User's Manual

1. Running environment

ProGeo-neo2.0 requires a Linux operation system(centos7) with Python(V3.7), Perl(V5.26) and Java(V1.7) installed.

2. External reference datasets

In order to run normally, several third-party software such as BWA, GATK, and ANNOVAR need extra databases. Here, we provided these files in the "reference_file" directory, such as "hg38.fasta". In addition, during annotating genetic variants, ANNOVAR software needs lots of databases including: refGene, ensGene, cytoband, avsnp147, dbnsfp30a, MT_ensGeneMrna, refGeneWithVerMrna, etc. of hg38, putting them into "humandb" folder for the sake of convenience.

3. Usage of ProGeo-neo2.0

Run the following codes before getting started.
cd ProGeo-neo2.0
bash start.sh
(Users with root privileges can ignore the following:)
chmod 755 software/bwa/bwa
chmod 755 software/samtools/samtools
chmod 755 software/bcftools/bcftools
pip install numpy
pip install pandas
pip install future
pip install pyomo
pip install pysam
pip install matplotlib
pip install tables

3.1. WGS/WES processing

3.1.1. Software installation and configuration

(1) BWA
cd software/bwa
make
(2) SAMtools
cd software/samtools
./configure
make && make install
(3) GATK
Download and install gatk via anaconda or miniconda:
conda config –add channels bioconda
conda install -c bioconda gatk4
(4) ANNOVAR

Download and install Annovar to ProGeo-neo2.0/software/annovar, then execute the following command:
chmod 755 software/annovar/convert2annovar.pl
chmod 755 software/annovar/table_annovar.pl
chmod 755 software/annovar/annotate_variation.pl
chmod 755 software/annovar/coding_change.pl
(5) Include bwa, samtools, bcftools, gatk, and cbc in your PATH environment variable. Add HDF5's lib directory to your LD_LBRARY_PATH.

3.1.2. Samples processing

(1) Sample files placement
Create two directories and input WGS (Whole Genome Sequencing) or WES (Whole Exome Sequencing) data files in FASTA format, including tumor samples and normal samples under those directories respectively. Test files can be downloaded by start.sh.
(2) Run scripts
python wes_mutation_peptides.py /path/to/wes-tumor /path/to/wes-normal
eg:
python wes_mutation_peptides.py test/wes/tumor test/wes/normal
(3) Get result files
Temporary files and final result files generated during data processing will be placed under **outfile1** and **outfile-wes** directories respectively.

3.2. RNA processing

3.2.1 Software installation and configuration

(1) STAR-Fusion
cd STAR-Fusion- v1.9.1
make && make install
(2) OptiType
tar -zxvf OptiType.tar.gz
In the 'OptiType' directory edit the script config.ini'.

```
[mapping]

# Absolute path to RazerS3 binary, and number of threads to use for mapping

razers3=/path/to/razers3
threads=16

[ilp]

# A Pyomo-supported ILP solver. The solver must be globally accessible in the
# environment OptiType is run, so make sure to include it in PATH.
# Note: this is NOT a path to the solver binary, but a keyword argument for
# Pyomo. Examples: glpk, cplex, cbc.

solver=cbc
threads=1
```

(3) HDF5
cd hdf5
./configure
make & make install

(4) HLAminer
Download HLAminer and be sure that the file path of HLAminer.pl is /path/to/software/HLAminer/HLAminer_1.4/bin/HLAminer.pl.
(5) Kallisto
Download and install kallisto via anaconda or miniconda:
conda install kallisto
(6) Include STAR-Fusion, hdf5, razers3, and blast in your PATH environment variable.
3.2.2 Samples processing
(1) Sample files placement
Create a directory and input the RNA-seq data sample files in FASTA format into it. Test files can be downloaded by start.sh.
(2) Run scripts
python RNA_seq_mutation.py /path/to/rna
eg:
python RNA_seq_mutation.py test/rna
notes:
HLA types are required for predicting neoantigens. You can either enter your own types or infer HLA types by using the following commands.
Input "y" if users need HLA types from RNA sequences when the system prompts: "Predicting HLA class I types from next-generation sequencing data: (y/n)?" or "Predicting HLA class II types from next-generation sequencing data: (y/n)?", otherwise input "n".
(3) Get result files
Temporary files and final result files generated during data processing will be placed under the **outfile2** and the **outfile-rna** respectively. It should be noted that the predicted HLA types will be stored in .tsv files under the outfile-rna/hla directory, contributing to subsequent neoantigen screening strategies. In addition, the synthetic mutant long peptides which will be saved in the Varsequence.fasta will be used for mass spectrometry library construction later.

3.3. Neoantigen prediction
3.3.1. Software installation and configuration
(1) Download NetMHCpan and NetMHCIIpan and include them in your PATH environment variable.
3.3.2. Neoantigen prediction and data processing
(1) Run scripts
python neoantigen_prediction.py
notes:
Input the HLA types predicted in 3.2 or other types that the user interested in when the system prompts:
"please input an HLA class I allele like 'HLA-A03:01' or multiple alleles like 'HLA-A03:01,HLA-B07:02,HLA-B35:03':" and "please input an HLA class II allele like 'DRB1_0101' or multiple alleles like 'DRB1_0102,DRB1_0301,DQB1_0101':".

(2) Get result files

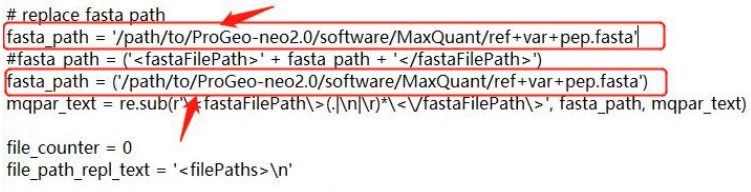
The predicted HLA class I and HLA class II candidate neoantigens will be stored in the MHC-I and MHC-II catalogs under the **outfile-candidate-neoantigens** directory, with the temporary files under the **outfile3** directory.

3.4. Mass spectrometry filtration

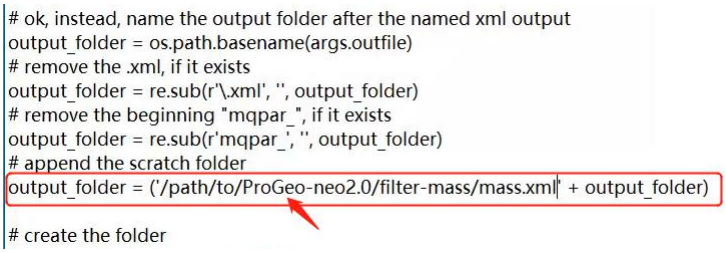
3.4.1. Software installation and configuration

(1) Edit the script software/gen_mqpar.py.

Edit lines 25 and 27:



Edit line 79:



(2) mono

cd software/mono-1.1.7.4

./configure --prxfix=path/to/software

make && make install

(2) MaxQuant

Download MaxQuant to ProGeo-neo2.0/software/MaxQuant.

Reference method:

In order to generate the customized protein sequence database, protein sequences with missense mutation sites can be generated by substituting the mutant amino acid in normal protein sequences and all mutant sequences were appended to the normal protein and cRAP fasta file. Here, we only provide mutant protein sequences (Varsequence.fasta) based on RNASeq data, users can add other reference protein sequences as needed. MaxQuant software will automatically build the inverse library.

3.4.2. Data processing

(1) Sample files placement

Create a directory and input the mass spectrometry files into it. Test files can be downloaded by start.sh.

(2) Run scripts

python MS_filtration.py /path/to/mass

eg:

python MS_filtration.py test/mass

(3) Get result files

The result peptides filtered by this step will be stored in the **filter-mass/Maxpep.txt** file and be used for further neoantigen filtration.

3.5. Neoantigen filtration

3.5.2. Data processing

(1) Run scripts

python neoantigen_filtration.py

(3) Get result files

After different levels of strict threshold filtration, the final candidate neoantigens will be saved in the MHC-I and MHC-II directories under the **outfile-filter-neoantigens** directory.

4. Required Software Downloads

Some of the third-party software needed in ProGeo-neo v2.0 has been downloaded and placed in the "software" directory, while others need to be downloaded and installed by the user. You need to make sure that each software is available and installed in the correct path. The software download path is shown in Table 1.

Table 1. Summarizes the needed software and download links

Software	Download address
Trimmomatic (v0.38) [1]	http://www.usadellab.org/cms/index.php?page=trimmomatic
FastQC (v0.11.5) [2]	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
BWA-0.7.17 [3]	http://bio-bwa.sourceforge.net/
SAMtools-1.10 [4]	https://github.com/SAMtools
GATK4.2.0.0 [5]	https://software.broadinstitute.org/gatk/download/
Annovar [6]	http://annovar.openbioinformatics.org/en/latest/user-guide/download/
OptiType-1.3.5 [7]	https://github.com/FRED-2/OptiType
HLAminer-1.4.0 [8]	https://github.com/warrenlr/HLAminer
NetMHCpan-4.1 [9]	http://www.cbs.dtu.dk/services/NetMHCpan-4.1/
NetMHCIIpan-4.0 [10]	https://services.healthtech.dtu.dk/services/NetMHCIIpan-4.0/
MaxQuant [11]	http://www.coxdocs.org/doku.php?id=MaxQuant:start
Kallisto-0.46.2 [12]	https://github.com/pachterlab/kallisto
STAR-Fusion-1.9 [13]	https://github.com/STAR-Fusion/STAR-Fusion/releases
Blast [14]	https://blast.ncbi.nlm.nih.gov/Blast.cgi

Reference:

1. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) 2014, 30, 2114-2120, doi:10.1093/bioinformatics/btu170.
2. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Reference Source. 2010.
3. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler

transform. Bioinformatics (Oxford, England) 2010, 26, 589-595, doi:10.1093/bioinformatics/btp698.
4. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England) 2011, 27, 2987-2993, doi:10.1093/bioinformatics/btr509.
5. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 2010, 20, 1297-1303, doi:10.1101/gr.107524.110.
6. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 2010, 38, e164, doi:10.1093/nar/gkq603.
7. Szolek, A.; Schubert, B.; Mohr, C.; Sturm, M.; Feldhahn, M.; Kohlbacher, O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics (Oxford, England) 2014, 30, 3310-3316, doi:10.1093/bioinformatics/btu548.
8. Warren, R.L.; Choe, G.; Freeman, D.J.; Castellarin, M.; Munro, S.; Moore, R.; Holt, R.A. Derivation of HLA types from shotgun sequence datasets. Genome medicine 2012, 4, 95, doi:10.1186/gm396.
9. Jurtz, V.; Paul, S.; Andreatta, M.; Marcatili, P.; Peters, B.; Nielsen, M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. Journal of immunology (Baltimore, Md. : 1950) 2017, 199, 3360-3368, doi:10.4049/jimmunol.1700893.
10. Karosiene, E.; Rasmussen, M.; Blicher, T.; Lund, O.; Buus, S.; Nielsen, M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics 2013, 65, 711-724, doi:10.1007/s00251-013-0720-y.
11. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nature protocols 2016, 11, 2301-2319, doi:10.1038/nprot.2016.136.
12. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology 2016, 34, 525-527, doi:10.1038/nbt.3519.
13. Haas, B.; Dobin, A.; Stransky, N.; Bo, L.; Xiao, Y.; Tickle, T.; Bankapur, A.; Ganote, C.; Doak, T.; Pochet, N. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. 2017.
14. McGinnis, S.; Madden, T.L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic acids research 2004, 32, W20-25, doi:10.1093/nar/gkh435.