IMPACT-Pipeline Documentation

Release 1.0.1

Ronak H Shah, Donavan Cheng, Ahmet Zehir, Aijaz Syed, Raghu C

Mar 08, 2017

Contents

Author Ronak H Shah, Donavan Cheng, Ahmet Zehir, Aijazuddin Syed, Raghu Chandramohan, Meera Prasad et.al

Contact rons.shah@gmail.com

Source code http://github.com/rhshah/IMPACT-Pipeline

License Apache License 2.0

IMPACT-Pipeline is a Perl/Python/R command-line software toolkit to process data, call somatic aberrations (SNV,INDELS,SV,CNV) generated by NGS based MSK-IMPACT assay. It is designed for use with hybrid capture, including both whole-exome and custom target panels, and short-read sequencing platforms such as Illumina. Contents:

CHAPTER 1

Installation

```
git clone --recursive https://github.com/rhshah/IMPACT-Pipeline.git
```

You can also download the release file in zip, tar.gz format and do:

```
wget https://github.com/rhshah/IMPACT-Pipeline/archive/1.0.1.tar.gz
tar xzvf 1.0.1.tar.gz
```

CHAPTER 2

Requirements

Note:

This will only run on any Linux system that has an SGE or LSF cluster.

Please see template.conf file in the configuration folder.

perl v5.20.2

python v2.7.8

R v3.1.2

Required Perl Packages

Note:

For **MSKCC_DMP_Logger** is wriiten internally. To get this please contact us.

Getopt::Long Getopt::Long

IO::File IO::File

Tie::Autotie Tie::Autotie 'Tie::IxHash'

Statistics::Lite Statistics::Lite qw(:all)

MSKCC_DMP_Logger Contact Us

Required Python Packages

pandas Pandas

Required R Packages

knitr knitr

ggplot2 ggplot2

plyr plyr

reshape reshape

scales scales

xtable xtable

RColorBrewer RColorBrewer

gplots gplots

googleVis googleVis

Purpose and Tools Used

> **Note:**
>
> For Timmomatic we are using a custom old version which uses cutadapt v1.1 internally. To get this please contact us.

Trimming Trimmomatic

Alignment BWA v0.7.5a

Somatic SNV calling MuTect v1.1.4

Somatic INDEL calling SomaticIndelDetector in GATK v2.3-9

Somatic INDEL calling PINDEL v0.2.5a7

Indel Realignment ABRA v0.92

Mark Duplicates and Various Statistics Picard Tools v1.96

Base Quality Recalibration and Find Covered Intervals GATK v3.3.0

Genotyping Position SAMTOOLS v0.1.19

Somatic Structural Variant Framework IMPACT-SV v1.0.1

Inside the config file

There are three sections:

Section 1	Section 2	Section 3
Locations	Parameters	Versions

All of this section start with > sign.

Inside each of the section here are the things that need to be set:

Locations

ZCAT Location of the `zcat` program on linux

TMPDIR Set the temporary directory for all tools please set somthing other then `/tmp`

JAVA_1_6 Set JAVA version 1.6

JAVA_1_7 Set JAVA version 1.7

GATK_SomaticIndel Path to GATK somatic indel detector (GATK version 2.3-9)

GATK Path to GATK (GATK version 3.3.0)

Reference Path to fasta referece file to be used (GRCh37)

Refseq Path to refgene file to be used

PICARD Path to picard tools (Picard version 1.19)

Mutect Path to MuTect (MuTect version 1.1.4)

BaitInterval Bait file to be used for analysis. Please make the interval file based on Picard's HSmetrics tool format.

TargetInterval Target file to be used for analysis. Please make the interval file based on Picard's HSmetrics tool format.

ABRA Path to ABRA (ABRA version 0.92)

TargetRegionLIST Target File in GATK's List format.

PINDELBIN Path to pindel installation (Pindel version 0.2.5a7)

SVpipeline Path to structural variant framework (only required if running the structural variant detection framework)

CAT Location of `cat` program on Linux

PYTHON Location of `python` program on Linux (Python version 2.7.8)

TrimGalore Path to trimgalore tool

PERL Location of `perl` program on Linux (Perl version 5.20.2)

BWA Path to bwa tool

GeneInterval Gene interval file

GeneIntervalAnn Gene interval annotated file

GeneCoord Path to Gene Coordinate file

TilingInterval = Path to tiling intervals

TilingIntervalAnn Path to tiling intervals - annotated for cytoband, for copy number

FingerPrintInterval Path to FingerPrint Interval file

dbSNP Path to db snp vcf file

COSMIC Path to cosmic vcf (version 0.68)

Mills_1000G_Indels Path to Mills 1000G Indels

dbSNP_bitset Path to dbsnp bitset file

AnnotateAssessFilterVariants Path to Annotate Assess and Filter variants script

LoessNormalization Path ot Loess Normalization for copynumber

GCBiasFile Path to GCbias file for copy number

HistNormDir Path to Historiacal Normal dir for Copy number

BestCopyNumber Path to Copy number script

NormVsNormCopyNumber Path to Normal vs. Normal Copy number script

StdNormalLoess_TM Standard Normals for copy number analysis - FFPE for tumor samples#

StdNormalLoess_NVN Standard Normals for copy number normal vs normal analysis

AllMetrics Path to all metrics R script

SAMTOOLS Path to samtools

BEDTOOLS Path to bedtools

GenotypeAllele Path to Genotype allele script

CosmicHotspotVcf Path to cosmic hotspot vcf

Annovar Path to Annovar script

Annovar_db Path to Annovar DB

Canonical_refFlat_file Path to canonical reflat file

IGVtools Path to IGV tools

TranslationFolder Path to translation folder

HotSpot_mutations Path to hotspot mutations for 2 tiered filtering

clinicalExons ListOfClinicalExon

Validated_Exons File with List Of Clinically Validated Exons

Tumor_supressor_list Path to list of tumor supressor genes

Canonical_Exon_Interval_table_with_aa Path to exon interval table

Canonical_Exon_Interval_list Path to canonical exon interval table for DoC

NormalVariantsVCF Path to compiled variants found in mixed normals

QSUB Path to qsub for SGE

BSUB Path to bsub for LSF

RHOME Path to R bin directory

RLIBS Path to R library directory

RSYNC Path to `rsyn` on system

BarcodeKey Path to barcode key file

AdaptorKey Path to adaptor key file

StandardNormalsDirectory Directory where the standard normals are stored

Parameters

Set the parameters to different file/folders/values required by the IMPACT pipeline

StdNormalForMutationCalling Path to standard normal to be used for mutation calling

ListOfFiles File of Files(FOF) for different steps for the pipeline (only required when the process dont start from merging fastq)

Process Which process to run the pipeline on (can be 1,2,3,4,5,6,7 independently or continuous combination in ascending order)

FastqSource Where are the fastq file from (can be `GCL` or `DMP`)

MAPQ Mapping Quality Threshold (Used by DMP-IMPACT:0.2)0

BASQ Base Quality Threshold (Used by DMP-IMPACT:0.2)

MergeDinucleotide Flag to Merge di-nucleotide mutation(can be 1(True) or 2(False))

MoveFiles Flag to Move file in folders (can be 1(True) or 2(False))

DeleteIntermediateFiles Flag ti Delete Intermediate Files (can be 1(True) or 2(False))

TNfreqRatio_MutectStdFilter TN freq Ratio for mutect std filter (Used by DMP-IMPACT:5)

TNfreqRatio_SomIndelStdFilter TN freq Ratio for SID std filter (Used by DMP-IMPACT:5)

VF_threshold_hotspot Variant Frequency threshold for SNV hotspot (Used by DMP-IMPACT:0.01)

AD_SomIndelSTDFilter Allele Depth Threshold for SID standard filter (Used by DMP-IMPACT:5)

DP_SomIndelSTDFilter Total Depth Threshold for SID standard filter (Used by DMP-IMPACT:0)

VF_SomIndelSTDilter Variant Frequency Threshold for SID standard filter (Used by DMP-IMPACT:0.01)

AD_MutectSTDFilter Allele Depth Threshold for Mutect standard filter (Used by DMP-IMPACT:5)

DP_MutectSTDFilter Total Depth Threshold for Mutect standard filter (Used by DMP-IMPACT:0)

VF_MutectSTDFilter Variant Frequency Threshold for Mutect standard filter (Used by DMP-IMPACT:0.01)

TNfreqRatio_AnnotationFilter Tumor to Normal frequency ratio therehold for Annotation (Used by DMP-IMPACT:5)

PON_AD_Threshold Panel of Normal Allele Depth Threshold (Used by DMP-IMPACT:3)

PON_TPVF_Threshold Panel of Normal TPVF Threshold (Used by DMP-IMPACT:10)

Pindel_Min_Indel_Len Minimum Length of INDEL called by PINDEL(Used by DMP-IMPACT:25)

Pindel_Max_Indel_Len Maximum Length of INDEL called by PINDEL (Used by DMP-IMPACT:2000)

MAFthreshold_AnnotationFilter Maf threshold for Annotation (Used by DMP-IMPACT:0.01)

minimumDPforSNV Minimum Total Depth for Novel SNVs (Used by DMP-IMPACT:20)

minimumADforSNV Minimum Allele Depth for Novel SNVs (Used by DMP-IMPACT:10)

minimumVFforSNV Minimum Variant Frequency for Novel SNVs (Used by DMP-IMPACT:0.05)

minimumDPforSNVhs Minimum Total Depth for Hotspot SNVs (Used by DMP-IMPACT:20)

minimumADforSNVhs Minimum Allele Depth for Hotspot SNVs (Used by DMP-IMPACT:8)

minimumVFforSNVhs Minimum Variant Frequency for Hotspot SNVs (Used by DMP-IMPACT:0.02)

minimumDPforINDEL Minimum Total Depth for Novel INDELs (Used by DMP-IMPACT:20)

minimumADforINDEL Minimum Allele Depth for Novel INDELs (Used by DMP-IMPACT:10)

minimumVFforINDEL Minimum Variant Frequency for Novel INDELs (Used by DMP-IMPACT:0.05)

minimumDPforINDELhs Minimum Total Depth for Hotspot INDELs (Used by DMP-IMPACT:20)

minimumADforINDELhs Minimum Allele Depth for Hotspot INDELs (Used by DMP-IMPACT:8)

minimumVFforINDELhs Minimum Variant Frequnecy for Hotspot INDELs (Used by DMP-IMPACT:0.02)

occurrencePercent Minimum Percentage For Occurrence In Other Normals (Used by DMP-IMPACT:0.2)

Coverage_threshold_darwin_report Coverage threshold for darwin reports(good coverage vs bad coverage) (Used by DMP-IMPACT:100)

QUEUE_NAME Name of the queue on the SGE or LSF

CLUSTER Flag for what cluster to be used (can `SGE` or `LSF`)

runABRA Flag to whether use ABRA or GATK indel realignment(can be 1(True) or 2(False))

Versions

Note:

This section is just to print what version of things you are using so you can have all the dependencies with the respective versions listed here.

Inside the version there are version that are being used for each tool. This is just for consistency in reports.

Description for title_file.txt

Headers for this tab-delimited file should be exactly with this names:

Barcode Has to start with bc and end with any number [for example: bc01 or bc101 should match the **adaptor & barcode** file mentioned in configuration file

Pool Can be any string **joined by "-"** and **not "_"** and all entries should be from same pool

Sample_ID Can be any string **joined by "-"** and **not "_"**

Collab_ID Can be any string or –

Note:

Patient with multiple samples should have **same Patient_ID**

Patient_ID Can be any string **joined by "-"** and **not "_"**

Class Can be Tumor or Normal.

Sample_type Can be any string or –

Input_ng Can be any float or –

Library_yield Can be float or –

Pool_input Can be float or –

Bait_version Can be any string or –

Gender Can be any Male/Female or –

PatientName Can be any string or –

MAccession Can be any string or –

Extracted_DNA_Yield Can be a float or –

For analysis to start the **outputDirectory** will be required to have this file with `title_file.txt` as the name or this file needs to be present in the **configuration** file with either `title_file.txt` as then name or `Pool_title.txt` as the name where **Pool** is the string used above for that category.

Description for SampleSheet.csv

This is a comma separated file is created by the illumina sequencer and it is used to merge the fastq files.

Headers for this tab-delimited file should be exactly with this names:

FCID Flowcell ID (required)

Lane Lane Number, this is used to merge the fastq files across lanes (required)

SampleID Sample ID, this is used to merge the files (required)

SampleRef Sample Reference is from [example:HUMAN]

Index Index used to sequence the sample (require)

Description Description of the samples

Control Can be any string or –

Recipe Can be any string or –

Operator Can be any string or –

SampleProject Can be any string or –

For analysis to start the **outputDirectory** will be required to have this file with `SampleSheet.csv` as the name or this file needs to be present in the **configuration** file with `SampleSheet.csv` as the name.

Description for adaptor file in the configuration file

The adaptor file is the tab-delimited file with two columns:

1. Barcode Key to which the adaptor belongs which should always start with `bc`
2. Adaptor sequence itself

There is **no header** in this file.

For Example:

bc01	GATCGGAAGAGCACACGTCTGAACTCCAGTCACAACGTGATATCTCGTATGCCGTCTTCTGCTTG

Description for barcode file in the configuration file

The barcode file is the tab-delimited file with two columns:

1. Barcode Sequece
2. Barcode Number that sequence represent.

There is **a header** in this file.

For Example:

Sequence	TruSeqBarcode
AACGTGAT	bc01

CHAPTER 3

Usage

Quick Usage

RunIlluminaProcess.pl [options]

–config | -c S Path to configration file(required)

—svConfig | -sc S Path to structural variant configration file(optional)

–symLinkFlag | -sf I Flag for Keeping or removing the symolic links(1:Remove;2:Keep)(default:2)

—dataDirectory | -d S Path where all the files to be processed are located (required)

–outputDirectory | -o S Path where all the output files will be written (required)

Assuming you have setup the configuration file properly and you have SampleSheet.csv and title_file.txt in the **dataDirectory** you can run:

```
nohup perl RunIlluminaProcess.pl -c configuration.txt -sc configuration_sc.txt -d /
↪path/to/fastq/files -o /path/to/output/directory
```

Detailed Usage

The behaviour of the prgram depends on the inputs in the configuration file:

In the configuration file the **Process** variable in section **>Parameters** tells pipeline following:

What does each number represent

Process	Things Pipeline will do
1	Merge Fastq
2	Trimming, Mapping & sorting of SAM file giving you a BAM file
3	Mark Duplicates, Indel Realignment, Base Quality Recalibration
4	Metrics Calculation, QC Report Genaration and launching IMPACT-SV if given -sc flag specified
5	Variant Calling
6	Variant Filtering and Genotyping
7	Variant Annotation and Variant Filtering

Using different Process to run Pipeline

1. To run the complete pipeline. Set the following in the configuration file:

   Process 1,2,3,4,5,6,7

2. To run from **Process 1**. Set the following in the configuration file:

   ListOfFilesListOfFiles fastq.list #where fastq.list contains all the fastq files to be proces, this needs to be an even number as it automatically pairs them.

   Process 2,3,4,5,6,7

3. To run from the **Process 3 to 7**. Set the following in the configuration file:

   ListOfFiles SortedBam.list (where SortedBam.list contains all the sorted bam files from Process 2 to be processed)

   Process 3,4,5,6,7

4. To run from the **Process 4 to 7**. Set the following in the configuration file:

   ListOfFiles RecalibratedBam.list (where Recalibrated.list contains all the recalibrated bam files from Process 3 to be processed)

   Process 4,5,6,7

Note:

For this to be sucessfull you should have the files from step 4 in the **outputDirectory**

5. To run from the **Process 5 to 7**. Set the following in the configuration file:

   ListOfFiles RecalibratedBam.list (where Recalibrated.list contains all the recalibrated bam files from Process 3 to be processed)

   Process 5,6,7

Note:

For this to be sucessfull you should have the files from step 5 in the **outputDirectory**

6. To run from the **Process 6 to 7**. Set the following in the configuration file:

   ListOfFiles RecalibratedBam.list (where Recalibrated.list contains all the recalibrated bam files from Process 3 to be processed)

Process 6,7

> **Note:**
>
> For this to be sucessfull you should have the files from step 6 in the **outputDirectory**

7. To run from the **Process 7**. Set the following in the configuration file:

    ListOfFiles RecalibratedBam.list (where Recalibrated.list contains all the recalibrated bam files from Process 3 to be processed)

    Process 7

If you want to run each Process separetly that is also possible but you need to make sure that files from previous procss are present in the **outputDirectory**

Shell Script to run pipeline

There is also a **helper shell script (Run_Pipeline_Example.sh)** in the bin directory which will help to run the framework. Which looks like this:

> **Note:**
>
> Please comment out the lines using (#) according to the cluster type and analysis type.

```
##Run_Pipeline_Example.sh
#author:Ronak H Shah
#v1.0.1
##Path where the fastq are stored
export DATADIR=<Path for Data Directory>
##Path where the output shoud be written
export OUTDIR=<Path To Output Directory>
##Path to the IMPACT-Pipeline script
export PipelineScript=<Path to IMPACT-Pipeline Script>
##Path to Perl installation
export Perl=<Path to Perl>
##Project associated with the Run
export ProjectName=<ProjectName>
##Path to working directory where you will write the LSF/SGE outputs
export WorkingDir=<Path to write sge/lsf files>
##Path to configfile for running main IMPACT pipeline
export CONFIGFILE=<Path To Pipeline Configuration File>
##Path to structural variants pipeline configuration file
export SV_ConfigFile=<Path to SV detection configuration file>

##Run both IMPACT-Pipeline & SV Process on LSF
echo bsub -q sol -cwd ${WorkingDir} -J ${ProjectName} -e${ProjectName}.stderr -o $
↪{ProjectName}.stdout -We 24:00 -R "rusage[mem=2]" -M 4 \"${Perl} ${PipelineScript} -
↪c ${CONFIGFILE} -sc {$SV_ConfigFile} -d ${DATADIR} -o ${OUTDIR}\"
bsub -q sol -cwd ${WorkingDir} -J ${PoolName} -e${ProjectName}.stderr -o $
↪{ProjectName}.stdout -We 24:00 -R "rusage[mem=2]" -M 4 \"${Perl} ${PipelineScript} -
↪c ${CONFIGFILE} -sc {$SV_ConfigFile} -d ${DATADIR} -o ${OUTDIR}\"
##Run IMPACT-Pipeline on LSF
echo bsub -q sol -cwd ${WorkingDir} -J ${ProjectName} -e${ProjectName}.stderr -o $
↪{ProjectName}.stdout -We 24:00 -R "rusage[mem=2]" -M 4 \"${Perl} ${PipelineScript} -
↪c ${CONFIGFILE} -d ${DATADIR} -o ${OUTDIR}\"
```

```
bsub -q sol -cwd ${WorkingDir} -J ${PoolName} -e${ProjectName}.stderr -o $
↪{ProjectName}.stdout -We 24:00 -R "rusage[mem=2]" -M 4 \"${Perl} ${PipelineScript} -
↪c ${CONFIGFILE} -d ${DATADIR} -o ${OUTDIR}\"

##Run both IMPACT-Pipeline & SV Process on SGE
echo qsub -q test.q -wd ${WorkingDir} -N ${ProjectName} -l hvmem=2G,virtual_free=2G -
↪pe smp 1 -b y \"${Perl} ${PipelineScript} -c ${CONFIGFILE} -sc ${SV_ConfigFile} -d $
↪{DATADIR} -o ${OUTDIR}\"
qsub -q test.q -wd ${WorkingDir} -N ${ProjectName} -l hvmem=2G,virtual_free=2G -pe␣
↪smp 1 -b y \"${Perl} ${PipelineScript} -c ${CONFIGFILE} -sc ${SV_ConfigFile} -d $
↪{DATADIR} -o ${OUTDIR}\"
##Run both IMPACT-Pipeline on SGE
echo qsub -q test.q -wd ${WorkingDir} -N ${ProjectName} -l hvmem=2G,virtual_free=2G -
↪pe smp 1 -b y \"${Perl} ${PipelineScript} -c ${CONFIGFILE} -d ${DATADIR} -o $
↪{OUTDIR}\"
qsub -q test.q -wd ${WorkingDir} -N ${ProjectName} -l hvmem=2G,virtual_free=2G -pe␣
↪smp 1 -b y \"${Perl} ${PipelineScript} -c ${CONFIGFILE} -d ${DATADIR} -o ${OUTDIR}\"
```

CHAPTER 4

Description of sub-scripts

Script in the bin folder

Compiling QC metrics, Generating QC-Report and running copynumber analysis

dmp_compile_qc_metrics.pl [options]

–bamList | -i S File of files having list of all bam files (required)

—titleFile | -t S tab-delimited title file for the samples (required and submit with full path)

–AllMetrics | -am S Path to AllMetrics script (required)

—LoessNorm | -ln S Path to Loess Normalization Script (required)

–BestCN | -cn S Path to Best Copy Number Script (required)

—GCBias | -gcb S Path to GC bias file (required)

–HistNorm | -his S Path to Directory with all historical normal files (required)

—queue | -q S Name of the Sun Grd Engine Queue where the pipeline needs to run (required)

--qsub S Path to qsub executable for SGE(default:None,optional)

—bsub S Path to bsub executable for LSF(default:None,required)

–metricsScript | -ms S Name of the script used to generate .html and .pdf files

—outdir | -o S Path where all the output files will be written (optional) [default:cwd]

Genotyping Variants across multiple samples

dmp_genotype_allele.pl [options]

–FilteredMutationVcfFile | -fmv S vcf file describing details about the mutations (required)

—BamFile | -bam S bam file to be used for genotyping (required)

–RefFile | -rf S Path to genome reference file (required)

—samtools | -s S Path to samtools (required)

–bedtools | -b S Path to bedtools (required)

—MinBaseQualit | -mbq I Min. Base Quality Threshold (optional;default:5)

–MinMappingQuality | -mmq I Min. Mapping Quality Threshold (optional;default:5)

—deleteUnwantedFiles | -d I 2=>To delete files 1=> To keep files (default:2,optional)

–outdir | -o S Path where all the output files will be written (optional;default:current working directory)

—outFile | -of S Name of the allele depth output file (optional;default:BamFame-.bam+_mpileup.alleledepth)

–bamId | -bi S Bam Id to be used (optional;default:bamfile name)

—queue | -q S Name of the SGE / LSF Queue where the pipeline needs to run (required)

--qsub S Path to qsub executable for SGE(default:None,optional)

—bsub S Path to bsub executable for LSF(default:None,required)

–mpileUpOutFile | -mof S Name of samtools mpileup output file (optional;default:BamFile-.bam+.mpileup)

—typeOfSample | -tos S Type of Sample (optional;default:Tumor;canbe Tumor or Normal)

Running Indel Realignment using ABRA

usage: Run_AbraRealignment.py [options]

Run ABRA Indel Realignment

arguments:

-h, --help show this help message and exit

-i BamFile.list, –bamList BamFile.list Full path to the tumor bam files as a fof.

-p PatientID, --patientId PatientID Id of the Patient for which the bam files are to be realigned

-v, --verbose make lots of noise [default]

-t 5, –threads 5 Number of Threads to be used to run ABRA -d, –mdp Threshold for downsampling depth to run ABRA -k [43 [43 ...]], –kmers [43 [43 ...]]

Number of k-mers to be used to run ABRA; Multiple k-mers are separated by space

-temp /somepath/tmpdir, –temporaryDirectory /somepath/tmpdir Full Path to temporary directory

-r /somepath/Homo_Sapeins_hg19.fasta, –referenceFile /somepath/Homo_Sapeins_hg19.fasta Full Path to the reference file with the bwa index.

-a /somepath/ABRA.jar, –abraJar /somepath/ABRA.jar Full Path to the ABRA jar file.

-tr /somepath/targetRegion.bed, –targetRegion /somepath/targetRegion.bed Full Path to the target region bed file

-j /somepath/java, –javaPATH /somepath/java Path to java executable.

-b /somepath/bin, –bwaPATH /somepath/bin Path to the bin of bwa executable.

-q all.q or clin.q, –queue all.q or clin.q Name of the SGE queue

-o /somepath/output, –outDir /somepath/output Full Path to the output dir.

-qsub /somepath/qsub, –qsubPath /somepath/qsub Full Path to the qsub executables of SGE.

-bsub /somepath/bsub, –bsubPath /somepath/bsub Full Path to the bsub executables of LSF.

Call Indels > 25bp using Pindel

usage: Run_Pindel.py [options]

Run Pindel for Long Indels & MNPS (32bp-350bp)

optional arguments:

-h, --help show this help message and exit

-i pindel.conf, –pindelConfig pindel.conf Full path to the pindel configuration

-pId PatientID, –patientId PatientID Id of the Patient for which the bam files are to be realigned

-v, --verbose make lots of noise [default]

-t 5, –threads 5 Number of Threads to be used to run Pindel -r /somepath/Homo_Sapeins_hg19.fasta, –referenceFile /somepath/Homo_Sapeins_hg19.fasta

Full Path to the reference file with the bwa index.

-p /somepath/pindel/bin, –pindelDir /somepath/pindel/bin Full Path to the Pindel executables.

-chr ALL, –chromosomes ALL Which chr/fragment. Pindel will process reads for one chromosome each time. ChrName must be the same as in reference sequence and in read file.

-q all.q or clin.q, –queue all.q or clin.q Name of the SGE queue

-o /somepath/output, –outDir /somepath/output Full Path to the output dir.

-op TumorID, –outPrefix TumorID Id of the Tumor bam file which will be used as the prefix for Pindel output files

-qsub /somepath/qsub, –qsubPath /somepath/qsub Full Path to the qsub executables of SGE.

-bsub /somepath/bsub, –bsubPath /somepath/bsub Full Path to the bsub executables of LSF.

Script in the support-scripts folder

Calculate intervals from bam file that have some minimum coverage

usage: Run_FindCoveredInterval.py [options]

This will run find covered interval program from GATK.

optional arguments:

-h, --help show this help message and exit

-i BamFile.list, –bamList BamFile.list Full path to the tumor bam files as a fof.

-of OutFilePrefix, –outFilePrefix OutFilePrefix Output Covered Interval File Prefix for the bam files.

-v, --verbose make lots of noise [default]

-t 5, –threads 5 Number of Threads to be used to run FindCoveredIntervals

-dp 20, –totaldepth 20 Total depth threshold

-mbq 20, –minbasequality 20 Threshold for minimum base quality for Running Find Covered Interval

-mmq 20, –minmappingquality 20 Threshold for minimum mapping quality for Running Find Covered Interval

-r /somepath/Homo_Sapeins_hg19.fasta, –referenceFile /somepath/Homo_Sapeins_hg19.fasta Full Path to the reference file with the bwa index.

-g /somepath/GenomeAnalysisTK.jar, –gatkJar /somepath/GenomeAnalysisTK.jar Full Path to the GATK jar file.

-j /somepath/java, –javaPATH /somepath/java Path to java executable.

-q all.q or clin.q, –queue all.q or clin.q Name of the SGE queue

-o /somepath/output, –outDir /somepath/output Full Path to the output dir.

-qsub /somepath/qsub, –qsubPath /somepath/qsub Full Path to the qsub executables of SGE.

-bsub /somepath/bsub, –bsubPath /somepath/bsub Full Path to the bsub executables of LSF.

Annotating variants in merged variant file

dmp_annotate_variants.pl [options]

–SomaticMutIndelFile | -si S File containing mutations (required and submit with full path,Ex:/SomePath/Some_SomaticMutIndel.txt)

—ConfigurationFile | -c S Configuration file that contains the locations for the programs and the databases (required and submit with full path)

–titleFile | -t S tab-delimited title file for the samples (required and submit with full path)

—outdir | -o S Path where all the output files will be written (optional;default:cwd)

–exonCoverageFile | -ec S Path where the all exon coverage file is located (full path]

—geneCoverageFile | -gc S Path where the gene coverage file is located (full path)]

—deleteUnwantedFiles | -d I 2=>To delete files 1=> To keep files (default:2,optional)

Filter variants after annotation

dmp_filter_genotyped_variants.pl [options]

–input | -i S File containing mutations with genotype information (required)

—hotspots | -h S File containing the list of hotspots (required)

–clinicalExons | -ce S File containing the list of clinical exons (required)

—titleFile | -t S Title file (required)

–minimumDPforSNVs | -dp_snv I Minumum accepted DP for novel SNVs (default: 20)

—minimumADforSNVs | -ad_snv I Minimum accepted AD for novel SNVs (default: 10)

–minimumVFforSNVs | -vf_snv F Minimum accepted VF for novel SNVs (default: 0.05)

—minimumDPforSNVhotspot | -dp_snvHS I Minumum accepted DP for Hotspot SNVs (default: 20)

–minimumADforSNVhotspot | -ad_snvHS I Minimum accepted AD for Hotspot SNVs (default: 8)

—minimumVFforSNVhotspot | -vf_snvHS F Minimum accepted VF for Hotspot SNVs (default: 0.02)

–minimumDPforINDELs | -dp_indel I Minumum accepted DP for novel INDELs (default: 20)

—minimumADforINDELs | -ad_indel I Minimum accepted AD for novel INDELs (default: 10)

–minimumVFforINDELs | -vf_indel I Minimum accepted VF for novel INDELs (default: 0.05)

—minimumDPforINDELhotspot | -dp_indelHS F Minumum accepted DP for Hotspot INDELs (default: 20)

–minimumADforINDELhotspot | -ad_indelHS I Minimum accepted AD for Hotspot INDELs (default: 8)

—minimumVFforINDELhotspot | -vf_indelHS F Minimum accepted VF for Hotspot INDELs (default: 0.02)

–minimumOccurrencePercent | -occurrence S Minimum accepted value of occurrence in other normals, in percent (default: 20)

—TNfreqRatioThreshold | -tn_ratio S Minimum value for VFt/VFn value (default: 5)

–MAFthreshold | -mt F Minimum accepted MAF values for unmatched variant calls (default : 0.01)

Filter indels from SomaticIndelDetector before genotyping

dmp_filter_indel.pl [options]

–IndelTxtFile|t S tab-delimted Indel file describing details about the mutations (required)

—IndelVcfFile|v S VCF format Indel file describing details about the mutations (required)

–sampleName|s S Name of the sample (required)

—totaldepth|dp I Tumor total depth threshold for Somatic Indel Detector(default:0,optional)

–alleledepth|ad I Tumor Allele depth threshold for Somatic Indel Detector(default:3,optional)

—variantfreq|vf F Tumor variant frequency threshold for Somatic Indel Detector(default:0.01,optional)

–TNratio|tnr I Tumor-Normal variant frequency ratio threshold for Somatic Indel Detector(default:5,optional)

—outdir|o S Path where all the output files will be written (optional) default:current working directory

Filter snv from MuTect before genotyping

dmp_filter_mutect.pl [options]

–MutationTxtFile|t S tab-delimted Mutect file describing details about the mutations (required)

–MutationVcfFile|v S VCF format Mutect file describing details about the mutations (required)

—sampleName|s S Name of the sample (required)

–totaldepth|dp I Tumor total depth threshold for Mutect(default:0,optional).

—alleledepth|ad I Tumor Allele depth threshold for Mutect(default:3,optional).

–variantfreq|vf F Tumor variant frequency threshold for Mutect(default:0.01,optional).

—TNratio|tnr I Tumor-Normal variant frequency ratio threshold for Mutect(default:5,optional).

–outdir|o S Path where all the output files will be written (optional) default:current working directory

Filter indels from PINDEL before genotyping

usage: dmp_filter_pindel.py [options]

Filter Indels from the output of pindel

optional arguments:

-h, --help show this help message and exit

-v, --verbose make lots of noise [default]

-i SomeID.vcf, -inputVcf SomeID.vcf Input vcf freebayes file which needs to be filtered

-tsn SomeName, –tsampleName SomeName Name of the tumor Sample

-dp 0, –totaldepth 0 Tumor total depth threshold

-ad 3, –alleledepth 3 Tumor allele depth threshold

-tnr 5, –tnRatio 5 Tumor-Normal variant frequency ratio threshold -vf 0.01, –variantfrequency 0.01

Tumor variant frequency threshold

-o /somepath/output, –outDir /somepath/output Full Path to the output dir.

-min 25, –min_var_len 25 Minimum length of the Indels

-max 500, –max_var_len 500 Max length of the Indels

CHAPTER 5

File Description

Here is small description for each of the important output files.

QC Metrics Files

Proj_*_All_Metrics.pdf

PDF file with graphical representation of all calculated quality and performance metrics for all samples in project. Numerical values are contained in the text files below.

Proj_*_ALL_basequalities.txt

Illumina base quality score by cycle (following GATK base quality recalibration), displayed for each sample.

Proj_*_ALL_Canonical_exoncoverage.txt

Mean coverage for each target interval corresponding to exons in canonical transcripts, displayed for each sample. Coverage is computed for reads with mapping quality > 20.

Proj_*_ALL_exoncoverage.txt

Mean coverage for each target interval corresponding to all protein-coding exons in all transcripts (including flanking splice sites), displayed for each sample. Coverage is computed for reads with mapping quality > 20.

Proj_*_ALL_exonnomapqcoverage.txt

Mean coverage for each target interval corresponding to all protein-coding exons in all transcripts (including flanking splice sites), displayed for each sample. Coverage is computed for reads regardless of mapping quality (i.e., including reads mapping to multiple locations).

Proj_*_ALL_FPavgHom.txt

Average minor allele fraction across all tiling SNPs that are homozygous in the given sample. This is an indication of the degree of contamination from unrelated DNA. (The IMPACT and HemePACT panels contain >1,000 "tiling SNPs" where both alleles are common in the population.)

Proj_*_ALL_FPCResultsUnMatch.txt

Pairs of samples from different individuals with concordant fingerprint genotypes. (The number represents the fraction of tiling SNPs homozygous in one sample which are homozygous for the alternate allele in the other sample.)

Proj_*_ALL_FPCResultsUnMismatch.txt

Pairs of samples from the same individual with discordant fingerprint genotypes. (The number represents the fraction of tiling SNPs homozygous in one sample which are homozygous for the alternate allele in the other sample.)

Proj_*_ALL_FPCsummary.txt

Matrix of all pairwise concordance values based on fingerprint genotypes. (The number represents the fraction of tiling SNPs homozygous in one sample which are homozygous for the alternate allele in the other sample.)

Proj_*_ALL_FPhet.txt

Fraction of all tiling SNPs that are heterozygous in the given sample. Samples with >0.50 heterozygous SNPs may be contaminated with unrelated DNA. (The IMPACT and HemePACT panels contain >1,000 "tiling SNPs" where both alleles are common in the population.)

Proj_*_ALL_FPsummary.txt

For each tiling SNP in each sample: the observed allele counts of each base, the inferred genotype, and the minor allele fraction. (The IMPACT and HemePACT panels contain >1,000 "tiling SNPs" where both alleles are common in the population.)

Proj_*_ALL_gcbias.txt

For each sample, the average coverage for target intervals in different bins of GC content (from 25-30% to 80-85%).

Proj_*_ALL_genecoverage.txt

Mean coverage for each target gene (across all target exons), displayed for each sample. Coverage is computed for reads with mapping quality > 20.

Proj_*_ALL_genotypehotspotnormals.txt

Sites of known somatic mutation hotspots (COSMIC) with variants detected in a normal sample. Allele counts are also shown for the matched tumor sample as well as any other tumor where the corresponding variant was detected.

Proj_*_ALL_HSmetrics.txt

Quality and performance metrics for hybrid selection, calculated by Picard.

Proj_*_ALL_insertsizemetrics.txt

Numerical values for histogram of library insert size distribution, displayed for each sample.

Proj_*_ALL_intervalnomapqcoverage_loess.txt

Mean coverage for each target interval, displayed for each sample. Coverage is computed for reads regardless of mapping quality (i.e., including reads mapping to multiple locations). Coverage is then normalized according to a Loess normalization to adjust for biases in GC content of target intervals.

Proj_*_ALL_intervalnomapqcoverage.txt

Mean coverage for each target interval, displayed for each sample. Coverage is computed for reads regardless of mapping quality (i.e., including reads mapping to multiple locations).

Proj_*_ALL_orgbasequalities.txt

Illumina base quality score by cycle (prior to GATK base quality recalibration), displayed for each sample.

Copy Number Files

Proj_*_ALL_copynumber.seg

Segmented copy number (following %GC-based loess normalization and CBS-based segmentation) for all samples. This file can be loaded in IGV.

Proj_*_copynumber_segclusp.genes.txt

Gene-level copy number (following %GC-based loess normalization and CBS-based segmentation) for all samples. Each gene is assigned the copy number ratio (fold-change) for the segment on which it falls. If a gene spans multiple segments, it is assigned the average copy number ratio. The normal sample selected for copy number normalization is shown for each sample. Both fold-change and a p-value (indicative of the signal-to-noise ratio) are given.

Proj_*_copynumber_segclusp.intragenic.txt

List of genes with putative intragenic copy number deletions. Individual targets (exons) are listed down the right-most column. The copy number ratios for the exons in each gene are grouped into two clusters. (Cluster membership is labeled as "1" or "2".) Genes where exons with lower copy number are consecutive are candidates for intragenic deletions, though this analysis tends to overcall events.

Proj_*_copynumber_segclusp.pdf

Plots of copy number profiles for all tumors. Each data point represents a target interval (blue = exon; red = tiling SNP). Each tumor is normalized against a diploid normal, either from the same project or from a historical panel of normals; the specific normal used for normalization is shown. Listed below each plot are the 20 highest-level amplifications (left) and deletions (right), though not all are significant. Significantly altered genes (fold-change > 2 and p-value < 0.05) are marked with an asterisk.

Proj_*_copynumber_segclusp.probes.txt

Target-level copy number (following %GC-based loess normalization) for all samples. The normal sample selected for copy number normalization is shown for each sample. "fc" = fold change; "lr" = log ratio.

Proj_*_discrete_CNA.txt

Matrix of significant copy number alterations by gene (rows are genes; columns are tumors). 2 = amplification; 0 = neutral; -2 = deletion.

Proj_*_loessnorm.pdf

Plots showing best loess normalization fit to adjust for coverage bias related to %GC content of target intervals.

Structural Variant Files

Proj_*_AllAnnotatedSVs.txt

Annotated output from DELLY rearrangement detection algorithm. Genomic coordinates and gene annotations are provided for both breakpoints. The distance between breakpoints (SV_LENGTH) and orientation of fragments that are joined (Conection_Type) are shown. Also provided are the number of supporting paired reads and split reads and the inferred derived sequence at the breakpoint (when possible). This analysis tends to overcall events—as the vast majority of rearrangements are false positives, manual review and further filtering are necessary. Events listed (in Site2Description) as "Protein fusion: in frame" should be prioritized.

Proj_*_AllAnnotatedSVs.xlsx

Same information as Proj_*_AllAnnotatedCVs.txt (above) in Excel file format.

Per Sample Files

s_*_Proj_*_copynumber.seg

Segmented copy number for each individual sample. Each segment is listed with the genomic coordinates, number of target intervals, and copy number log ratio.

s_*_Proj_*_mrg_cl_aln_srt_MD_IR_BR.canonical.exon.covg.sample_interval_summary

Coverage statistics (total, mean, quartiles) for each exon of a canonical transcript in each individual sample. Coverage is calculated for reads with mapping quality > 20.

s_*_Proj_*_mrg_cl_aln_srt_MD_IR_BR.gene_nomapq.covg.sample_interval_summary

Coverage statistics (total, mean, quartiles) for each exon of any target transcript in each individual sample. Coverage is computed for reads regardless of mapping quality (i.e., including reads mapping to multiple locations).

s_*_Proj_*_mrg_cl_aln_srt_MD_IR_BR.gene.covg.sample_interval_summary

Coverage statistics (total, mean, quartiles) for each exon of any target transcript in each individual sample. Coverage is calculated for reads with mapping quality > 20.

s_*_Proj_*_mrg_cl_aln_srt_MD_IR_BR.target.covg

Coverage statistics (mean covearage and %GC content) for all target intervals (coding exons and tiling SNPs) in each individual sample. Coverage is calculated for reads with mapping quality > 20.

s_*_Proj_*_mrg_cl_aln_srt_MD_IR_BR.tiling_nomapq.covg.sample_interval_summary

Coverage statistics (mean covearage and %GC content) for just tiling intervals (tiling SNPs and fingerprint SNPs) in each individual sample. Coverage is computed for reads regardless of mapping quality (i.e., including reads mapping to multiple locations).

s_*_Proj_*_mrg_cl_aln_srt_MD_IR_BR.tiling.covg.sample_interval_summary

Coverage statistics (mean covearage and %GC content) for just tiling intervals (tiling SNPs and fingerprint SNPs) in each individual sample. Coverage is calculated for reads with mapping quality > 20.

Sample Info File

Proj_*_title_file.txt

Sample IDs (provided by CMO and Investigators) and limited metadata (including sample type, input DNA amount, library yield, and capture platform) for all samples in the project.

Variant Files

annotated_exonic_variants.txt

Mutations and indels (and genomic annotations and allele counts) called in all tumors in the project. The normal sample used for mutation calling is displayed in the second column (usually either the matched normal or a pool of unmatched normal). Presence in dbSNP, COSMIC, and the 1000 genomes project (depicted as the minor allele fraction in the population) is shown. For tumors called against an unmatched pool of normal, variants present in >0.01 of the 1000 genomes project (and not in COSMIC) are automatically filtered out. Allele counts across a panel of historical normal are shown to indicate potential systematic artifacts. The rightmost columns display the stats for every variant in every sample, including a panel of historical normals (usually labeled with a capital M or S). DP = depth of coverage at the variant site; RD = counts of reference allele; AD = counts of alternate (mutant) allele; VF = variant frequency.

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotated.txt

Intermediate variants file displaying all mutations and indels called prior to any filtering. Stats are included for every variant in every sample in the project (as described above).

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedExonic.txt

Intermediate variants files displaying all non-silent mutations and indels called in exonic (plus splice site) regions of target genes.

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedExonic.Dropped.txt

Intermediate variants files displaying all non-silent mutations and indels called in exonic (plus splice site) regions of target genes. The file labeled “Dropped” includes candidate mutations that were subsequently rejected based on empirical filters (e.g., present in historical normal or insufficient coverage, read support, or variant frequency).

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedExonic.Filtered.txt

Intermediate variants files displaying all non-silent mutations and indels called in exonic (plus splice site) regions of target genes. The file labeled “Filtered” includes candidate mutations that survived all filters.

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedSilent.txt

Intermediate variants files displaying all silent mutations called in exonic (plus splice site) regions of target genes.

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedSilent.Dropped.txt

Intermediate variants files displaying all silent mutations called in exonic (plus splice site) regions of target genes. The file labeled “Dropped” includes candidate mutations that were subsequently rejected based on empirical filters (e.g., present in historical normal or insufficient coverage, read support, or variant frequency).

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedSilent.Filtered.txt

Intermediate variants files displaying all silent mutations called in exonic (plus splice site) regions of target genes. The file labeled "Filtered" includes candidate mutations that survived all filters.

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedNonPanelExonic.txt

Intermediate variants files displaying all non-silent mutations and indels called in exonic (plus splice site) regions of off-target genes.

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedNonPanelExonic.Dropped.txt

Intermediate variants files displaying all non-silent mutations and indels called in exonic (plus splice site) regions of off-target genes. The file labeled "Dropped" includes candidate mutations that were subsequently rejected based on empirical filters (e.g., present in historical normal or insufficient coverage, read support, or variant frequency).

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedNonPanelExonic.Filtered.txt

Intermediate variants files displaying all non-silent mutations and indels called in exonic (plus splice site) regions of off-target genes. The file labeled "Filtered" includes candidate mutations that survived all filters.

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedNonPanelSilent.txt

Intermediate variants files displaying all silent mutations and indels called in off-target regions, including introns, intergenic regions, and off-target genes.

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedNonPanelSilent.Dropped.txt

Intermediate variants files displaying all silent mutations and indels called in off-target regions, including introns, intergenic regions, and off-target genes. The file labeled "Dropped" includes candidate mutations that were subsequently rejected based on empirical filters (e.g., present in historical normal or insufficient coverage, read support, or variant frequency).

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotatedNonPanelSilent.Filtered.txt

Intermediate variants files displaying all silent mutations and indels called in off-target regions, including introns, intergenic regions, and off-target genes. The file labeled "Filtered" includes candidate mutations that survived all filters.

Proj_*_AllSomaticMutIndel_withAlleleDepth_annovarAnnotated_All_Filtered.txt

Aggregation of all mutations and indels called (on-target and off-target) that survived all empirical filters (in the "Filtered" files above).

Proj_*_AllSomaticMutIndel_withAlleleDepth.txt

Candidate mutations called in all samples, prior to genomic annotation.

Proj_*_AllSomaticMutIndel_withAlleleDepth_mergedDNP.txt

Candidate mutations called, prior to genomic annotation, with adjacent SNVs with similar coverage and allele frequencies merged into dinucleotide substitutions.

CHAPTER 6

Citation

If you use this software in a publication, please cite our paper

Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology.