# rMATS-DVR User Manual

## Introduction

rMATS-DVR is a convenient and user-friendly program to streamline the discovery of DVRs (Differential Variants in RNA) between two RNA-seq sample groups with replicates. rMATS-DVR combines a stringent GATK-based pipeline for calling SNVs including SNPs and RNA editing events in RNA-seq reads, with our rigorous rMATS statistical model for identifying differential isoform ratios using RNA-seq sequence count data with replicates.

## Availability

rMATS-DVR is a free software, which can be downloaded from https://github.com/Xinglab/rMATS-DVR.

## Requirements

1. Install Python 2.6.x or Python 2.7.x and corresponding versions of NumPy and SciPy.
2. Install Java v1.8 (https://java.com/en/download/manual.jsp#lin).
3. Add the Python and the Java directories to the $PATH environment variable.
4. Download Picard v2.6.0 from https://github.com/broadinstitute/picard/releases/tag/2.6.0.
5. Download GATK v3.6 from https://software.broadinstitute.org/gatk/download/.
6. Download SAMtools v1.3.1 (https://github.com/samtools/samtools/releases/tag/1.3.1) and add it to $PATH environment variable.

Note: We have only tested rMATS-DVR in the linux platform.

## Installation

1. Create soft links of Picard and GATK into rMATS-DVR folder.

```
- cd rMATS-DVR
- ln -s /path/to/picard/picard.jar ./
- ln -s /path/to/GATK/GenomeAnalysisTK.jar ./
```

2. Then the source code can be directly called from Python.

## Required and optional external files

All the external files based on human hg19 genome can be downloaded from http://www.mimg.ucla.edu/faculty/xing/public_data/rMATS-DVR/hg19_resource.tar.gz.

To decompress and extract the files:

```
tar -xvf hg19_resource.tar.gz
```

Alternatively, users can also prepare the external files under the following instructions:

1) Genome (required): we highly recommend the users use the genome sequence (together with a dictionary file such as "ucsc.hg19.dict" and index file such as "ucsc.hg19.fasta.fai") from the GTAK bundle (https://software.broadinstitute.org/gatk/download/bundle). Alternatively, please follow the instructions in GATK (https://software.broadinstitute.org/gatk/guide/article?id=1204) to prepare the reference genome in the proper format.

2) Known SNPs (optional but highly recommended): SNP annotation in the VCF format. The dbSNP annotation of human can be downloaded from the GTAK bundle (https://software.broadinstitute.org/gatk/download/bundle). Alternatively, please follow the instructions in GATK (https://software.broadinstitute.org/gatk/guide/article?id=1204) to prepare the valid VCF file.

3) Known RNA editing sites (optional): table delimited txt file with the first two columns as chromosomes and coordinates. The other columns are ignored. Header is optional. Users can download the file from the RADAR dababase (http://rnaedit.com/download/).

4) Genome-wide repeat elements (optional): RepeatMasker Genomic Datasets downloaded from http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html. For example, hg19.fa.out.gz.

5) Gene annotation (optional): the gene annotation is in the GenePred (extended) format (see https://genome.ucsc.edu/FAQ/FAQformat.html for detail). We recommend users to download it from UCSC (http://hgdownload.soe.ucsc.edu/downloads.html). For example, one can download the hg19 RefSeq gene from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz.

**Run rMATS-DVR in one step**

In the one step mode, rMATS-DVR will first calibrate the bam files one by one and then calculate the Differential Variants of RNA using all samples. Mapping with STAR is highly recommended.

Usage:

```
python rMATS-DVR.py --sample1 S1_rep_1.bam[,S1_rep_2.bam][,...,S1_rep_n.bam] --sample2 S2_rep_1.bam[,S2_rep_2.bam][,...,S2_rep_n.bam] --label S1,S2 --genome hg19.fa --output /Path/to/output/S1_vs_S2 [--known dbSNP147.vcf] [--editing RADAR2.txt] [--repeat repeats.txt] [--gene RefSeq.txt] [--minQ 20] [--minDP 5] [--thread 1] [--diff 0.0001] [--merge] [--ReadStranded] [--ReadPaired] [--skipBamCalibration] [--KeepTemp]
```

Required Parameters:

-h, --help            Show this help message and exit.

--sample1   <str>        Bam (or sam) files of sample 1, replicates are separated by comma.

--sample2   <str>        Bam (or sam) files of sample 2, replicates are separated by comma.

--label     <str>        Lables of sample 1 and sample 2, separated by comma, e.g. Sample1,Sample2.

--output    <str>        Path and prefix of the output file.

--genome    <str>        Genome sequence in the fasta format.

Optional Parameters:

--known     <str>        Known SNPs in the VCF format.

--editing   <str>        Known RNA editing sites.

--repeat    <str>        Repeat elements annotation.

--gene      <str>        Gene annotation.

--minQ      <int>        Minimum variant quality. [20]

--minDP     <int>        Minimum mean read coverage of both samples. [5]

--thread    <int>        Number of processors. [1]

--diff      <float>      Required level difference between the two samples. [0.0001]

--ReadStranded RNA-seq reads are Illumina strand-specific reads. Disable by default.

--ReadPaired RNA-seq reads are paired-end reads. Disable by default.

--merge Merge the counts of all replicates. Enable by default when there are less than 2 replicates in at least one sample group.

--skipBamCalibration Skip the step of calibrating the bam files. Enable it when the input bam files have already been calibrated using bam_calibration.py (see below). Disable by default.

--KeepTemp Keep the temporary files. Disable by default.

## Run rMATS-DVR in two steps

When there are a large number of replicates, the one step mode, which calibrate bam files one by one, may take long time. In these cases, we recommend users to run the bam calibration for all bam files in parallel at the first step. Then the users can run rMATS-DVR.py with --skipBamCalibration.

Usage:

```
python bam_calibration.py --bam sample.bam --output /Path/to/output/prefix --genome hg19.fa [--known dbSNP147.vcf] [--KeepTemp]
```

Parameters:

-h, --help Show this help message and exit.

--bam <str> Input bam (or sam) file.

--output <str> Path and prefix of the output file.

--genome <str> Genome sequence in the fasta format.

--known <str> [Optional] Known SNPs in the VCF format.

--KeepTemp [Optional] Keep the temporary files. Disable by default.

## Output

The final output files are in "Prefix_rMATS-DVR_results" folder, including "rMATS-DVR_Result.txt" and "rMATS-DVR_Result_summary.txt".

"rMATS-DVR_Result.txt" provides the variant information, read counts, P value, FDR, gene location, and multiple annotations based on the known databases. "rMATS-DVR_Result_summary.txt" summarizes the frequencies of all types of total variants and DVRs respectively, substratified into known SNPs, known RNA editing sites, and novel variants. All other files are temporary files.

1) rMATS-DVR_Result.txt

Chroms: Chromosome of variant.
Site: 1-based coordinates of variant.
Ref_allele: reference allele.
Alt_allele: alternative allele.
RNA-seqStrand: RNA strand from which the RNA-seq reads are originated. Only valid when --ReadStranded is applied in rMATS-DVR.
Variant_quality: GATK reported Phred-scaled quality score of the variant.
Sample1_Alt: read counts of the alternative allele in sample 1, replicates are separated by comma.
Sample1_Ref: read counts of the reference allele in sample 1, replicates are separated by comma.
Sample2_Alt: read counts of the alternative allele in sample 2, replicates are separated by comma.
Sample2_Ref: read counts of the reference allele in sample 2, replicates are separated by comma.
Pvalue: P value of differential allelic count ratios between the two sample groups.
FDR: Benjamini-Hochberg corrected FDR of the above P value.
Sample1_Alt_allele_fraction: fraction of the alternative allele counts in sample 1, replicates are separated by comma.
Sample2_Alt_allele_fraction: fraction of the alternative allele counts in sample 2, replicates are separated by comma.
Alt_allele_fraction_diff: average (Sample1_Alt_allele_fraction) - average (Sample2_Alt_allele_fraction).
Genename: name of the gene in which the variant is located.
Strand: strand of the gene.
Ref_onSense: reference allele on the sense strand.
Alt_onSense: alternative allele on the sense strand.
Location: location of the variant on gene.
KnownSNP: rs ID of the known SNP (dbSNP) hit.
KnownRNAediting: boolean variable to show whether the variant has a hit in the known RNA editing database.
RepeatName: name of repeat element which covers the variant.
RepeatName: family of repeat element which covers the variant.

2) rMATS-DVR_Result_summary.txt

Type (Ref-Alt) on sense strand: type of variants in the format of reference allele-alternative allele on the sense strand.

All Variants: frequency of each type of all called variants.

All DVRs (FDR<0.05): frequency of each type of all variants with FDR < 0.05.

SNP DVRs: frequency of each type of all known SNPs with FDR < 0.05.

RNA editing DVRs: frequency of each type of all known RNA editings with FDR < 0.05.

Novel DVRs: frequency of each type of all novel variants with FDR < 0.05.

**Run with the least input**

For species without SNP or transcript annotation, users can run rMATS-DVR with only the RNA-seq alignments and the corresponding genome. In that case, rMAT-DVR still reports the variant information and the DVR information (column 1-15) in rMATS-DVR_Result.txt. It is also helpful to provide a *de novo* transcript assembly in the GenePred (extended) format as previously described.

**Contacts**

Yi Xing: yxing@ucla.edu

Jinkai Wang: jinkwang@ucla.edu

If you find a bug or mistake in this project, we would like to know about it. Before you send us the bug report, please check the following:

Are you using the latest version? The bug you find may have already been fixed.

Check that your inputs are in the correct formats and you have selected the correct options.

Please reduce your input to the smallest possible size that still produces the bug; we will need your input data to reproduce the problem, and the smaller you can make it, the easier it will be.

**Copyright and License Information**

Authors: Jinkai Wang and Yi Xing