README file for EPIQ version 0.1.0

This source code directory contains an implementation of EPIQ. EPIQ is a quantification algorithm for any type of isotopic labeling with nominal mass spacing, such as SILAC and *in vitro* chemical labeling schemes.

MS-GF+ PSM identification by an automated script

The PSM identification step of EPIQ is done by MS-GF+. Currently, MS-GF+ is not fully integrated to EPIQ implementation. Instead, MS-GF+ could be run automatically by utilizing a **Snakefile** under **RunMSGF** directory. To run **Snakefile**,

1. Prepare a Unix machine with 'gawk' and 'snakemake' installed. You can find the newest version of snakemake from
https://snakemake.readthedocs.io/en/stable/getting_started/installation.html

2. Set your MS-GF+ location, fasta file location, input file location, and input file extension (mzML, mgf, etc.) at the top of the **Snakefile**.

```
MSGFPATH = "/casa/yeon/bin/msgf+/MSGFPlus.jar"
LABELS = "L0 L1 L2 L3 L4 L5".split()
DB = "/casa/yeon/p/DEAQ/db/cRAP_added/hsa_UP000005640/hsa_UP000005640_cRAP_added.fasta"
INPUTFILEEXT = "mzML"
FILELOCATION = "../mzML_MSConvert/"
```

3. Make MS-GF+ modification file for each channel. Mass shift by label should be set as the fixed modification. Below examples are MS-GF+ modification files for DE-6plex channel 1 and channel 2. For DE-6plex and SILAC-6plex, you can use example modification files in **RunMSGF/DE6_modfiles** and **RunMSGF/SILAC6_modfiles**, respectively.

```
# for channel 1
NumMods=2
O1,M,opt,any,Oxidation

C2H3N1O1,C,fix,any,Carbamidomethyl

56.0626,K,fix,any,Diethyl
56.0626,*,fix,N-term,Diethyl

# for channel 2
NumMods=2
O1,M,opt,any,Oxidation

C2H3N1O1,C,fix,any,Carbamidomethyl

58.075154,K,fix,any,Diethyl
58.075154,*,fix,N-term,Diethyl
```

4. Set your modification file name form in rule run_msgf+.

```
rule run_msgfplus:
    input: '../mzML_MSConvert/{{sample}}.{}'.format(INPUTFILEEXT),
           DB.replace('.fasta', '.revCat.fasta'),
           'modfiles/Mod_StatC_IAA_DynKNterm_DE_{label}.txt',
```

The '{label}' is replaced by one of the channels, for example in 6-plex, one of the 'L0', 'L1', 'L2', 'L3', 'L4' and 'L5'. Thus, '`modfiles/Mod_StatC_IAA_DynKNterm_DE_{label}.txt`' is replaced by one of the below files under `modifiles/` directory that are made in step 3.

`Mod_StatC_IAA_DynKNterm_DE_L0.txt, Mod_StatC_IAA_DynKNterm_DE_L1.txt, Mod_StatC_IAA_DynKNterm_DE_L2.txt, Mod_StatC_IAA_DynKNterm_DE_L3.txt, Mod_StatC_IAA_DynKNterm_DE_L4.txt, and Mod_StatC_IAA_DynKNterm_DE_L5.txt`

5. Run **Snakefile** according to the snakemake manual.

MS-GF+ PSM identification by manual

If the prerequisite for Snakemake is not available, you can run MS-GF+ manually.

1. Run MS-GF+ for individual channel specifying the modification. For instance, for 6-plex system, you need to run MS-GF+ 6 times. For recommended MS-GF+ parameters, please refer to the rule_msgfplus part in **RunMSGF/Snakefile**.

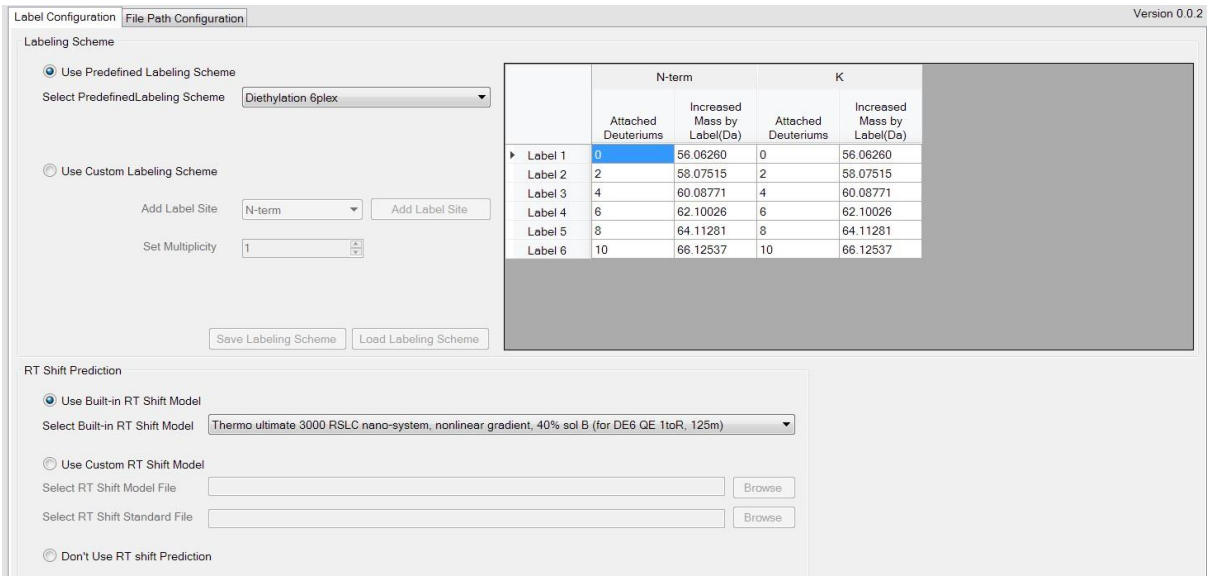
2. To run EPIQ, the MS-GF+ results should be converted into tsv file format, which are merged into a single tsv file representing identification results of a single spectrum file. To do so, convert MS-GF+ results into tsv files (please follow MS-GF+ instruction). The converted tsv files should be merged, and the tsv header should appear only once at the first line of the merged file. This can be done manually by copy-and-paste using spreadsheet software (such as Excel). If the size of tsv file is so large that the spreadsheet software cannot handle it, you can use the following shell command on Unix shell or Windows Cygwin:

```
% gawk 'FNR==1 && NR!=1{next;}{print}' CHANNEL1.tsv CHANNEL2.tsv
CHANNEL3.tsv CHANNEL4.tsv CHANNEL5.tsv CHANNEL6.tsv > MERGED.tsv
```

Running EPIQ by the graphical user interface

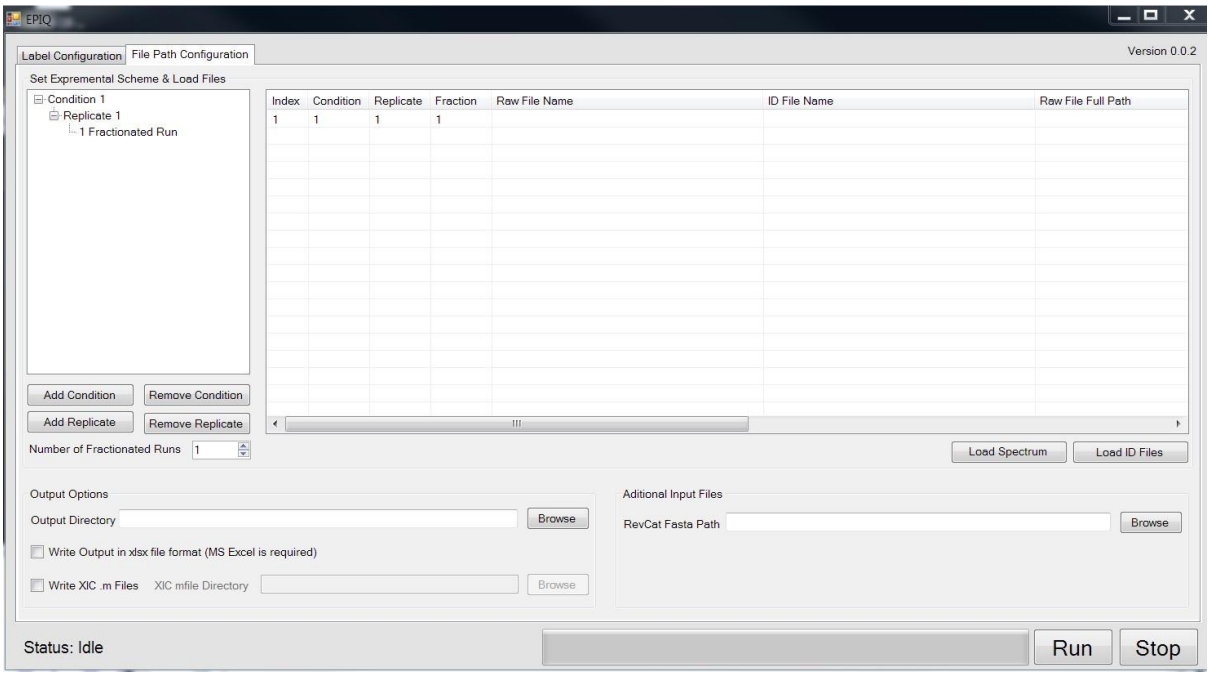
The graphical user interface of EPIQ can be found at **Executable/EPIQgui.exe**. Please aware that this executable file was only tested on x64 windows7 / windows10 machines. You can also find a video tutorial of EPIQgui.exe on **EPIQgui_movie.mp4** or youtube (https://youtu.be/_j7yXA_ZLyY).

1. Label configuration



a. Choose labeling scheme from the drop-down box. Custom Labeling Scheme option is under development.
b. Choose RT shift prediction model from the drop-down box. Custom RT shift model option is also available if you have appropriate RT shift prediction model files. The examples of these files can be found at **RtModels** directory. However, we are still developing RT shift prediction model training pipeline for end users. To turn off RT shift prediction, select Don't Use RT shift prediction option.

2. File path configuration



a. Set experimental schemes in the left upper panel. Add condition, replicate, fractionated runs to fit your design. ‘Remove condition’ and ‘Remove Replicate’ functions are under development.
b. Load spectrum files (mzML or raw) and MS-GF+ identification files (tsv) on right upper table.
c. Set output directory.
d. revCat.fasta file generated by MS-GF+ is required for EPIQ. To find this file, go the directory where you put the fasta file for MS-GF+. If you have run MS-GF+ successfully, revCat.fasta file will be generated on the same directory by MS-GF+. Please load that file on ‘Additional Input Files’ section.

3. Press ‘Run’ button and wait

The progress bar at the bottom is not working at this moment. However, you can still see the progress from the console window, which appears as a pop-up when you press the ‘Run’ button. When EPIQ finishes running, a pop-up window with a message ‘Done!’ will show up.