WHITE PAPER
Ion RNA-Seq

Methods, tools, and pipelines for analysis of Ion PGM™ Sequencer miRNA and gene expression data

Introduction

High-resolution measurements of transcriptional activity and organization have been made possible in recent years through massively parallel RNA sequencing (RNA-Seq) technologies. The level of precision enabled by this technology has been demonstrated to be superior to traditional platforms for measuring gene expression in a cell type or tissue [1,2]. Furthermore, with increased sequencing depth, other more specialized analyses such as identification of novel splice variants [3], transcript fusions in cancer [4], and microRNAs [5] (miRNA) are possible. The depth needed for a given experiment is informed by the biological question at hand and complexity of the transcriptome being measured.

RNA-Seq experiments typically produce vast data files that introduce potential bottlenecks when attempting to extract and interpret gene expression information from a given sample. This problem is exacerbated when bioinformatics personnel are few or not available. Much of this can be mitigated through the use of standardized file formats and automated analytical pipelines that are modular and flexible.

In this white paper, we describe methods and bioinformatics tools that have been logically connected into automated pipelines used in the analysis of miRNA and whole transcriptome (WT) sequencing data generated on the Ion PGM™ Sequencer. Dataflow for these pipelines begins with raw sequences from the Ion PGM™ Sequencer and ends with statistical reporting and mapped counts assigned to each gene or miRNA. Methods for quality assessment and quantification are discussed along with descriptions of individual mapping programs and standard file formats, which are used to help ensure interoperability between known and tested third-party tools. Count outputs from these pipelines enable downstream applications such as differential gene expression analysis and miRNA profiling. Through the use of these types of pipelines, a reproducible and accurate representation of RNA-Seq library content can be achieved.

Implementation

At a high level, both the WT and miRNA pipelines have five fundamental tasks to perform once sequences (reads) are acquired from a dedicated Ion PGM™ Torrent Server.

1. 3′ quality/adapter trimming preprocessing
2. Quality control/assessment
3. Mapping to a reference genome and or transcripts
4. Counting mapped reads
5. Generation of global mapping statistics

Whole transcriptome RNA-Seq analysis pipeline

Attached to each Ion PGM™ Sequencer is a dedicated Torrent Server running the Torrent Suite software that processes raw signal produced by the PGM™ Sequencer and calls bases with associated per-base quality values [6]. These data values are written to a standard FASTQ [7] file, then copied to an analysis server where it serves as the only input required by the WT pipeline. At this point, reads have already been trimmed at their 3′ ends of regions of trailing low quality, and adapter (P1) sequences have been clipped by the Torrent Suite software [8]. With larger insert libraries, adapter clipping is not often necessary as the read does not extend far enough given the number of cycles configured for a given run.

Preprocessing

While the Torrent Suite is very effective for 3′ quality trimming and adapter clipping using default parameters, occasionally residual adapter and low-quality bases remain at 3′ ends of reads. For this reason, we add one additional 3′ end trimming function as an initial step in the WT pipeline (Figure 1). It should be noted that adapter and quality trimming are critical since the proxy for mRNA and miRNA quantitation is alignment to a known sequence reference and, ideally, each RNA fragment should map to a genomic location from which it was transcribed. Low quality and portions of adapter sequence flanking an RNA insert can lead to mapping errors such as a misalignment, or can prevent the read from mapping completely.

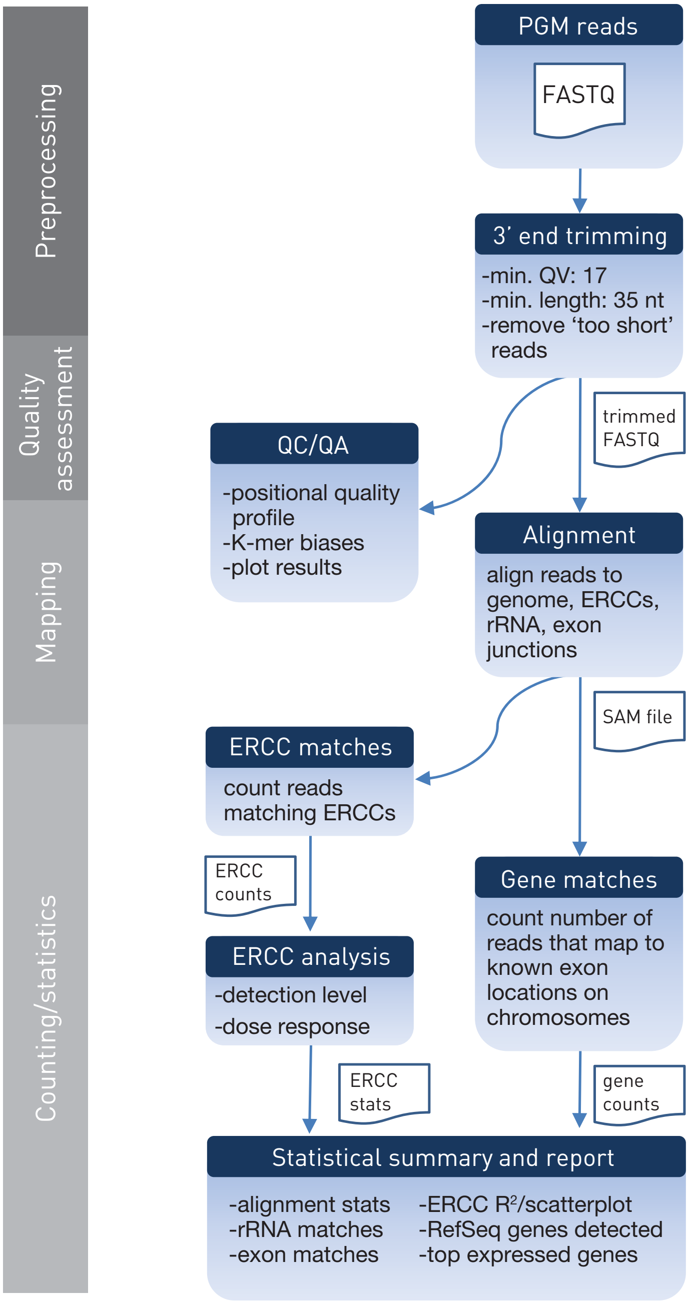

Figure 1. Whole transciptome workflow for quantifying RefSeq genes and ERCC spike-in controls. Final outputs include counts tables and pertinent global statistics.

This preprocessing step is accomplished through the use of the FASTX-toolkit [9] available from the Hannon lab (CSHL). This software suite includes a set of tools for processing and evaluating FASTQ files. Specifically, the *fastq_quality_trimmer* tool is applied such that sequences below a minimum Phred [10] quality value (QV) 17 scanning from 5' to the 3' end of the read are trimmed. If the read falls below 35 bases after trimming, it is removed from further analysis to ensure higher specificity when aligning to a genome reference.

Quality control and assessment

After this preprocessing step, trimmed reads in FASTQ format are used as input to generate plots that may be utilized for future quality control/assessment. A program called FastQC [11] uses a FASTQ file as input, calculates global statistics based on QVs and nucleotide composition, and subsequently produces several plots that provide a high-level view of the sequence quality and potential biases detectable in a RNA-Seq library. Specific elements in this tool set include plots of positional QV across the length of all reads, QV distribution, positional nucleotide composition, and relative k-mer enrichment over read length. A scenario highlighting how this analysis could prove very useful is when fewer reads are mapping to the reference genome than expected. After reviewing the FastQC report, it is observed that the distribution of QVs is shifted much lower than normal indicating a problem may have occurred at steps upstream from the pipeline.

Mapping

Running in parallel to the FastQC step, the trimmed FASTQ reads are aligned to human reference genome using the TMAP mapping program [12]. This mapper is ideal for aligning reads of variable length and includes three algorithms that may be run individually (*map1, map2,* and *map3*) or together (*mapall*). For the purposes of the WT pipeline, we use the *mapall* parameter with seed lengths of 18 nucleotides and employ the default number of allowable mismatches per seed. We have found that for an entire alignment of a read to a reference, allowing for approximately 10% mismatches will preserve "alignability" and specificity. Run time for this step is typically just over two hours for a FASTQ file containing 2 million reads and utilizing 8 threads (-*n* 8 TMAP parameter) on a multi-core analysis server. The output of TMAP (and most other aligners for RNA-Seq) is a Sequence Alignment Map or SAM [13] formatted file. This file contains pertinent information about the alignment of each read such as chromosomal location, chromosomal strand, number of mismatches, alignment score, and insertions/deletions that occurred between the read and a reference sequence. A binary and more compact version of a SAM file, or BAM file, may be generated from a SAM file using SamTools [14].

The genome reference contains the chromosomal sequences available from UCSC [15], ribosomal RNAs (rRNA), known and putative exon-exon junctions, and the sequences of 92 synthetic RNA spike-in controls (ERCC) [16], which will be discussed later. Ribosomal RNA sequences are included in the reference so that rRNA in a library may be quantified for the purposes of evaluating

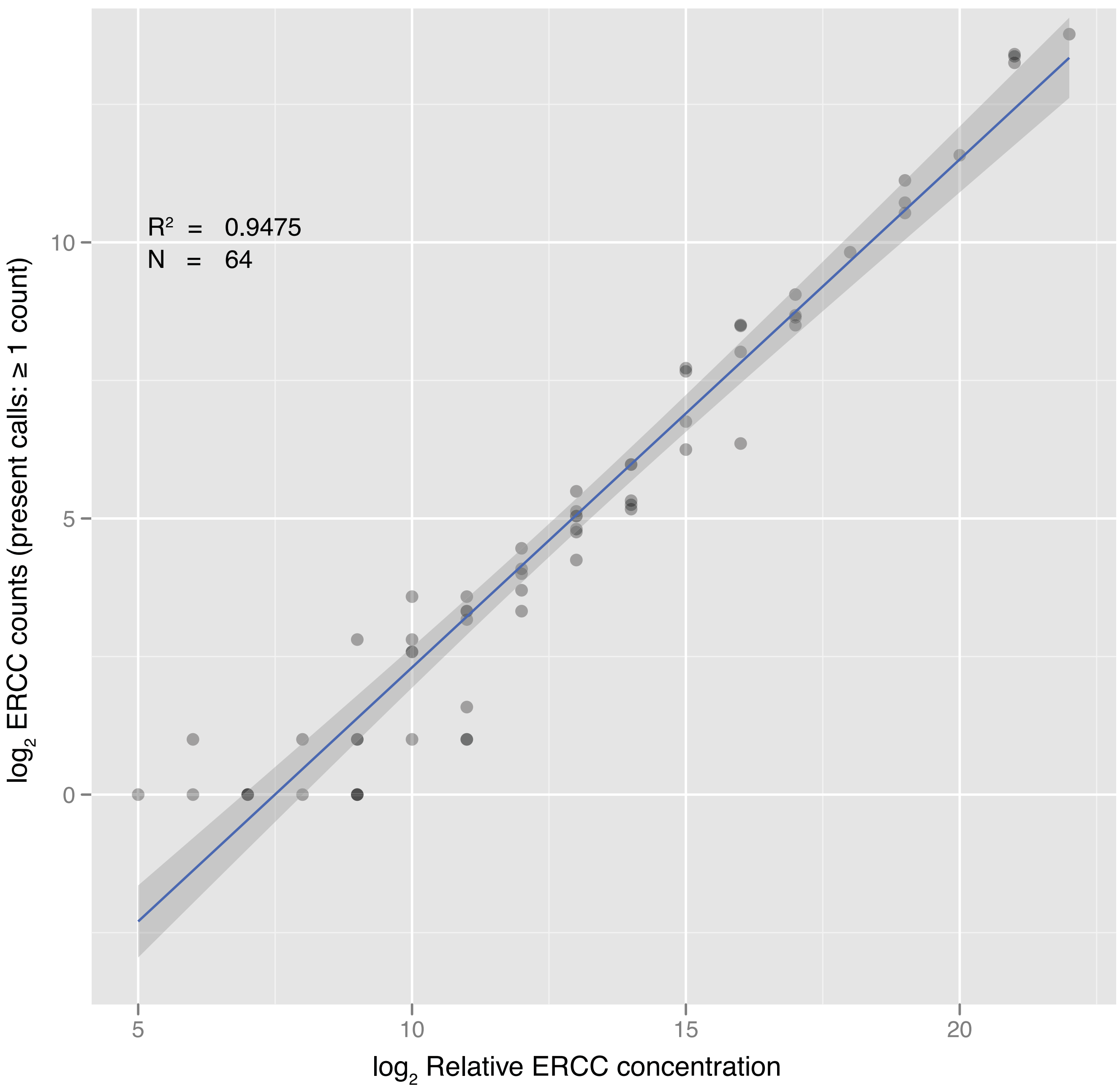

Figure 2. ERCC spike-in control dose-response plot generated from the whole transcriptome pipeline. A sample ERCC dose-response scatter plot and linear regression statistics from an RNA-Seq library prepared from HeLa cell poly(A)-enriched RNA with Ambion® ERCC RNA Spike-In Mix (4456740) added and sequenced on an Ion 318™ Chip. This plot is automatically generated by the whole transcriptome pipeline.

sensitivity and/or the efficiency of rRNA depletion or poly(A) selection techniques. RNA-Seq reads that map to span splice junctions in the genome may result in suboptimal partial alignments or not map at all due to the presence of introns. To mitigate this issue, a sequence reference of all possible combinations of known and putative exon-exon junctions was constructed and included in the mapping reference. The number of reads mapping to this reference tends to increase with mean alignment length suggesting that the likelihood of crossing a splice junction increases with read length.

Counting

After the alignment step, mapped read counts per gene and ERCC are obtained through the use of the *htseq-count.py* script included the HTSeq [17] Python package. This script will extract chromosomal coordinates of each mapped read in a SAM file and detect overlaps with known RefSeq

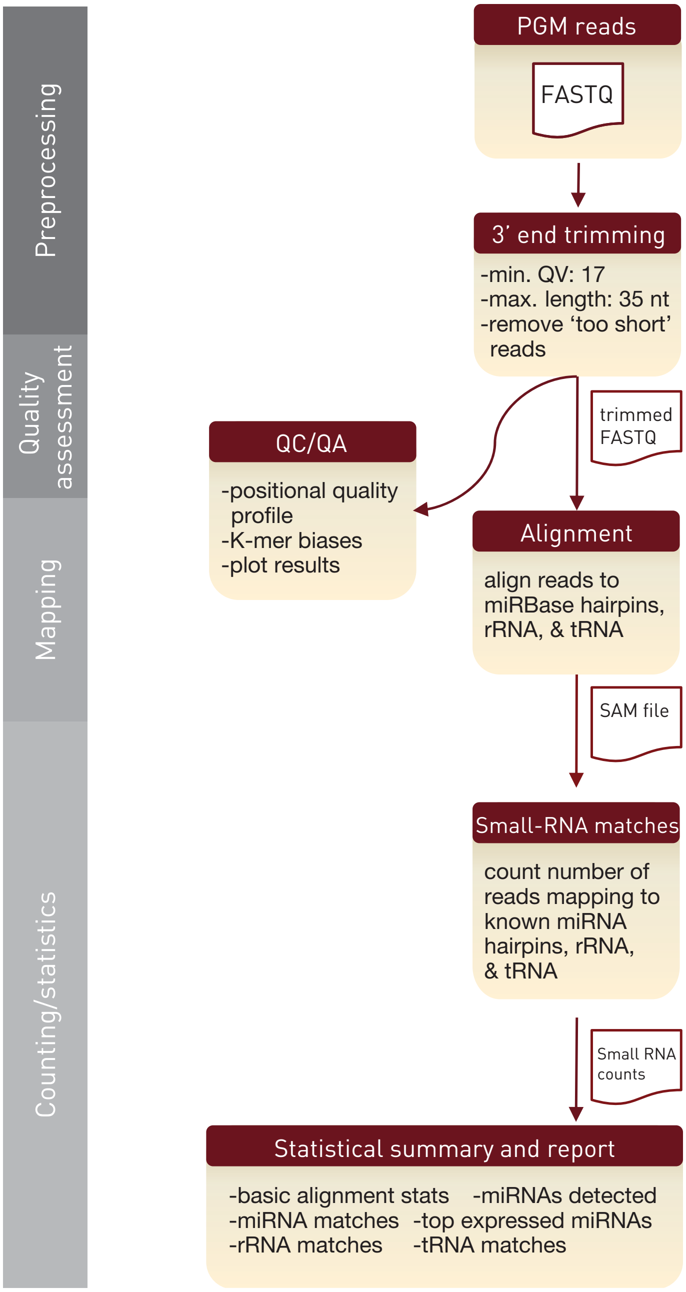

Figure 3. Small RNA workflow for quantifying miRNA transcripts. Final outputs include counts tables and pertinent global statistics.

exon features that are annotated in a GTF (gene transfer format) [18] file referred to as a RefGene GTF. This may be downloaded as a table from the UCSC Genome Browser website. For counting read matches to ERCCs, we simply added 92 entries to the RefGene GTF—one for each transcript—then again run *htseq-count.py* to count reads mapping to each ERCC. Counts to genes and ERCCs are generated in two distinct steps, each run in parallel and resulting in a table of counts.

Analysis of external RNA spike-in controls

Dynamic range, sensitivity, and variability may be examined with the use of 92 ERCC external RNA spike-in controls. These transcripts are polyadenylated, unlabeled RNAs which have been certified and tested by the National Institute of Standards and Technology (NIST) as a means to evaluate RNA measurement systems for performance and control sources of variability. These transcripts have been balanced for GC content to closely represent characteristics of endogenous eukaryotic mRNAs and lengths range from 250 to 2,000 nucleotides. The ERCC pool of transcripts is configured in known titrations designed to represent a large dynamic range of expression levels. The table of ERCC counts from the HTSeq step along with known relative concentrations are sent to an R [19] script which then generates a linear regression model. From this, a dose-response curve is plotted in $\log_2$ space (Figure 2) and R^2 slope and sample size are extracted. A typical, high-quality sequenced library will have an R^2 above 0.9 and a sample size between 60 and 70. The sample size is defined as the number of ERCC transcripts detected with 1 or more counts.

Statistical summary and report

At this point, all major analysis components in the WT pipeline have completed and need to be summarized. A Perl script is used to parse the SAM file and calculate basic mapping statistics such as total number of mapped reads, reads mapping to exons, exon junctions, rRNAs, and reads which map to the genome but do not overlap with known feature annotations. In addition, the number RefSeq genes detected at several minimum count thresholds are reported as a measure of sensitivity and library complexity along with the genes with the highest expression are extracted. These summary statistics are formatted for clarity and written to single report file which resides inside an Ion PGM™ Sequencer–specific directory along with all plots, program output files and counts tables.

miRNA-Seq analysis pipeline

There are many workflow similarities between the WT and miRNA analysis pipelines, and they share the basic high-level functionality. The following sections describing individual components of the miRNA-Seq analysis pipeline (Figure 3) will focus on where the two pipelines differ and why.

Preprocessing

The 3′ end trimming based on quality uses a QV threshold of 17 like the WT pipeline. However, the mean length of mature miRNAs that comprise a small RNA library is 22 with a range between 18 and 32 nucleotides. Thus, minimum sequence length after trimming is set to

17 bases. At this minimum length, high mapping specificity is achievable, since the mapping reference is relatively minuscule (178 kb) and fewer locations exist where reads can ambiguously align. If the read does happen to fall below 17 bases after trimming, it is removed from further analysis.

The detection and trimming of adapter sequences is more crucial with small RNA libraries. For a typical small RNA run on the Ion PGM™ Sequencer, reads are usually longer than the mature miRNA insert and thus the P1 adapter constitutes the last several bases in a read. As mentioned before, this can complicate and or inhibit mappability of these reads, but is well handled by the Torrent Suite in most cases. The *fastq_quality_trimmer* tool includes a parameter to set the maximum length of each read. We set this to 35 bases, thereby maximizing the number alignments possible to the small RNA reference and leaving little or no adapter sequence to compromise mapability. Moreover, the aligner used in this pipeline has a local alignment step which will find the best alignment of the read to the reference even if it is not a contiguous match.

Mapping

Once FastQC has been run on the trimmed FASTQ file for miRNA-Seq libraries, reads are mapped using the SHRiMP [20] short read aligner to a small RNA reference containing miRNA hairpin sequences, rRNAs, tRNAs, and 3' adapter sequence. SHRiMP was used since it has been tested and optimized for miRNA mapping and can be run using a special miRNA mode. As with TMAP, each alignment is initialized with a seeding step. Additionally, a recent paper [21] comparing several aligners for the purposes of quantitative miRNA expression analysis found that SHRiMP was one of two aligners that showed the highest sensitivity in mapping a published synthetic miRNA dataset. While TMAP has shown promising results in this area, extensive testing and optimization has yet to be performed.

The primary mapping reference sequences for this pipeline are miRNA hairpins. These are precursor miRNAs 60–90 nucleotides in length and are available from the RFAM miRBase [22] database in FASTA format. This repository is frequently updated and includes supporting evidence from RNA-Seq experiments. Precursor sequences are chosen for mapping rather than mature sequences since non-canonical mature miRNA sequences, termed isomiRs (first described by Morin et al. [23]), may be present in a sample due to variability in biogenesis.

Also included in the mapping reference are tRNAs and rRNAs. While mappings to these RNAs are typically less frequent than miRNAs, the level of these RNA species can provide insight into library preparation procedures and optimizations; especially in terms of how the RNA sample was size selected. Adapter sequences are included to accommodate the situation where the entire read is only 3' adapter, indicating that there is no RNA insert attached to a bead and the rare event that the sequence was not effectively removed by Torrent Suite preprocessing.

Output from SHRiMP can either be in the larger 'pretty' alignment format or can be written directly to a SAM file. The 'pretty' format is reminiscent of default BLAST alignment format [24] and is useful if one intends to view alignments nucleotide by nucleotide. Given the vastness of these files and incompatibility with downstream applications, this format is generally regarded as impractical.

Counting

Since all mapping is to a reference of RNA transcripts and not a genome, counts to individual transcripts may be calculated directly from the SAM file. This is accomplished by a simple Perl script, which parses the SAM header for transcript identifiers and then scans the remainder of the file to tally counts to each of those RNAs. The resulting counts file contains a column of IDs and counts for each miRNA precursor, tRNA and rRNA.

Statistical summary and report

As with the WT pipeline, counts and basic mapping statistics are compiled with a Perl summary script that processes the SAM file. In addition, counts of matches to adapter only (no insert) and number of miRBase miRNAs detected are reported.

Discussion

Downstream applications of RNA-Seq count data

Counts tables serve as quantitative measure of gene/miRNA expression. Many methods for examining differential expression, normalization, and profiling have emerged as RNA-Seq has increased in popularity and with technological advances of deep-sequencing platforms. Specifically for comparing gene expression, a table of gene counts from two experiments having multiple biological replicates can be compared to estimate differential expression while taking into account of biological variability using the R package DESeq [25]. While DESeq has generally been applied to whole transciptome count data, Cordero et al. [21] reports that DESeq also shows very good sensitivity and specificity when examining miRNA-Seq count data.

Since the SAM (and BAM) formats have become so versatile, many tools are now available for visualization of read coverage across genomes and transcripts that use this format for input. The Integrative Genomics Viewer (IGV) [26] and Tablet [27] are among the most popular.

Commercial software packages from Partek [28] and CLC bio [29], which combine some of the pipeline steps described above with various types of visualizations and in-depth transcriptome analysis, are also available.

Pipeline automation and jobs scheduling

Analysis pipelines such as these lend themselves well to additional automation to easily integrate into a high-throughput environment where several Ion PGM™ Sequencer runs are to be processed in a particular day. A simple 'wrapper' Perl or Python script can easily execute each pipeline step in series. An even more efficient option, if a UNIX-based compute cluster environment is available, would be to submit each step as task in a job scheduler such as the Portable Batch System (PBS) [30]. In this system, steps may be configured to run in parallel, and jobs will run from a queue as computational resources are available.

Processing time and hardware requirements

For either pipeline, the rate-limiting step is the alignment component. Processing times are approximately one hour and ten minutes for the WT and small RNA pipelines respectively, per million reads. Each pipeline is run such that aligner uses 8 threads allowing for parallel processing on a multi-core CPU. Memory requirements scale with the size of the reference genome to which is being mapped to. For the WT pipeline described here (which includes the human genome, ERCC and rRNA transcripts, plus putative and novel exons junction sequences), approximately 16 GB of RAM is required by TMAP while only 1.5 GB is required for SHRiMP. For the description of the hardware configuration used for running these pipelines, please see Appendix A.

References

1. Wang Z, Gerstein M, Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63.
2. Mortazavi A, Williams BA, McCue K, et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
3. Sultan M, Schulz MH, Richard H, et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321:956–960.
4. Smith TM, Olson NE, Smith D (2010) Making cancer transcriptome sequencing assays practical for the research and clinical scientist. *Genome Biology* (Supp 1), p.39.
5. Ryu S, Joshi N, McDonnell K, et al. (2011) Discovery of novel human breast cancer microRNAs from deep sequencing data by analysis of pri-microRNA secondary structures. PLoS One 6:e16403. doi:10.1371/journal.pone.0016403.
6. Torrent Suite technical note: The Per-Base Quality Score System (http://lifetech-it.hosted.jivesoftware.com/docs/DOC-2306)
7. FASTQ format description from Wikipedia: (http://en.wikipedia.org/wiki/FASTQ_format)
8. Torrent Suite technical note: Filtering and Trimming (http://lifetech-it.hosted.jivesoftware.com/docs/DOC-2305)
9. FASTX-Toolkit: (http://hannonlab.cshl.edu/fastx_toolkit/)
10. Phred quality score description from Wikipedia: (http://en.wikipedia.org/wiki/Phred_quality_score)
11. FastQC sequencing quality control program: (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/)
12. TMAP: The Ion Torrent flow sequence mapping program: (http://lifetech-it.hosted.jivesoftware.com/docs/DOC-2101)
13. SAM Format Specification (v1.4-r985): (http://samtools.sourceforge.net/SAM1.pdf)
14. Li H, Handsaker B, Wysoker A, et al. (2009) 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
15. UCSC human genome reference: (http://hgdownload.cse.ucsc.edu/downloads.html#human)
16. Jiang L, Schlesinger F, Davis CA, et al. (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res* 21:1543–1551.
17. HTSeq overview: (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html)
18. GTF/GFF file format: (http://uswest.ensembl.org/info/website/upload/gff.html)
19. R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0: (http://www.R-project.org/)
20. David M, Dzamba M, Lister D, et al. (2011) SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics* 27:1011–1012.
21. Cordero F, Beccuti M, Arigoni M, et al. (2012) Optimizing a massive parallel sequencing workflow for quantitative miRNA expression analysis. *PLoS One* 7:e31630. doi:10.1371/journal.pone.0031630.
22. Kozomara A, Griffiths-Jones S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* (database issue) 39:D152–D157.
23. Morin RD, O'Connor MD, Griffith M, et al. (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* 18:610–621.
24. BLASTN alignment format: (http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastall/blastall_node83.html)
25. Anders S., Huber W. (2010) Differential expression analysis for sequence count data. *Genome Biol* 11:R106.
26. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. (2011) Integrative genomics viewer. *Nat Biotechnol* 29:24–26.
27. Milne I, Bayer M, Cardle L, et al. (2010) Tablet--next generation sequence assembly visualization. *Bioinformatics* 26:401–402.
28. Partek Genomics Suite: (http://www.partek.com/ngs#rnaseq)
29. CLC bio: (http://www.clcbio.com/index.php?id=1243)
30. Portable Batch System (PBS): (http://en.wikipedia.org/wiki/Portable_Batch_System)

Appendix A

Pipeline software versions utilized

Torrent Server v2.2.1

FASTX-Toolkit v0.0.13

FastQC v0.9.5 (requires Java Runtime Environment)

TMAP v0.1.3

SHRiMP v2.1.0

R v2.12.2

HTSeq v0.5.3p1

Perl v5.8.8

Python v2.6

Pipeline hardware configuration

CPU: Intel® Xeon® E5540 4 chips @ 2.53 GHz, 4 cores per chip

RAM: 64 GB

Storage array: 72 hard drives, 1 TB each

For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.

Printed in the USA. **CO25176 0512**

Headquarters
5791 Van Allen Way | Carlsbad, CA 92008 USA | Phone +1.760.603.7200 | Toll Free in the USA 800.955.6288
www.lifetechnologies.com

life technologies™