TREAT User Guide, version 1.0

Department of Biomedical Statistics and Informatics, Mayo Clinic

Mar 20 2011

Contents

1. Introduction
2. Requirements
    a. Treat Workflow
    b. Examples:
        a. Fastq as Input
        b. Bam as Input
        c. Variants as Input

Introduction:
TREAT, Targeted RE-sequencing Annotation Tool, offers an end-to-end solution for analyzing and interpreting targeted re-sequencing data. TREAT comprises of different modules which includes sequence alignment, variant calling, variant annotation, variant filtering (default), and visualization. In addition, an Amazon Cloud Image of the TREAT is also provided for researchers with no access to local bioinformatics infrastructures. The source code and the executables of TREAT are available for download.
http://mayoresearch.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm

Platform: Currently, TREAT only analyzes Illumina GA/Hiseq data properly. It is not appropriate for 454 or Solid Data.

Requirements:
To use TREAT at least the following two

- If BAM (Li et al., 2009) is an input then the reference sequence file used to align the reads is needed, we actually recommend starting with fastq as input so as to get accurate results.
- The treat workflow uses various scripting language so having PERL, JAVA, and shell scripting in your environment is necessary
- Workflow uses different next gen sequencing tools like Samtools (Li et al., 2009) and Picard

TREAT WORKFLOW

The Treat Workflow has four modules:

1. Alignment
2. Variant
3. Annotation
4. All

Before running the Treat Workflow you need to manually create three configuration file,

1. Configuration file
   In the configuration file you need to specify the path for the different tools in a specific manner. The identifier for each tool must be same as the example configuration file in the example folder as the workflow uses the tool identifier to get the path for the specific tool. Provide full path to all the tools and references.
   Example:

```
## path to all the scripts and reference genomes and tools

script_path=path_to_treat_workflow_files
bwa_ref=path_to_bwa_ref
bowtie_ref=path_to_bowtie_ref
human_ref=path_to_ref
sift_ref=path_to_sift_ref
dbsnp_ref=path_to+dbSNP_ref
bwa_path=path_to_bwa
bowtie_path=path_to_bowtie
java_path=path_to_java
samtools_path=path_to_samtools
gatk_path=path_to_GTAK
sift_path=path_to_SIFT
OnTarget=OnTarget_file
```

```
bed_path=path_to_bedtools
dbsnp_rsids=dbsnp_rsids_file
snvmix_path=path_to_SNVmix tool
kgenome=path to kgenome data
picard_path=path_to_picard
GeneIdMap=geneid to gene name mapping file
hapmap=path to hapmap data
ref_flat=ref flat file
codon_ref=codon usage file
UCSC_tracks=ucsc_tracks reference file
```

2. Sample info file
   This file has information about the samples and name of the fastq files or BAM files or variant file depends on what module you want to run.

   For ALL or Alignment module

```
Option 1: One Paired End sample per lane

        ##Here sample name follows '=' sign and then read1 and read2 are tab separated (just
        specify the name of the file)
        ##specify the number of samples
        ##lane Information is: separated

        sampleA=NameOfRead1ForSampleA NameOfRead2ForSampleA
        sampleB=NameOfRead1ForSampleB NameOfRead2ForSampleB
        ...
        lanes=1:2:3
        num=XX

Option 2: One Single End sample per lane

        ## if its a Single read per lane then

        sampleA=NameOfReadForSampleA
        sampleB=NameOfReadForSampleB
        ...
        lanes=1:2:3
        num=XX

Option 3: Multiple Lanes per sample for Paired End

        ## If there are multiple lanes per sample then specify multiple lanes on different rows
        but the sample name should be same for multiple rows
        ## Lane Information for same sample should be, separated

        sampleA=NameOfRead1ForSampleA NameOfRead2ForSampleA
        sampleA=NameOfRead1ForSampleA NameOfRead2ForSampleA
        ....
        lanes=1,2:3:4
        num=XX
Option 4: Multiple lanes per sample for Single End

        ## If there are multiple lanes per sample then specify multiple lanes on different rows
        but the sample name should be same for multiple rows
        ## lane Information for same sample should be , separated

        sampleA=NameOfReadForSampleA
        sampleA=NameOfReadForSampleA
        ....
        lanes=1,2:3:4
        num=XX

        NOTE:
        Please specify only one option for each run
```

   For Variant Module

```
sampleA=BAMforSampleA
sampleB=BAmforSmapleB
.....
lanes=1:2
num=XX (number of samples)
```

For Annotation Module

```
snv_sampleA=nameOfFile
indel_sampleA=nameOfFile
.....
lanes=1:2
num=XX (number of samples)
```

3. Run info file

```
## information about the run and parameter user want to use to run TREAT
tool=type_of_analysis
Date=mm/dd/yyyy
Aligner=BWA/BOWTIE
SNV_caller=SNVmix/GATK
paired=1/0
ReadLength=XXX
Disease=XXX
queue=queue for the cluster
variant_type=BOTH/SNV/INDEL
PI=lastname_firstname_lanID
output_dir=output directory location
input_dir=path to input directory
email=lastname.firstname@mayo.edu
sampleNames=sampleA:sampleB
config=path to configuration file
sample_info=path to sample info file
analysis=mayo/all/variant/annotation
output_folder=name for the output folder
center=MAYO/TGEN/BMC/XXX
platform=illumina
GenomeBuild=hg18/hg19
SampleInformation=Free Text for HTML report (information about the samples)
```

Limitations:

- Sample name should not start with a number and dot (.) is not permitted in the sample name
- If user have multiple BAM files for same sample, then user need to preprocess the BAM to make one BAM per sample, which can be done using Samtools merge module

```
samtools merge <out.bam> <in1.bam> <in2.bam> [...]
```

NOTE:

All the reference files are pre processed, so we assume that you downloaded all the reference files from our website.

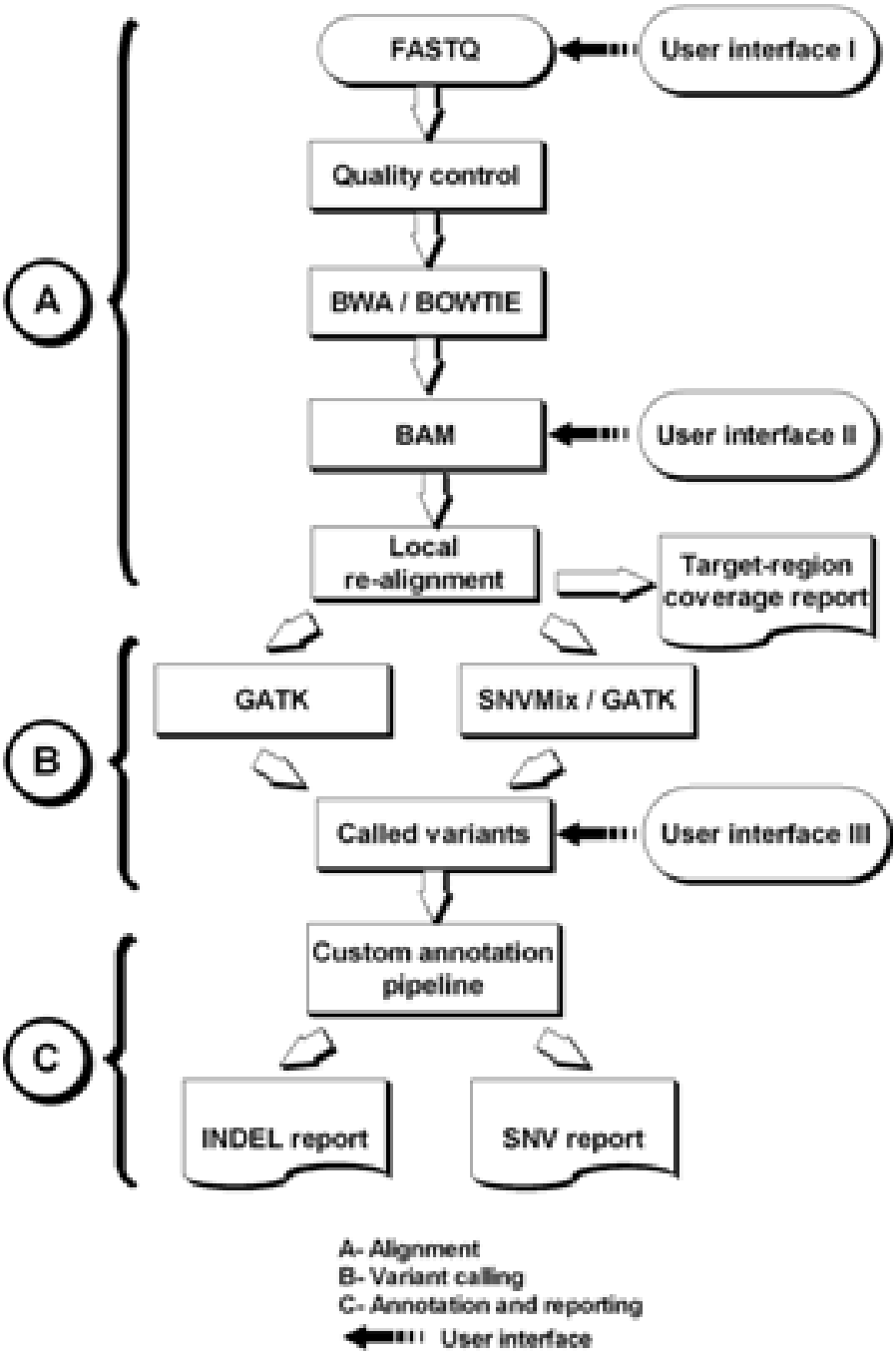


Figure 1: TREAT workflow

How to run TREAT

TREAT is just a command line tool. Run treat.sh to get the usage information.

```
$ /path_to_workflow/treat.sh
Usage: <ToolInfo/RunInfoFile>
(Please specify full path to the folder and directory)
```

Alignment:

Before aligner your reads we suggest to use FASTQC to get QC information for the reads. FASTQC gives out an HTML report which has different flags for each read and has some suggestions. This suggestions and warnings are just a consensus from the general trend so this is not your project specific.
We have two inbuilt aligner to choose to do the sequence alignment i.e. BWA (Li and Durbin, 2009) and BOWTIE (REF). Both the aligner fast light weighted tool that aligns relatively short sequences to the reference target sequence. Both the aligners are mapped aligner allows mismatches which is Indel for INDEL detection.
This module takes Fastq's as input, all in one folder. There is no specific name required for fastq; you can name your fastq files in any manner. Specify the aligner from above two for the analysis. The output will be sorted and indexed BAM, which can be used to visualize using IGV. The output also contains and HTML report with some basic information about the run, Metadata and mapping information.

For alignment we use BWA and BOWTIE;
Parameters for BWA are

```
-l INT    seed length [32]
-t INT    number of threads [4]
```

Standard parameters for excellent performance

Parameters for BOWTIE

```
-S/--sam      write hits in SAM format
-m <int>      suppress all alignments if > <int> exist (def: no limit)     [1]
--best hits guaranteed best stratum; ties broken by quality
 --strata     hits in sub-optimal strata aren't reported (requires --best)
--quieT       validation stringency
```

We assume that you have all the reference files and three files prepared to run the workflow; configuration, sample info and run info.

TREAT is just a command line tool. Run treat.sh to get the usage information.

```
$ /path_to_workflow/treat.sh
Usage: <ToolInfo/RunInfoFile>
(Please specify full path to the folder and directory)
```

Variant:

The input for this module is BAM for each sample. We need all the BAM's should be in one folder. We don't require the BAM's should be sorted or Indexed as in the workflow we sort and add the read group and platform information to the bam for further processing.
We assume that you have all the reference files and three files prepared to run the workflow; configuration, sample info and run info.

TREAT is just a command line tool. Run treat.sh to get the usage information.

```
$ /path_to_workflow/treat.sh
Usage: <ToolInfo/RunInfoFile>
(Please specify full path to the folder and directory)
```

Variant module includes realignment of input BAMs or aligned BAM from TREAT workflow, to get the appropriate calling using variant callers. In this module we have two options to choose for SNV calling i.e. GATK or SNVmix

Filtering Criterion:

- Min Mapping quality – 20
- Min base quality – 20
- Filter out reference Homozygous calls
- Filter out variant calls with probability less than 0.8 (SNVmix specific)

The output from the workflow is two annotated reports per samples with links to the visualization and different annotations for a specific variant call. First file is On Target raw file and second is filtered file with default filtering criterion. There is a HTML report which gives out some Meta data information provided by user and Sample table with specific information on each row for the variant call.

Annotation:

The input for this module and variant calls text tab delimited files. For SNVs the workflow requires four tab delimited columns namely chromosome, Position, Reference and Alternate allele.

Example:

```
chr1   100    A      T
chr1   1000   C      T
chr2   100    G      C
……..
```

For INDELs the workflow requires GATK indel format which has four tab delimited columns namely chromosome, start, stop and information about the IDNEL.

Example:

```
chr1   100    100    +A:34/45
chr1   1000   1001   -T:23/34
```

……..

In this format if there is an insertion then start = stop. For deletion start ≠ stop, stop=start + #Bases.
Last column starts with '+' or '-' for insertion or deletion respectively followed by base/s inserted or deleted. Last two numbers from the example row are indel supported reads and read depth for the INDEL.

We assume that you have all the reference files and three files prepared to run the workflow; configuration, sample info and run info.

TREAT is just a command line tool. Run treat.sh to get the usage information.

```
$ /path_to_workflow/treat.sh
Usage: <ToolInfo/RunInfoFile>
(Please specify full path to the folder and directory)
```

ALL:
This is the better approach to analyze the data. If you have bams then also you can use this module by converting the BAMs using Picard module. The Input for this module is fastq files for all the sample need to analyze in one folder as input folder. There is no specific naming convention required for the files.
A minimal FASTQ file might look like this:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Fastq's from the Illumina Sequencer has sequence Identifier like

```
@HWUSI-EAS100R:6:73:941:1973#0/1
@HWUSI-EAS100R:6:73:941:1973#0/2
```

The output is reports per sample and a merged report for INDELs and SNVs. There is a filtered report for variants with default filtering. A detailed HTML report is also generated with information about the samples (Meta data given by user). The report gives per sample statistics which is broken down in a manner which make sense to the investigator.

Sample HTML Report

Contents

- 1 Project Title
- 2 Initial meeting/email and Timeline
- 3 Project Description
  - 3.1 Background
  - 3.2 Study design
- 4 Analysis plan
- 5 Received Data
  - 5.1 Gerald folder
  - 5.2 Sample Summary
  - 5.3 Lane results summary
- 6 Results Summary
  - 6.1 QC steps
  - 6.2 Statistics based on per sample analysis
  - 6.3 Percent coverage of target region
- 7 Results and Conclusions
- 8 Results Delivered

I. Project Title :

NGS Bioinformatics for Exome sequencing

II. Initial Meetings and Time line:

Item	Date
Email(transfer complete date)	2/11/2011
Deadline	2/19/2011
Completed	2/19/2011
Results Delivered	2/19/2011

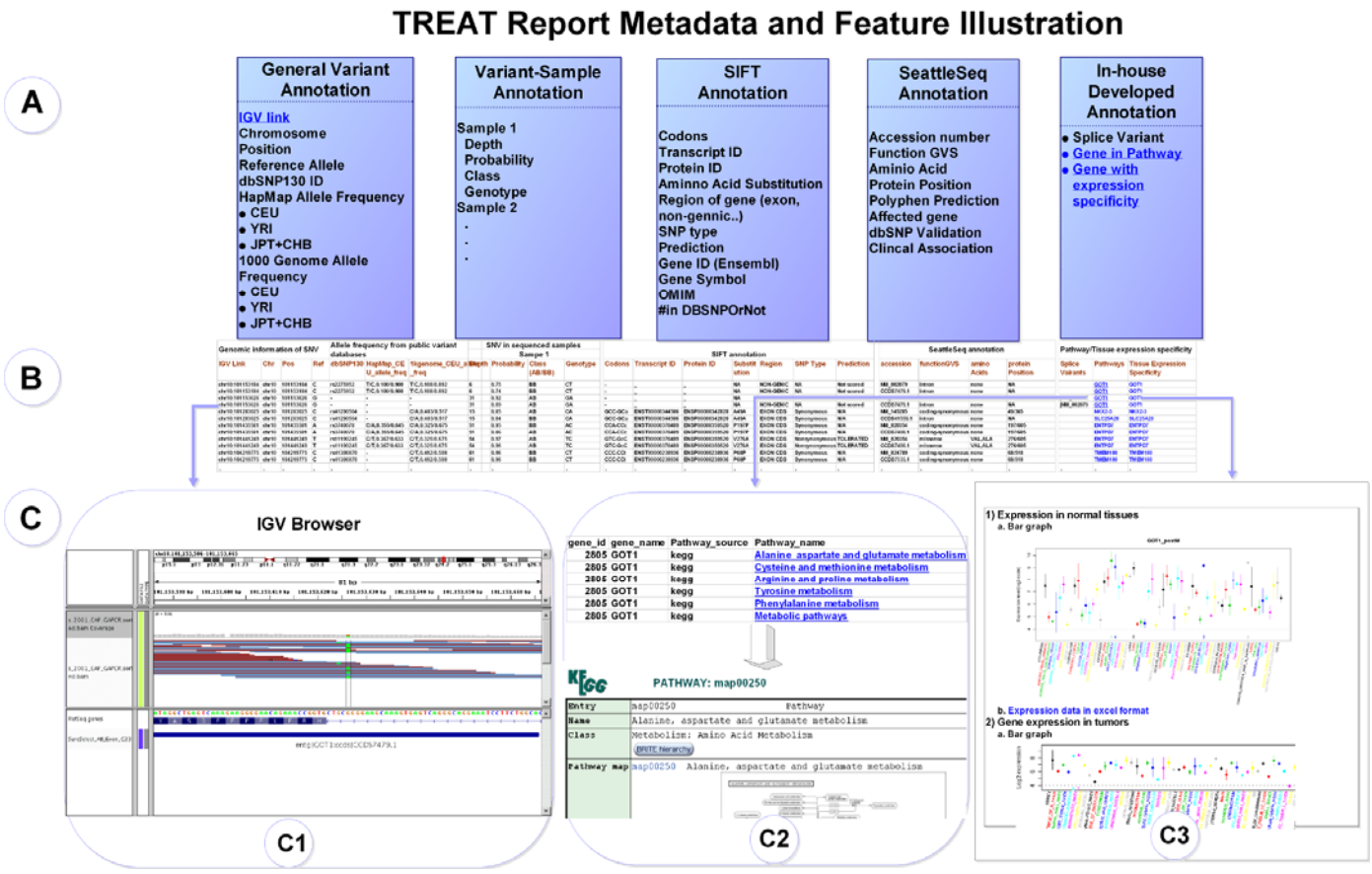
III. Project Description

Sample Statistics Table

2.Statistics based on per Sample Analysis

	s_53031	s_54378	s_58664	s_110459	s_241025	s_245640	s_247358	s_295098
Total Number of Reads	189332484	180697110	182829704	172789060	162442986	161042734	185082166	187612962
Number of Mapped Reads	184213635	176084481	177487119	167557539	157548919	156637183	180093947	182139733
Percentage of Mapped Reads	97.3%	97.4%	97.1%	97.0%	97.0%	97.3%	97.3%	97.1%
Mapped Reads on Target	121780001	116997288	116832483	108068893	103135732	100818634	115273262	118758202
Percentage of Mapped Reads on Target	66.1%	66.4%	65.8%	64.5%	65.5%	64.4%	64.0%	65.2%
Coverage at 10X	199862	199206	200022	199624	199034	199645	200217	199681
Coverage at 20X	193634	192844	193733	192949	192212	192943	193873	193734
Coverage at 30X	188584	187672	188690	187538	186661	187317	188781	188848
Coverage at 40X	184049	182870	183982	182689	181477	182232	184226	184558
Total SNVs	7271893	7391570	7157974	6828478	6718386	6611872	7462016	7565287
Filtered SNVs	567775	543962	648772	565266	512278	549750	657389	627539
On-target Filtered SNVs	39725	38462	47971	38544	38315	38421	38948	38536
Transition to Transversion ratio	2.61	2.64	2.66	2.64	2.65	2.62	2.65	2.64
In dbSNP130	35927	35077	40444	35348	35115	35139	35414	35081
Not in dbSNP130	3798	3385	7527	3196	3200	3282	3534	3455
Total INDELs	25823	25482	29167	24904	22700	24865	28386	28248
On-target INDELs	2307	2253	2689	2245	2288	2260	2267	2293
INDELs leading to frameshift	225	209	230	211	199	199	214	226
INDELs in coding region	105	114	119	106	113	104	102	107
INDELs in splice sites	25	20	29	21	26	17	20	22

Expected Reports from the Workflow using Variant, annotation or all module



Examples:

FASTQ as input

```
@R0211812_0096:1:1:1223:2139#0/1
NGTGAGCCACAGCACCCAGCCTGCTTCATCTGTTAAATAAGGTTTACAGCATCTGCAACCCTTAACTCCTTGGTTTATGGTTATTCCATAACCCA
AACAA
+R0211812_0096:1:1:1223:2139#0/1
BSUUJPTXNTPVWPW\[U]\c_cccccPcc[][[GIIKGGVWXX[SVXVY_X___BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
BBBBB
```