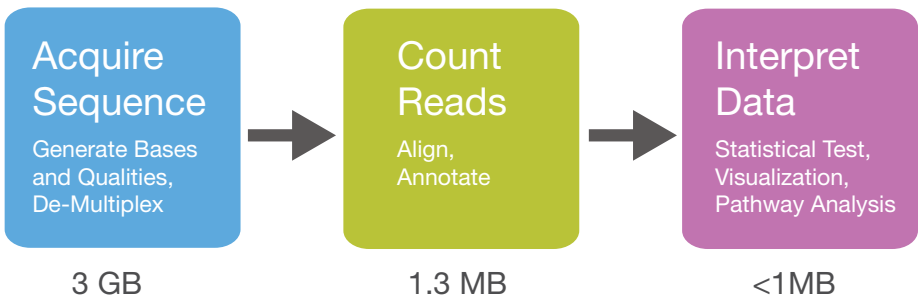
RNA-Seq Data Analysis Workflow

Once the domain of bioinformatics experts, RNA sequencing data analysis tools are more accessible than ever. A simplified overview of the data analysis steps and the output from a typical RNA-Seq experiment is presented in Figure 1. Thanks to improvements in on-instrument processing, Illumina's high-throughput sequencers produce quality-scored base calls during the run with minimal user intervention. This step is equivalent to scanning a microarray. The sequencing output files (compressed FASTQ files) are the input for secondary analysis. Reads are aligned to an annotated reference genome, and those aligning to exons, genes and splice junctions are counted. Optionally, variants (cSNPs, indels) can be called after the alignment step. The final steps are data visualization and interpretation, consisting of calculating gene- and transcript-expression and reporting differential expression. Depending on the analysis tools used, novel splice junctions and isoforms, rare transcript discovery and allele-specific expression can be examined within the same RNA-Seq dataset, beyond what is achievable with microarrays. Additional statistical tests and pathway analyses can aid in biological interpretation, detailed in the next section.

Demystifying Data Analysis

Tools for aligning and counting cDNA reads were once considered to be complex, compute-intensive, and under-supported. In recent years, software packages for RNA-Seq data analysis have become more accessible and user-friendly, detailed in Table 1. Specifically, the free, open-source Tuxedo Suite comprising Bowtie, TopHat, and

Figure 1: RNA-Seq Data Analysis



Illumina RNA-Seq data analysis overview of a typical run of 2 x 50 bp at 30 million reads. Primary analysis occurs on-instrument, generating bases and qualities output as compressed FASTQ files of approximately 3 GB. These are put into secondary analysis, for aligning to reference genome, counting, and normalizing, generating about 1.3 MB. cSNPs, indels, and splice junctions are identified, and gene- and transcript-level expression and differential expression levels are calculated. Data output from these tertiary analyses vary, but are usually less than 1 MB.

Cufflinks, has been widely adopted as the analysis toolset of choice for differential gene expression, discovery of splice forms, and fast alignment of short reads[2,3]. The usability of these tools has been greatly simplified by browser-based access to the Tuxedo Suite through free resources such as the Galaxy Project[4,5] from Penn State and Emory University. Galaxy is an academic initiative with a thriving user base, web-based documentation, and public forums by which community questions are addressed. The need for expensive localized hardware has been reduced through the use of shared servers and on cloud-based solutions such as Amazon's EC2 framework[6].

For the average biologist without extensive bioinformatics training, this means that the RNA-Seq alignment and counting process boils down to obtaining access to Galaxy, producing and uploading FASTQ files, and selecting from a limited number of alignment and counting parameters for Bowtie, TopHat and Cufflinks. A list of counts for RefSeq genes can be obtained within a few hours.

Partek, one of the most commonly-used gene expression microarray analysis tools, has added dedicated functionality for processing sequencing based measurements, including models of variance that account for counting statistics, tools to visualize read-level data, and models to predict and visualize RNA processing. Importantly, Partek recognizes native TopHat output. Several other excellent tool suites are available for interpreting RNA-Seq data and finding patterns in gene expression experiments. Commercial software providers such as CLC Bio, GeoSpiza, GenomeQuest, and Strand Scientific Intelligence have developed workflows for analyzing RNA-Seq data.

Further interpretation of biological pathways using RNA-Seq data can be conducted with Ingenuity's IPA software. Links and descriptions for both open source and commercial tools are detailed in Table 1. For a complete list of third-party analysis tools compatible with Illumina RNA-Seq data, go to IlluminaConnect[7].

Conclusion

The complexity of gene expression and regulation demands the highest sensitivity for measuring transcripts and detecting changes in abundance and structure. The number of recently published RNA-Seq studies confirms that it is the preferred tool for the deepest levels of transcriptome investigation. A simplified data analysis workflow, coupled with improvements in throughput and cost, makes RNA-Seq an accessible and affordable solution for transcriptome research.

References

1. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression of mRNA-Seq experiments BMC Bioinformatics 11:94.
2. Trapnell C, Pachter L, and S Salzberg (2009) TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 25(9) 1105–1111.
3. Langmead B, Trapnell C, and S Salzberg. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
4. Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. 2010 Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86.
5. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, et al. "Galaxy: a web-based genome analysis tool for experimentalists". Current Protocols in Molecular Biology. 2010 Jan; Chapter 19:Unit 19.10.1-21.
6. http://aws.amazon.com/ec2
7. illuminaConnect. http://www.illumina.com/software/illumina_connect.ilmn

Your Path to Success Starts Here

Visit www.illumina.com/rna for protocols, kits, and publications highlighting Illumina's solutions for RNA profiling and discovery.

Table 1: RNA-Seq Analysis Tools

Tool or Suite	Description	Availability	Link
Galaxy	Free form-based access to Bowtie, TopHat, and Cufflinks. Web browser client.	Academic/Open Source	galaxy.psu.edu
GenePattern	Free form-based access to Bowtie, TopHat, and Cufflinks. Local client.	Academic/Closed Source	www.broadinstitute.org/cancer/software/genepattern
Partek	Advanced statistics and interactive visualization for microarray and sequencing data.	Commercial	www.partek.com
CLC Bio	Software analyzing and visualizing sequencing data	Commercial	www.clcbio.com
GeoSpiza	Cloud-based analysis for microarray and sequencing data	Commercial	www.geospiza.com
GenomeQuest	Software for sequence data management	Commercial	www.genomequest.com
Avandis NGS	Software for sequence data analysis and management	Commercial	www.avadis-ngs.com
Ingenuity IPA	Software for biological pathway analysis	Commercial	www.ingenuity.com/products/pathways_analysis.html

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2011 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, HiScan, TruSeq, Eco, and MiSeq are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 470-2011-003 Current as of 14 April 2011

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA