GenePattern

RNASeQC Documentation

Description: Calculates metrics on aligned RNA-seq data.

Author: David S. 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.

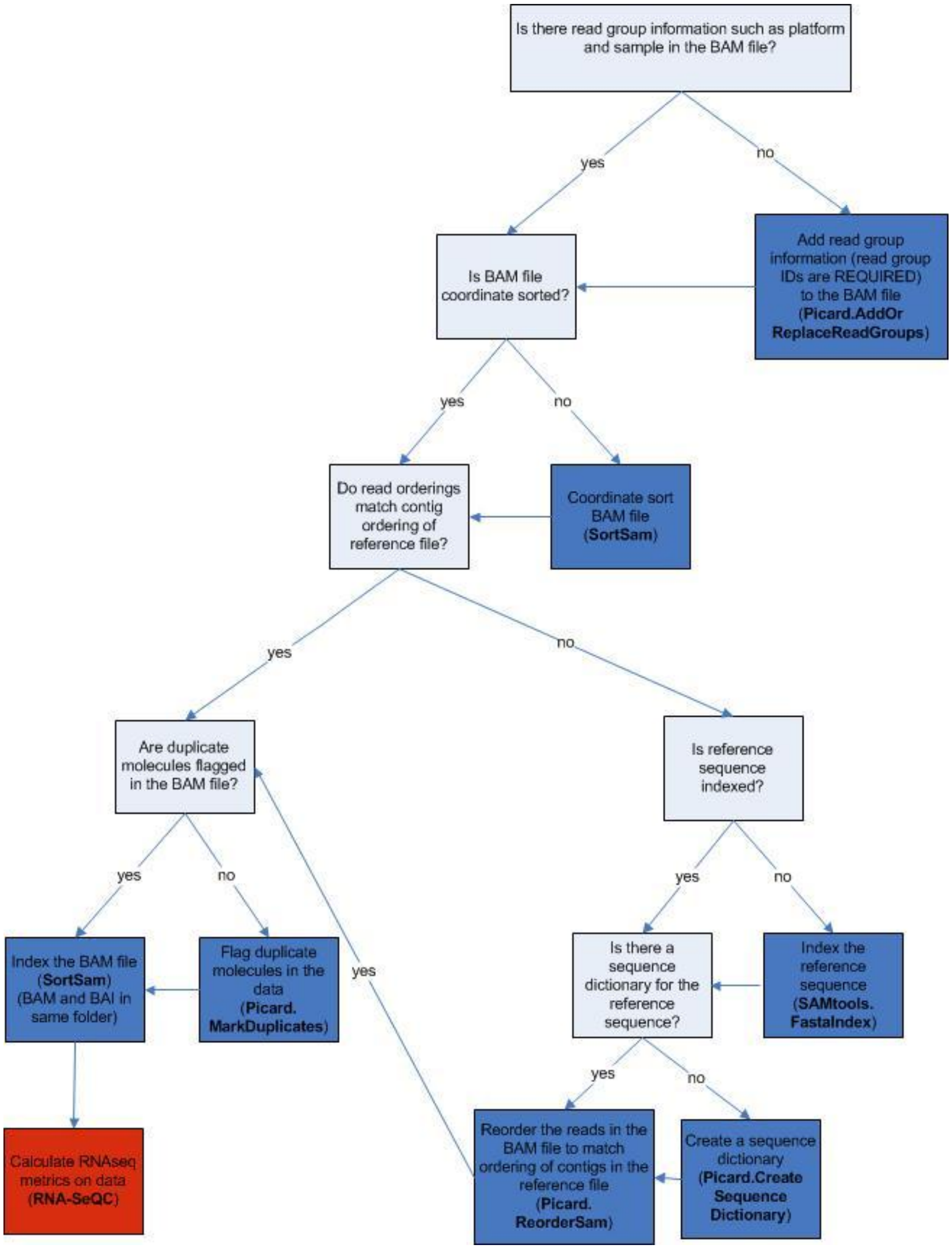
For more information on the BAM format, which is a binary form of the SAM format, see the SAM file specification here: http://samtools.sourceforge.net/.

Usage

The RNA-seq data must be preprocessed in a particular manner for the RNASeQC 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 RNASeQC 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

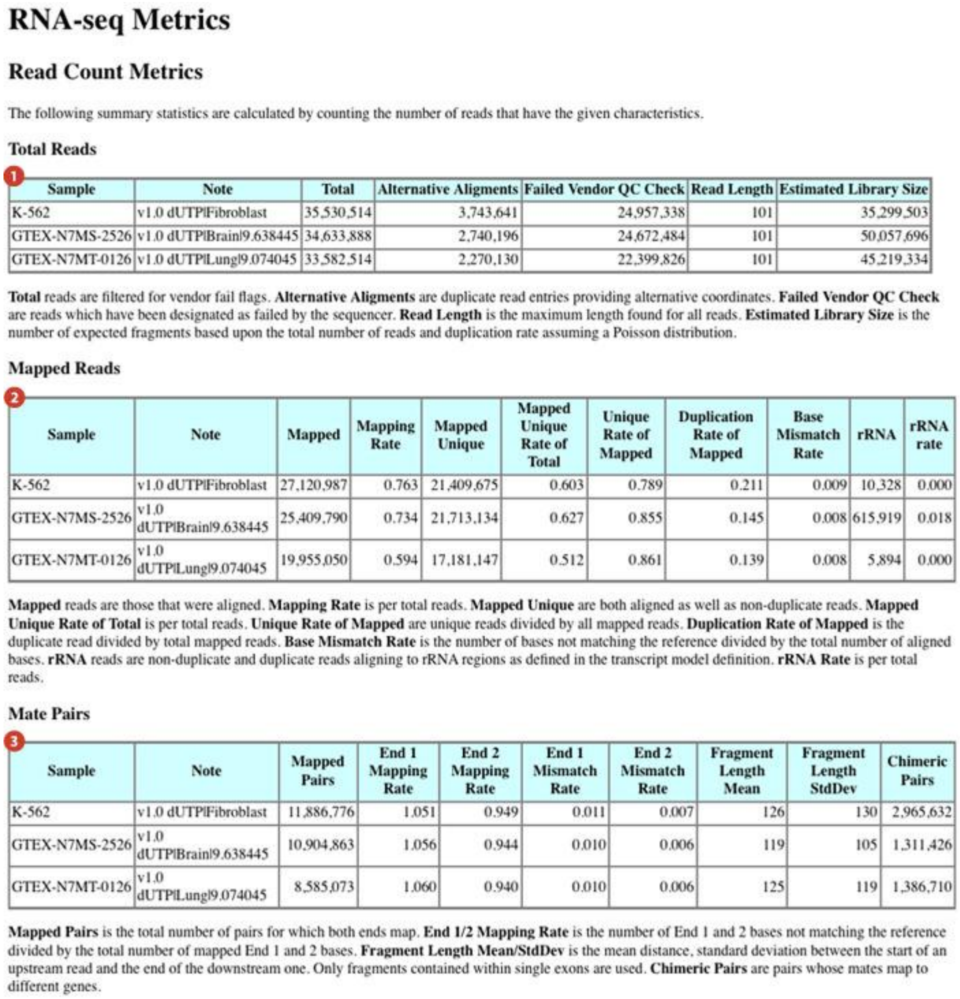
Output Files

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.

GenePattern

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	Alternative Aligments	Failed Vendor QC Check	Read Length	Estimated Library Size
K-562	v1.0 dUTP|Fibroblast	35,530,514	3,743,641	24,957,338	101	35,299,503
GTEX-N7MS-2526	v1.0 dUTP|Brain|9.638445	34,633,888	2,740,196	24,672,484	101	50,057,696
GTEX-N7MT-0126	v1.0 dUTP|Lung|9.074045	33,582,514	2,270,130	22,399,826	101	45,219,334

Total reads are filtered for vendor fail flags. **Alternative Aligments** are duplicate read entries providing alternative coordinates. **Failed Vendor QC Check** are reads which have been designated as failed by the sequencer. **Read Length** is the maximum length found for all 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

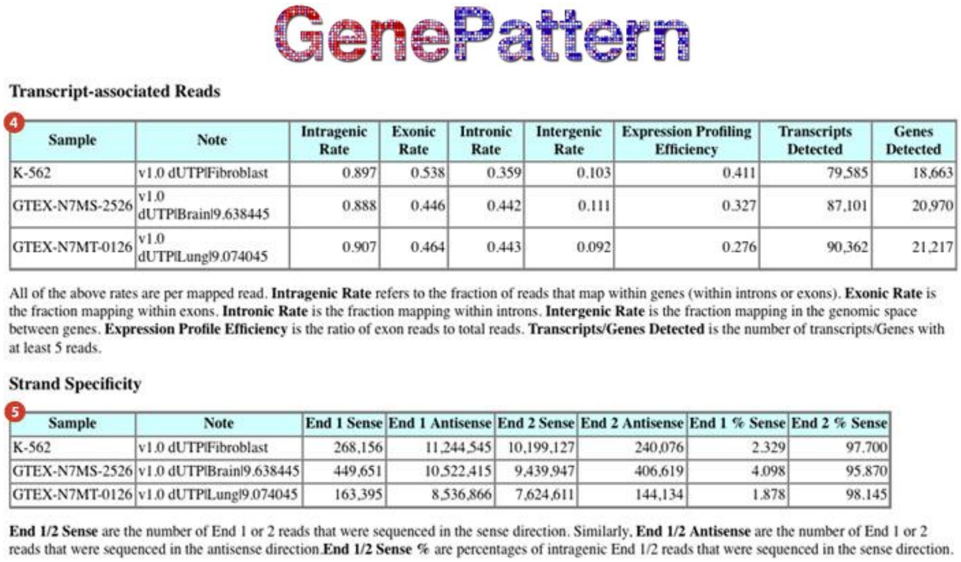
Sample	Note	Mapped	Mapping Rate	Mapped Unique	Mapped Unique Rate of Total	Unique Rate of Mapped	Duplication Rate of Mapped	Base Mismatch Rate	rRNA	rRNA rate
K-562	v1.0 dUTP|Fibroblast	27,120,987	0.763	21,409,675	0.603	0.789	0.211	0.009	10,328	0.000
GTEX-N7MS-2526	v1.0 dUTP|Brain|9.638445	25,409,790	0.734	21,713,134	0.627	0.855	0.145	0.008	615,919	0.018
GTEX-N7MT-0126	v1.0 dUTP|Lung|9.074045	19,955,050	0.594	17,181,147	0.512	0.861	0.139	0.008	5,894	0.000

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 of Total** is per total reads. **Unique Rate of Mapped** are unique reads divided by all mapped reads. **Duplication Rate of Mapped** is the duplicate read divided by total mapped reads. **Base Mismatch Rate** is the number of bases not matching the reference divided by the total number of aligned bases. **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.

Mate Pairs

Sample	Note	Mapped Pairs	End 1 Mapping Rate	End 2 Mapping Rate	End 1 Mismatch Rate	End 2 Mismatch Rate	Fragment Length Mean	Fragment Length StdDev	Chimeric Pairs
K-562	v1.0 dUTP|Fibroblast	11,886,776	1.051	0.949	0.011	0.007	126	130	2,965,632
GTEX-N7MS-2526	v1.0 dUTP|Brain|9.638445	10,904,863	1.056	0.944	0.010	0.006	119	105	1,311,426
GTEX-N7MT-0126	v1.0 dUTP|Lung|9.074045	8,585,073	1.060	0.940	0.010	0.006	125	119	1,386,710

Mapped Pairs is the total number of pairs for which both ends map. **End 1/2 Mapping Rate** is the number of End 1 and 2 bases not matching the reference divided by the total number of mapped End 1 and 2 bases. **Fragment Length Mean/StdDev** is the mean distance, standard deviation between the start of an upstream read and the end of the downstream one. Only fragments contained within single exons are used. **Chimeric Pairs** are pairs whose mates map to different genes.



GenePattern

Transcript-associated Reads

4

Sample	Note	Intragenic Rate	Exonic Rate	Intronic Rate	Intergenic Rate	Expression Profiling Efficiency	Transcripts Detected	Genes Detected
K-562	v1.0 dUTP¦Fibroblast	0.897	0.538	0.359	0.103	0.411	79,585	18,663
GTEX-N7MS-2526	v1.0 dUTP¦Brain¦9.638445	0.888	0.446	0.442	0.111	0.327	87,101	20,970
GTEX-N7MT-0126	v1.0 dUTP¦Lung¦9.074045	0.907	0.464	0.443	0.092	0.276	90,362	21,217

All of the above rates are per mapped read. **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. **Transcripts/Genes Detected** is the number of transcripts/Genes with at least 5 reads.

Strand Specificity

5

Sample	Note	End 1 Sense	End 1 Antisense	End 2 Sense	End 2 Antisense	End 1 % Sense	End 2 % Sense
K-562	v1.0 dUTP¦Fibroblast	268,156	11,244,545	10,199,127	240,076	2.329	97.700
GTEX-N7MS-2526	v1.0 dUTP¦Brain¦9.638445	449,651	10,522,415	9,439,947	406,619	4.098	95.870
GTEX-N7MT-0126	v1.0 dUTP¦Lung¦9.074045	163,395	8,536,866	7,624,611	144,134	1.878	98.145

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 to exclude reads with vendor fail or alternative alignment flags) • Alternative Alignments: Duplicate read entries providing alternative coordinates • Failed Vendor QC Check: Reads that have been designated as failed by the sequencer • Read Length: Maximum detected read length found • 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 of Total: Ratio of mapping of reads that were aligned and were not duplicates to total reads • Unique Rate of Mapped: Ratio of unique reads to all mapped reads • Duplication Rate of Mapped: Ratio of the number of duplicate reads to total mapped reads • Base Mismatch Rate: Ratio of the number of bases not matching the reference to the total number of bases aligned • 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	• Mapped Pairs: Total number of pairs for which both ends map • End 1 Mapping Rate: Ratio of the number of End 1 bases not matching the reference to the total number of mapped End 1 bases • End 2 Mapping Rate: Ratio of the number of End 2 bases not matching the reference to the total number of mapped End 2 bases • Fragment Length Mean and StdDev: The fragment length is the distance between the start of an upstream read and the end of the downstream pair mate • Chimeric Pairs: Pairs whose mates map to different genes
4	• 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 Transcripts Detected: Total number of transcripts with at least 5 exon mapping reads† • Genes Detected: Total number of genes with at least 5 exon mapping reads†

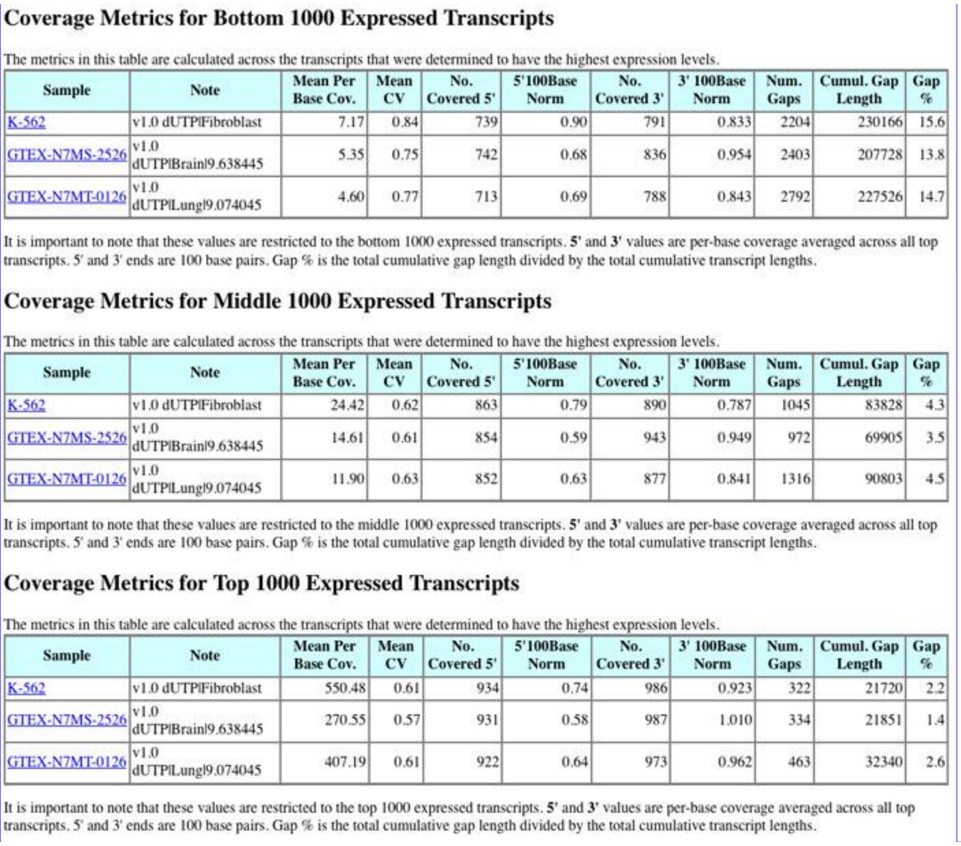
5	• End 1 Sense: Number of End 1 reads that were sequenced in the sense direction • End 1 Antisense: Number of End 1 reads that were sequenced in the antisense direction • End 2 Sense: Number of End 2 reads that were sequenced in the sense direction • End 2 Antisense: Number of reads that were sequenced 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

†Five reads were chosen as a default because this ensures a reasonable standard error.

[need new image here]

Comment [A1]: Need new version with correlation included.



Coverage Metrics for Bottom 1000 Expressed Transcripts

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

Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'100Base Norm	No. Covered 3'	3' 100Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
K-562	v1.0 dUTP|Fibroblast	7.17	0.84	739	0.90	791	0.833	2204	230166	15.6
GTEX-N7MS-2526	v1.0 dUTP|Brain|9.638445	5.35	0.75	742	0.68	836	0.954	2403	207728	13.8
GTEX-N7MT-0126	v1.0 dUTP|Lung|9.074045	4.60	0.77	713	0.69	788	0.843	2792	227526	14.7

It is important to note that these values are restricted to the bottom 1000 expressed transcripts. **5'** and **3'** values are per-base coverage averaged across all top transcripts. 5' and 3' ends are 100 base pairs. Gap % is the total cumulative gap length divided by the total cumulative transcript lengths.

Coverage Metrics for Middle 1000 Expressed Transcripts

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

Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'100Base Norm	No. Covered 3'	3' 100Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
K-562	v1.0 dUTP|Fibroblast	24.42	0.62	863	0.79	890	0.787	1045	83828	4.3
GTEX-N7MS-2526	v1.0 dUTP|Brain|9.638445	14.61	0.61	854	0.59	943	0.949	972	69905	3.5
GTEX-N7MT-0126	v1.0 dUTP|Lung|9.074045	11.90	0.63	852	0.63	877	0.841	1316	90803	4.5

It is important to note that these values are restricted to the middle 1000 expressed transcripts. **5'** and **3'** values are per-base coverage averaged across all top transcripts. 5' and 3' ends are 100 base pairs. Gap % is the total cumulative gap length divided by the total cumulative transcript lengths.

Coverage Metrics for Top 1000 Expressed Transcripts

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

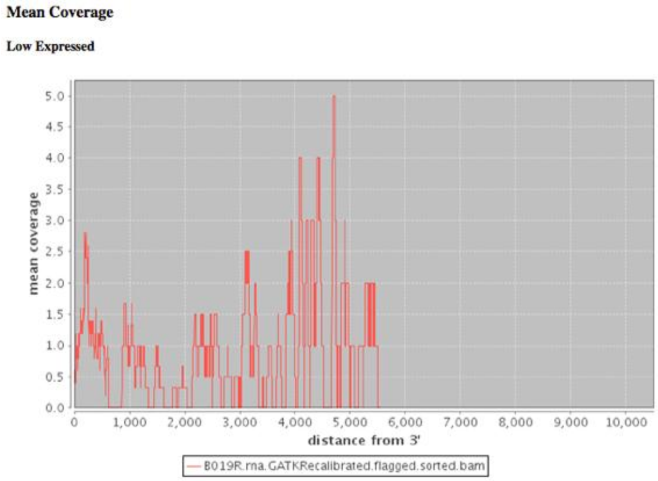
Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'100Base Norm	No. Covered 3'	3' 100Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
K-562	v1.0 dUTP|Fibroblast	550.48	0.61	934	0.74	986	0.923	322	21720	2.2
GTEX-N7MS-2526	v1.0 dUTP|Brain|9.638445	270.55	0.57	931	0.58	987	1.010	334	21851	1.4
GTEX-N7MT-0126	v1.0 dUTP|Lung|9.074045	407.19	0.61	922	0.64	973	0.962	463	32340	2.6

It is important to note that these values are restricted to the top 1000 expressed transcripts. **5'** and **3'** values are per-base coverage averaged across all top transcripts. 5' and 3' ends are 100 base pairs. Gap % is the total cumulative gap length divided by the total cumulative transcript lengths.

1	Spearman and Pearson Correlation Coefficients of the RNA-seq log (RPKM) values to the reference expression profile provided. (Will only be calculated and displayed on this page if a correlation comparison file is provided at runtime.) If multiple samples are provded, a correlation matrix is displayed, comparing samples among themselves.

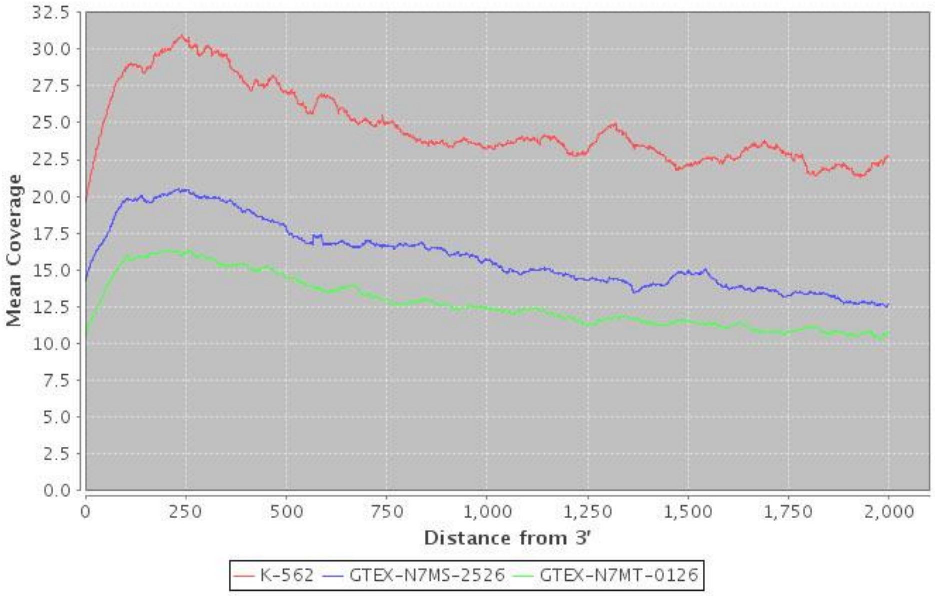
2	The coverage metrics are generated separately on three sets of transcripts. The set size (n) is defined by the user (n=1000 in the example). The bottom n transcripts are those transcripts taht have the lowest expression values (RPKM) but are non-zero. The middle n are taken from the transcripts surrounding the median non-zero expression levels. The top n are taken as the highest expressed transcripts. For each set, the following metrics are computed: • mean coverage per base: coverage is averaged per base across each transcript, and averaged again across all transcripts • mean coefficient of variation: standard deviation in base coverage 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

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

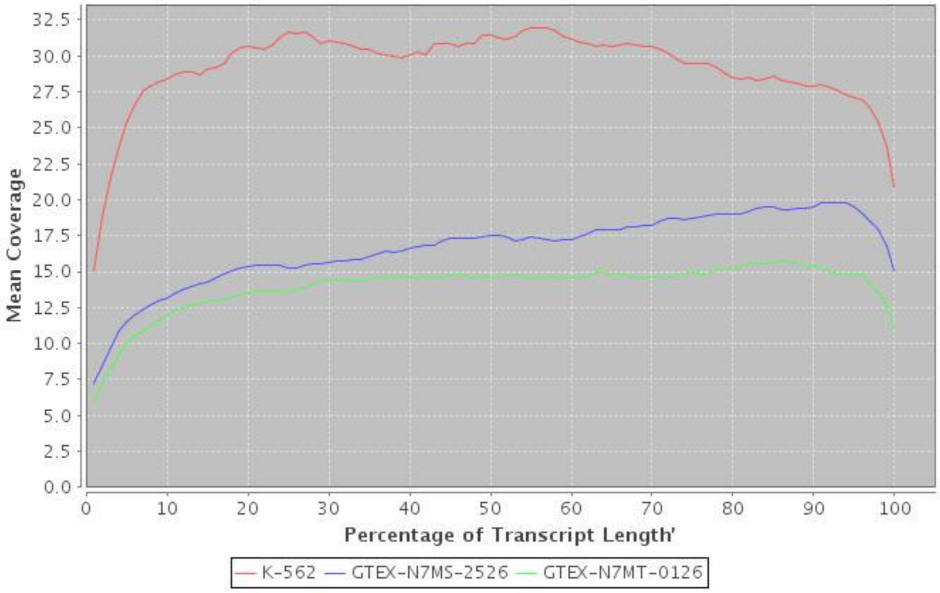


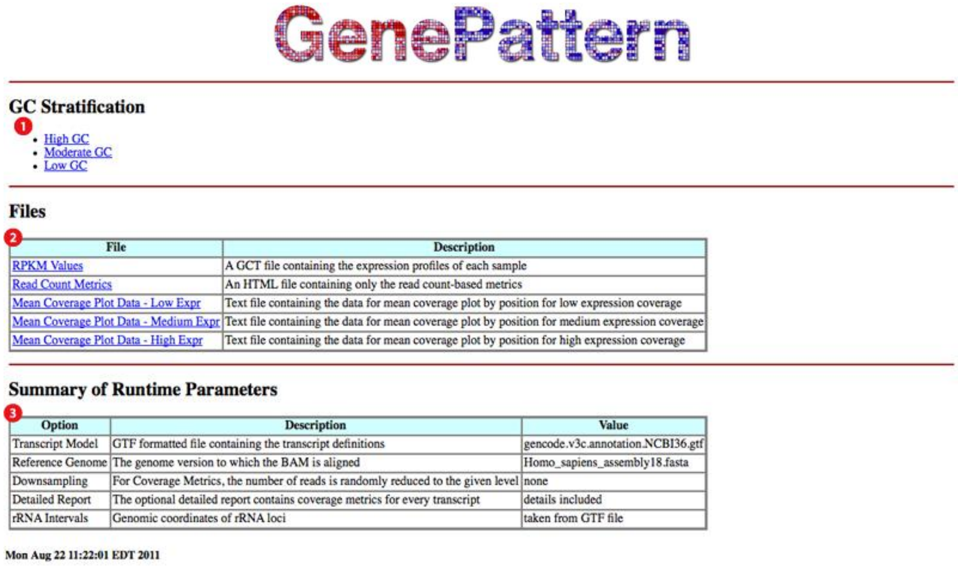
The following plot shows the mean coverage for the mid-range-expressed transcripts over the distance from the 3' end.

GenePattern



The following plot shows the mean coverage for expressed transcripts from 5' to 3' end, with the lengths of transcripts normalized to 1-100.





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	GC Stratification: The entire report is generated for 3 transcript subsets corresponding to different levels of GC content. The default setting of high (>62%), low (<38%), and moderate (everything in between) was chosen such that the lower and upper sets correspond to the lower and upper quartiles of transcripts by GC content in the human transcript set. Links are provided to each of the sub-reports: • 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: depth of coverage for each base
    - 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)