

SMAP: A Pipeline for Sample Matching in Proteogenomics

Version 1.0.0

March 2021

Contents

1. Introduction	3
1.1 Software requirement.....	3
1.2 Contact information	3
1.3 License	3
2. How to run SMAP (standard alone version)	4
2.1 Download the pipeline.....	4
2.2 Run SMAP program	4
2.3 Input data	4
2.3.1 A variant peptide table.....	4
2.3.2 A genotype in VCF format	5
2.3.3 Output files	5
2.4 Cloud-based SMAP	7
2.4.1 Introduction.....	7
2.4.1 Input data.....	7
2.4.2 Data Summary & Filtering.....	8
2.4.2.1 summary of variant peptide data	8
2.4.3.3 Distributions and filtering	9
2.4.3 sample cluster.....	11
2.4.4 summary of Genotype data	12
2.4.5 Filtering of Genotype data	13
2.4.6 Allele frequency	15
2.4.7 Run SMAP.....	15
3.References.....	17

1. Introduction

SMAP is a pipeline designed for verifying and correcting sample identity for a large mass spectrometry (MS)-based proteomics project. SMAP takes a variant peptide data that can be generated using the proteogenomics approach. The program then infers allelic information for each sample based on its expression level of the variant peptides. The program finally aligns the MS-based proteomic samples with genomic information (i.e., genotypic data) by using two discriminant scores.

1.1 Software requirement

SMAP has both standard alone and cloud-based versions. The standard alone version supports all 64-bit operating systems. The program is written by a combination of Perl and R. The minimum required Perl version should be Perl 5.6 or R 3.1.0.

1.2 Contact information

For any questions, please contact Xusheng Wang (xusheng.wang@und.edu)

1.3 License

This program is free software. You can redistribute and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

2. How to run SMAP (standard alone version)

2.1 Download the pipeline

The pipeline could be downloaded from <https://github.com/XWangLab/SMAP>

2.2 Run SMAP program

After installing SMAP program, you can run the program using the following command.

```
perl SMAP.pl -vf variant_peptide_table[file] -g genotype[file] -o result[file]
```

<code>--variant_peptide, -vf</code>	(A file containing quantitative values of variant peptides; required)
<code>--genotype, -g</code>	(A genotype file used sample verification; required)
<code>--output, -o</code>	(An output filename; required)
<code>--plex, -p</code>	(Multiplex number of the isobaric labeling approach)
<code>--fold_change, -fc</code>	(Signal to Noise ratio (optional; default is 3))
<code>--noise_level, -nl</code>	(The upper threshold of a noise level)
<code>--version, -h</code>	(Print version)
<code>--help, -h</code>	(Print help)
<code>--licence, -l</code>	(Print licencejump -s (search))

2.3 Input data

2.3.1 A variant peptide table

The variant peptide table uses the following format:

Column 1: Peptide ID

Column 2: Gene/Protein

Column 3: Peptide Spectrum Match (PSM)

Column 4: SNP ID **MUST MATCH GENOTYPE SNP ID

Column 5-N: Sample Peptide Quantification (One column per sample)

An example of the variant peptide table

Peptide	Gene	PSM	SNP	2015-1341	...	2016-965	Internal standard	group
VSNEEKVR	CAPZA1	b20_f39.15855.1.3	chr1:113162494:G:A	53788.04	...	83146.90	46477.36	nonzero
HWQQFYFLSTR	FBXO2	b20_f36.35042.1.3	chr1:11710561:T:G	25447.82	...	15590.47	19626.55	nonzero
SIEDLLR	PDE4DIP	b20_f22.28382.1.2	chr1:144877111:G:T	13161.86	...	10127.43	8410.05	nonzero

2.3.2 A genotype in VCF format

SMAP also takes a genotype in VCF format.

An example of the genotype data

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	2014-2194	2014-2195	2014-2196
1	949608	chr1:949608:G:A	G	A	.	.	PR	GT	0/1	0/1	0/1
1	2441358	chr1:2441358:T:C	T	C	.	.	PR	GT	0/0	0/0	0/0
10	115644040	chr10:115644040:G:A	G	A	.	.	.	GT	0/1	0/0	0/1

2.3.3 Output files

SMAP generates a final report and several intermediate results.

The final report contains four columns, including Sample ID, Inferred ID, CSore and DeltaCScore.

An example of the final report

Sample ID	Inferred ID	CSore	DeltaCScore
2015-1341	2015-1341	4.22	0.70
2015-737	2015-737	4.03	0.56
2015-804	2015-804	3.70	0.59
2015-42	2015-37	3.14	0.51
2015-1555	2015-1555	2.91	0.54
2015-244	2015-244	2.62	0.44
2015-735	2015-735	2.53	0.43
2014-2200	2015-857	2.52	0.48
2016-958	2016-958	1.39	0.03
2016-965	2016-965	1.27	0.03
Internal standard	2015-1339	1.71	0.00

In addition, the program also generates three intermediate files, including sample-specific genotypes and inferred genotypes.

An example of sample-specific genotype

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	2014-2194	2014-2195	2014-2196
1	949608	chr1:949608:G:A	G	A	.	.	PR	GT	H	H	H
1	2441358	chr1:2441358:T:C	T	C	.	.	PR	GT	T	T	T
10	115644040	chr10:115644040:G:A	G	A	.	.	.	GT	C	H	H

An example of inferred genotypes

SNP	2015-1341	2015-737	2015-804	2015-42	2015-1555	2015-244	2015-735	2014-2200	2016-958	2016-965	Internal standard
chr11:75298468:A:C	A	C	A	A	A	A	A	A	A	A	A
chr5:140503474:C:G	H	H	H	H	H	C	H	H	C	C	C
chr19:40408821:C:G	C	C	C	C	C	H	G	C	C	C	C

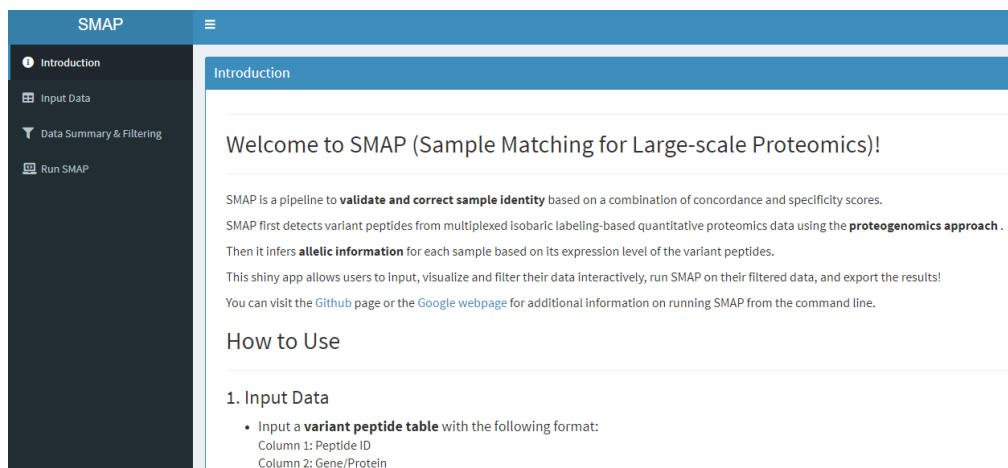
2.4 Cloud-based SMAP

The cloud-based SMAP is built with R shiny. It can be found:

<https://smap.shinyapps.io/smap/>

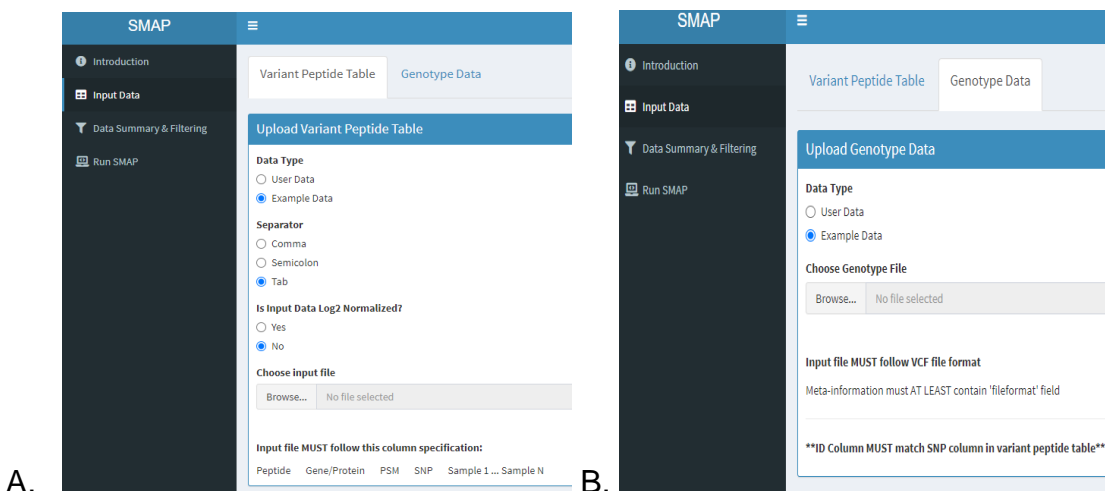
Once the website is loaded, the main page of SMAP is shown as below:

2.4.1 Introduction



2.4.1 Input data

User can upload data using “Browse” buttons in variant peptide table and genotype data menus. The format of both files can be found in the section 2.3.



For the variant peptide table, the data will be converted into log2 scale if the data is not log2 transformed.

Log2 transformed data (Preview)

Log2 Normalized Data Preview:

Show 10 entries

Peptide	Gene	PSM	SNP	X2015.1341	X2015.737	X2015.804
1 LSELEVANK	SASS6	b20_f10.31266.1.2	1:100575933:G:A	15.3598181800299	14.3807207917533	15.3875434867934
2 LSELEVANK	SASS6	b20_f10.31266.1.2	1:100575933:G:A	17.1630866029913	17.193266324361	17.3299292727448
3 LSELEVANK	SASS6	b20_f10.31266.1.2	1:100575933:G:A	19.662040004271	19.8150297810864	19.6405019018649
4 HLLNSATDPFNR	UBE4B	b20_f27.25011.1.3	1:10239569:C:G	14.2290380836841	14.192925396187	14.2412994647833
5 HLLNSATDPFNR	UBE4B	b20_f27.25011.1.3	1:10239569:C:G	13.0461088618742	13.3305607808431	13.3391705183224
6 HLLNSATDPFNR	UBE4B	b20_f27.25011.1.3	1:10239569:C:G	14.5386248521218	14.641089876631	14.7385620303866
7 VAFLEPAGPGDQNGK	AMIGO1	b20_f23.35257.1.3	1:110050180:C:T	15.22175023083	14.8661062667659	14.7171462380352
8 VAFLEPAGPGDQNGK	AMIGO1	b20_f23.35257.1.3	1:110050180:C:T	13.9707404361759	14.1844042830958	13.8717050633081
9 AVEEELDTEDRPAWNSK	SLC6A17	b20_f21.41028.1.3	1:110709720:G:A	16.5326101675361	16.3405895284774	15.795113614519
10 AVEEELDTEDRPAWNSK	SLC6A17	b20_f21.41028.1.3	1:110709720:G:A	13.1868087950015	13.0554820803755	12.7705691277804

Showing 1 to 10 of 100 entries

2.4.2 Data Summary & Filtering

2.4.2.1 summary of variant peptide data

1. Select Data Summary & Filter tab on left.
2. Select Variant peptide table then select Summary, set the number of groups in your dataset. Default is 30, then the window will expand to show intensity distribution for all peptides, other parameters will be listed in the left, such as number of rows, number of peptides, number of genes, number of PSMs, and number of SNP.

Variant peptide data summary (Preview)

Variant Peptide Table

Genotype Data

Summary

Distributions and Filtering

Sample Cluster

Variant Peptide Data Summary

Total Rows: 1340

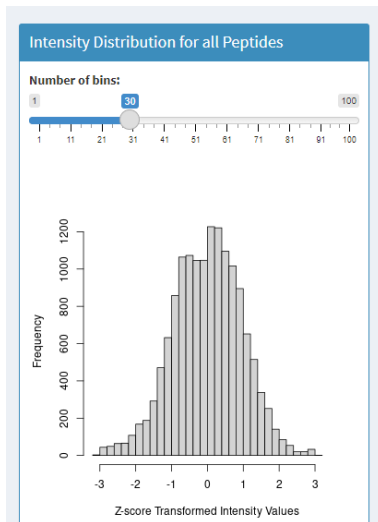
Peptide #: 670

Gene #: 605

PSM #: 670

SNP #: 670

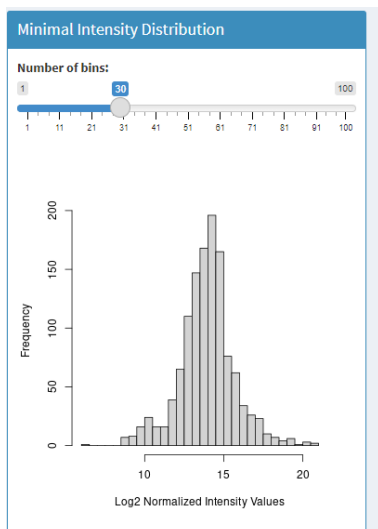
Intensity distribution for all peptides (Preview)



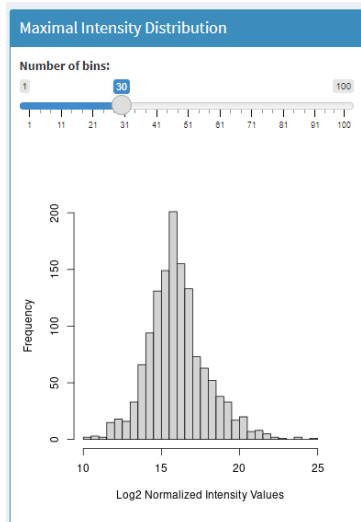
2.4.3.3 Distributions and filtering

The parameters are included the minimal expression value in variant peptide (default is 30); the maximal expression value in variant peptide (default is 30); and the ration between Maximal and minimal values (default is 30).

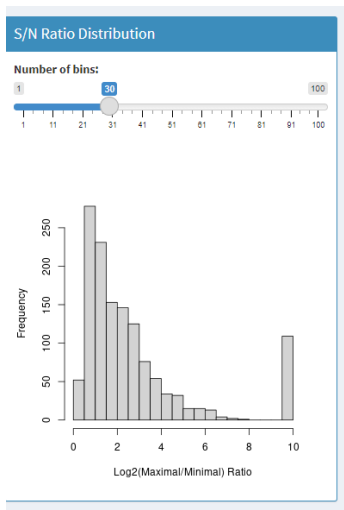
Minimal intensity distribution (Preview)



Maximal intensity distribution (Preview)



S/N ratio distribution (Preview)



Variant peptide data filtering

User could modify the parameters, such as the minimal expression value in variant , the maximal expression value in variant peptide; and the ration between Maximal and minimal values. Then download the filtered data.

Filtering parameters and summary (Preview)

Variant Peptide Data Filtering

Minimal Intensity Distribution (Intensity of reference peptides)

☐ No Filtering

☒ Input Filter:

Default value (if left blank above) is 17.4495 (Mean + 2SD)

Maximal Intensity Distribution

☐ No Filtering

☒ Input Filter:

Default value (if left blank above) is 12.254 (Mean - 2SD)

S/N Ratio Distribution

☐ No Filtering

☒ Input Filter:

Default value (if left blank above) is 3

Filter Summary

Minimal Filter Value:

17.4494601218967

Maximal Filter Value:

12.2538735669953

S/N Ratio Filter Value:

3

Total Rows After Filtering:

337

Unique Peptides After Filtering:

158

Save Filtered Output

Enter name for download file:

Filtered_Variant_Pepti

Separator

☐ Comma

☒ Tab

Download

Filtered data table (Preview)

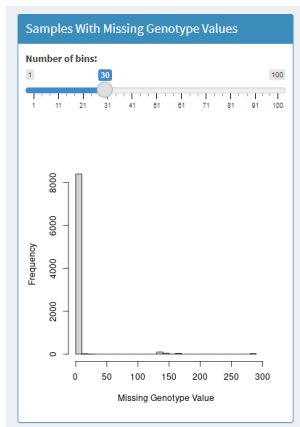
Filtered Data Table (First 100 Rows)							
Show <div>10</div> entries							
	Peptide	Gene	PSM	SNP	X2015.1341	X2015.737	X2015.804
2	LSELEVANK	SASS6	b20_f10.31266.1.2	1:100575933:G:A	17.1630866029913	17.193266324361	17.3299292727448
3	LSELEVANK	SASS6	b20_f10.31266.1.2	1:100575933:G:A	19.662040004271	19.8150297810864	19.6405019018649
9	AVEEELDTEDRPAWNSK	SLC6A17	b20_f21.41028.1.3	1:110709720:G:A	16.5326101675361	16.3405895284774	15.795113614519
11	AVEEELDTEDRPAWNSK	SLC6A17	b20_f21.41028.1.3	1:110709720:G:A	17.1993997914999	16.9417161448804	16.0950718399063
18	DTEGGPKKEESPV	SLC16A1	b20_f06.9552.1.2	1:113456546:A:T	16.5681853838081	13.6554824238151	16.3063000712424
25	GPGAEGSGSGSPEK	MINDY1	b20_f21.3392.1.2	1:150970577:G:T	15.0716211541836	15.7246334865449	14.8391156841471
26	GPGAEGSGSGSPEK	MINDY1	b20_f21.3392.1.2	1:150970577:G:T	16.641243025577	16.0170426108615	16.1441884062865
31	LGEHLDPSPR	NB2	b20_f01.7243.1.2	1:154541971:T:G	15.0104504793921	12.9871170700085	15.5233078448957
37	LGEHLDPSPR	NB2	b20_f01.7243.1.2	1:154541971:T:G	15.8254107545848	13.8947268342845	15.971939782644
44	VNEAYGFR	BCAN	b20_f01.19706.1.2	1:156616814:C:G	13.1989913203918	13.345712937097	13.4017328495226
Showing 1 to 10 of 100 entries							

2.4.3 sample cluster

1. Select Data Summary & Filter tab on left.
2. Select Variant peptide table then select Sample Cluster. Then the window will explore PCA plot for the test samples.

PCA plot (Preview)

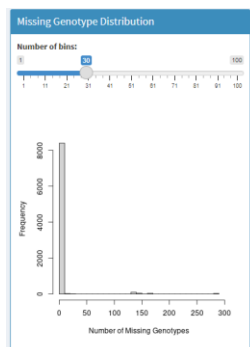
Samples with missing genotype values (Preview)



2.4.5 Filtering of Genotype data

1. Select Data Summary & Filter tab on left.
2. Select Genotype Data then select Filtering, input the number of missing genotypes. Then the window will explore the filtered VCF data and the number of total SNPs after filtering and number of SNPs with at least one missing genotype after filtering.

Missing genotype distribution (Preview)



Genotype filtering and summary (Preview)

Genotype Data Filtering

Filter Out SNPs With Missing Genotype Values

☒ No Filter

☐ Filter

Filter Out SNPs That Do NOT Appear in Filtered Variant Peptide Table

☒ No Filter/Input VCF is Already Filtered

☐ Filter

***This filter will be applied when SMAP is run, but can be set here to see/download filtered vcf*

Filtering summary


Total SNPs After Filtering:
8651

Total SNPs with At Least One Missing Genotype After Filtering: 388

Save Filtered Output

Enter name for download file:

Separator
☐ Comma
☒ Tab

 Download

Filtered VCF preview (Preview)

Filtered VCF Preview (First 100 lines)

Show 10 entries

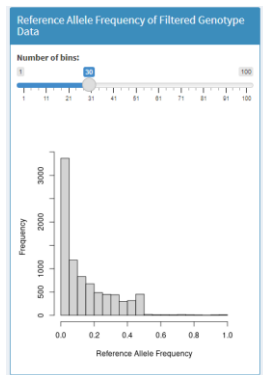
	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	X2014.2194	X2014.2195	X2014.2196
1	1	13273	1:13273;G:C	G	C	.	.	PR	GT	0/0	0/1	0/0
2	1	14599	1:14599;T:A	T	A	.	.	PR	GT	0/0	0/0	1/1
3	1	14604	1:14604;A:G	A	G	.	.	PR	GT	0/0	0/0	1/1
4	1	47159	1:47159;T:C	T	C	.	.	PR	GT	0/0	0/0	0/0
5	1	49298	1:49298;C:T	C	T	.	.	PR	GT	0/0	0/0	0/0
6	1	49554	1:49554;A:G	A	G	.	.	PR	GT	0/0	0/0	0/0
7	1	52238	1:52238;G:T	G	T	.	.	PR	GT	0/0	0/0	0/0
8	1	52253	1:52253;C:G	C	G	.	.	PR	GT	0/0	0/0	0/0
9	1	54490	1:54490;G:A	G	A	.	.	PR	GT	0/0	0/1	0/0
10	1	58814	1:58814;G:A	G	A	.	.	PR	GT	0/0	0/0	0/0

<

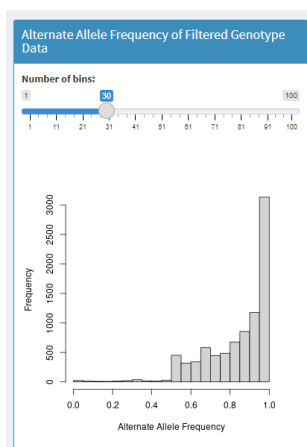
Showing 1 to 10 of 100 entries

2.4.6 Allele frequency

Reference allele frequency of filtered genotype data (Preview)

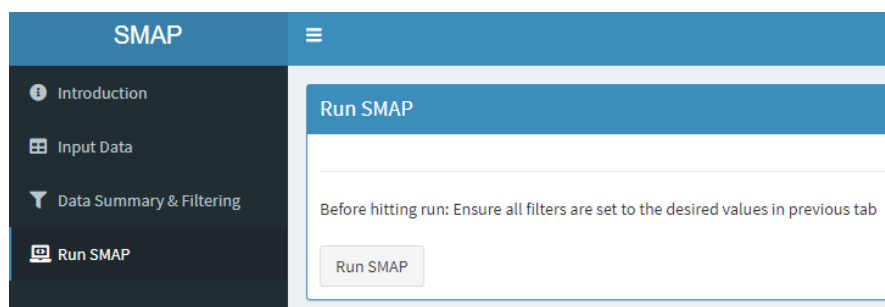


Alternate allele frequency of filtered genotype data (Preview)

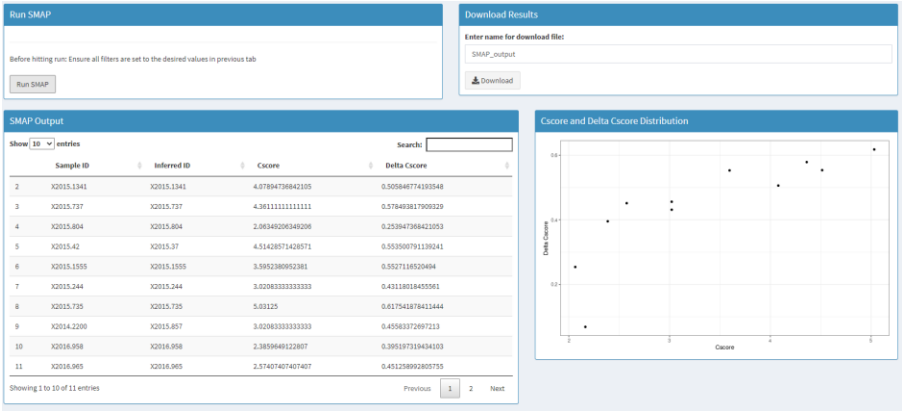


2.4.7 Run SMAP

1. Select Run SMAP on the left.
2. Note: before hitting run, ensure all filters are set to the desired values in previous tab.



After running SMAP, user could find the score table in the left and a plot in the right to distribute all of scores. User could download the score table using the tab in the upright.



3. References

1. Junmin Peng, J.E.E., Carson C Thoreen, Larry J Licklider, Steven P Gygi.(2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis the yeast proteome.pdf>. J Proteome Res *2*, 43-50.
2. Li, Y., Wang, X., Cho, J.H., Shaw, T.I., Wu, Z., Bai, B., Wang, H.,Zhou, S., Beach, T.G., Wu, G.*, et al.* (2016). JUMPg: An Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells. J Proteome Res *15*, 2309-2320.
3. UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic acids research *49*, D480-D489.
4. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.Nucleic acids research *38*, e164.
5. Wang, X., Li, Y., Wu, Z., Wang, H., Tan, H., and Peng, J. (2014). JUMP:a tag-based database search tool for peptide identification with high sensitivity and accuracy. Mol Cell Proteomics *13*, 3663-3673.