Treat Workflow scripts and Files Description

Treat workflow is executed using one standalone script which in turn calls different scripts to get to the final results. To execute the TREAT it requires three reference files. Example reference file can be found in TREAT scripts folder named examples.

- Configuration file (path to various tools and reference datasets)

```
## 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 references
```

- Run information file (meta data information about the project, how many samples need to be processed, parameters user want to use in the workflow)

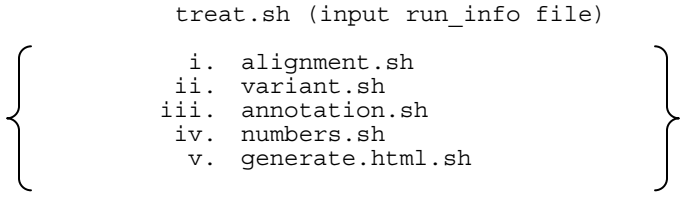
```
## information about the run and parameter user want to use to run TREAT

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)
```

- Sample information file (names of the files user want to use in the workflow, as this tool is independent of the naming convention of the input files)

There are different formats for the sample info files, depends on which module user wants to use, refer to the user manual for all the sample info formats.

```
$script_path/treat.sh
Usage:
```

1. <run_info file>



All the steps are interdependent, and each is executed after the completion of the last one executed.

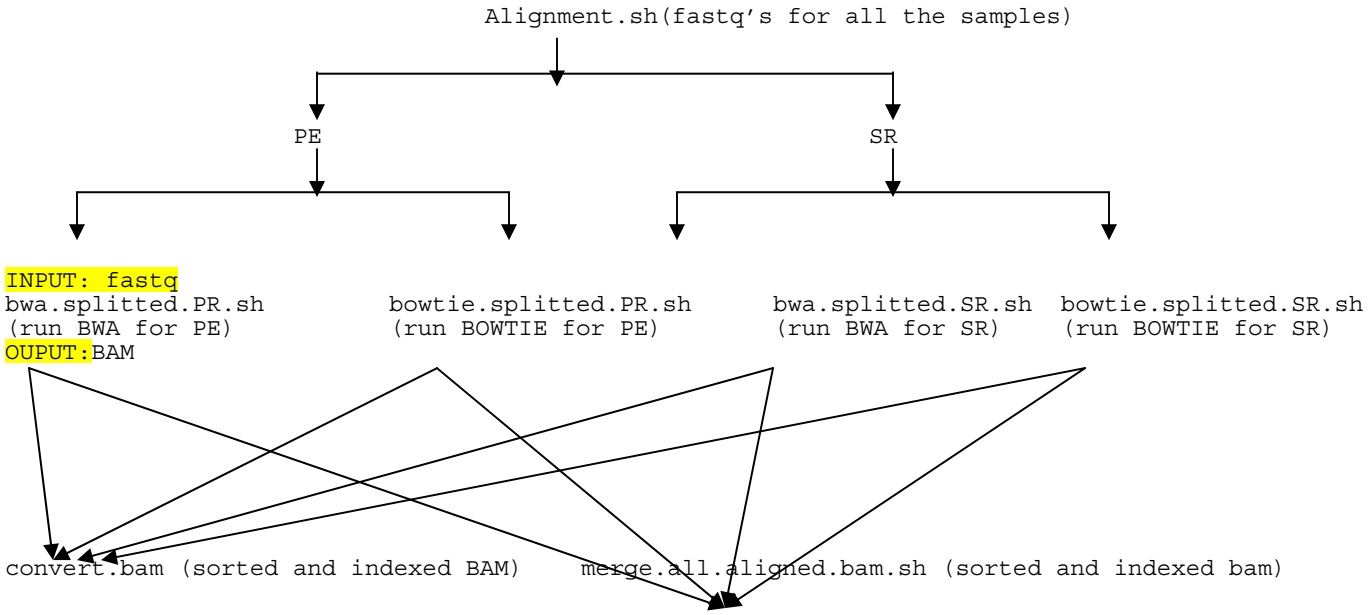
The Workflow is divided into four (4) modules:

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

ALIGNMENT

Input: FASTQ
Example for fastq:

```
@R0211812_0096:1:1:1223:2139#0/1
NGTGAGCCACAGCACCCAGCCTGCTTCATCTGTTAAATAAGGTTTACAGCATCTGCAACCCTTAACTCCTTGGTTTATGGTTATTCCATAACCCAAACAA
+R0211812_0096:1:1:1223:2139#0/1
BSUUJPTXNTPVWPW\[U]\c_cccccPcc[][[GIIKGGVWXX[SVXVY_X___BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```



Output: Sorted BAM & mapping numbers for the BAM

Note:

- BAM includes all the reads (mapped or unmapped), which we achieve for longer term. BAM is binary and compressed file sorted in coordinate order. There is Samtools command to extract header of Samtools which contains information about the BAM.

```
/projects/bsi/bictools/apps/alignment/samtools/samtools-0.1.12a/samtools view -H $bam
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:247249719
@SQ     SN:chrX LN:154913754
@SQ     SN:chrY LN:57772954
@PG     ID:bwa  PN:bwa  VN:0.5.9-r16
@PG     ID:BWA
@RG     ID:6-HBPN-Land-JL       SM:6-HBPN-Land-JL
LB:/data1/bsi/refdata/bictools/sequence/human/ncbi/36.49/indexed/BWA_Indexed_reference_including_ChrM/allChr.fa
PL:illumina     CN:MAYO
```

(GenomeBuild, aligner used, sample name, platform information etc.)

- Fastq's can be extracted from BAM using Picard tool

```
$ /data2/bsi/RandD/exome/bam2fastq/bam2fastq.sh
Usage: for PE: bam2fastq.sh <input bam>(provide full path) <name for read1> <name for read2>
                                for SR: bam2fastq.sh <input bam>(provide full path) <name for read1>
```

Folder to look at:
`Output_folder/alignement/sample/`

Each sample folder should includes important file

- Sorted bam (*.bam) - aligned BAM file (no filtering)

- Indexed bam (*.bai) - indexed BAM

Important script we use to add platform, read group, sample. GenomeBuild information to the sam files

```
Perl $script_path/add.read_group.platform.pl -i input_sam -o output_sam -r $sample -s $sample -p $platform -a $Aligner -c $center -l $GenomeBuild
```

Output: will add the above said information to the sam and which can be use further

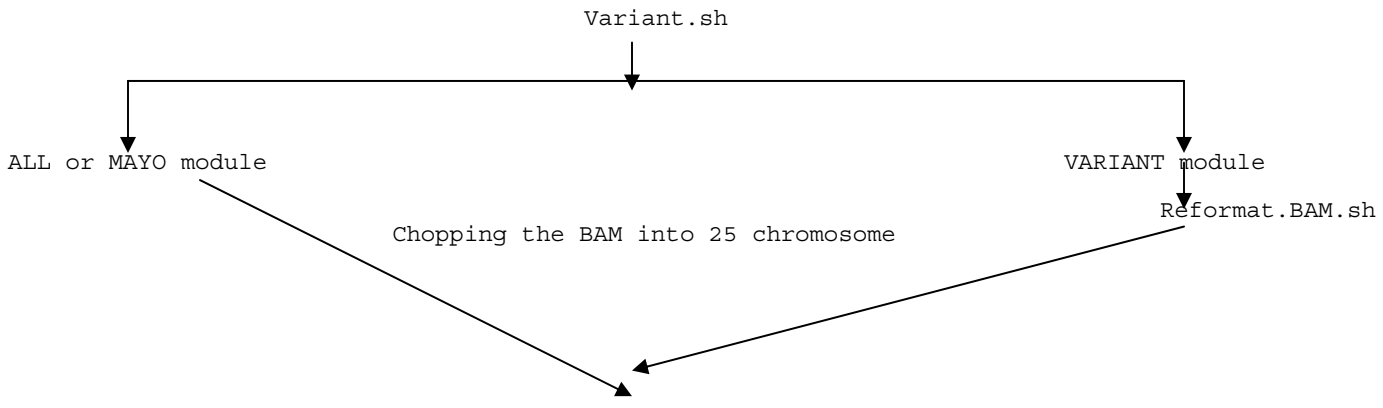
Tools used for this module:

- SAMTOOLS
- BWA
- BOWTIE
- PERL SCRIPTING
- SHELL SCRIPTING

VARIANT

Input: BAM

This module includes Realignment, Variant calling, On Target calling. This is the slowest step in the workflow, so we chop the input BAM into 25 chromosomes (Homosapiens) and run them in parallel to get the results faster.



variant_per_chr.sh (parallel processing)

- Realign.sh (realignment using GATK)

Input: per Chr BAM
Output: realigned and recalibrated BAM (Q>= 20 as a filter to BAM's to call variants)

- Call.indels.sh (calling Indels using GATK)

Input: per Chr realigned and recalibrated BAM
Output: VCF output INDEL calls (per Chr)

- Call.snvs.sh (calling SNVs using SNVmix/GATK)

Input: per Chr realigned and recalibrated BAM
Output: outputfile (per Chr)

- Format.varaints.sh (format the variants in tab delimited format)

Input: output files from call.snvs and call.indels to format in tab delimited format.
Output: Tab delimited format with required columns

- OnTarget.SNV.sh (intersect the SNV with target kit provided)

Input: SNV formatted file
Output: OnTraget SNVs

- OnTarget.INDEL.sh (intersect the INDEL with target kit provided)

Input: INDEL formatted file
Output: ONTarget INDELs

- OnTarget.BAM.sh (intersect the BAM with target kit provided)

Input: realigned and recalibrated BAM
Output: OnTarget BAM

- OnTarget.PILEUP.sh (intersect the PILEUP with target kit provided)

Input: pileup from realigned and recalibrated BAM
Output: ON Target pileup to get coverage numbers

OUTPUT: variant calls (On Target SNVs and INDELs)

Folder to look for realigned and recalibrated BAM

```
Output_folder/realigned_data/sample
        Realigned BAM for IGV visualization
```

FOR VARIANT CALLS

```
Output_folder/variants/SNV
Output_folder/variants/INDEL
(all the variants calls are per chromosome)
```

RAW variant calls:

- SNV `$sample.chr*.snps.raw.gatk.vcf / $sample.chr*.snps.raw.snvmix`
- INDEL `$sample.chr*.indel.gatk.vcf`

FILTERED variant calls:

- SNV `$sample.chr*.raw.snvs`
  - SNVmix
    - Probability greater or equal to 0.8 (>=0.8)
    - Discard reference homozygous calls
  - GATK
    - Discard reference homozygous calls
    - Quality filers
- INDEL `$sample.chr*.raw.indels`

OnTarget Variant calls

```
Output_folder/OnTarget/
```

- SNV `$sample.chr*.raw.snvs.bed.i.ToMerge`
- INDEL `$sample.chr*.raw.indels.bed.i.ToMerge`

FORMAT for the SNV files

GATK

```
Chr Position  Ref    Alt    GenotypeClass Alt-SupportedReads  Ref-SupportedReads  ReadDepth     Quality
Example:
chr1    4770    A       G       GG      4       0       4       345.63
chr1    4793    A       G       GG      5       0       5       34.67
chr1    42101   T       G       GG      3       0       3       880.56
chr1    53560   T       C       TC      18      35      53      1003.32
```

SNVmix

```
Chr Position  Ref    Alt    GenotypeClass Alt-SupportedReads  Ref-SupportedReads  ReadDepth     Probability
Example:
chr1    4770    A       G       GG      4       0       4       0.93
chr1    4793    A       G       GG      5       0       5       0.96
chr1    42101   T       G       GG      3       0       3       0.88
chr1    53560   T       C       TC      18      35      53      1.00
```

FORMAT for the INDEL files

GATK

```
Chr     Start   Stop    INDEL information
Example:
chr1    53598   53601   -CTA:12/36
chr1    1578607 1578610 -GCG:90/122
chr1    1637509 1637509 +T:18/18
chr1    1637753 1637753 +TTTCTT:70/70
```

Tools used for this module:

- GATK
- SNVmix
- SAMTOOLS
- PERL scripting
- SHELL scripting
- JAVA 64BIT
- BED TOOLS
- PICARD

NOTE:

- **SAMTOOL has a bug that even we sort the file as coordinate sort it won't change the sort tag to coordinate. So we use PICARD to fix that header tag.**
- **SNVmix accepts the pileup which can be generated using few parameters**
  `$samtools/samtools pileup -s -f $ref bam > pileup`
- **SNVmix applies quality filters**
  - Min mapping quality 20
  - Minimum base quality 20
- **GATK quality filters (BOTH SNVs AND INDELs)**
  - Min mapping quality 20
  - Min base quality 20
  - Number of threads 4

IMPORTANT SCRIPTS

To reformat the external Bam

```
perl $script_path/add.read_group.platform.pl -i input_sam -o output_sam -r $sample -s $sample -p $platform -a
$Aligner -c $center -l $GenomeBuild
```

To parse the VCF SNV from GATK

```
perl $script_path/parse.vcf.SNV.pl input_snv_vcf > output_file
```

To parse the VCF INDEL from GATK

```
perl $script_path/parse.vcf.INDEL.pl input_indel_vcf> output_file
```

To parse the SNVmix SNV output

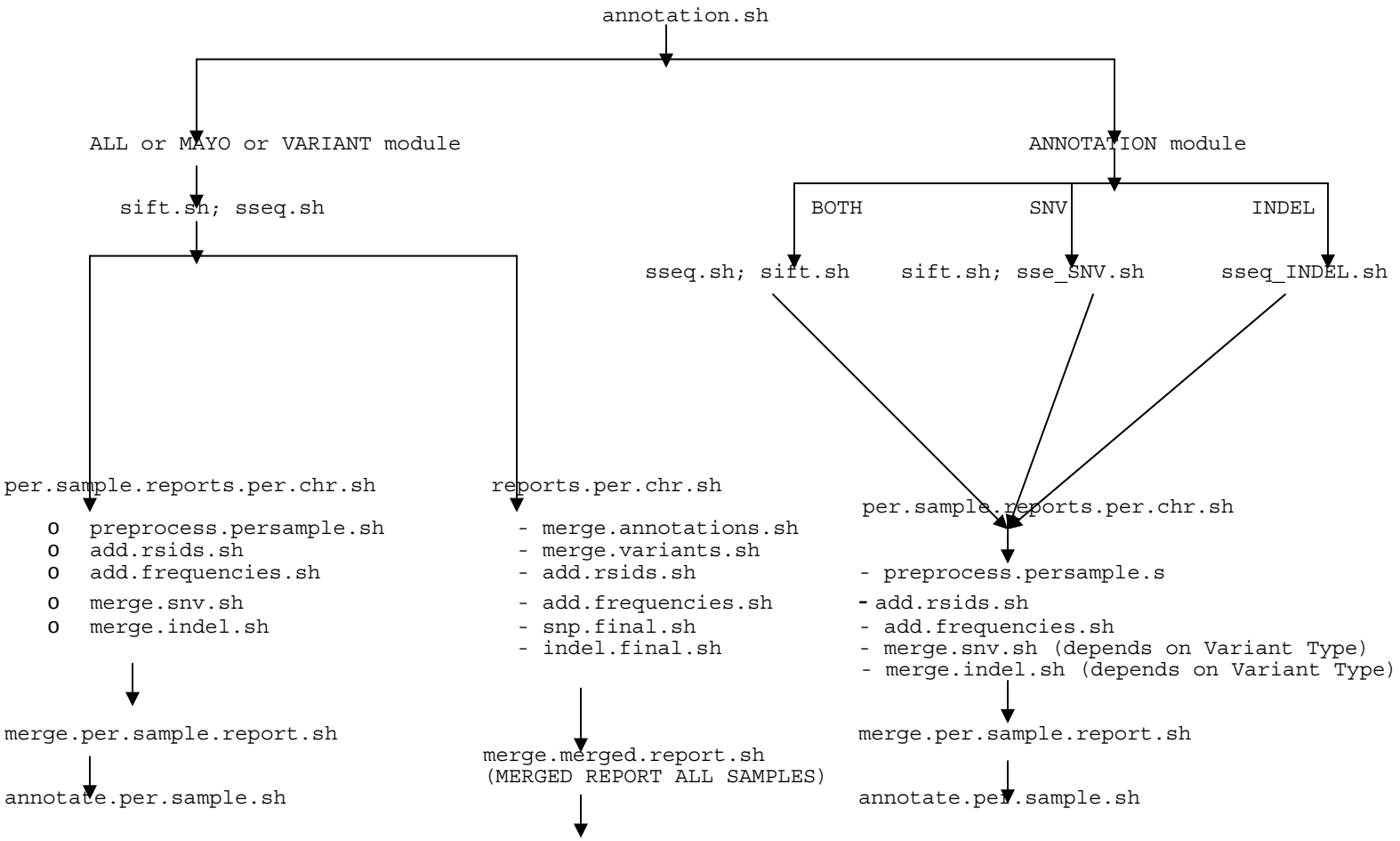
```
perl $script_path/parse.snvmix.to.snvs.pl input_snvmix_file > output_file
```

To add rsIDs to the SNV on file

```
perl $script_path/add_dbsnp_ucsc.pl
Raw parameters:
Usage: /projects/bsi/bictools/scripts/dev/exome/TREAT1.0_hold/add_dbsnp_ucsc.pl [-i input file containing chr
and pos in tab delimited format -full path]  [-b pos base in input file, 1 for 1-based, 0 for 0-based] [-s
dbSNP UCSChg##dbSNP### file use chromStart 0-based, current rsids only] [-c column# of chr in the input file]
[-p column# of pos ] [-o output file path] [-r chromosome ]
```

ANNOTATION

Input: OnTarget Variant calls (per Chr)



(PER SAMPLE REPORTS) (PER SAMPLE REPORTS)

variant.distance.sh
(Spilce variants +/- from Exons)

Output: Per sample and Merged Variant reports

Folders to look for:

Output_folder/Reports per Sample
(Per sample reports are available)

- SNV
  - $sample.SNV.cleaned_annot.xls (All the OnTarget variant calls)
  - $sample.SNV.cleaned_annot_filtered.xls (default filtered variant calls)
- INDEL
  - $sample.INDEL.cleaned_annot.xls
  - $sample.INDEL.cleaned_annot_filtered.xl

Output_folder/Reports
(Merged reports)

- SNV
  - SNV.cleaned_annot.xls
  - SNV.cleaned_annot_filtered.xls
- INDEL
  - INDEL.cleaned_annot.xls
  - INDEL.cleaned_annot_filtered.xls
- Variant Distance
  - variantLocation_SNVs
  - variantLocation_INDELs

Tools used for this module:

1. JAVA scripts
2. perl and shell scripting
3. SIFT
4. SSEQ web submission script
5. JAVA 64 Bit

Important scripts used:

- Error with SIFT that it flips the input allele to something , so this script help us to discard those annotation

```
perl $script_path/sift.inconsistent.pl $id $sift/$snv_file.sift $sift
```

- To add MAYO annotation to the reports

```
java -jar  /projects/bsi/bictools/scripts/dev/exome/TREAT2.0/exome_annot.jar
Usage: java -jar exome_annot.jar annotate <Folder with SNV and INDEL File (must end with INDEL.report and
SNV.report)>java -jar exome_annot.jar tissue_files
```

- To get the variant distance

```
java -jar  /projects/bsi/bictools/scripts/dev/exome/TREAT2.0/exonvariantlocation.jar
USAGE: java -jar exonvariantlocation.jar <ref_seq_file> <variant_file> <variant_type(snp or indel)>
```

- To add dbsnp id to the list of variants (using reference dbSNP file)

```
$script_path/add.rsids.sh
Usage: <TempReportDir> <configuration file> <variant file><chromosome>
```

- To add allele frequencies from different populations from Hapmap and 1kgenome

```
$script_path/add.frequencies.sh
Usage: <TempReportDir> <configuration file> <variant file with rsids ><chromosome>
```

- To preprocess the input file so as to merge the input variants and annoattaions

```
$script_path/preprocess.sh
Usage:<sample> <sift dir> <sseq dir> <configuration file> <tempReport dir> <input variant
folder><email><analysis type><run info><reports per sample> <input variant folder><chromosome>
```

- To merge per sample SNV and INDEL files with all the annotation

```
$script_path/merge.snv
Usage:<Tempreports><configuration><sample name><chromosome><sseq dir><sift dir><variant file>
$script_path/merge.indel.sh
Usage <TempReports> <config> <sample> <which_chr> <sseq> <indel_file><input_dir>
```

- All the processing is done on per chr variant per sample so to merge all the chr in one file

```
$script_path/merge.per.sample.report.sh
Usage: <config> <output_dir> <TempReports> <sample> <run_info>
```

- To annotate using MAYO annoattaions for per sample SNV and INDEL reports

```
$script_path/annotate.per.sample.sh
Usage : <config> < output_dir>
```

- To get variant splice distance

```
$script_path/variant.distance.sh
usage:<config><TempReports><output_dir>
```

- To merge all the annotations from different samples so that we can use same annotations with the merged report

```
$script_path/merge.annotations.sh
Usage: <sift dir><sseq dir><config><chromosome>
```

- To merge all the variant calls for all the samples

```
$script_path/ merge.variants.sh
Usage: <output dir><temp reports><config><chromossome><run_info>
```

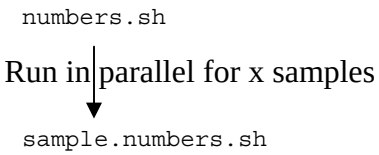
- To merge all the annotation with the variant call (merged report)

```
$script_path/snp.final.sh
Usage: <Tempreports dir><sift dir><sseq dir><config><chromosome><variant file>
$script_path/indel.final.sh
Usage:<Tempreports><sseq dir><configuration file><chromsome><variant file>
```

- To add MAYO annotation to the merged report

```
$script_path/merge.merged.report.sh
Usage: <output_dir><TempReports><config>
```

NUMBERS



Folder to look for:

Output_folder/numbers
There are two files per sample

1. $sample.out
2. $sample.coverage.out

Example:
Total Reads
2915968
Mapped Reads
2838835
……..
Each number is preceded by a description and files are used to create HTML report

Tools used for this module:

1. shell and perl scripting

Important Scripts:

To get the value for transition to transversion ratio
perl $script path/transition.transversion.pl
To get one annotation call per variant from Seattle seq (INDELS) with default priority list provided
perl $script_path/to.parse.sseq.result.indel.per.sample.pl

frameshift > coding > splice-5=splice3 > others

To get one annotation call per variant from Seattle seq (SNVs) with default priority list provided
perl $script_path/to.parse.sseq.result.per.sample.pl

nonsense > missense > coding-synonymous > coding-notMod3 > others

To generate HTML report

$script path/generate.html.sh $output dir $run info
This script generates HTMl report and send out an email to the person running the TREAT stating that the workflow is complete.

To clean the space on secondary space
$script_path/cleanspace.sh $output_dir