User Guide: Somatic Workflow 1.0.

Version 1.0

Somatic workflow User Guide

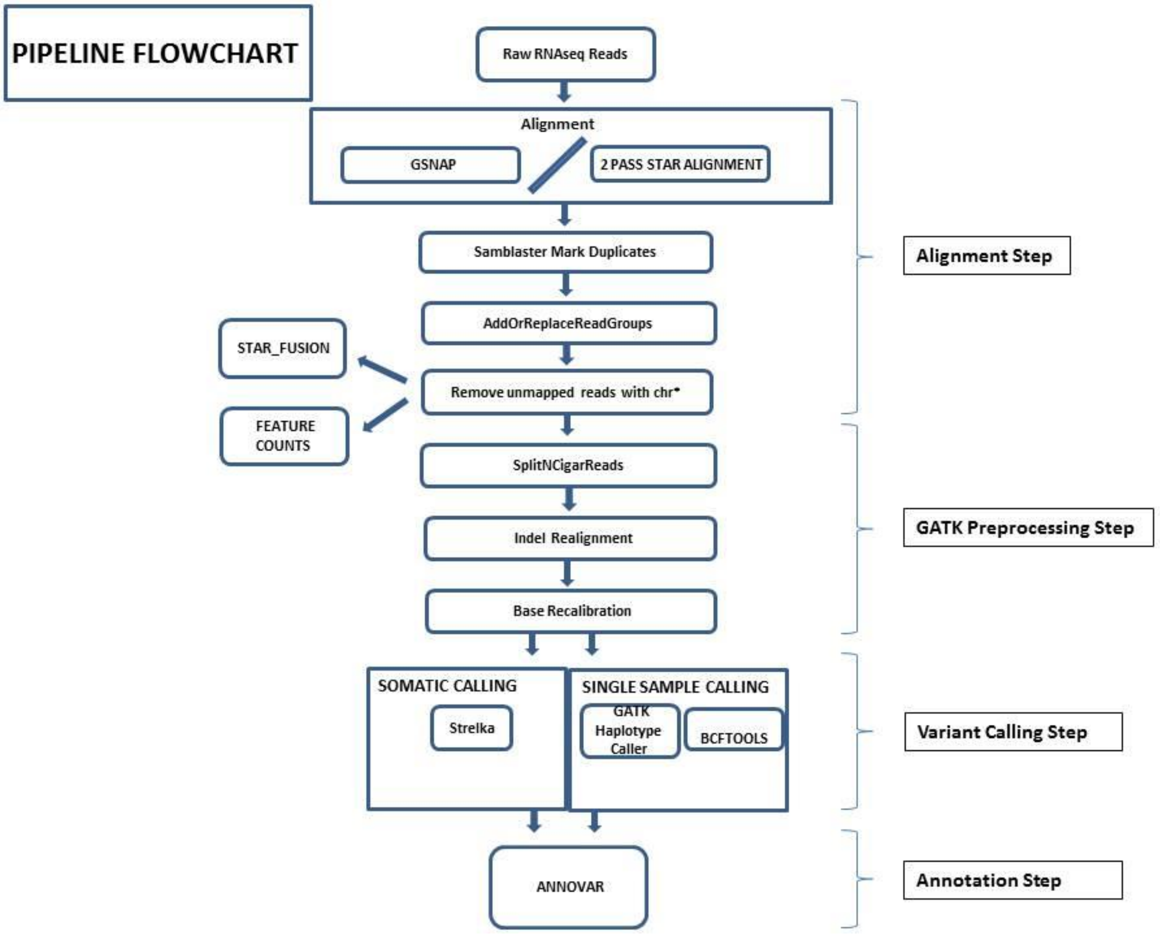
Table of Contents

Somatic workflow User Guide

Introduction:

The purpose of the pipeline is

1. To align the RNASEQ fastq files using GSNAP and STAR
2. Marking Duplicates using SAMBLASTER and GATK preprocessing (SplitNCigar, indel realignment, base recalibration)
3. Single sample variant/Indel calling using GATK
4. Somatic calling using STRELKA
5. Fusion calling using STAR-FUSION
6. Gene counts using FeatureCounts



Somatic workflow User Guide

List of scripts:

The scripts in this workflow:

(1) SomaticCaller.py: This is the main wrapper script does the following tasks
  (i) Creates the folder structure
  (ii) Checks the config files and input files for validity, permissions and formats
  (iii) Submit all the jobs
(2) GSNAP.sh: This script does the alignment using the GSNAP, marking duplicates using samblaster, separating the unmapped reads and adding read group information using picard tool
(3) STAR.sh: This script does the alignment using the STAR, marking duplicates using samblaster, separating the unmapped reads and adding read group information using picard tool
(4) GATK_PREPROCESS.sh: This script does the following
  - GATK SplitNCigarReads step
  - GATK Indel Realignment step
  - GATK Reclibration step
(5) GATK_CALLER.sh: This script calls the variants and indels using GATK in each sample
(6) VARSCAN_SINGLESAMPLE.sh: This script calls the variants and indels using VARSCAN in each sample
(7) VARSCAN_SOMATIC.sh: Somatic variant and indel calling using VARSCAN SOMATIC module
(8) STRELKA_SOMATIC.sh: Somatic variant and indel calling using STRELKA SOMATIC module
(9) CLEANUP.sh: Removing the unnecessary files
(10) shared_functions.sh: Some of the common functions shared across the scripts
(11) BCFTOOLS.sh: This script calls the variants and indels using BCFTOOLS in each sample
(12) FEATURECOUNTS.sh: This script generates gene counts file
(13) STAR_FUSION.sh: This script calls the gene fusions in each sample

SomaticCaller.py:

This is the main wrapper which does the following task:

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
```

RUN_INFO PARAMETERS

SINGLE_FASTQ=/data2/labdev/mgf/dev/naresh/MGF/RNASEQ_DELETION/INDEL_SOMATICSNV/Workflow/1.0/sampleconfigfiles/normal_fastq.txt [Path to FASTQ files

Column0: Ouptut file name(optional)

Column1 : Path to Read1 fastq file

Column 2: Path to Read2 fastq file]

PROCESSDIR=/data2/labdev/mgf/dev/naresh/MGF/RNASEQ_DELETION/INDEL_SOMATICSNV/Workflow/1.0/TMP2[Path to process and output directory]

PAIRED_FASTQ=/data2/labdev/mgf/dev/naresh/MGF/RNASEQ_DELETION/INDEL_SOMATICSNV/Workflow/1.0/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']

EMAIL=prodduturi.naresh@mayo.edu [Email]

RUNID=MYFIRSTRUN [Run ID]

ALIGNERS=GSNAP,STAR [Aligners: You can specify one or both aligners]

CALLERS=GATK,VARSCAN,VARSCAN_SOMATIC,STRELKA_SOMATIC, STAR_FUSION,FEATURECOUNTS [Callers: You can specify one or many callers. "VARSCAN_SOMATIC" and "STRELKA_SOMATIC" will not work when SOMATIC_FASTQ=NA]

VARSCAN_STRELKA_NO_GATK_PREPROCESS=YES [GATK preprocessing step is failing for some of the samples aligned using GSNAP. So optionally Varscan and Strelka can be run directly on the GSNAP aligned bam file without GATK preprocessing]

COMPRESS_VARIANT_OUTPUT=YES["YES: if you want to gzip the output vcf files]

TOOL_INFO PARAMETERS

PYTHON=/usr/local/biotools/python/3.4.3/bin/python3[Path to python]

SAMBLASTER=/data5/bsi/bictools/src/samblaster/0.1.22/samblaster[Path to samblaster]

SAMBLASTER_OPTIONS=" "[Use -r option to remove duplicate reads from aligner output bam files]

SAMTOOLS=/data5/bsi/bictools/alignment/samtools/samtools-1.2/samtools

```
PICARD=/data5/bsi/bictools/alignment/picard/1.140/picard.jar

PICARD_ARG_OPTION="SO=coordinate RGID=group1 RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=sample1"

JAVA=/usr/java/jdk1.7.0_03/bin/java

GATK=/data5/bsi/bictools/alignment/gatk/3.5/GenomeAnalysisTK.jar

GATK_KEY=/projects/bsi/bictools/apps/alignment/GenomeAnalysisTK/3.1-1/Hossain.Asif_mayo.edu.key

REF_GENOME=/data5/bsi/refdata-new/app/gatk_bundle/human/2.8/b37/processed/2015_11_04/chr1-22XYM.fa

GATK_BASE_RECALIBRATION_KNOWNSITES="-knownSites /data2/bsi/reference/annotation/1KGenome/1000G_phase1.snps.high_confidence.hg19.vcf.gz -knownSites /data2/bsi/reference/annotation/dbSNP/hg19/dbsnp_137.hg19.vcf.gz -knownSites /data2/bsi/reference/annotation/1KGenome/Mills_and_1000G_gold_standard.indels.hg19.vcf.gz"

GSNAP=/data5/bsi/bictools/src/gsnap/2015-09-29/bin/gsnap

STAR=/data5/bsi/bictools/src/star/2.4.2a/bin/Linux_x86_64/STAR

GSNAP_QUEUE=4-days

GSNAP_MEM=30G

STAR_QUEUE=lg-mem

STAR_MEM=50G

GATK_QUEUE=4-days

GATK_MEM=30G

VARSCAN_QUEUE=4-days

VARSCAN_MEM=30G

BCFTOOLS_QUEUE=4-days

BCFTOOLS_MEM=30G

STRELKA_QUEUE=4-days

STRELKA_MEM=30G

STAR_OPTION="--runThreadN 4"
```

```
STAR_OPTION_STEP2="--chimSegmentMin 12 --chimJunctionOverhangMin 12 --alignSJDBoverhangMin 10 --alignMatesGapMax 200000 --alignIntronMax 200000 --limitBAMsortRAM 31532137230 --outSAMstrandField intronMotif --outSAMtype BAM Unsorted"

STAR_REF="/data5/bsi/refdata-new/app/gatk_bundle/human/2.8/b37/processed/2015_11_04/STAR_alignment"

GSNAP_OPTION="-t 4 -A sam -D /data5/bsi/refdata-new/app/gatk_bundle/human/2.8/b37/processed/2015_11_04/GSNAP/ -d GSNAP --use-splicing=/data5/bsi/refdata-new/app/gatk_bundle/human/2.8/b37/processed/2015_11_04/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_JAVA_OPTION="-XX:CompileThreshold=1000 -XX:ReservedCodeCacheSize=128m -Xmx20g -Xms5g"

GATK_JAVA_OPTION="-XX:CompileThreshold=1000 -XX:ReservedCodeCacheSize=128m -Xmx20g -Xms5g"

STAR_QUEUE=4-days

STAR_MEM=30G

SH=/bin/bash

GATK_SPLITNCIGAR_OPT="-RMQF 255 -RMQT 60"

BAMTOOLS=/projects/bsi/bictools/apps/alignment/bamtools/bin/bamtools

GATK_JAVA_OPTION="-XX:CompileThreshold=1000 -XX:ReservedCodeCacheSize=128m -Xmx20g -Xms5g"

GATK_HAPLOTYPE_CALLER_OPTION=" -dontUseSoftClippedBases -stand_call_conf 20.0  -ERCIS 50 -pcrModel HOSTILE  -stand_emit_conf 20.0  -mmq 20 -L /projects/bsi/bictools/apps/variant_detection/rvboost/RVboost_0.1/resources/coding.bed "[Haplotype caller parameters: variants are restricted to coding region, for whole region remove –L option]

QSUB=/home/oge/ge2011.11/bin/linux-x64/qsub

VARSCAN=/data5/bsi/bictools/src/varscan/2.4.0/VarScan.v2.4.0.jar

STRELKA_WORKFLOW=/data5/bsi/bictools/src/strelka/1.0.14

STRELKA_CONFIG=/data2/labdev/mgf/dev/tools/strelka/strelka_workflow-1.0.14/demo/strelka_demo_config.ini[Strelka parameters]
```

PERL=/usr/local/biotools/perl/5.16.2-centos6/bin/perl

WORKFLOW_PATH=/data2/labdev/mgf/dev/naresh/MGF/RNASEQ_DELETION/INDEL_SOMATICSNV/Workflow/1.0

VARSCAN_FILTER_OPTIONS="-–min-reads2 4 –-min-var-freq 0.15 –-p-value 0.05"

```
[     OPTIONS:
        --min-coverage  Minimum read depth at a position to make a call [8]
        --min-reads2    Minimum supporting reads at a position to call variants [2]
        --min-avg-qual  Minimum base quality at a position to count a read [15]
        --min-var-freq  Minimum variant allele frequency threshold [0.01]
        --p-value       Default p-value threshold for calling variants [99e-02]
]
```

DEBUG=NO[Temp files deleted if DEBUG=NO]

NOVOSORT=/projects/bsi/bictools/apps/alignment/novoalign/3.02.04/novosort

NOVOSORT_PARAM=" --ram 12G --tmpcompression 0 --threads 4 -f"

REMOVE_DUP_READS="TRUE"

ANNOVAR=/data5/bsi/bictools/src/annovar/2015_06

ANNOVAR_OPTION="/data5/bsi/refdata-new/app/annovar/human/latest/downloaded/2015_05_01 -buildver hg19 -remove -protocol refGene -operation g -nastring ." [Annovar options]

#ANNOVAR_OPTION="/data2/labdev/mgf/dev/references/hg19/ANNOVAR_humandb/ -buildver hg19 -remove -protocol ensGene,tfbsConsSites,cytoBand,targetScanS,genomicSuperDups,dgvMerged,gwasCatalog,wgEncodeBroadHmmGm12878HMM,ALL.sites.2012_04,snp138,ljb23_sift,esp6500si_all,exac03,gerp++gt2,clinvar_20140211,cosmic68 -operation g,r,r,r,r,r,r,r,f,f,f,f,f,f,f,f -nastring ."

ANNOVAR_QUEUE=1-day

ANNOVAR_MEM=30G

BCFTOOLS=/data5/bsi/bictools/src/bcftools/1.2/bcftools

BCFTOOLS_OPTIONS=""

SAMTOOLS_BCFTOOLS_OPTIONS=" -q 20"

DELETE_BAM_POST_GATK_PROCESS=NO[Delete the gatk realign, recaliber bam files]

STAR_FUSION=/data5/bsi/bictools/src/STAR-Fusion/STAR-Fusion

STAR_FUSION_CTAT_LIB=/data2/labdev/mgf/dev/references/hg19/Hg19_CTAT_resource_lib

```
FEATURECOUNTS=/projects/bsi/bictools/apps/alignment/subread/1.4.4/featureCounts

FEATURECOUNTS_OPTION="-t exon -g gene_name -a /data2/bsi/RandD/MAPRSeq/2.0/refs/Ensemble_GeneExon_hg19.mod.gtf"[Feature counts options]
```

GSNAP.sh:

This script does the alignment using the GSNAP, marking duplicates using samblaster and adding read group information using picard tool

```
sh GSNAP.sh -h

Options specified: -h

####################################################################################

##	script to run gsnap

## Script Options:

##	-c	<configfile>	-	(REQUIRED)	required config file

##	-f	<fastqfile>	-	(REQUIRED)	required file with fullpath to fastq(each line should contain <SAMPNAME> <FASTQ1> <FASTQ2>)

##	-r	<rundir>	-	(REQUIRED)	rundir

##	-e	<email>	-	(REQUIRED)	email

##	-i	<runid>	-	(REQUIRED)	runid

##	-h	- Display this usage/help text (No arg)

############################################################################################
```

STAR.sh:

This script does the alignment using the STAR, marking duplicates using samblaster and adding read group information using picard tool

```
sh STAR.sh -h

Options specified: -h

######################################################################

##      script to run star

## Script Options:

##      -c      <configfile>      -      (REQUIRED)      required config file

##      -f      <fastqfile>      -      (REQUIRED)      required file with path to fastq(each line should contain <FASTQ1> <FASTQ2>)

##      -r      <rundir>      -      (REQUIRED)      rundir

##      -e      <email>      -      (REQUIRED)      email

##      -i      <runid>      -      (REQUIRED)      runid

##      -h      - Display this usage/help text (No arg)

#############################################################################
```

GATK_PREPROCESS.sh:

This script does the following

- GATK SplitNCigarReads step
- GATK Indel Realignment step
- GATK Reclibration step

```
sh GATK_PREPROCESS.sh -h

Options specified: -h

#########################################################################################

##	script to run gsnap

## Script Options:

##	-c	<configfile>	-	(REQUIRED)	required config file

##	-f	<BamfilesPath>	-	(REQUIRED)	required file with fullpath to bamfile(each line should contain <BamFile>)

##	-r	<rundir>	-	(REQUIRED)	rundir

##	-e	<email>	-	(REQUIRED)	email

##	-i	<runid>	-	(REQUIRED)	runid

##	-h	- Display this usage/help text (No arg)

##################################################################################################
```

GATK_CALLER.sh:

This script calls the variants and indels using GATK in each sample (Annovar annotation optional)

```
sh GATK_CALLER.sh -h

Options specified: -h

##############################################################################

##      script to run gsnap

## Script Options:

##      -c      <configfile>      -       (REQUIRED)      required config file

##      -f      <BamfilesPath>      -       (REQUIRED)      required file with path to bam files(each line
should contain <Bam file>)

##      -r      <rundir>      -       (REQUIRED)      rundir

##      -e      <email>      -       (REQUIRED)      email

##      -i      <runid>      -       (REQUIRED)      runid

##      -h      - Display this usage/help text (No arg)

######################################################################################
```

VARSCAN_SINGLESAMPLE.sh:

This script calls the variants and indels using VARSCAN in each sample(Annovar annotation optional)

```
sh VARSCAN_SINGLESAMPLE.sh -h

Options specified: -h

#############################################################################

##	script to run gsnap

## Script Options:

##	-c	<configfile>	-	(REQUIRED)	required config file

##	-f	<BamfilePath>	-	(REQUIRED)	required file with fullpath to Bamfiles(each line should contain <BAMFILE>)

##	-r	<rundir>	-	(REQUIRED)	rundir

##	-e	<email>	-	(REQUIRED)	email

##	-i	<runid>	-	(REQUIRED)	runid

##	-h	- Display this usage/help text (No arg)

#####################################################################################
```

VARSCAN_SOMATIC.sh:

Somatic variant and indel calling using VARSCAN SOMATIC module(Annovar annotation optional)

```
sh VARSCAN_SOMATIC.sh -h

Options specified: -h

#######################################################################################

##	script to run gsnap

## Script Options:

##	-c	<configfile>	-	(REQUIRED)	required config file

##	-f	<BamfilePath>	-	(REQUIRED)	required file with fullpath to Bamfiles(each line should contain <NORMAL BAMFILE>)

##	-s	<BamfilePath>	-	(REQUIRED)	required file with fullpath to Bamfiles(each line should contain <TUMOR BAMFILE>)

##	-r	<rundir>	-	(REQUIRED)	rundir

##	-e	<email>	-	(REQUIRED)	email

##	-i	<runid>	-	(REQUIRED)	runid

##	-h	- Display this usage/help text (No arg)

################################################################################################
```

STRELKA_SOMATIC.sh:

Somatic variant and indel calling using STRELKA SOMATIC module(Annovar annotation optional)

```
sh STRELKA_SOMATIC.sh -h

Options specified: -h

##########################################################################

##      script to run gsnap

## Script Options:

##      -c      <configfile>    -       (REQUIRED)      required config file

##      -f      <BamfilePath>   -       (REQUIRED)      required file with fullpath to Bamfiles(each line should contain <NORMAL BAMFILE>)

##      -s      <BamfilePath>   -       (REQUIRED)      required file with fullpath to Bamfiles(each line should contain <TUMOR BAMFILE>)

##      -r      <rundir>        -       (REQUIRED)      rundir

##      -e      <email>         -       (REQUIRED)      email

##      -i      <runid>         -       (REQUIRED)      runid

##      -h      - Display this usage/help text (No arg)

#################################################################################
```

BCFTOOLS.sh:

This script calls the variants and indels using BCFTOOLS in each sample(Annovar annotation optional)

```
sh BCFTOOLS.sh -h

Options specified: -h

#############################################################################

##      script to run gsnap

## Script Options:

##      -c      <configfile>      -      (REQUIRED)      required config file

##      -f      <fastqfile>      -      (REQUIRED)      required file with fullpath to fastq(each line should contain
<SAMPNAME> <FASTQ1> <FASTQ2>)

##      -r      <rundir>      -      (REQUIRED)      rundir

##      -e      <email>      -      (REQUIRED)      email

##      -i      <runid>      -      (REQUIRED)      runid

##      -h      - Display this usage/help text (No arg)

#####################################################################################
```

STAR_FUSION.sh:

(1) This script calls the gene fusions in each sample

```
sh STAR_FUSION.sh -h

##########################################################################

##      script to run star fusion

## Script Options:

##      -c      <configfile>      -      (REQUIRED)      required config file

##      -f      <fastqfile>      -      (REQUIRED)      required file with path to fastq(each line should contain
<FASTQ1> <FASTQ2>)

##      -r      <rundir>      -      (REQUIRED)      rundir

##      -e      <email>      -      (REQUIRED)      email

##      -i      <runid>      -      (REQUIRED)      runid

##      -h      - Display this usage/help text (No arg)

##################################################################################
```

FEATURECOUNTS.sh: This script generates gene counts file

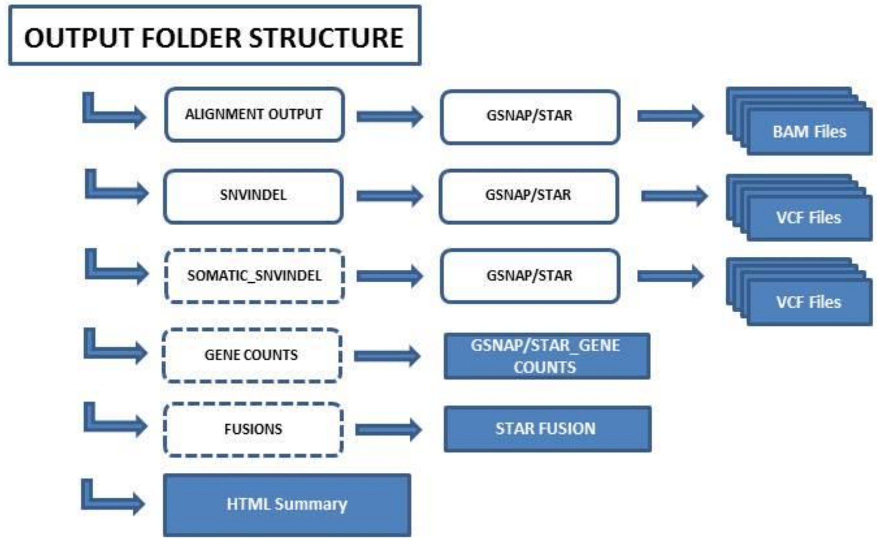
sh FEATURECOUNTS.sh

Options specified:

```
######################################################################

##	script to run feature counts

## Script Options:

##	-c	<configfile>	-	(REQUIRED)	required config file

##	-f	<BamfilesPath>	-	(REQUIRED)	required file with path to bam files(each line should contain <Bam file>)

##	-r	<rundir>	-	(REQUIRED)	rundir

##	-e	<email>	-	(REQUIRED)	email

##	-i	<runid>	-	(REQUIRED)	runid

##	-h	- Display this usage/help text (No arg)

#############################################################################
```

Output Files:

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



ALIGNMENT OUTPUT:

The directory contains the bam files aligned using aligners (GSNAP/STAR)

NORMAL

Raw output bam files: <Outfile Name>.GSNAP.RAW.bam

Raw output bam files: <Outfile Name>.STAR.RAW.bam

GATK processed output bam files: <Outfile Name>.GSNAP. gatkin.splitNC.realign.recaliber.bam

GATK processed output bam files: <Outfile Name>.STAR. gatkin.splitNC.realign.recaliber.bam

TUMOR (SOMATIC MODE : if supplied SOMATIC_FASTQ)

Raw output bam files: <Outfile Name>. tumor.GSNAP.RAW.bam

Raw output bam files: <Outfile Name>. tumor.STAR.RAW.bam

GATK processed output bam files: <Outfile Name>. tumor.GSNAP. gatkin.splitNC.realign.recaliber.bam

GATK processed output bam files: <Outfile Name>. tumor.STAR. gatkin.splitNC.realign.recaliber.bam

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

FUSIONS:

Fusion output files are created in this directory

GENECOUNTS:

FeatureCounts output files are created in this directory

SGE_JOBID_COMMAND.txt:

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

SNVINDEL OUTPUT:

Varscan Output

```
OUTPUT
          Tab-delimited SNP calls with the following columns:
          Chrom               chromosome name
          Position  position (1-based)
          Ref                 reference allele at this position
          Cons                Consensus genotype of sample in IUPAC format.
          Reads1              reads supporting reference allele
          Reads2              reads supporting variant allele
          VarFreq             frequency of variant allele by read count
          Strands1  strands on which reference allele was observed
          Strands2  strands on which variant allele was observed
          Qual1               average base quality of reference-supporting read bases
          Qual2               average base quality of variant-supporting read bases
          Pvalue              Significance of variant read count vs. expected baseline error
          MapQual1  Average map quality of ref reads (only useful if in pileup)
          MapQual2  Average map quality of var reads (only useful if in pileup)
          Reads1PlusNumber of reference-supporting reads on + strand
          Reads1Minus         Number of reference-supporting reads on - strand
          Reads2PlusNumber of variant-supporting reads on + strand
          Reads2Minus         Number of variant-supporting reads on - strand
          VarAllele Most frequent non-reference allele observed
```