User Guide: PanMutsRx 1.0.

The users should have some basic bioinformatics and linux skills to use this Workflow. Users can run the workflow on a cluster [Tested on Open Grid Engine not tested on SGE and PBS] or on single machine (High memory > 40GB is required for one or more tools used in the workflow)

Version 1.1

PanMutsRx User Guide

Table of Contents

PANMUTSRX overview:

PANMUTSRX is a comprehensive pipeline to detect various mutations (snvs, intermediate indels, fusions, abnormal expression) from RNA-seq that are potentially actionable in clinic. The pipeline performs following functions:

1. Aligning the RNASEQ fastq files using STAR or/and GSNAP (sensitive for indel detection)
2. Marking Duplicates using SAMBLASTER and preprocessing aligned reads for variant calling using GATK (SplitNCigar, indel realignment, base recalibration) or OPOSSUM
3. Conducting variant/Indel calling using GATK or/and BCFTOOLS for non-paired samples (tumors)
4. Detecting somatic mutations from paired tumor/normal desing using STRELKA
5. Performing fusion detection using STAR FUSION
6. Quantifying gene expression by FeatureCounts

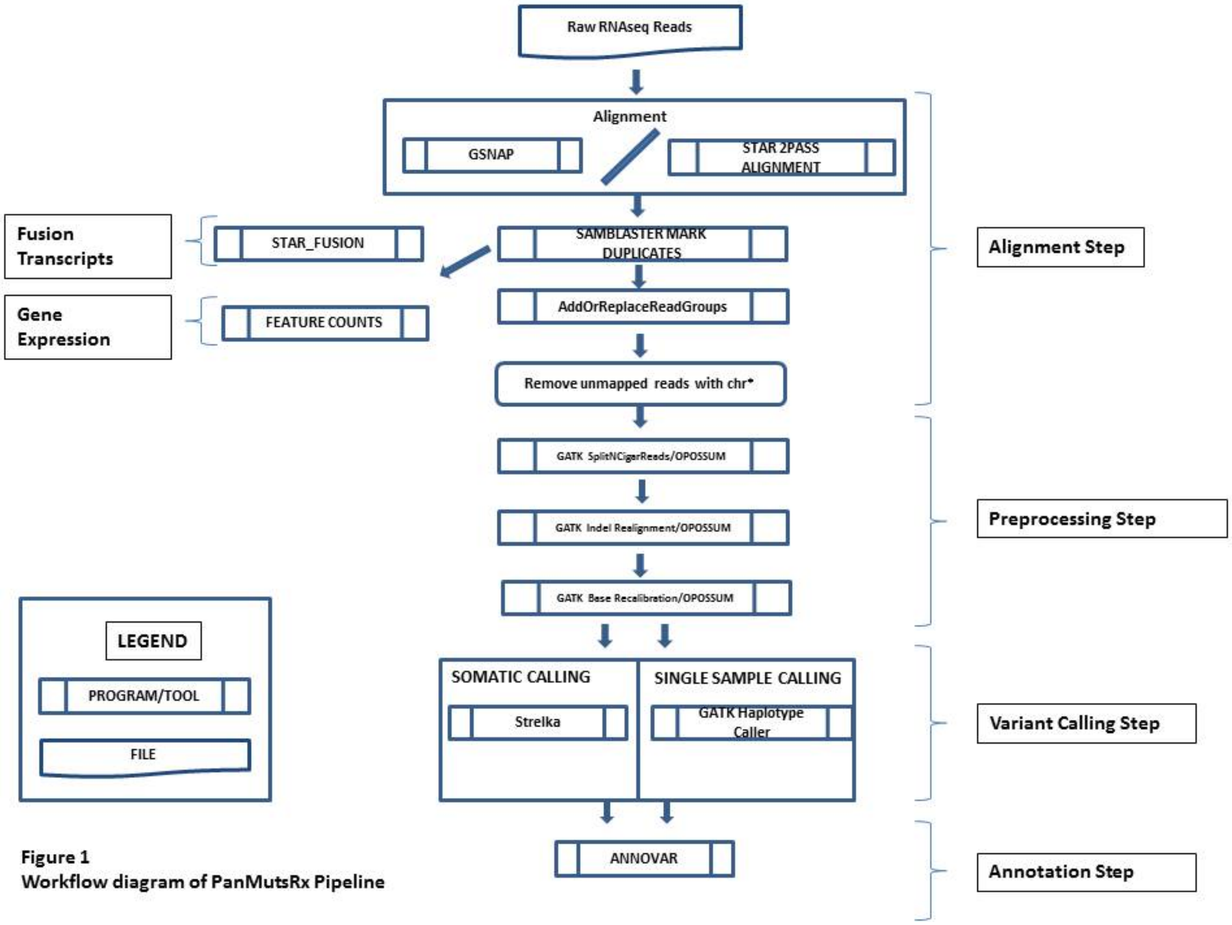

Figure 1
Workflow diagram of PanMutsRx Pipeline

Virtual Machine Version:

Virtual machine image is provided for the users to run the pipeline easily but with limited features. Vitual machine pipeline version will be using single thread to run, so it can only be run on small fastq files with around 1 million reads.The virtual image size is around 12 Gb and required 30 GB of physical space to install.

Steps to use virtual machine

(i) Install the “Oracle VM Virtual Box” from here https://www.virtualbox.org/wiki/Downloads

(ii) Download the virtual image file “Workflow.ova” file from http://bioinformaticstools.mayo.edu/research/

(iii) Open the “Oracle VM Virtual Box” and import the Ova file
File -> Import Applicance -> “select the downloaded ova file”-> press “Next”->press “Import”

(iv) Once the import process is complete, we can prose “start ->” to login in to the virtual machine

(v) Username : “workflow” (no need to enter username) Enter the Password : “workflow”

(vi) Now you are in the virtual machine, open the “terminal”

(vii) All the necessary scripts and dependencies are installed and sample fastq files are placed within the workflow folder

Sample run

(i) Navigate to folder “/home/workflow/Workflow_run_chr22/1.1/samplefastq_vm”

(ii) The input files and config file are already set up for you to run.
Simply execute the script “SomaticCaller_singlemachine.sh”
/bin/bash SomaticCaller_singlemachine.sh
It takes 20 mins to run default parameters (STAR ALIGNMENT with feature counts, GATK variant calling and fusion calling for a sample with 1 million reads)
Results folder is located in the same folder “process_dir”

PANMUTSRX installation:

This enclosed script in the package will install tools required for the workflow. Some basic tools are required to install these tools and their paths should be provided in the "./source_package/tool_install_config.txt".

System requirements: Perl "5.16.2", Java >=1.8, python>=3.4.3,python2>=2.7.10, and QSUB (optional) for cluster environment. Also make sure system libraries like libncurses, zlibc are installed and in the system path.

Make changes to the configuration files according to your system environment. The sample is provided below.

```
cd <downloaded program directory>
cat ./source_package/tool_install_config.txt
```

```
TAR=/bin/tar
UNZIP=/usr/bin/unzip
BZIP2=/usr/bin/bzip2
JAVA=/usr/local/biotools/java/jdk1.8.0_20/bin/java
PERL=/usr/local/biotools/perl/5.16.2-centos6/bin/perl
PYTHON=/usr/local/biotools/python/3.4.3/bin/python3
SH=/bin/bash
QSUB=/home/oge/ge2011.11/bin/linux-x64/qsub ("NA" if you want the run the workflow on a single machine)
PYTHON2=/usr/local/biotools/python/2.7.10/bin/python
PYTHON2_LB_LIB=/usr/local/biotools/python/2.7.10/lib
```

Install necessary tools: After modifying the configuration file, run the script "sh INSTALL_TOOLS.sh" from the main package directory with following options:

```
bash INSTALL_TOOLS.sh
Options specified:
##########################################################################
##      script to install tools
## Script Options:
##      -s      <source code directory>     -       (REQUIRED)      required source code directory
##      -d      <install directory>     -       (REQUIRED)      required path to directory for installation
##      -t      <tool install config file>      -       (REQUIRED)      required path to tool info config file
##      -w      <workflow script directory>     -       (REQUIRED)      required path to downloaded workflow scripts directory
##      -h      - Display this usage/help text (No arg)
#################################################################################
```

If the installation script fails to install some tools(you can see list of tools not installed properly in the log output), please manually inspect and install the necessary tool s. Also remove the word "(PLEASE INSTALL PROPERLY)" from the output tool info file generated by the script. After the installation, a workflow "toolinfo" file will be automatically created in the installed tool directory. However, this "toolinfo" file is not enough for running the workflow, you need to run the "PREPARE_REF_FILES.sh" script as well to make it work. The reference directory requires approximately 4GB of space and few mins to run.

Preparing reference files for different tools:

This script will download and format the required reference files for different tools in the workflow. I'm using many hyperlinks in this script and you may need to make sure they are active and working links. The reference directory requires approximately 130GB of space and few hours to run(because we need to download huge files from external sources).

You need to run the install tool script "INSTALL_TOOLS.sh" before running this script. All the paths to reference files downloaded and processed will be appended to tool info file created by "INSTALL_TOOLS.sh".

```
bash PREPARE_REF_FILES.sh
Options specified:
#########################################################################
##      script to install tools
## Script Options:
##      -r      <ref file directory>      -      (REQUIRED)      required path to directory for ref files download
##      -t      <tool install config file>      -      (REQUIRED)      required path to tool info config file
##      -h      - Display this usage/help text (No arg)
```

Sample execution

```
sh PREPARE_REF_FILES.sh \
 -r /home/usr/Workflow_dir/ref_files/  \
-t /home/usr/Workflow_dir/install_dir/TOOL_INFO.txt \
```

Quick start – run the workflow

SomaticCaller.py:

This is the main wrapper script which a user needs to run in order to start the workflow and this script does the following tasks:

a. Creates the folder structure

b. Checks the config files and input files for validity, permissions and formats

c. Submit all the jobs

```
/usr/local/biotools/python/3.4.3/bin/python3 SomaticCaller.py --help
You are running Somatic caller Workflow 1.0
usage: SomaticCaller.py [-h] -r RUN_INFO -t TOOL_INFO

optional arguments:
  -h, --help            show this help message and exit
  -r RUN_INFO, --run_info RUN_INFO
                        Run information file
  -t TOOL_INFO, --tool_info TOOL_INFO
                        Tool information file
```

Input files: These are the input files which are required

(i) run info file : Information about input sample fastq files and tools to run

(ii) tool info file : Path to installed tools , reference files and individual tool parameters

Sample Execution:

[example script provided

Single machine mode: “./sampleconfigfiles/sample_cluster_config/SomaticCaller_cluster.sh”

Cluster mode: “./sampleconfigfiles/sample_cluster_config/SomaticCaller_cluster.sh”]

```
/usr/local/biotools/python/3.4.3/bin/python3  \
/home/usr/Workflow_dir/SomaticCaller.py \
-r /home/usr/Workflow_dir/sampleconfigfiles/runinfo2.txt \
-t /home/usr/Workflow_dir/install_tools/TOOL_INFO1.txt
```

RUN_INFO PARAMETERS

[Explaination in the red font]
Sample run info file is provided in the “sampleconfigfiles” directory

SINGLE_FASTQ=/home/usr/Workflow_dir/sampleconfigfiles/normal_fastq.txt [Path to FASTQ files
Column0: Output file name(optional)
Column1 : Path to Read1 fastq file
Column 2: Path to Read2 fastq file] [example file provided in “/sampleconfigfiles/sample_cluster_config/STAR_FASTQ_NORMAL.txt”]
PROCESSDIR=/home/usr/Workflow_dir/working_directory [Path to process and output directory]
PAIRED_FASTQ=/home/usr/Workflow_dir/sampleconfigfiles/tumor_fastq.txt [Path to TUMOR FASTQ files
Column0: Ouptut file name(optional)
Column1 : Path to Read1 fastq file
Column 2: Path to Read2 fastq file][The fastq files order should be same for ‘SINGLE_FASTQ’ and ‘PAIRED_FASTQ’] [example file provided in “/sampleconfigfiles/sample_cluster_config/STAR_FASTQ_TUMOR.txt”]
EMAIL=prodduturi.naresh@mayo.edu [Email]
RUNID=MYFIRSTRUN [Run ID]
ALIGNERS=GSNAP,STAR [Aligners: You can specify one or both aligners]
CALLERS= OPOSSUM,GATK,STRELKA_SOMATIC,STAR_FUSION,FEATURECOUNTS [Callers: You can specify one or many callers. “STRELKA_SOMATIC” will not work when SOMATIC_FASTQ=NA, “STAR FUSION” will only work if you specify star aligner. OPOSSUM can be chosen instead of GATK preprocess for variant calling; default is GATL preprocess]

SINGLE SAMPLE MODE

You should provide the file with locations to fastq file samples to “SINGLE_FASTQ” parameter in the run info file. You can get only single sample variant calls in this mode.If you want to get somatic variant calls you should run below paired sample mode.

PAIRED SAMPLES (SOMATIC MODE):

If you want to run the case samples (Ex: cancer tissue) along with controls (Ex: normal tissue) then you should assign the case samples file list to “PAIRED_FASTQ” parameter in the run info file and control samples file list to “SINGLE_FASTQ”. The number of lines in both the files should be same and sample order should be same. Multiple case samples can have same normal sample. Somatic calls can only be called in this mode.

TOOL_INFO PARAMETERS

[Explaination in the red font, do not]

You need to provide the tool info file generated by scripts “INSTALL_TOOLS.sh” and “PREPARE_REF_FILES.sh”.
For reference sample run info file is provided in the “sampleconfigfiles” directory.
Please carefully review the tool info file for individual tool parameter and you can change them in this file.
Here are the parameters

JAVA=/usr/local/biotools/java/jdk1.8.0_20/bin/java[Path to Java 1.8]
PYTHON=/usr/local/biotools/python/3.4.3/bin/python3[Path to Python 3.4.3]
PERL=/usr/local/biotools/perl/5.16.2-centos6/bin/perl[Path to Perl 5.16.2]
SH=/bin/bash [Path to bash or shell]
WORKFLOW_PATH=/home/usr/Workflow_dir/ [Path to workflow directory]
STAR_FUSION_PERL_PACKAGE=/home/usr/Workflow_dir/install_tools/STARFUSION/PERLPACKAGE/lib/lib/site_perl/5.16.2/x86_64-linux/[Path to star fusion perl package]
QSUB=/home/oge/ge2011.11/bin/linux-x64/qsub [Path to QSUB or if you want to run on single machine then mention this option as “NA”]
STAR=/home/usr/Workflow_dir/install_tools/STAR/STAR-2.5.2b/bin/Linux_x86_64/STAR[Path to STAR aligner]
GSNAP=/home/usr/Workflow_dir/install_tools/GSNAP/gmap-2014-12-29/bin/gsnap[Path to GSNAP aligner]
SAMTOOLS=/home/usr/Workflow_dir/install_tools/SAMTOOLS/samtools-1.3.1/samtools [Path to Samtools]
HTSDIR=/home/usr/Workflow_dir/install_tools/HTSDIR/htslib-1.3.2/bin[Path to HTS tool directory]
FEATURECOUNTS=/home/usr/Workflow_dir/install_tools/FEATURECOUNTS/subread-1.5.1-Linux-x86_64/bin/featureCounts[Path toFeatureCounts]
BCFTOOLS=/home/usr/Workflow_dir/install_tools/BCFTOOLS/bcftools-1.3.1/bin/bcftools [Path to BCFTools]
SAMBLASTER=/home/usr/Workflow_dir/install_tools/SAMBLASTER/samblaster-master/samblaster [Path to Samblaster]
SAMBAMBA=/home/usr/Workflow_dir/install_tools/SAMBAMBA/sambamba_v0.6.4 [Path toSAMBAMBA]
PICARD=/home/usr/Workflow_dir/install_tools/PICARD/picard.jar [Path to Picard jar]
GATK=/home/usr/Workflow_dir/install_tools/GATK/GenomeAnalysisTK.jar [Path toGATK]
ANNOVAR=/home/usr/Workflow_dir/install_tools/ANNOVAR/annovar [Path to ANNOVAR]
STRELKA_WORKFLOW=/home/usr/Workflow_dir/install_tools/STRELKA/strelka_workflow-1.0.15 [Path to STRELKA tool directory]
STAR_FUSION=/home/usr/Workflow_dir/install_tools/STARFUSION/STAR-Fusion/STAR-Fusion [Path to STAR FUSION tool]
BEDTOOLS=/home/usr/Workflow_dir/install_tools/BEDTOOLS/bedtools2/bin/[Path to Bedtools directory]
SAMBLASTER_OPTIONS=" " [SAMBLASTER OPTION; default is none]

```
PICARD_ARG_OPTION="SO=coordinate RGID=group1 RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=sample1" [Picard tool read group information options ]
STAR_OPTION="--runThreadN 4" [STAR threads information; default 4 threads ]
STAR_OPTION_STEP2="--chimSegmentMin 12 --chimJunctionOverhangMin 12 --alignSJDBoverhangMin 10 --alignMatesGapMax 200000 --alignIntronMax 200000 --limitBAMsortRAM 31532137230 --outSAMstrandField intronMotif --outSAMtype BAM Unsorted" [STAR aligner options ]
GSNAP_JAVA_OPTION="-XX:CompileThreshold=1000 -XX:ReservedCodeCacheSize=128m -Xmx20g -Xms5g" [GSNAP JAVA Options]
GATK_JAVA_OPTION="-XX:CompileThreshold=1000 -XX:ReservedCodeCacheSize=128m -Xmx20g -Xms5g"[GATK JAVA Options]
GATK_SPLITNCIGAR_OPT="-RMQF 255 -RMQT 60" [GATK splitNCigar step parameters]
DEBUG=NO [If you need to investigate the intermediate files in each step because of failure of the workflow run then set DEBUG=YES]
ANNOVAR_QUEUE=1-day[Cluster Parameters for each tool:ANNOVAR Cluster Queue]
ANNOVAR_MEM=30G[Cluster Parameters for each tool:ANNOVAR Cluster Memory]
BCFTOOLS_OPTIONS=""[BCFTOOLS tool parameters]
SAMTOOLS_BCFTOOLS_OPTIONS=" -q 20"[Samtools parameter in the BCFTOOLS]
DELETE_BAM_POST_GATK_PROCESS=NO[If “YES” GATK processed BAMFILES will be deleted]
SAMBAMBA_PARAM="  -m 12GB -t 4 "[SAMBAMBA parameters which included memory]
GSNAP_QUEUE=4-days[Cluster Parameters for each tool:GSNAP Cluster Queue]
GSNAP_MEM=30G[Cluster Parameters for each tool:GSNAP Cluster Memory]
STAR_QUEUE=lg-mem[Cluster Parameters for each tool:STAR Cluster Queue]
STAR_MEM=50G[Cluster Parameters for each tool:STAR Cluster Memory]
GATK_QUEUE=4-days[Cluster Parameters for each tool:GATK Cluster Queue]
GATK_MEM=30G[Cluster Parameters for each tool:GATK Cluster Memory]
BCFTOOLS_QUEUE=4-days[Cluster Parameters for each tool:BCFTOOLS Cluster Queue]
BCFTOOLS_MEM=30G[Cluster Parameters for each tool:BCFTOOLS Cluster Memory]
STRELKA_QUEUE=4-days[Cluster Parameters for each tool:STRELKA Cluster Queue]
STRELKA_MEM=30G[Cluster Parameters for each tool: STRELKA Cluster Memory]
STAR_QUEUE=4-days[Cluster Parameters for each tool:STAR Cluster Queue]
STAR_MEM=30G[Cluster Parameters for each tool:STAR Cluster Memory]
STRELKA_CONFIG=/home/usr/Workflow_dir/install_tools/STRELKA/strelka_workflow-1.0.15/strelka_demo_config.ini[Path to Strelka tool config file]
REF_GENOME=/home/usr/Workflow_dir/ref_files/GRCh37_gencode_v19_CTAT_lib_July272016/ref_genome.fa[Path to human reference file hg 19]
GATK_BASE_RECALIBRATION_KNOWNSITES="-knownSites /home/usr/Workflow_dir/ref_files/1000G_phase1.snps.high_confidence.hg19.sites.vcf.gz -knownSites /home/usr/Workflow_dir/ref_files/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz -knownSites /home/usr/Workflow_dir/ref_files/dbsnp_138.hg19.excluding_sites_after_129.vcf.gz"[GATK base calibration known variants files]
STAR_REF=/home/usr/Workflow_dir/ref_files/STAR[Path to STAR alignment reference files]
```

```
GSNAP_OPTION="-t 4 -A sam -D /home/usr/Workflow_dir/ref_files/GSNAP -d GSNAP --use-splicing=/home/usr/Workflow_dir/ref_files/GSNAP/GSNAP.maps/gencode.v19.splicesites.iit -N 1 --read-group-id=group1 --read-group-name=sample1 --read-group-library=lib1 --read-group-platform=illumina"[GSNAP alignment parameters including reference files]
GATK_HAPLOTYPE_CALLER_OPTION=" -dontUseSoftClippedBases -stand_call_conf 20.0  -ERCIS 50 -pcrModel HOSTILE  -stand_emit_conf 20.0  -mmq 20 -L /home/usr/Workflow_dir/ref_files/coding.bed "[GATK Haplotype caller variant discovery step option]
ANNOVAR_OPTION="/home/usr/Workflow_dir/ref_files/ANNOVAR_humandb/ -buildver hg19 -remove -protocol refGene -operation g -nastring ." [Annovar options including the reference files]
STAR_FUSION_CTAT_LIB=/home/usr/Workflow_dir/ref_files/GRCh37_gencode_v19_CTAT_lib_July272016[STAR FUSION tool reference files]
FEATURECOUNTS_OPTION="-t exon -g gene_name -a /home/usr/Workflow_dir/ref_files/Homo_sapiens.GRCh37.75.gtf" [Feature count tool options including gtf file]
STARFUSION_MITO_FILTER="YES" [This option is used to filter Mitochondria chromosome fusions]
OPOSSUM_OPTIONS="  --SoftClipsExist True" [OPOSSUM option to specify STAR/GSNAP BAM as input]
OPOSSUM_PYTHON_PACKAGES=/data2/bsi/staff_analysis/m081429/Somatic_workflow_paper/install_tools/OPOSSUM/PYTHON_PACKAGES [OPOSSUM path to python packages]
OPOSSUM=/data2/bsi/staff_analysis/m081429/Somatic_workflow_paper/install_tools/OPOSSUM/Opossum-master/Opossum.py [Path to OPOSSUM]
PYTHON2=/usr/local/biotools/python/2.7.10/bin/python[Path to OPOSSUM python2 version]
PYTHON2_LB_LIB=/usr/local/biotools/python/2.7.10/lib[Path to OPOSSUM python2 lib]
```

Output Files

The process directory specified in the runinfo config contain following directories and output files

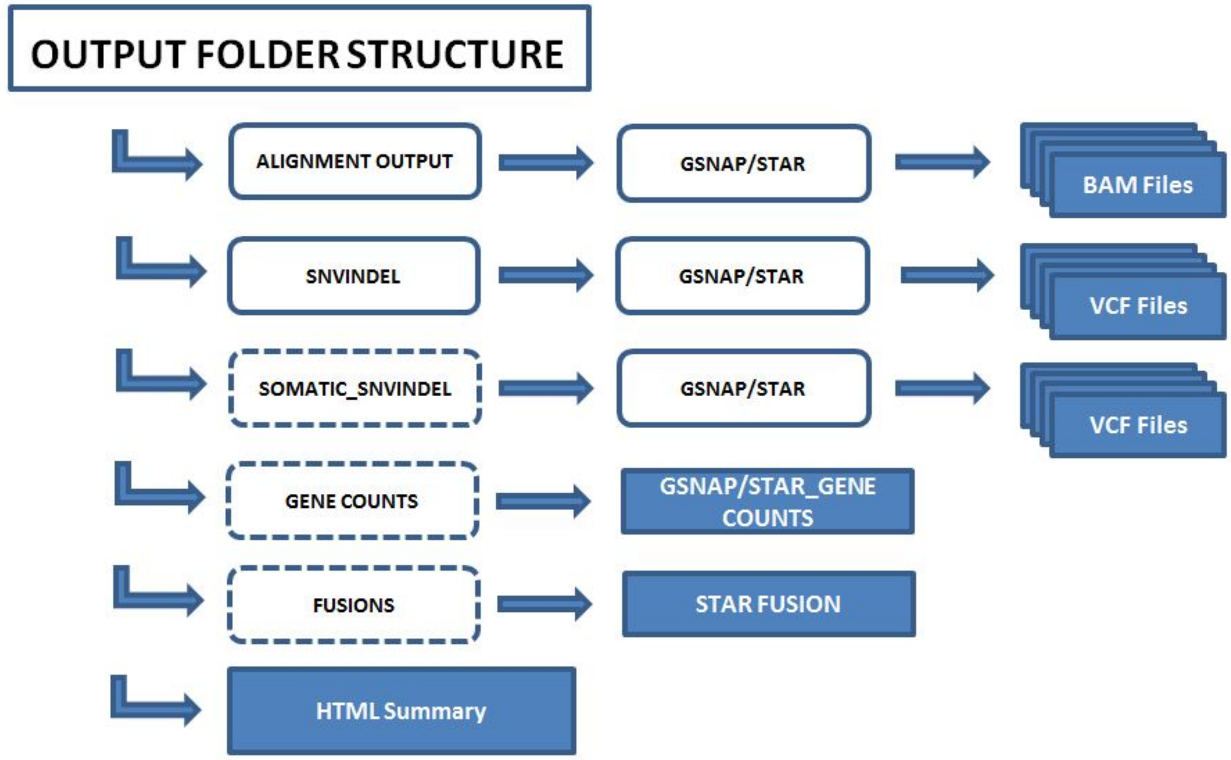

Figure 2 : Output folder structure from eSNVIndel Pipeline

HTML FILE

The HTML file shows the workflow and output layout and the summary statistics for alignments, number of variants, number of fusions and paths to output files.

ALIGNMENT OUTPUT

The directory contains sub directories “STAR” and or “GSNAP” accordingly and the sub directories contains the bam files aligned using aligners (GSNAP/STAR)

CONFIG

The input config files and config files created in the processed will be copied in to this directory

LOG

All the cluster log files are created in this directory

SGE_JOBID_COMMAND.txt

This file contains list of the cluster job commands for different steps

SNVINDEL OUTPUT(SINGLE SAMPLE VARIANT CALLING)[VCF Files]

This folder contains variant/INDEL output files using GATK/BCFTOOLS in single sample mode

Normal GATK output file: <Outfile Name>.<aligner>. GATK.Filtered.ANNOVAR.vcf

Tumor GATK output file: <Outfile Name>.<tumor>.<aligner>. GATK.Filtered.ANNOVAR.vcf

SOMATIC_SNVINDEL OUTPUT[VCF Files]

This folder contains variant/INDEL output files using STRELKA in somatic calling mode

SOMATIC STRELKA RESULTS

Strelka snp file: <Outfile Name>.tumor.<aligner>.gatkin.strelka.all.somatic.snvs.vcf

Strelka indel file: <Outfile Name>.tumor.<aligner>.gatkin.strelka.all.somatic.indels.vcf

FILTERED SOMATIC STRELKA RESULTS

Strelka snp file: <Outfile Name>.tumor.<aligner>.gatkin.strelka. passed.somatic.snvs.vcf

Strelka indel file: <Outfile Name>.tumor.<aligner>.gatkin.strelka. passed.somatic.indels.vcf

CLUSTER MODE/ SINGLE MACHINE

Users can run the workflow on a cluster[Tested on Open Grid Engine not tested on SGE and PBS] or on single machine (High memory > 45GB is required for one or more tools used in the workflow)

Use “QSUB=NA” in the tool info file to run on a single machine mode

DEBUGGING THE RUN

If you encounter any error in any job of the workflow and if you need to investigate the intermediate files in each step then set DEBUG=YES in the tool info file.