NeoFlow2 development document

1. Overview

   There are six main modules in current version. Four of them have been encoded in NextFlow and other two ones are still in Shell.

   a. `neoflow_hlatyping.nf`
   b. `neoflow_db.nf`
   c. `neoflow_msms.nf`
   d. `neoflow_rna_variant_calling.sh`
   e. `neoflow_rna_expression.sh`
   f. `neoflow_neoantigen.nf`

2. `HLA-typing`

   In this module, Optitype is applied to predict HLA type of one sample. The `neoflow_hlatyping.nf` file in Github needs .fastq file(s) as input. But most of CPTAC projects only provide us .bam file. We could convert it to .fastq by samtools.

   samtools fastq -1 $fastq1 -2 $fastq2 $bam

   The inputs of this process are: 1) wxs reads data (fastq or bam). This file is usually in GDC which can be downloaded by UUID; 2) HLA reference database.

   The output of this process is a text file of predicated HLA types of one sample. This file will be used in neoantigen predication process.

3. `Customized database building`

   In this module, annovar is applied firstly to annotate somatic mutations. Then, customprodbj is performed to generate customized database. For labeled experiment data, such as TMT or iTraq, we should combine mutations of all samples in one experiment together and generate one database for one experiment. Thus, users need provide a mapping file.

   The inputs of this process are: 1) mapping information between samples and experiments; 2) somatic vcf files; 3) annovar installation path; 4) reference information.

   The outputs of this process are: 1) one fasta file of customized database. This file will be used in msms searching process; 2) one txt file of variant proteins and sample information.

4. `MSMS searching`

   In this module, msms searching, FDR controlling and PepQuery validation are performed. AutoRT is an optional validation.

   The inputs of this process are: 1) customized database generated in last step; 2) mgf file of one experiment.

   The output of this process is validated variant peptides.

   Please note that AutoRT validation requires GPU for computing.

5. `RNA variant calling`

   In this module, RNA-seq variant calling is performed based on GATK Best Practices Workflow.

   The inputs of this process are: 1) RNA reads data (in fastq); 2) reference data.

   The output is one text file containing all variant information (merge-varInfo.txt).

6. `RNA expression`

   In this module, RSEM is applied to get RNA level expression result.

   The input of this process is RNA reads data (in fastq).

   The output is one text file of expression result.

7. `Neoantigen prediction and combination of different evidences`

   In this module, all possible neoepitopes will be predicated by NetMHCPan. And the predication results will be mapped with results of RNA variant calling, RNA expression and proteome identification. Finally, final report is generated.

   The inputs of this process are: 1) customized database fasta file; 2) one txt file of variant proteins and sample information; 3) HLA types predication results; 4) validated variant peptides; 5) RNA variant calling results; 6) RNA expression results; 7) reference database.