**Automatic Pipeline for Panel Sequence Data**

1. Objectives.

The pipeline is developed with aims to conveniently and efficiently pre-process and analyze sequence data from targeted panels to identify somatic variants, such as SNVs, short InDels, CNVs, and internal tandem duplicates (ITDs). The pipeline also includes quality control for the sequences and produces summary statistics and plots for quality evaluations.

1. Methods.

The pipeline applies existing tools on the sequence data, and works on Linux environments with cluster computer facilities.

2.1. Pre-processing.

Pre-processing of the raw sequence data applies BWA, Picard, and GATK tools to align reads on the reference genome, sort the reads, mark duplicates, and recalibrate base quality scores.

2.2. SNV and short InDel.

SNVs and short InDels are decided as a union of variants called by VarDict, SNVer, LoFreq, and are annotated with Funcotator. SNV-InDels are filtered to exclude germline variants and variants with low pathogenic impacts.

2.3. CNV.

CNACS and CNVkit generate two sets of CNV list, each tool excludes possible germline CNVs with panel of normals created from normal samples.

2.4. ITD.

VarDict, Pindel, ScanITD, and getITD are applied on the reads from the targeted regions on FLT3 gene.

2.5. QC.

FASTQuick gives summary statistics and illustrations for sequence quality evaluations.

1. File organizations.

All work is saved under the folder:

/home/projects/cu\_10184/projects/PTH

3.1. Sequence data (fastq).

The bait and target files of the panels are saved under:

/home/projects/cu\_10184/projects/PTH/PanelSeqData/Bait\_Target

The sequence files are saved by batch:

/home/projects/cu\_10184/projects/PTH/PanelSeqData/[batch\_name]/fastq

3.2. Meta data.

Panel information for batches, and patient and tissue information for samples are saved under:

/home/projects/cu\_10184/projects/PTH/Meta

BatchInfo.txt SampleInfo.txt

3.3. Reference files.

Reference files that are more specific to our study and our data, such as known variants identified from clinical study (REDCap and Horizon), filtering schemes decided by our data, panel of normals computed from our normal samples, annotation for specific regions on FLT3 gene for ITD identification, are saved under:

/home/projects/cu\_10184/projects/PTH/Reference

3.4. Pipeline scripts.

Pipeline scripts are saved under the folder:

/home/projects/cu\_10184/projects/PTH/Code

3.5. Files from running by a batch.

All files generated from running the pipeline on a batch are saved under:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]

Log files and error files are under:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/log

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/error

Some intermediate files and final result files are under:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Result

Quality control files are under:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/QC

1. Commands and outputs.

4.1. Full pipeline.

To run the full pipeline:

sh /home/projects/cu\_10184/projects/PTH/Code/Primary/Ensemble/Ensemble.sh -d PTH -b [batch\_name] -p [panel\_name] -t 8

-d: The name of a project that is located at:

/home/projects/cu\_10184/projects/[project\_name]

-b: The name of the batch.

-p: The name of the panel that is used for the batch. If the input is empty, it will try to find bait and target information from:

/home/projects/cu\_10184/projects/PTH/PanelSeqData/Bait\_Target

-t: The number of cores used by each job.

To run the full pipeline for multiple batches:

sh /home/projects/cu\_10184/projects/PTH/Code/Primary/Ensemble: vim SubmitJobs\_AllBatches.sh

The outputs:

* BAM files from alignment:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/BAM

* SNV\_InDel:

Output from VarDict:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/SNV\_InDel/VarDict/vcf

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/SNV\_InDel/VarDict/maf

Output from SNVer:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/SNV\_InDel/SNVer/vcf

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/SNV\_InDel/SNVer/maf

Output from LoFreq:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/SNV\_InDel/LoFreq/vcf

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/SNV\_InDel/LoFreq/maf

Combined outputs from VarDict, SNVer, and LoFreq:

(Long format table has one row with one variant called by any of three callers, and can have multiple rows for the same variant if it is called by multiple callers. Wide format table reshaped Long format table, and if a variant is called by multiple callers, it takes only one row and additional information from more callers are saved in additional columns.)

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Results/SNV\_InDel/AllVariants/CalledVariants\_Long

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Results/SNV\_InDel/AllVariants/CalledVariants\_Wide

Filtered variant list:

(The combined variant list is filtered with three filtering schemes to include only potential pathogenic variants.)

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Results/SNV\_InDel/Filtered/Long

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Results/SNV\_InDel/Filtered/Medium

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Results/SNV\_InDel/Filtered/Short

* CNV:

Output from CNVkit:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/CNV/CNVkit

Output from CNACS:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/CNV/CNACS

* ITD:

Output from VarDict:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/ITD/VarDict

Output from Pindel:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/ITD/Pindel

Output from ScanITD:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/ITD/ScanITD

Output from getITD:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/ITD/getITD

IGV on FLT3 gene:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Lock/ITD/IGV

Combined outputs from VarDict, Pindel, ScanITD, getITD:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Result/ITD/Table

Zoom in IGV in the region of ITD:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Result/ITD/IGV

* QC:

Quality evaluation on the sequence data:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/QC/FASTQuick

4.2. Pipeline for filtering of SNV-InDel.

The main pipeline performs filtering of SNV-InDel to identify potential pathogenic variants with three fixed schemes - Long, Medium, Short.

In order to perform filtering with a new scheme with a new set of thresholds for filtering, please first create a new set of reference files for the new filtering scheme by running:

sh /home/projects/cu\_10184/projects/PTH/Code/Primary/FilteringScheme/NewScheme/Create\_FilteringReferences\_NewScheme.sh -f [thresh\_maf\_norm=0.05] -l [thresh\_dp\_low] -h [thresh\_dp\_high] -a [thresh\_n\_alt] [thresh\_maf\_db] [thresh\_maf\_norm] [thresh\_silhouette] [thresh\_lower\_cluster\_center]

-f: The name of the new filtering scheme. The reference files for the filtering scheme will be saved under a folder with the filtering scheme name as folder name. (default: NewScheme)

-l: The threshold for DP to decide variants as technical error due to low DP. (default: 200)

-h: The quantile threshold for DP to decide variants as technical error due to high DP. (default: 0.9975)

-t: The threshold for the minimum number of alternative reads. (default: 4)

-p: The threshold for population allele frequency obtained from public databases to decide polymorphism. (default: 0.01)

-n: The threshold for population allele frequency estimated from our normal samples to decide polymorphism. (default: 0.05)

-s: The threshold for silhouette score to decide if the two clusters can be considered as clearly separated when we decide region specific technical errors. (default: 0.8)

-c: The threshold for the center of the lower cluster to decide if the lower cluster is technical error when we decide region specific technical errors. (default: 0.08)

The reference files for the new filtering scheme are saved under:

/home/projects/cu\_10184/projects/PTH/Reference/Filtering/[NewScheme]

Then, perform filtering with the new scheme by running:

sh /home/projects/cu\_10184/projects/PTH/Code/Primary/SNV\_InDel/Filter\_NewScheme/Filter.sh -d [project\_name] -b [batch\_name] -f [FileringScheme]

-d: The project name, for example, PTH.

-b: The batch name, for example, Primary\_001

-f: The name of the filtering scheme with reference files already created.

The variants filtered by the new scheme are saved under:

/home/projects/cu\_10184/projects/PTH/BatchWork/[batch\_name]/Results/SNV\_InDel/Filtered/[NewScheme]

To perform filtering for multiple batches:

sh /home/projects/cu\_10184/projects/PTH/Code/Primary/SNV\_InDel/Filter\_NewScheme/Submit\_Jobs\_AllBatches\_Filter.sh

4.3. CNV.

To rerun only CNVkit:

sh /home/projects/cu\_10184/projects/PTH/Code/Primary/CNV/CNVkit.sh -d [project\_name] -b [batch\_name]

To run CNVkit for multiple batches

sh /home/projects/cu\_10184/projects/PTH/Code/Primary/CNV/Submit\_Jobs\_AllBatches\_CNVkit.sh

To rerun only CNACS:

sh /home/projects/cu\_10184/projects/PTH/Code/Primary/CNV/CNACS.sh -d [project\_name] -b [batch\_name]

4.4. ITD.

To rerun only ITD identification:

sh /home/projects/cu\_10184/projects/PTH/Code/Primary/ITD/Ensemble/ITD.sh -d PTH -b [batch\_name] -p [panel\_name] -t 8

4.4. QC.

To rerun FASTQuick:

To run verifyBamID:

To run Someliar: