

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
3 Cloud-based SMAP	7
3.1 Introduction	7
3.2 Input data	7
3.3 Data Summary & Filtering	8
3.4 Run SMAP	10
4 References.....	11

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/UND-Wanglab/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
VSNEEKVR	CAPZA1	b20_f39.15855.1.3	chr1:113162494:G:A	53788.04	...	83146.90	46477.36
HWQQFYFLSTR	FBXO2	b20_f36.35042.1.3	chr1:11710561:T:G	25447.82	...	15590.47	19626.55
SIEDLLR	PDE4DIP	b20_f22.28382.1.2	chr1:144877111:G:T	13161.86	...	10127.43	8410.05

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

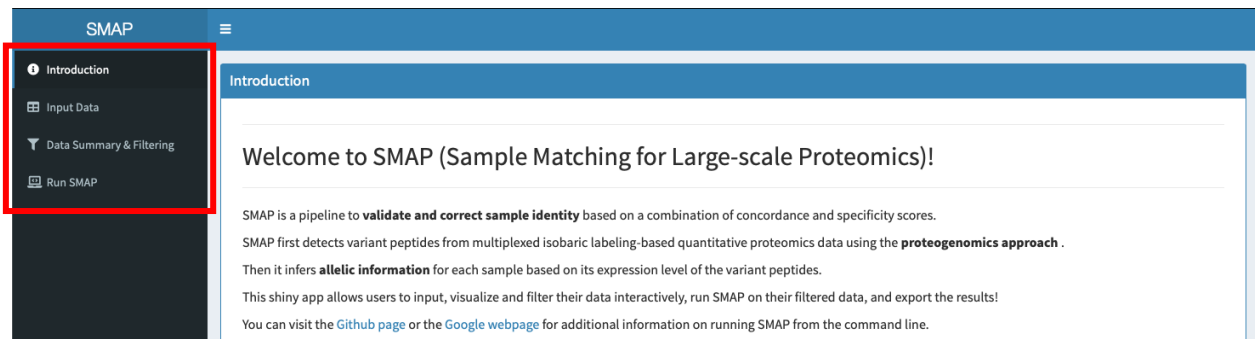
3 Cloud-based SMAP

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

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

3.1 Introduction

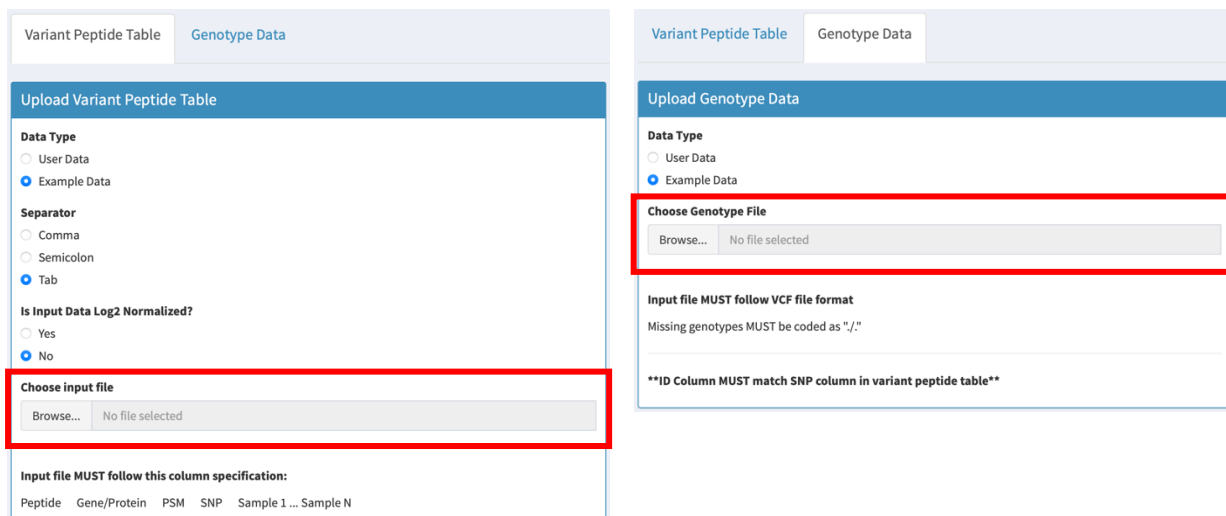
Navigation through the webpage is done by clicking on any of the four tabs at the left.



3.2 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.

- The cloud based SMAP application accepts .vcf files with any length of meta-information (including none).
- Variant peptide data will be converted into log2 scale if it is not already.

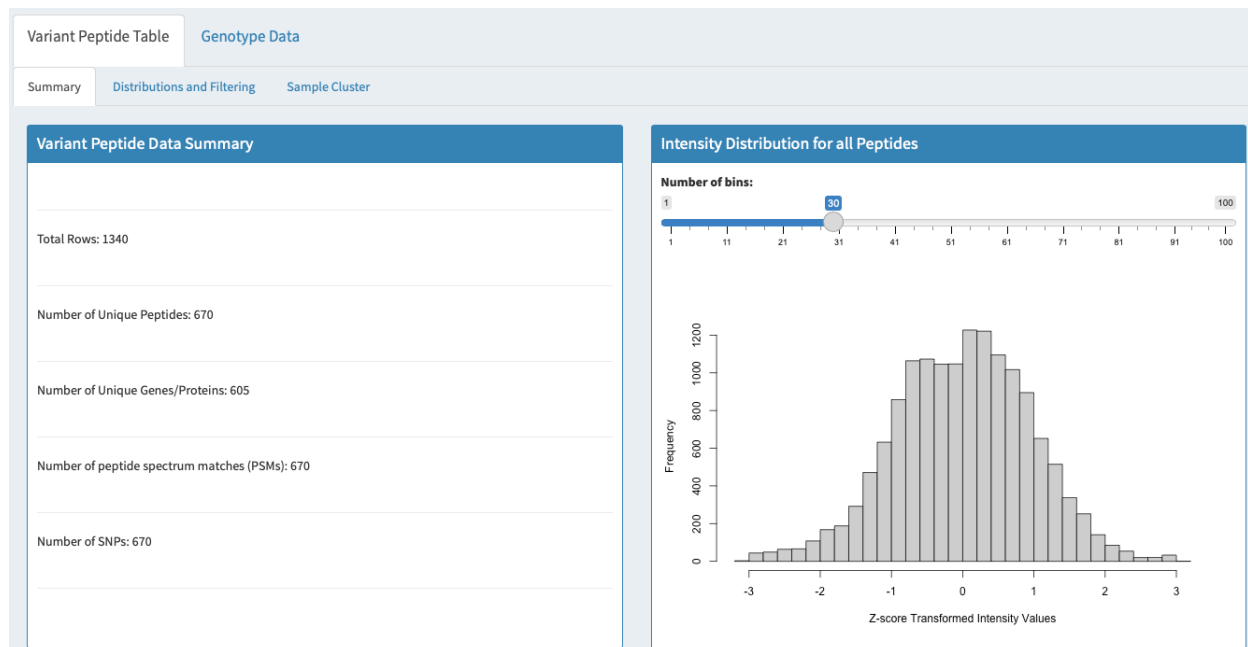
The image contains two side-by-side screenshots of the SMAP web application. The left screenshot shows the 'Upload Variant Peptide Table' form. It has two tabs: 'Variant Peptide Table' (selected) and 'Genotype Data'. The form includes sections for 'Data Type' (User Data, Example Data), 'Separator' (Comma, Semicolon, Tab), and 'Is Input Data Log2 Normalized?' (Yes, No). A red box highlights the 'Choose input file' section, which contains a 'Browse...' button and 'No file selected' text. Below this, it states 'Input file MUST follow this column specification:' followed by a table with columns: Peptide, Gene/Protein, PSM, SNP, Sample 1 ..., Sample N. The right screenshot shows the 'Upload Genotype Data' form. It has two tabs: 'Variant Peptide Table' and 'Genotype Data' (selected). The form includes a 'Data Type' section (User Data, Example Data) and a 'Choose Genotype File' section. A red box highlights the 'Choose Genotype File' section, which contains a 'Browse...' button and 'No file selected' text. Below this, it states 'Input file MUST follow VCF file format' and 'Missing genotypes MUST be coded as "/."'. At the bottom, it says '**ID Column MUST match SNP column in variant peptide table**'.

3.3 Data Summary & Filtering

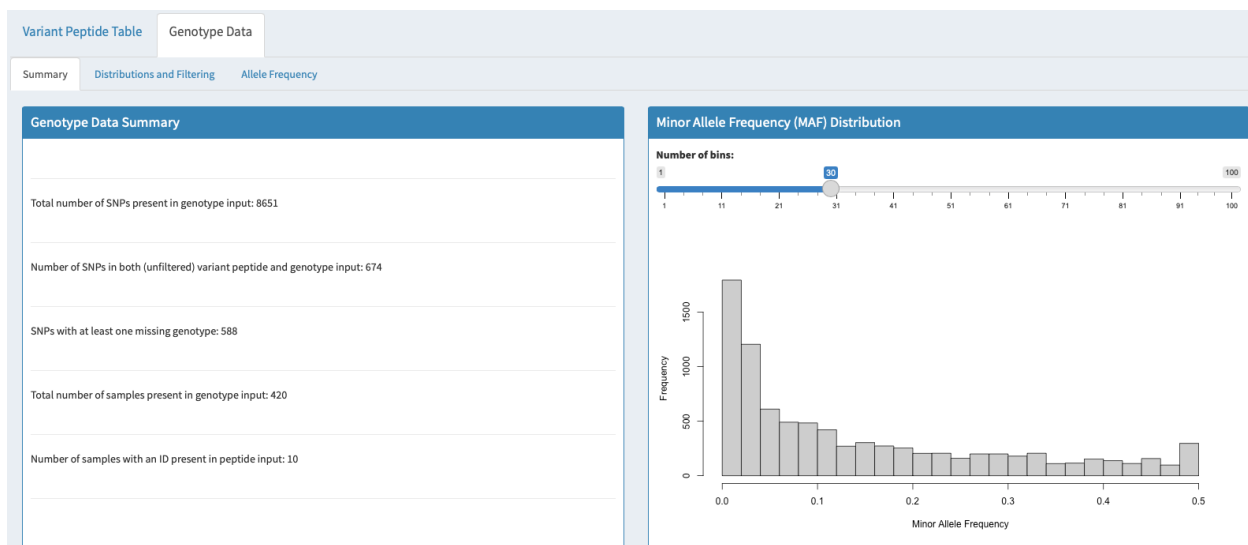
For both variant peptide and genotype data, SMAP provides summary values and relevant distributions for the input files.

- Large genotype files may take a few moments to load

Variant Peptide Data Summary (Example):



Genotype Data Summary (Example):



Users have the options to set each variant peptide filtering parameter (minimal intensity, maximal intensity, signal/noise ratio) and the genotype filtering parameter (number of missing genotypes tolerated per SNP) based on the data distributions.

Variant Peptide Data Filtering

Minimal Intensity Distribution (intensity of reference peptides)

☐ No Filtering

☒ Input Filter:

Default value (if left blank above) is **17.45** (Mean + 2SD)

Maximal Intensity Distribution

☐ No Filtering

☒ Input Filter:

Default value (if left blank above) is **12.25** (Mean - 2SD)

S/N Ratio Distribution

☐ No Filtering

☒ Input Filter:

Default value (if left blank above) is **3**

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***

Default parameters are set (and selected if no user input is given) as follows:

- Minimal and maximal intensity filters are set based on the means and standard deviations of the minimal and maximal peptide distributions.
- Signal/noise ratio is always set at the default of 3.
- Number of missing genotypes tolerated filter is OFF at default, but the user can set this filter if they have a large amount of missing data.

3.4 Run SMAP

After selecting the desired filters, SMAP is by clicking the “Run SMAP” button.

SMAP

Introduction

Input Data

Data Summary & Filtering

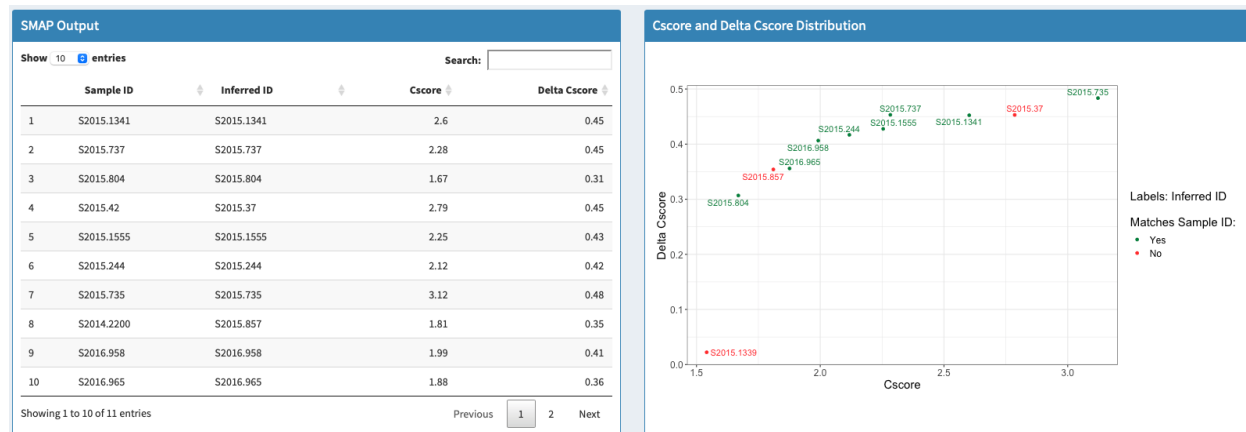
Run SMAP

Run SMAP

Before hitting run: Ensure all filters are set to the desired values in previous tab

Run SMAP

After running SMAP, a table output will be generated displaying the variant peptide sample IDs and their matched genotype IDs. The Cscore and Delta Cscore for each match is also reported and graphed.



Users can download the results table by entering a desired file name and clicking download at the top right:

Download Results

Enter name for download file:

SMAP_output

Download

4 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,2003(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(7), 2309-2320.
3. UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic acids research,49(D1), 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(16), 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(12), 3663-3673.