RNA-SeQC Documentation

Description: Calculates metrics on aligned RNA-seq data.

Author: David Deluca (Broad Institute), gp-help@broadinstitute.org

Summary

This module calculates standard RNA-seq related metrics, such as depth of coverage, ribosomal RNA contamination, continuity of coverage, and GC bias. Required input includes a BAM file or a zipped set of BAM files, and a reference genome in FASTA format.

Metrics can include:

- total read number, number of unique reads, and number of duplicate reads
- duplication rate (number of duplicates/total reads)
- number of reads mapped/aligned and mapping rate (mapped reads/total reads)
- number of unique reads mapped and mapped unique rate (mapped unique reads/mapped reads)
- reads that are mapped to rRNA regions and rRNA rate (reads mapped to rRNA regions/total reads)
- intragenic rate (reads mapped to intragenic regions/mapped unique reads)
- exonic rate (reads mapped to exonic regions/mapped unique reads)
- coding rate (reads mapped to coding regions/mapped unique reads)
- intergenic rate (reads mapped to intergenic regions/mapped unique reads)
- strand specificity metrics
- coverage metrics (particularly for the top expressed transcripts)
- RPKM: this metric quantifies transcript levels in reads per kilobase of exon model per million mapped reads (RPKM). The RPKM measure of read density reflects the molar concentration of a transcript in the starting sample by normalizing for RNA length and for the total read number in the measurement. This facilitates transparent comparison of transcript levels both within and between samples.

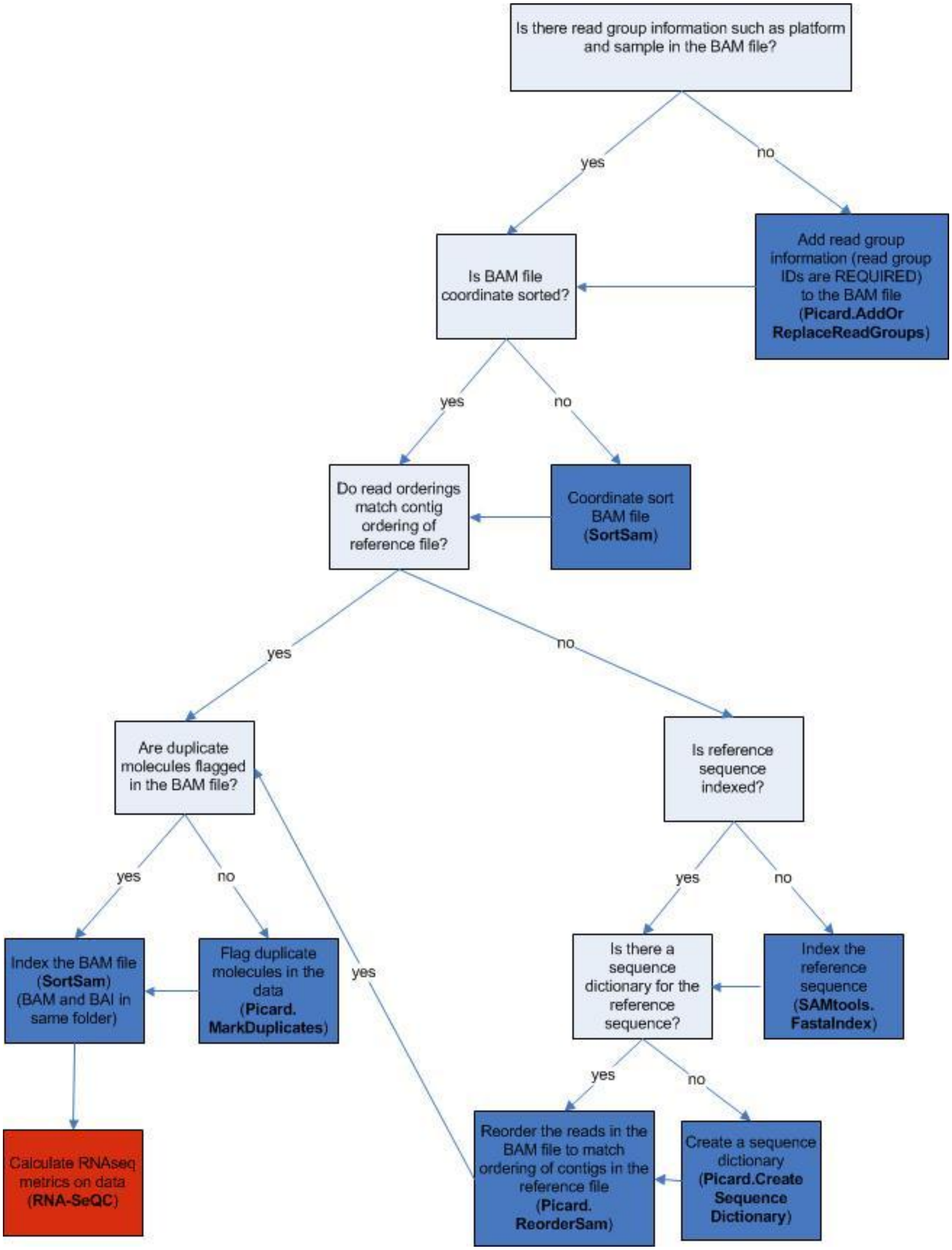
For more information on the BAM format, which is a binary form of the SAM format, see the SAM file specification here: http://samtools.sourceforge.net/.

Usage

The RNA-seq data must be preprocessed in a particular manner for the RNA-SeQC module to work on it correctly. The input BAM file must:

- be coordinate-sorted
- have read group information (Each read group **must** have an ID and contain the platform [PL] and sample [SM] tags; for the platform value, the module currently supports 454, LS454, Illumina, Solid, ABI_Solid, and CG [all values are case-sensitive]. Each read in the BAM file must be associated with exactly one read group.)
- be accompanied by an indexed reference sequence
- be accompanied by a sequence dictionary
- have duplicate reads flagged
- be indexed (if SortSam is used to index the BAM file, then the BAM and BAI files are located in the same folder)

The following decision tree illustrates the preprocessing that should be used for the BAM file before it can be run in the RNA-SeQC module.


Is there read group information such as platform and sample in the BAM file?
yes
no
Add read group information (read group IDs are REQUIRED) to the BAM file (Picard.AddOr ReplaceReadGroups)
Is BAM file coordinate sorted?
yes
no
Coordinate sort BAM file (SortSam)
Do read orderings match contig ordering of reference file?
yes
no
Are duplicate molecules flagged in the BAM file?
Is reference sequence indexed?
yes
no
yes
no
Index the BAM file (SortSam) (BAM and BAI in same folder)
Flag duplicate molecules in the data (Picard. MarkDuplicates)
yes
Is there a sequence dictionary for the reference sequence?
Index the reference sequence (SAMtools. FastaIndex)
yes
no
Calculate RNAseq metrics on data (RNA-SeQC)
Reorder the reads in the BAM file to match ordering of contigs in the reference file (Picard. ReorderSam)
Create a sequence dictionary (Picard.Create Sequence Dictionary)

References

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly MA. Framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. 2011 Apr; 43(5):491-498.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods.* 2010;7:709–715.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010 Sep; 20(9):1297-303. Epub 2010 Jul 19.

Picard tools. http://picard.sourceforge.net

Parameters

Name	**Description**
bam.files (required)	An indexed BAM file or zipped set of indexed BAM files to be analyzed. If you are supplying a single BAM file, it should be located in the same folder as its associated index file (BAI). (If SortSam is used to index the BAM file, then the BAM and BAI files are located in the same folder.) If you are supplying a zipped set of BAM files, the ZIP archive must also include the appropriate BAI files. The BAM file must have a proper BAM header with read groups. Each read group must contain the platform (PL) and sample (SM) tags. For the platform value, the module currently supports (case-sensitive): • 454 • LS454 • Illumina • Solid • ABI_Solid • CG (Cont'd next page)

bam.files (cont'd)	Each read in the BAM file must be associated with exactly one read group. The order of the reads in the BAM file must match the contig ordering of the reference. Unfortunately, many BAM files have headers that are sorted in some other order; lexicographical order is a common alternative. To reorder the reads in the BAM file to match the contig ordering in the reference file, use the Picard.ReorderSam module.
sample.info.file	A TXT format sample info file containing a sample ID, sample file name, and notes column in tab-delimited format. The sample ID is used to label the samples in the output results. The sample file name is the name of the BAM file(s) specified in the BAM file input parameter.
single.end	Whether the BAM file contains single end reads. Default: *yes*
annotation.gtf	A genome annotation to use. If the annotation you need is not in the drop-down list, you can upload an annotation GTF file in the *annotation.gtf.file* parameter. Either an annotation GTF must be specified here, or an annotation GTF file must be provided.
annotation.gtf.file	A file containing a genome annotation in GTF format. If the annotation file you need is not provided in *annotation.gtf*, you can upload an annotation GTF file here. Either an annotation GTF must be specified, or an annotation GTF file must be provided here. NOTE: The transcript_id and gene_id attributes are required in the GTF file.
reference.sequence (required)	The sequence for the reference genome in FASTA format. The reference sequence must have an index (.fai) and a sequence dictionary (.dict). All three files (FASTA, FAI, and DICT) must either be located in the same directory OR specified in the reference sequence index and dictionary parameters. NOTE: The contig names in the reference sequence should match the contig names in the BAM file(s).

reference.sequence.index	A file (FAI) containing the index for the reference sequence. If the FAI file or the DICT file is not in the same folder as the FASTA file, then you must specify this file. If the FASTA, FAI, and DICT files are all in the same folder, you do not need to specify this file.
reference.sequence.dictionary	A file (DICT) containing the dictionary for the reference sequence. If the FAI file or the DICT file is not in the same folder as the FASTA file, then you must specify this file. If the FASTA, FAI, and DICT files are all in the same folder, you do not need to specify this file.
num.genes (required)	The number of top-expressed genes for which to calculate metrics. Default: 1000 Running the default number of genes requires at least 3GB of memory available for processing. If you find that you run out of memory during a run, try reducing the number of genes.
transcript.type.field	Specifies the column of the GTF file in which the transcript type is specified. By default, the module looks for an attribute called transcript_type in the GTF in order to find transcripts labeled as rRNA. If the GTF file does not have an attribute called transcript_type, then you will need to include which column in the GTF file specifies whether the transcript is rRNA.
rRNA.interval.file	A file containing the genomic coordinates of rRNA. This file is in GATK format and uses the .list extension. The file contains one genomic coordinate per line in the following format: *chr:start-stop* If this file is not provided, the information is drawn from the annotation GTF file. Either an rRNA interval file *OR* an aligned rRNA file can be provided, but not both.
rRNA.aligned.file	A SAM file containing ribosomal RNA (rRNA) reads that is used to estimate rRNA content. If this file is not provided, the information is drawn from the annotation GTF file. Either an rRNA interval file *OR* an aligned rRNA file can be provided, but not both.

transcript.end.length (required)	The length of the 3’ or 5’ end of a transcript. Available values are 10, 50, and 100. Default: *50*
transcript.level.metrics	Whether to calculate transcript-level metrics in addition to sample-level metrics. Default: *no*
gc.content.file	A file containing GC content for each of the transcripts. The file must be tab-delimited with 2 columns containing transcript name and GC content. The transcript name must appear in the GTF file. If you provide a GC content file, you will get an additional section of results first stratified (ranked) by their GC content, and then metrics for the high-, middle-, and low-expressed transcripts in that ranking.
num.downsampling.reads	Perform downsampling on the given number of reads. It randomly samples the specified number of reads in all experimental samples when calculating metrics.
correlation.comparison.file	A GCT expression data file used to calculate the correlation between expression values. Only uses the first sample if the GCT file contains more than one sample. Note that the GCT file must contain the gene symbols that appear in the annotation GTF file.
output.prefix (required)	A prefix to use for the output file name.

Input Files

Required input files:

- an indexed, coordinate-sorted BAM file with read group information and duplicate reads flagged
- the index for the BAM file (.BAI) (if SortSam is used to index the BAM file, then the BAM and BAI files are located in the same folder)
- a reference sequence (FASTA)
- the index for the reference sequence (.FAI)
- a sequence dictionary for the reference sequence (.DICT)

Optional input files:

- sample info file containing a sample ID, sample file name, and notes column in tab-delimited format
- genome annotation file in GTF format (required if the annotation file is not available to be specified in the module)

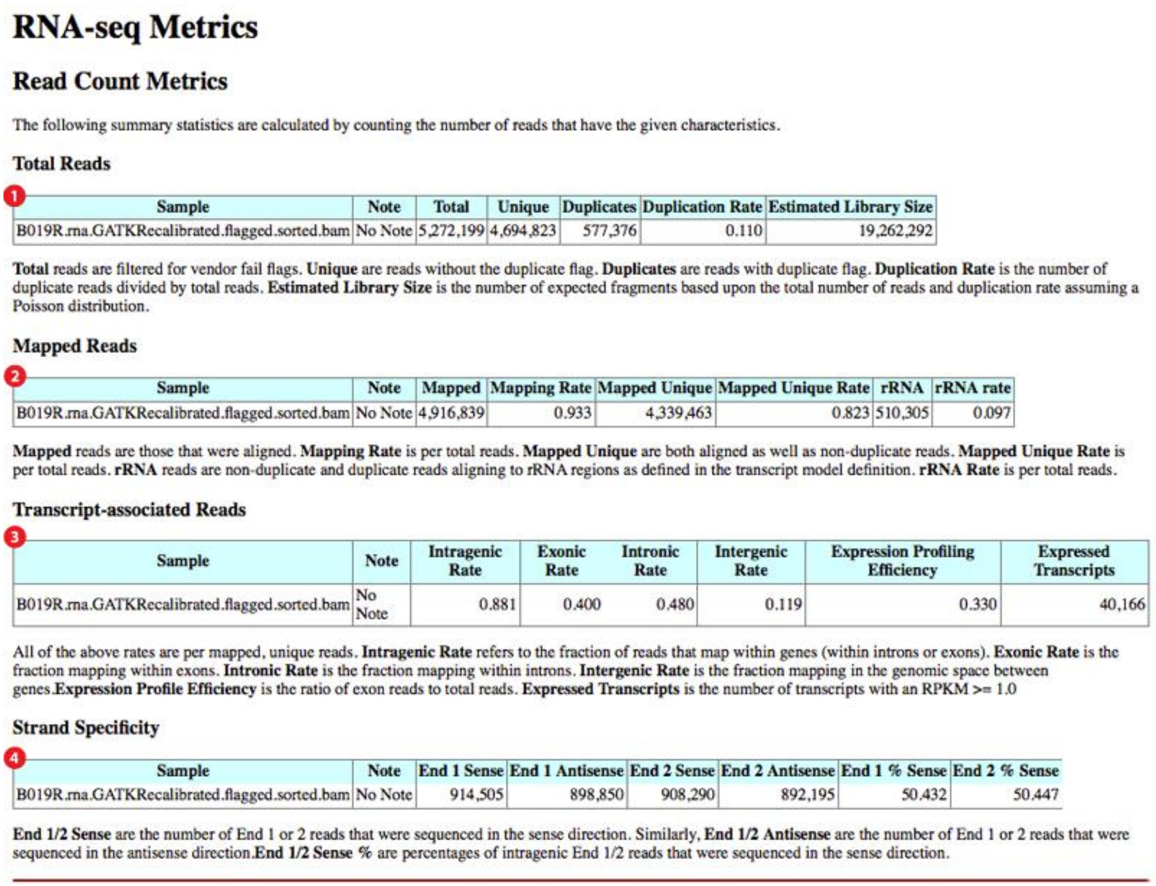
- one of either a file containing the genomic coordinates of rRNA, in GATK format (.LIST) or a SAM file containing ribosomal RNA (rRNA) reads that is used to estimate rRNA content
- a tab-delimited file containing GC content for each of the transcripts; must contain 2 columns with transcript name and GC content
- a GCT expression data file used to calculate the correlation between expression values

Output Files

1. ZIP archive

   The HTML report contains metrics stating the total number of reads, depth of coverage, etc. The report also links to specific metrics files. The archive contains a number of other files containing more details of metrics and statistics.
   The HTML report (index.html at the base level of the archive) contains the following information.

index.html



RNA-seq Metrics

Read Count Metrics

The following summary statistics are calculated by counting the number of reads that have the given characteristics.

Total Reads

Sample	Note	Total	Unique	Duplicates	Duplication Rate	Estimated Library Size
B019R.rna.GATKRecalibrated.flagged.sorted.bam	No Note	5,272,199	4,694,823	577,376	0.110	19,262,292

Total reads are filtered for vendor fail flags. **Unique** are reads without the duplicate flag. **Duplicates** are reads with duplicate flag. **Duplication Rate** is the number of duplicate reads divided by total reads. **Estimated Library Size** is the number of expected fragments based upon the total number of reads and duplication rate assuming a Poisson distribution.

Mapped Reads

Sample	Note	Mapped	Mapping Rate	Mapped Unique	Mapped Unique Rate	rRNA	rRNA rate
B019R.rna.GATKRecalibrated.flagged.sorted.bam	No Note	4,916,839	0.933	4,339,463	0.823	510,305	0.097

Mapped reads are those that were aligned. **Mapping Rate** is per total reads. **Mapped Unique** are both aligned as well as non-duplicate reads. **Mapped Unique Rate** is per total reads. **rRNA** reads are non-duplicate and duplicate reads aligning to rRNA regions as defined in the transcript model definition. **rRNA Rate** is per total reads.

Transcript-associated Reads

Sample	Note	Intragenic Rate	Exonic Rate	Intronic Rate	Intergenic Rate	Expression Profiling Efficiency	Expressed Transcripts
B019R.rna.GATKRecalibrated.flagged.sorted.bam	No Note	0.881	0.400	0.480	0.119	0.330	40,166

All of the above rates are per mapped, unique reads. **Intragenic Rate** refers to the fraction of reads that map within genes (within introns or exons). **Exonic Rate** is the fraction mapping within exons. **Intronic Rate** is the fraction mapping within introns. **Intergenic Rate** is the fraction mapping in the genomic space between genes.**Expression Profile Efficiency** is the ratio of exon reads to total reads. **Expressed Transcripts** is the number of transcripts with an RPKM >= 1.0

Strand Specificity

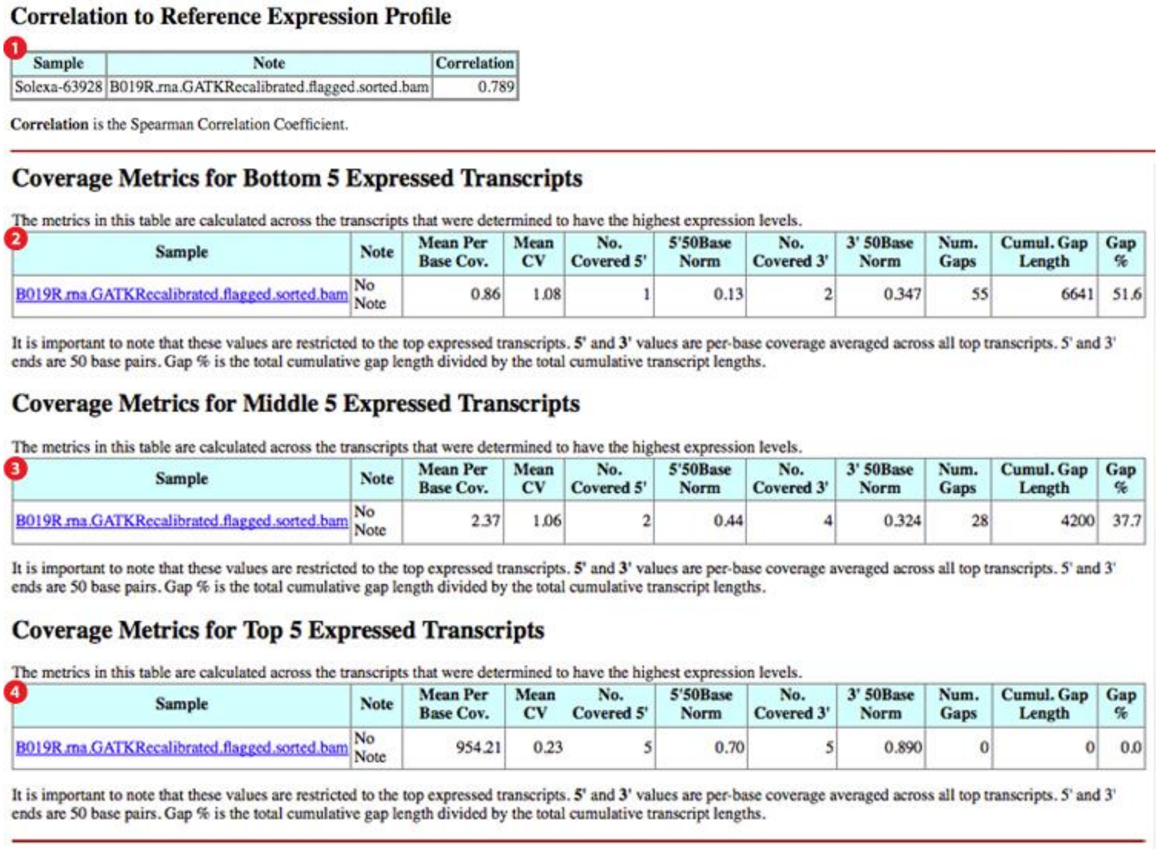
Sample	Note	End 1 Sense	End 1 Antisense	End 2 Sense	End 2 Antisense	End 1 % Sense	End 2 % Sense
B019R.rna.GATKRecalibrated.flagged.sorted.bam	No Note	914,505	898,850	908,290	892,195	50.432	50.447

End 1/2 Sense are the number of End 1 or 2 reads that were sequenced in the sense direction. Similarly, **End 1/2 Antisense** are the number of End 1 or 2 reads that were sequenced in the antisense direction.**End 1/2 Sense %** are percentages of intragenic End 1/2 reads that were sequenced in the sense direction.

1	• Total: Total reads (filtered for vendor fail flags) • Unique: Total reads without a duplicate flag • Duplicates: Total reads with a duplicate flag • Duplication Rate: Ratio of the number of duplicate reads to total reads • Estimated library size: Number of expected fragments based on the total reads and duplication rate assuming a Poisson distribution.

2	- Mapped: Total number of reads aligned/mapped - Mapping Rate: Ratio of total mapped reads to total reads - Mapped Unique: Number of reads that were aligned and did not have duplicate flags - Mapped Unique Rate: Ratio of mapping of reads that were aligned and were not duplicates to total reads - rRNA: Number of all reads (duplicate and non-duplicate) aligning to ribosomal RNA regions - rRNA Rate: Ratio of all reads aligned to rRNA regions to total reads
3	- Intragenic Rate: The fraction of reads that map within genes (within introns or exons) - Exonic Rate: The fraction of reads that map within exons - Intronic Rate: The fraction of reads that map within introns - Intergenic Rate: The fraction of reads that map to the genomic space between genes - Expression Profiling Efficiency: Ratio of exon reads to total reads - Expressed Transcripts: Total number of transcripts with a reads per kilobase of exon model per million mapped reads (RPKM) ≥1.0
4	- End 1 Sense: Number of End 1 reads that were sequenced in the sense direction - End 1 Antisense: Number of End 1 reads that were sequenced in the antisense direction - End 2 Sense: Number of End 2 reads that were sequenced in the sense direction - End 2 Antisense: Number of reads that were sequenced in the antisense direction - End 1 % Sense: Percentage of intragenic End 1 reads that were sequenced in the sense direction - End 2 % Sense: Percentage of intragenic End 2 reads that were sequenced in the sense direction

*Poisson distribution: the probability of a given number of events occurring in a fixed interval if these events occur with a known average rate



Correlation to Reference Expression Profile

1

Sample	Note	Correlation
Solexa-63928	B019R.rna.GATKRecalibrated.flagged.sorted.bam	0.789

Correlation is the Spearman Correlation Coefficient.

Coverage Metrics for Bottom 5 Expressed Transcripts

The metrics in this table are calculated across the transcripts that were determined to have the highest expression levels.

2

Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'50Base Norm	No. Covered 3'	3' 50Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
B019R.rna.GATKRecalibrated.flagged.sorted.bam	No Note	0.86	1.08	1	0.13	2	0.347	55	6641	51.6

It is important to note that these values are restricted to the top expressed transcripts. **5'** and **3'** values are per-base coverage averaged across all top transcripts. 5' and 3' ends are 50 base pairs. Gap % is the total cumulative gap length divided by the total cumulative transcript lengths.

Coverage Metrics for Middle 5 Expressed Transcripts

The metrics in this table are calculated across the transcripts that were determined to have the highest expression levels.

3

Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'50Base Norm	No. Covered 3'	3' 50Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
B019R.rna.GATKRecalibrated.flagged.sorted.bam	No Note	2.37	1.06	2	0.44	4	0.324	28	4200	37.7

It is important to note that these values are restricted to the top expressed transcripts. **5'** and **3'** values are per-base coverage averaged across all top transcripts. 5' and 3' ends are 50 base pairs. Gap % is the total cumulative gap length divided by the total cumulative transcript lengths.

Coverage Metrics for Top 5 Expressed Transcripts

The metrics in this table are calculated across the transcripts that were determined to have the highest expression levels.

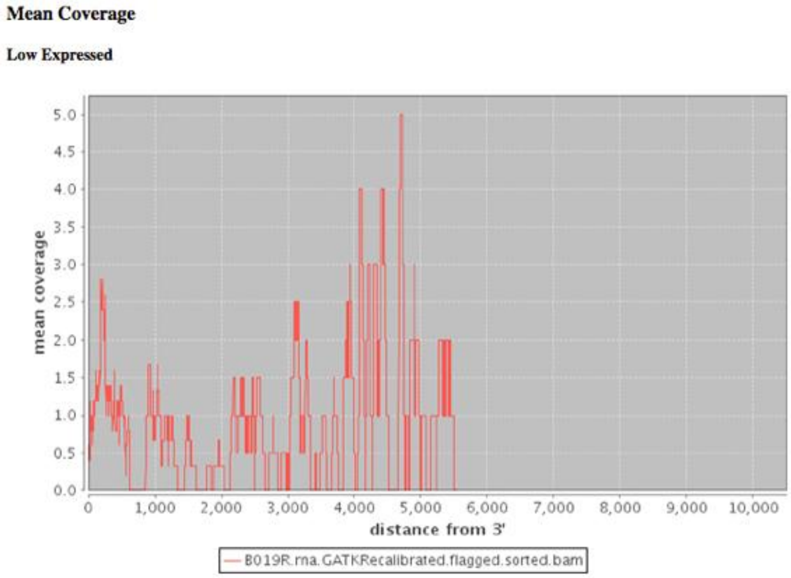
4

Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'50Base Norm	No. Covered 3'	3' 50Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
B019R.rna.GATKRecalibrated.flagged.sorted.bam	No Note	954.21	0.23	5	0.70	5	0.890	0	0	0.0

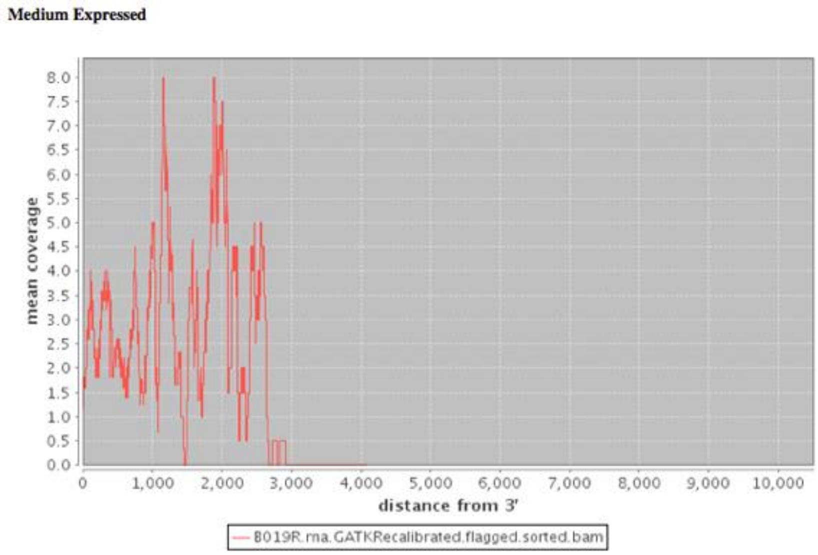
It is important to note that these values are restricted to the top expressed transcripts. **5'** and **3'** values are per-base coverage averaged across all top transcripts. 5' and 3' ends are 50 base pairs. Gap % is the total cumulative gap length divided by the total cumulative transcript lengths.

1	Spearman Correlation Coefficient of the data to the reference expression profile provided. (Will only be calculated and displayed on this page if a correlation comparison file is provided at runtime.)
2	Metrics for the 3 transcripts determined to have the highest expression levels in the lowest expressed transcripts: • mean coverage per base • mean coefficient of variation: standard deviation divided by mean coverage • number covered 5': the number of transcripts that have at least one read in their 5' end • 5' 50-based normalization: 50 (this number is the value for the *transcript end length* parameter) refers to the definition of how many bases are considered at the end; this value is the ratio between the coverage at the 5' end and the average coverage of the full transcript, averaged over all transcripts; to obtain this metric: 1. calculate the mean coverage of the transcript (every base has a coverage value, so the mean coverage is the average over all bases) 2. calculate the mean coverage of the 5' end of the transcript 3. calculate 5' coverage relative to the transcript's overall average coverage: 2/1 4. average the result from step #3 over all transcripts • number covered 3': the number of transcripts that have at least one read in their 3' end • 3' 50-base normalization: the ratio between the coverage at the 3' end and the average coverage of the full transcript, averaged over all transcripts • number of gaps: number of regions with ≥5 bases with zero coverage • cumulative gap length: cumulative length of gap regions • gap percentage: the total cumulative gap length divided by the total cumulative transcript lengths
3	Metrics for the 3 transcripts determined to have the highest expression levels in the middle expressed transcripts: mean coverage per base, mean coefficient of variation, number covered 5', 5' 50-based normalization, number covered 3', 3' 50-base normalization, number of gaps, cumulative gap length, gap percentage.
4	Metrics for the 3 transcripts determined to have the highest expression levels in the highest expressed transcripts: mean coverage per base, mean coefficient of variation, number covered 5', 5' 50-based normalization, number covered 3', 3' 50-base normalization, number of gaps, cumulative gap length, gap percentage.

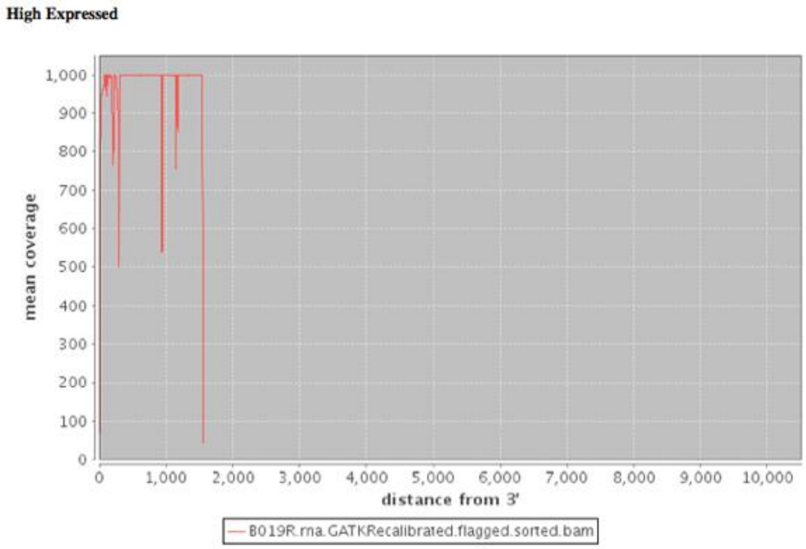
The following plot shows the mean coverage for the low-expressed transcripts over the distance from the 3' end.

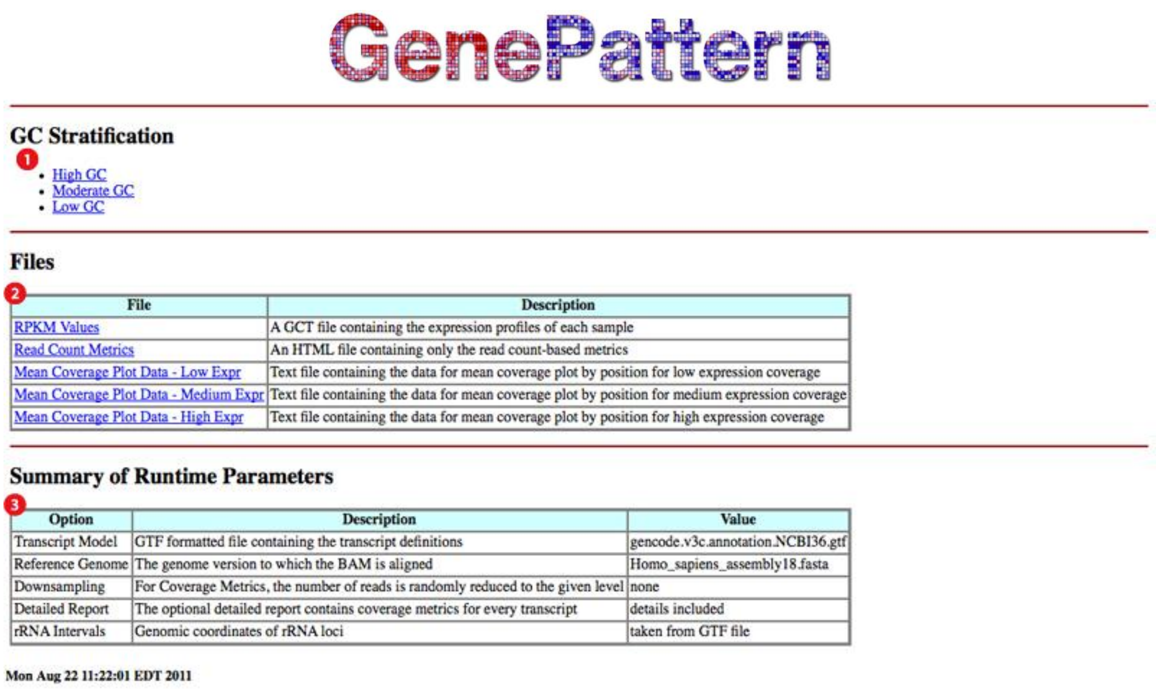


The following plot shows the mean coverage for the mid-range-expressed transcripts over the distance from the 3' end.



The following plot shows the mean coverage for the highly-expressed transcripts over the distance from the 3' end.





GC Stratification

- High GC
- Moderate GC
- Low GC

Files

File	Description
RPKM Values	A GCT file containing the expression profiles of each sample
Read Count Metrics	An HTML file containing only the read count-based metrics
Mean Coverage Plot Data - Low Expr	Text file containing the data for mean coverage plot by position for low expression coverage
Mean Coverage Plot Data - Medium Expr	Text file containing the data for mean coverage plot by position for medium expression coverage
Mean Coverage Plot Data - High Expr	Text file containing the data for mean coverage plot by position for high expression coverage

Summary of Runtime Parameters

Option	Description	Value
Transcript Model	GTF formatted file containing the transcript definitions	gencode.v3c.annotation.NCBI36.gtf
Reference Genome	The genome version to which the BAM is aligned	Homo_sapiens_assembly18.fasta
Downsampling	For Coverage Metrics, the number of reads is randomly reduced to the given level	none
Detailed Report	The optional detailed report contains coverage metrics for every transcript	details included
rRNA Intervals	Genomic coordinates of rRNA loci	taken from GTF file

Mon Aug 22 11:22:01 EDT 2011

1	Links to: • gc/high/index.html • gc/mid/index.html • gc/low/index.html (Will only be calculated and displayed on this page if a GC content file is provided at runtime.)
2	Links to: • exons.rpkm.gct • countMetrics.html • meanCoverage_low.txt • meanCoverage_med.txt • meanCoverage_high.txt
3	List of parameter choices for the run of the module.

All Result Files

The result files in the ZIP archive include:

- countMetrics.html
- exons.rpkm.gct
- index.html (detailed above)
- meanCoverage_high.png
- meanCoverage_high.txt
- meanCoverage_low.png
- meanCoverage_low.txt
- meanCoverage_medium.png
- meanConverage_medium.txt
- rRNA_intervals.list
- *<BAM file name>* folder
    - *<BAM file name>*.libraryComplexity.txt
    - *<BAM file name>*.metrics.tmp.txt
    - *<BAM file name>*.metrics.tmp.txt.rpkm.gct

- *<BAM file name>*.metrics.txt
- *highexpr* folder
  - *<BAM file name>* .DoCTranscripts
  - *<BAM file name>.*DoCTranscriptsSummary
  - *<BAM file name>*.transcripts.list
  - index.html
  - perBaseDoC.out
  - perBaseDoC.out.sample_statistics
  - perBaseDoC.out.sample_summary
  - *<transcript name>.*html (may be several; only created if transcript-level metrics are calculated)
  - *<transcript name>.*png (may be several; only created if transcript-level metrics are calculated and GNU Plots is available)
  - plots.html
- *lowexpr* folder
  - *<BAM file name>* .DoCTranscripts
  - *<BAM file name>.*DoCTranscriptsSummary
  - *<BAM file name>*.transcripts.list
  - index.html
  - perBaseDoC.out
  - perBaseDoC.out.sample_statistics
  - perBaseDoC.out.sample_summary
  - *<transcript name>.*html (may be several; only created if transcript-level metrics are calculated)
  - *<transcript name>.*png (may be several; only created if transcript-level metrics are calculated and GNU Plots is available)
  - plots.html
- *medexpr* folder
  - *<BAM file name>* .DoCTranscripts
  - *<BAM file name>.*DoCTranscriptsSummary
  - *<BAM file name>*.transcripts.list
  - index.html
  - perBaseDoC.out
  - perBaseDoC.out.sample_statistics
  - perBaseDoC.out.sample_summary
  - *<transcript name>.*html (may be several; only created if transcript-level metrics are calculated)
  - *<transcript name>.*png (may be several; only created if transcript-level metrics are calculated and GNU Plots is available)
  - plots.html

- *gc* folder: this information will only be calculated/output if a GC content file is provided at runtime; results are first stratified by their GC content, and then metrics for the high-, middle-, and low-expressed transcripts within that ranking
  - highgc.gtf
  - lowgc.gtf
  - medgc.gtf
  - *high* folder: contains a number of files regarding regions with high GC content
    - index.html
    - *meanCoverage_high.png*
    - *meanCoverage_high.txt*
    - *meanCoverage_low.png*
    - *meanCoverage_low.txt*
    - *meanCoverage_medium.png*

  - *meanCoverage_medium.txt*
  - *<sample name> folder*
    - highexpr folder: contains a number of files with high GC content and high expression
      - *<transcript name>.*html (may be several; only created if transcript-level metrics are calculated)
      - *<transcript name>.*png (may be several; only created if transcript-level metrics are calculated and GNU Plots is available)
      - index.html
      - perBaseDoC.out
      - perBaseDoC.out.sample_statistics
      - perBaseDoC.out.sample_summary
      - plots.html
    - *lowexpr* folder: contains a number of files with high GC content and low expression
      - *<transcript name>.*html (may be several; only created if transcript-level metrics are calculated)
      - *<transcript name>.*png (may be several; only created if transcript-level metrics are calculated and GNU Plots is available)
      - index.html
      - perBaseDoC.out
      - perBaseDoC.out.sample_statistics
      - perBaseDoC.out.sample_summary
      - plots.html
    - *medexpr* folder: contains a number of files with high GC content and medium expression
      - *<transcript name>.*html (may be several; only created if transcript-level metrics are calculated)
      - *<transcript name>.*png (may be several; only created if transcript-level metrics are calculated and GNU Plots is available)
      - index.html
      - perBaseDoC.out
      - perBaseDoC.out.sample_statistics
      - perBaseDoC.out.sample_summary
      - plots.html
- *low* folder: contains a number of files regarding regions with low GC content; file structure is the same as for the *high* folder
- *mid* folder: contains a number of files regarding regions with mid-range levels of GC content; file structure is the same as for the *high* folder

Example Data

Example input and output files are on the GenePattern FTP site:

- Input ZIP archive: ftp://ftp.broadinstitute.org/pub/genepattern/example_files/RNAseqMetrics/B019R.rna.GATKRecalibrated.flagged_input.zip
- Output ZIP archive: ftp://ftp.broadinstitute.org/pub/genepattern/example_files/RNAseqMetrics/B019R.rna.GATKRecalibrated.flagged_output.zip

Platform Dependencies

Module type:	RNA-seq
CPU type:	any
OS:	Mac OSX, Linux
Language:	Java (1.6 minimum)