

# Procedures for using PepQuery

## Contents

1	Introduction.....	1
2	Required resources.....	1
2.1	For web server .....	1
2.2	For standalone version .....	2
3	Using the web server version of PepQuery .....	2
3.1	Taking a novel peptide as input .....	2
3.2	Taking a novel protein sequence as input .....	6
3.3	Taking a novel DNA sequence as input .....	7
4	Using the standalone version of PepQuery .....	8
4.1	Taking a novel peptide as input .....	10
4.2	Taking a novel protein sequence as input .....	10
4.3	Taking multiple peptide sequences as input.....	10
4.4	Taking VCF/BED/GTF as input .....	13
5	Appendix.....	14
5.1	A novel peptide.....	14
5.2	A novel protein .....	14
5.3	A novel DNA sequence.....	14

## 1 Introduction

PepQuery ([www.pepquery.org/](http://www.pepquery.org/)) is a peptide-centric search engine for novel peptide identification and validation. It is available as a standalone application as well as a web-application. Users can find the detailed procedures for using PepQuery in this document. Completing all examples will take about less than 45 minutes.

## 2 Required resources

### 2.1 For web server

**Web server address:** <http://pepquery2.pepquery.org/> .

**Input data for testing:** please find the input sequences data in [appendix](#) section.

**Support web browser:**

Windows: Google Chrome 56.0 or later, Internet Explorer 10.

Mac: Google Chrome 56.0, Safari 10.0 or later.

We strongly recommend upgrading to the latest version of the supported browsers.

## 2.2 For standalone version

**Software download**: please download the latest version of PepQuery:

[http://pepquery.org/data/PepQuery\\_v1.0.0.tar.gz](http://pepquery.org/data/PepQuery_v1.0.0.tar.gz) .

**Java version**: 1.8 or later

**OS**: Windows, Linux and Mac OS

**Hardware**: 2 CPUs (more is better), 4 Gb memory

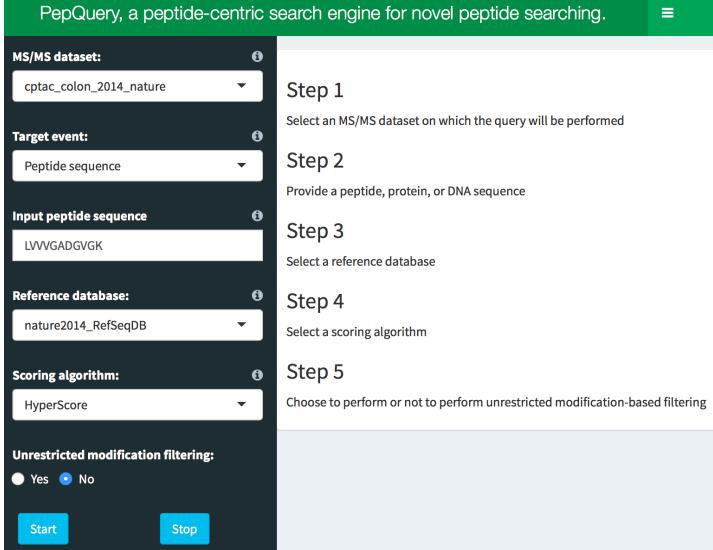
**Input data for testing**: The testing data is included in the software package

([http://pepquery.org/data/PepQuery\\_v1.0.0.tar.gz](http://pepquery.org/data/PepQuery_v1.0.0.tar.gz)).

## 3 Using the web server version of PepQuery

### 3.1 Taking a novel peptide as input

1. Open the web server: <http://pepquery2.pepquery.org/>. The web server looks like below:



The screenshot shows the PepQuery web interface. The header is green with the text "PepQuery, a peptide-centric search engine for novel peptide searching." and a menu icon. The main content area is dark grey with white text and input fields. On the left, there are five sections: "MS/MS dataset:" with a dropdown menu showing "cptac\_colon\_2014\_nature"; "Target event:" with a dropdown menu showing "Peptide sequence"; "Input peptide sequence" with a text input field containing "LWVGADGVGK"; "Reference database:" with a dropdown menu showing "nature2014\_RefSeqDB"; and "Scoring algorithm:" with a dropdown menu showing "HyperScore". Below these is a section for "Unrestricted modification filtering:" with radio buttons for "Yes" and "No", where "No" is selected. At the bottom left are "Start" and "Stop" buttons. On the right, there are five steps: "Step 1: Select an MS/MS dataset on which the query will be performed", "Step 2: Provide a peptide, protein, or DNA sequence", "Step 3: Select a reference database", "Step 4: Select a scoring algorithm", and "Step 5: Choose to perform or not to perform unrestricted modification-based filtering".

2. Set parameters. Please find detailed description of each parameter in the website of PepQuery: <http://pepquery.org/document.html> . Please set all the parameters like below:

The screenshot shows the PepQuery web interface. At the top, a green header reads "PepQuery, a peptide-centric search engine for novel peptide searching." Below this is a dark sidebar with several input fields: "MS/MS dataset:" with a dropdown menu showing "cptac\_colon\_2014\_nature"; "Target event:" with a dropdown menu showing "Peptide sequence"; "Input peptide sequence" with a text box containing "LWVGADGVGK"; "Reference database:" with a dropdown menu showing "nature2014\_RefSeqDB"; "Scoring algorithm:" with a dropdown menu showing "HyperScore"; and "Unrestricted modification filtering:" with radio buttons for "Yes" (selected) and "No". At the bottom of the sidebar are "Start" and "Stop" buttons. To the right of the sidebar, a light gray area contains a five-step process: "Step 1: Select an MS/MS dataset on which the query will be performed"; "Step 2: Provide a peptide, protein, or DNA sequence"; "Step 3: Select a reference database"; "Step 4: Select a scoring algorithm"; and "Step 5: Choose to perform or not to perform unrestricted modification-based filtering".

Please select “Yes” for unrestricted modification filtering so that unrestricted modification filtering is used.

The MS/MS dataset used in this testing is a large-scale colon tumor dataset which contains MS/MS data from 95 colon tumor samples. This MS/MS dataset contains 12,941,421 MS/MS spectra. Please go to this web page to find more information about this dataset if interested: <http://www.pepquery.org/document.html#webdataset>.

3. Click “Start” to submit the task. Analysis will be completed in about 30 seconds, and the search result will look like below:

PepQuery, a peptide-centric search engine for novel peptide searching.

**MS/MS dataset:** cptac\_colon\_2014\_nature

**Target event:** Peptide sequence

**Input peptide sequence:** LVVVGADGVGK

**Reference database:** nature2014\_RefSeqDB

**Scoring algorithm:** HyperScore

**Unrestricted modification filtering:**  Yes  No

Start Stop

### Identification overview

MS/MS searching parameters:

parameter	value
1 Enzyme	Trypsin
2 No. of missed cleavages	2
3 Fixed modifications	6
4 Variable modifications	107
5 Peptide tol. ±	20 ppm
6 MS/MS tol. ±	0.5 Da

### Identification result

The identification result is presented in the below table:

Show 10 entries Search:

	peptide	modification	n	spectrum	sample_name	charge	exp_mass	ppm	pep_mass	mz
1	LVVVGADGVGK	-	84	23981	TCGA-AG-A00Y	2	1012.587	4.748	1012.592	507.301
2	LVVVGADGVGK	-	84	30908	TCGA-AA-A01R	2	1012.595	-3.389	1012.592	507.305
3	LVVVGADGVGK	-	84	20357	TCGA-AA-A02O	2	1012.600	-7.789	1012.592	507.307
4	LVVVGADGVGK	-	84	20415	TCGA-AA-A02O	2	1012.592	-0.737	1012.592	507.303

Showing 1 to 4 of 4 entries

Download the identification result:

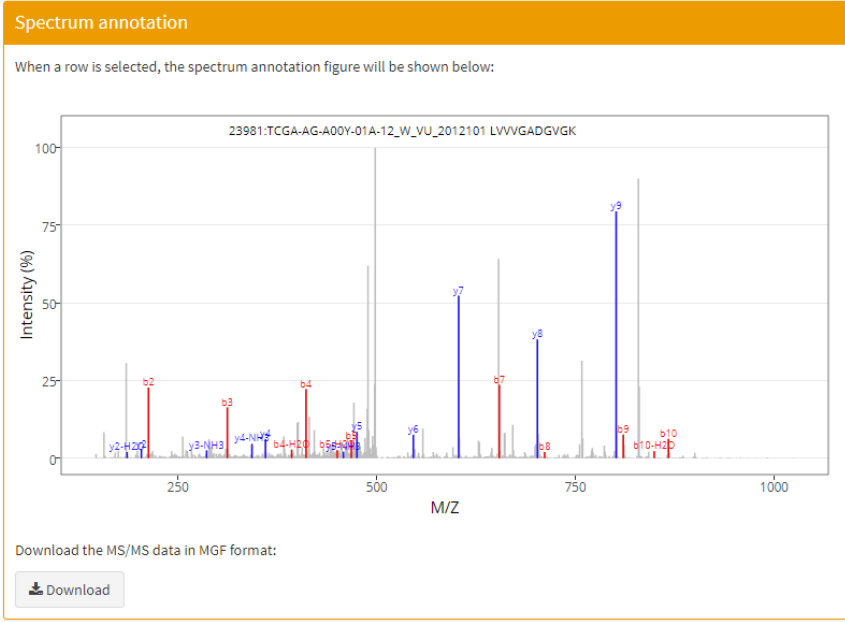
[Download](#)

Spectrum annotation

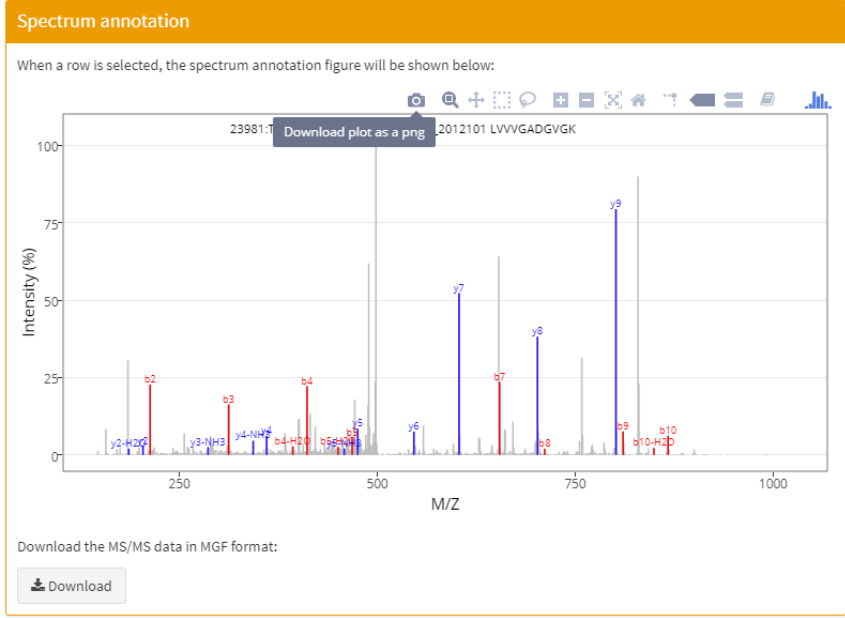
Sat Dec 9 14:13:28 2017 -> Your task is running ...

Sat Dec 9 14:13:42 2017 -> Congratulations. Your task is finished! Now you can go to the result page to see the result.

- From the result page, users will find a table listing a total of 4 spectra with confident match (p-value  $\leq 0.01$  and  $n_{\text{ptm}} = 0$ ,  $n_{\text{ptm}}$  is the number of better matched modification peptides when performing the unrestricted modification searching.) to the input novel peptide. If the value in the column of “Confident” is “Yes”, it means that the identification is confident with cutoff of p-value  $\leq 0.01$  and  $n_{\text{ptm}} = 0$ . Users can download the identification result table by clicking the “Download” button. When users click a row in the table, the annotated spectrum of this peptide-spectrum match (PSM) will be shown in the “Spectrum annotation” panel as shown below:



Click the “Download” button in this panel will download the MS/MS spectrum of this match in MGF format. Click the camera button as shown below will download the annotated spectrum in the png format:



Sample information for the identified spectra can be found in the panel of “**Sample information**” as shown below:

**Sample information**

ID	TCGA participant ID	Gender	Cancer	Histological.type	Tumor.site	Anatomic.organi	
1	TCGA-AG-A00Y-01A-12	TCGA-AG-A00Y	Male	Rectum	Rectal Adenocarcinoma	4 - rectum	Rectum

### 3.2 Taking a novel protein sequence as input

Please use the protein sequence in appendix section as an example. Because the input is a protein sequence, so users need to select the “Target event” as “Protein sequence”. **Actually, after users set the value of “Target event” as “Protein sequence”, the default sequence in the “Input protein sequence” is the testing novel protein sequence used in this document.** After setting the parameters as shown below, a click on the “Start” button will submit the job. **Please make sure that the value for parameter “Unrestricted modification filtering” is set as “Yes”.**

The screenshot displays the PepQuery web interface, a peptide-centric search engine for novel peptide searching. The interface is divided into two main sections: a dark-themed left sidebar for parameter selection and a light-themed right panel for step-by-step instructions.

**Left Sidebar (Parameters):**

- MS/MS dataset:** A dropdown menu set to "cptac\_colon\_2014\_nature".
- Target event:** A dropdown menu set to "Protein sequence".
- Input protein sequence:** A text input field containing the sequence: "MKKFFDSRRREQGSGSLGSSGGGGSTSGLGS GYIGRVFGIGRQQVTVDEVLAEGGFAIVFLVRTS".
- Reference database:** A dropdown menu set to "nature2014\_RefSeqDB".
- Scoring algorithm:** A dropdown menu set to "HyperScore".
- Unrestricted modification filtering:** Radio buttons for "Yes" (selected) and "No".
- At the bottom, there are "Start" and "Stop" buttons.

**Right Panel (Steps):**

- Step 1:** Select an MS/MS dataset on which the query will be performed.
- Step 2:** Provide a peptide, protein, or DNA sequence.
- Step 3:** Select a reference database.
- Step 4:** Select a scoring algorithm.
- Step 5:** Choose to perform or not to perform unrestricted modification-based filtering.

The job will take about 3 or 4 minutes. The result page will look like below. In total, 35 PSMs are identified, including 26 confident PSMs corresponding to 3 unique peptides (MGAQELLR, VFHELTQTDK and VENILLHDRK, highlighted in green color) and 9 PSMs that are not confident matches. The cutoff here is the same with the previous section ( $p\text{-value} \leq 0.01$  and  $n\text{-ptm} = 0$ ).

PepQuery, a peptide-centric search engine for novel peptide searching.

MS/MS dataset: cptac\_colon\_2014\_nature

Target event: Protein sequence

Input protein sequence: LHQCKTPIIHRDLKVENILLHRKVFHELTQDK MGAQELLR

Reference database: nature2014\_RefSeqDB

Scoring algorithm: HyperScore

Unrestricted modification filtering: Yes

Start Stop

Identification overview

MS/MS searching parameters:

parameter	value
1 Enzyme	Trypsin
2 No. of missed cleavages	2
3 Fixed modifications	6
4 Variable modifications	107
5 Peptide tol. ±	20 ppm
6 MS/MS tol. ±	0.5 Da

Identification result

The identification result is presented in the below table:

Show 10 entries

peptide	modification	n	spectrum	sample_name	charge	exp_mass	ppm	pep_mass
1 VFHELTQDK	-	63	23321	TCGA-AA-A000	3	1216.614	-4.000	1216.614
2 MGAQELLR	-	49	3844	TCGA-AA-A01R	2	916.484	-4.039	916.48
3 MGAQELLR	-	49	940	TCGA-AA-A00A	2	916.481	-1.508	916.48
4 MGAQELLR	-	49	987	TCGA-AA-A00A	2	916.483	-2.973	916.48
5 MGAQELLR	-	49	121545	TCGA-AF-2692	2	916.486	-6.370	916.48
6 MGAQELLR	-	49	149458	TCGA-AF-3913	2	916.486	-6.370	916.48
7 MGAQELLR	Oxidation of M@1[15.9949]	69	2711	TCGA-AA-A01R	2	916.486	-6.370	916.48
8 MGAQELLR	-	49	155142	TCGA-AG-3574	2	916.486	-6.370	916.48
9 MGAQELLR	-	49	121280	TCGA-AA-A02Y	2	916.486	-6.370	916.48
10 MGAQELLR	-	49	121339	TCGA-AA-A02Y	2	916.486	-6.370	916.48

### 3.3 Taking a novel DNA sequence as input

Please use the DNA sequence in appendix section as an example. Because the input is a novel DNA sequence, so users need to select the “Target event” as “DNA sequence”. **Actually, after users set the value of “Target event” as “DNA sequence”, the default sequence in the “Input DNA sequence” is the testing novel DNA sequence used in this document.** After setting the parameters as shown below, a click on the “Start” button will submit the job. **Please make sure that the value for parameter “Unrestricted modification filtering” is set as “Yes”.** Selecting 0 for “Frame” means translating the DNA sequence with the longest frame.

PepQuery, a peptide-centric search engine for novel peptide searching.

MS/MS dataset: cptac\_colon\_2014\_nature

Target event: DNA sequence

Input DNA sequence: GAGCCCGAGTTTGAGGCTGCAGTGAGTTGT GAGTCCAGCCTGGGTGACAGAGTGAGACTGT

Frame: 0 1 2 3 4 5 6

Reference database: nature2014\_RefSeqDB

Scoring algorithm: HyperScore

Unrestricted modification filtering: Yes

Start Stop

Step 1  
Select an MS/MS dataset on which the query will be performed

Step 2  
Provide a peptide, protein, or DNA sequence

Step 3  
Select a reference database

Step 4  
Select a scoring algorithm

Step 5  
Choose to perform or not to perform unrestricted modification-based filtering

The job will take about 2 or 3 minutes. The result page will look like below. Several matches can be found but none of them are confident matches. The cutoff here is the same with the previous section ( $p\text{-value} \leq 0.01$  and  $n_{\text{ptm}} = 0$ ).

The screenshot shows the PepQuery web interface. On the left is a control panel with the following settings:

- MS/MS dataset:** cptac\_colon\_2014\_nature
- Target event:** DNA sequence
- Input DNA sequence:** GAGCCAGGAGTTTGAGGCTGCAGTGAGTTGT, GAGTCCAGCCTGGGTGACAGAGTGAGACTG
- Frame:** 0
- Reference database:** nature2014\_RefSeqDB
- Scoring algorithm:** HyperScore
- Unrestricted modification filtering:** Yes

The main content area is divided into two sections:

### Identification overview

MS/MS searching parameters:

parameter	value
1 Enzyme	Trypsin
2 No. of missed cleavages	2
3 Fixed modifications	6
4 Variable modifications	107
5 Peptide tol. $\pm$	20 ppm
6 MS/MS tol. $\pm$	0.5 Da

### Identification result

The identification result is presented in the below table:

Show 10 entries Search:

	peptide	modification	n
1	TASCVAQAGVQWCDLGLSLQPPSEFKR	Carbamidomethylation of C@4[57.0215];Carbamidomethylation of C@14[57.0215]	32
2	TASCVAQAGVQWCDLGLSLQPPSEFKR	Carbamidomethylation of C@4[57.0215];Carbamidomethylation of C@14[57.0215]	188
3	EISHPNSEIFFFFK	-	72
4	EISHPNSEIFFFFK	-	72
5	RFCLSLPSSWDYR	Carbamidomethylation of C@3[57.0215]	163
6	RFCLSLPSSWDYR	Carbamidomethylation of C@3[57.0215]	
7	EISHPNSEIFFFFK	-	72

Showing 1 to 7 of 7 entries

Download the identification result:

## 4 Using the standalone version of PepQuery

We recommend PepQuery users to test the standalone version in Linux system. However, it should work in Windows. After download the standalone version of PepQuery from this link:

[http://www.pepquery.org/data/PepQuery\\_v1.0.0.tar.gz](http://www.pepquery.org/data/PepQuery_v1.0.0.tar.gz), please unzip it to a folder:

In Linux system, users can use the following command line to unzip:

```
tar xvfz PepQuery_v1.0.0.tar.gz
```

Then the folder of "PepQuery\_v1.0.0" looks like below:

```
|-- example/
| |-- create_proteindb_index.sh
| |-- iPRG2015/
| | |-- iPRG2015_no6.fasta
| | |-- JD_06232014_sample1-A.mgf
| | `-- peplist.txt
|-- run_input_a_peptide.sh
|-- run_input_a_protein.sh
|-- run_input_peptidelistfile.sh
|-- run_input_vcf.sh
|-- test_PGA.R
```



```

| `-- vcf
|   |-- Ref_Hsapiens_GRCh37
|   | |-- exon_anno.RData
|   | |-- hg19_refGeneCDS.fa
|   | |-- hg19_refGenePro.fa
|   | |-- ids.RData
|   | |-- procodingseq.RData
|   | |-- proseq.RData
|   | |-- splicemax.RData
|   | `-- txdb.sqlite
|   |-- TCGA-A6-3807-01A_f0106.mgf
|   `-- test.vcf
|-- lib/
|-- modification_list.txt
|-- pepquery.jar
`-- PDV/

```

The file pepquery.jar is the main program of PepQuery. The data in the example folder will be used by this testing procedures. One dataset (**iPRG2015**) is from the iPRG-2015 study (<http://pubs.acs.org/doi/abs/10.1021/acs.jproteome.6b00881>). The other one (**vcf**) is from the CPTAC study (<https://www.ncbi.nlm.nih.gov/pubmed/25043054>). The folder “PDV” contains the PDV software (<http://pdv.zhang-lab.org>) that can be used to visualize PepQuery results.

Firstly, in order to speed up the database searching, we can build a SQL-based index database for each reference protein database using the following command line (create\_proteindb\_index.sh):

```

# Change directory to the example folder
cd example/
# Please note that all the following command lines will be run in the example folder
java -Xmx2G -cp ../pepquery.jar main.java.util.createProDB -db iPRG2015/iPRG2015_no6.fasta -
fixMod 6 -varMod 107,142,143,15 -maxVar 3 -e 1 -c 2

```

But please note this step is optional. It will take about 4 minutes. The description of the above parameters is listed below:

Parameter	Description
-db	A reference protein database
-fixMod	Fixed modification. Each modification is represented by a number. Please find the complete modification list in file “modification_list.txt”. “6” is Carbamidomethylation of C.
-varMod	Variable modification. 15: Acetylation of peptide N-term; 107: Oxidation of M; 142: Deamidation of Q; 143: Deamidation of N

-maxVar	Max number of variable modifications per peptide
-e	Enzyme used for protein digestion. 1:Trypsin (default), 2:Trypsin (no P rule), 3:Arg-C, 4:Arg-C (no P rule), 5:Arg-N, 6:Glu-C, 7:Lys-C
-c	The max missed cleavages, default is 2

#### 4.1 Taking a novel peptide as input

Please find the detailed description for each parameter of PepQuery here:

<http://www.pepquery.org/document.html#saparameter> . Users can use “-pep” to specify the input peptide (run\_input\_a\_peptide.sh).

```
java -jar -Xmx2G ../pepquery.jar -db iPRG2015/iPRG2015_no6.fasta -fixMod 6 -varMod 107,142,143,15 -maxVar 3 -cpu 4 -fragmentMethod 1 -itol 0.02 -tol 10 -ms iPRG2015/JD_06232014_sample1-A.mgf -pep AVVQDPALKPLALVYGEATSR -m 1 -minScore 10 -n 1000 -um -o out
```

The above search will take about less than 10 seconds. One spectrum is identified for the input peptide with cutoff of p-value<=0.01 and n\_ptm==0. The final result is saved in the file out/psm\_rank.txt. Users can find the description for each column of this file here: <http://www.pepquery.org/document.html#saoutput>.

#### 4.2 Taking a novel protein sequence as input

Users can use “-i” and “-t” to specify a protein sequence as input (run\_input\_a\_protein.sh):

```
java -jar -Xmx2G ../pepquery.jar -db iPRG2015/iPRG2015_no6.fasta -fixMod 6 -varMod 107,142,143,15 -maxVar 3 -cpu 4 -fragmentMethod 1 -itol 0.02 -tol 10 -ms iPRG2015/JD_06232014_sample1-A.mgf -m 1 -minScore 10 -n 1000 -um -o out -i MSHHWGYGKHNGPEHWHKDFPIANGERQSPVDIDTKAVVQDPALKPLALVYGEATSRMVNNGHSFNV EYDDSQDKAVLKDGPLTGTYRLVQFHFHWGSSDDQGSEHTVDRKKYAAELHLVHWNTKYGDFGTAAQQP DGLAVVGVFLKVG DANPALQKVLDA LDISIKTKGKSTDFPNFDPG SLLPNVLDYWTYPGSLTTPPLLESVTWIV LKEPISVSSQ QMLKFRTLNFNAEGEPELLMLANWRPAQPLKNRQVRGFPK -t 1
```

The above search will take about 10 seconds. The final result is saved in the file out/psm\_rank.txt. In total, 7 PSMs for 6 unique peptides are identified with cutoff of p-value<=0.01 and n\_ptm==0. Users can also take a DNA sequence as input.

#### 4.3 Taking multiple peptide sequences as input

If users have multiple novel peptides, a file with multiple peptides can be used for parameter “-pep”. Below is an example (run\_input\_peptidelistfile.sh):

```
java -jar -Xmx2G ../pepquery.jar -db iPRG2015/iPRG2015_no6.fasta -fixMod 6 -varMod 107,142,143,15 -maxVar 3 -cpu 4 -fragmentMethod 1 -itol 0.02 -tol 10 -ms
```

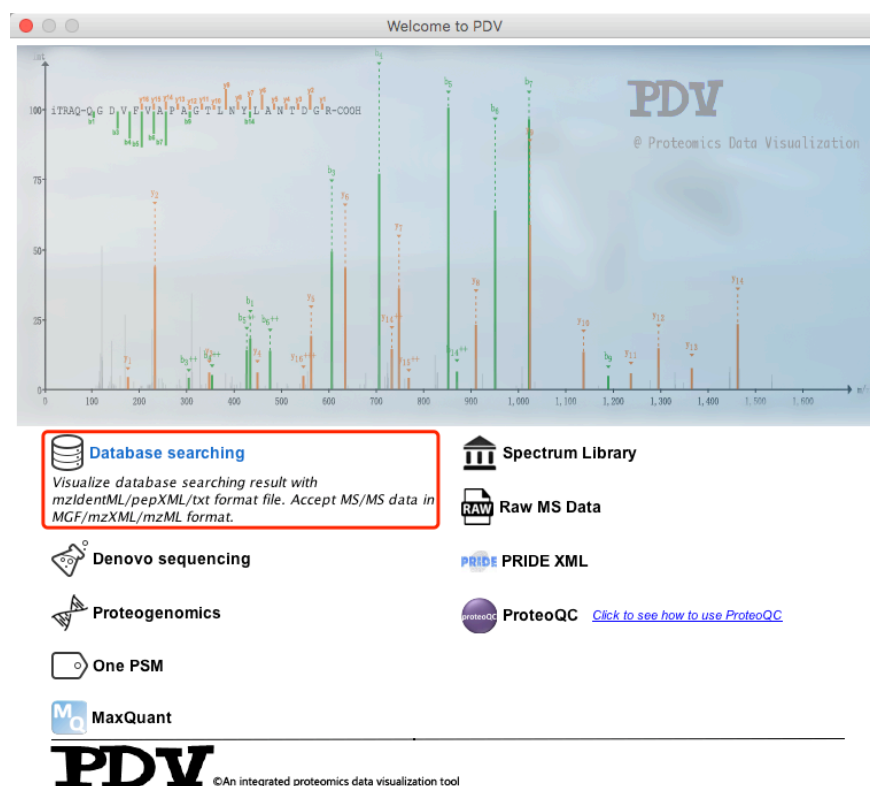
iPRG2015/JD\_06232014\_sample1-A.mgf -pep iPRG2015/peplist.txt -m 1 -minScore 10 -n 1000 -um -o out

In the input file “iPRG2015/peplist.txt”, it contains 4 peptides which do not exist in the reference protein database (iPRG2015/iPRG2015\_no6.fasta). Please note each row corresponds to one peptide. The content of file “iPRG2015/peplist.txt” looks like below:

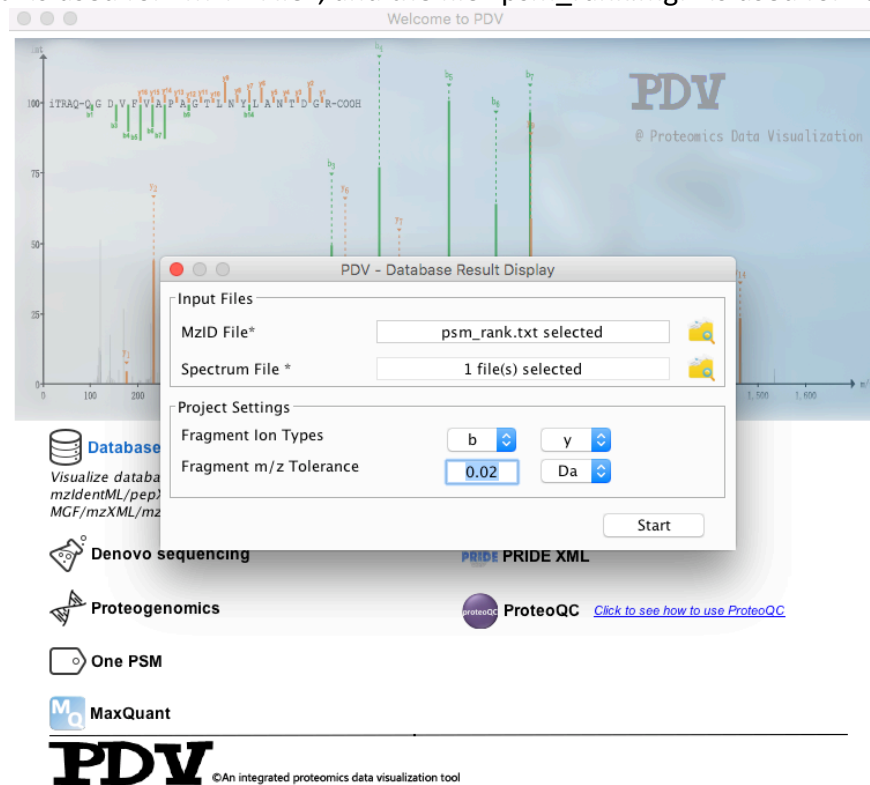
```
LLSYVDDEAFIRDVAK
AVVQDPALKPLALVYGEATSR
DFPIANGER
VG DANPALQK
```

The above search will take about 10 seconds. The final result is saved in the file out/psm\_rank.txt, including 5 spectra that are confidently matched to the 4 peptides with cutoff of p-value<=0.01 and n\_ptm==0. Please note that the result file (psm\_rank.txt) can be visualized by PDV (<http://pdv.zhang-lab.org>) using the following step by step procedure:

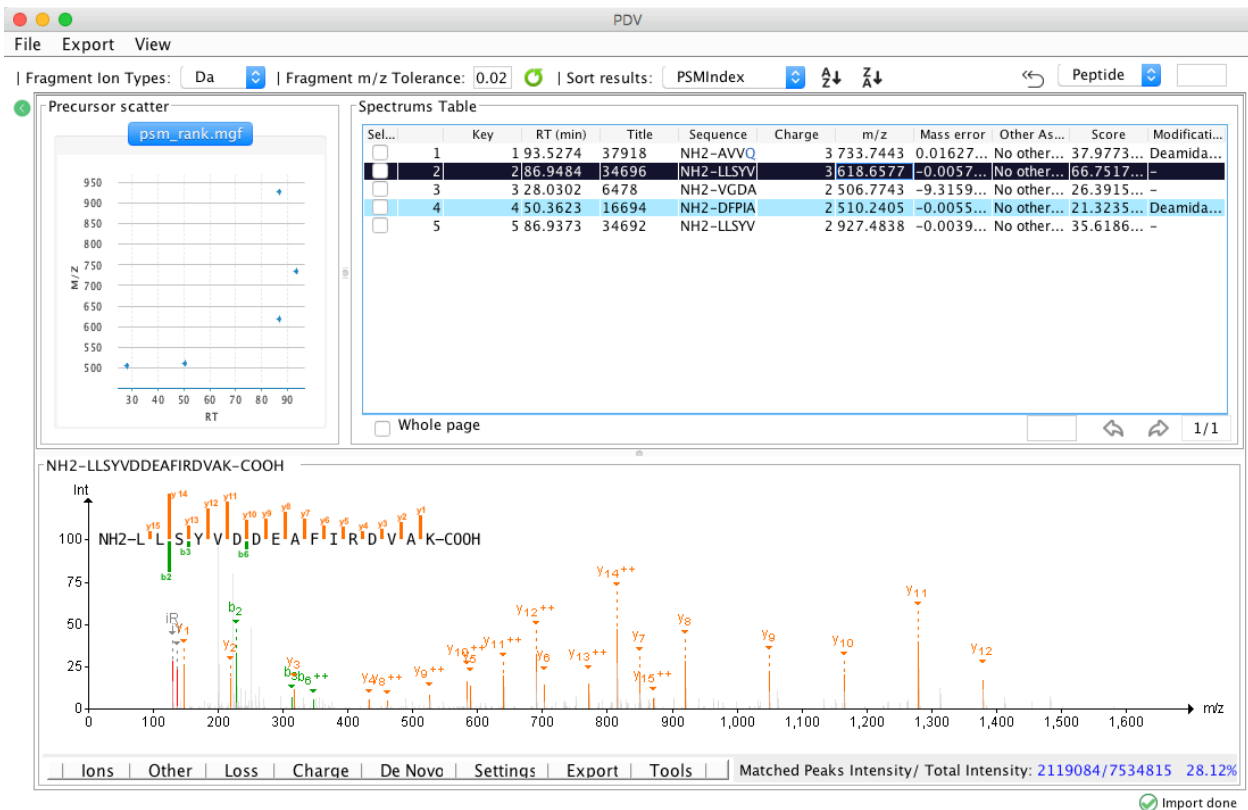
(1). Please launch PDV by double clicking the file “PDV-1.2.beta.jar” in the “PDV” folder. The main interface is shown below:



(2). Click “Database searching” in the above panel, then import data and set parameters as shown below. For visualization, only two files from the PepQuery output are required. The file “psm\_rank.txt” is used for “MzID File”, and the file “psm\_rank.mgf” is used for “Spectrum File”.



(3). Click on the “Start” button in the above panel. After about 1 or 2 seconds, users will see the visualization panel as shown below:



#### 4.4 Taking VCF/BED/GTF as input

Processing VCF/BED/GTF files is complicated. Currently, PepQuery relies on PGA (<http://bioconductor.org/packages/PGA/>) to translate the genomics information in VCF/BED/GTF to protein sequences. If users want to take VCF, BED or GTF as input, please install R (<https://www.r-project.org>) and the PGA package and the BSgenome.Hsapiens.UCSC.hg19 package (<http://bioconductor.org/packages/BSgenome.Hsapiens.UCSC.hg19/>). Please make sure that the R environment in users' computer is compatible with PGA. Then users need to prepare annotation data according to the instruction in the manual of PGA (<http://bioconductor.org/packages/devel/bioc/vignettes/PGA/inst/doc/PGA.pdf>). Users need to set parameters "-i", "-anno" and "-t" when they want to take VCF, BED or GTF as input.

After R, PGA and BSgenome.Hsapiens.UCSC.hg19 are installed in users' computer, please make sure that R has been added in the environment path. Please run the script test\_PGA.R in example folder to make sure PGA works well in users' computer:

Rscript test\_PGA.R

If the following three files are generated in the "out" folder, then it indicates PGA works well:

pga\_snv.fasta  
 pga\_snv.tab  
 pga\_txFinder.fasta

Below is an example to run PepQuery with taking VCF file as input (run\_input\_peptidelistfile.sh):

```
java -jar -Xmx2G ../pepquery.jar -i vcf/test.vcf -anno vcf/Ref_Hsapiens_GRCh37/ -t 3 -cpu 4 -db vcf/Ref_Hsapiens_GRCh37/hg19_refGenePro.fa -varMod 107 -fixMod 6 -tol 10 -itol 0.6 -o out -m 1 -ms vcf/TCGA-A6-3807-01A_f0106.mgf
```

The above search will take about 1 minutes. The final result is saved in the file out/psm\_rank.txt, there is no confident match with cutoff of p-value  $\leq 0.01$ .

## 5 Appendix

### 5.1 A novel peptide

LVVVGADGVGK

### 5.2 A novel protein

MKKFFDSRREQGSGSLGSGSSGGGGSTSLGSGYIGRVFGIGRQQVTVDEVLAEGGFVFLVRTSNGMKC  
ALKRMFVNNEHDLQVCKREIQIMRDLGSHKNIVGYIDSSINNVSSGDVWEVLILMDFCRGGQVNLMLNQR  
LQTGFTENEVLQIFCDTCEAVARLHQCKPIIHRDLKVENILLHDRKVFHELTQTDKMGAEQLLR

### 5.3 A novel DNA sequence

GAGCCAGGAGTTTGAGGCTGCAGTGAGTTGTGAGTCCAGCCTGGGTGACAGAGTGAGACACTGTCTC  
CAAAAATATAATAATAAAAATAAAGCCTGTATCTCAATGGGAAGATTTCTGCTAAGGCAGTTCAACTCAC  
TGAAAGAAGACTGCTGCATTAGGTTTCAACTTTAATGCTTTCTGTTACTTCTAGTAAAGGTTAAGTGATTTT  
TCCATTAGCTGTAGAAGTTTGGAGAGCTATTTACCAAGCCAGATCAATGATTTAAAAATTATTGGAAATT  
CATCTAAGAATCAAGTCTGAATGCCAATTCTATTGCGGTTGATTAGGTGTGATATTCTTTAAAGTTCAG  
GAATATTGGCAGTAAAAAATGAGCAGCTACTTTTCAATACTTTGTCCTTTTTTGGTGGTCTCTGCCTATTT  
CAATGTCTCTGATCAAAGATAAATAATTGGCACTGTGCCAAGGTTTGGTTTTCCAATAAGGTTTCAAC  
TGTGCCAGAAACCTATGTCCTTCACTTTGGTGGATGCTAAATGTTATTCTAAGAATATGCTTTTTCCCAAT  
TCTCCTTCTGATTTTTATGTATTAGTGGATGCAAATTGTCTTTCTAGTTGAATGAATAATTTTCGGCTAAT  
GCACGTGGAACCTTGCACCCAGATTCTTCCCATGGTCATTATCAAGTGAAGCCCTCAAAAACATGAGCG  
AAGAGCCTAGAAATACTCAGGGAGATTTCTCACCCCACTCAGAAATTTTTTTTTTTTTTTTCAAGACGGC  
GTCTTGCTCTGTCGCTCAGGCGGGAGTGCAGTGGTGTGATCTCGGCTCACTGCAACCTCCATTTTCCGAG  
TTCAAGCGATTCTGCCTCAGCCTCCCGAGTAGCTGGGATTATAGGCACACACCACCGCGCCTGGCTAATT  
TTTGATTCTTAGTAGAGATGGGGTTTACCATGTTGGCCAG