**SMAP: A Pipeline for Sample Matching in Proteogenomics**

**Version 1.0.0**

March 2021

**Contents**

[1. Introduction 4](#_Toc70002387)

[1.1 Software requirement 4](#_Toc70002388)

[1.2 Contact information 4](#_Toc70002389)

[1.3 License 4](#_Toc70002390)

[2. How to run SMAP (standard alone version) 5](#_Toc70002391)

[2.1 Download the pipeline 5](#_Toc70002392)

[2.2 Run SMAP program 5](#_Toc70002394)

[2.3 Input data 5](#_Toc70002395)

[2.3.1 A variant peptide table 5](#_Toc70002396)

[2.3.2 A genotype in VCF format 6](#_Toc70002397)

[2.3.3 Output files 6](#_Toc70002398)

[2.4 Cloud-based SMAP 8](#_Toc70002401)

[2.4.1 Introduction 8](#_Toc70002402)

[2.4.1 Input data 8](#_Toc70002403)

[2.4.2 Data Summary & Filtering 9](#_Toc70002405)

[2.4.2.1 summary of variant peptide data 9](#_Toc70002406)

[2.4.3.3 Distributions and filtering 10](#_Toc70002409)

[2.4.3 sample cluster 12](#_Toc70002416)

[2.4.4 summary of Genotype data 13](#_Toc70002418)

[2.4.5 Filtering of Genotype data 14](#_Toc70002421)

[2.4.6 Allele frequency 16](#_Toc70002425)

[2.4.7 Run SMAP 16](#_Toc70002428)

[3.References 18](#_Toc70002429)

# 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]*

*--variant\_peptide,-vf (A file containing quantitative values of variant peptides; required)*

*--genotype, -g (A genotype file used sample verification; required )*

*--output, -o (An output filename; required)*

*--plex, -p (Multiplex number of the isobaric labeling approach)*

*--fold\_change, -fc (Signal to Noise ratio (optional; default is 3))*

*--noise\_level, -nl (The upper threshold of a noise level)*

*--version, -h (Print version)*

*--help, -h (Print help)*

*--licence, -l (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, CScore 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

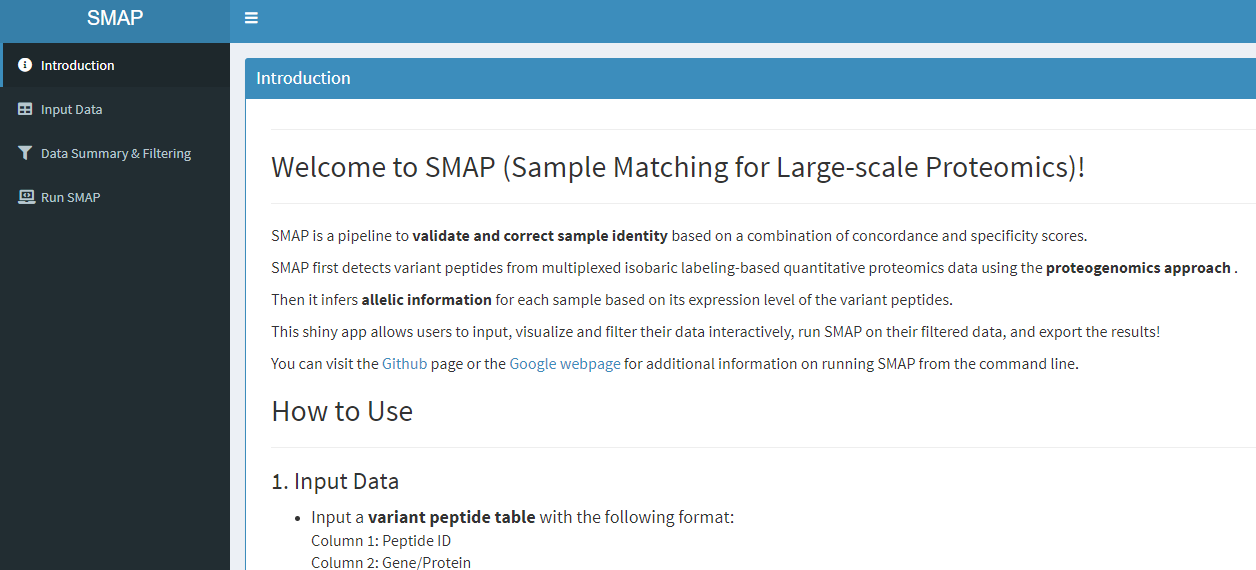
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SNP | 2015-1341 | 2015-737 | 2015-804 | 2015-42 | 2015-1555 | 2015-244 | 2015-735 | 2014-2200 | 2016-958 | 2016-965 | Internal  standard |
| chr11:75298468:A:C | A | C | A | A | A | A | A | 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/>

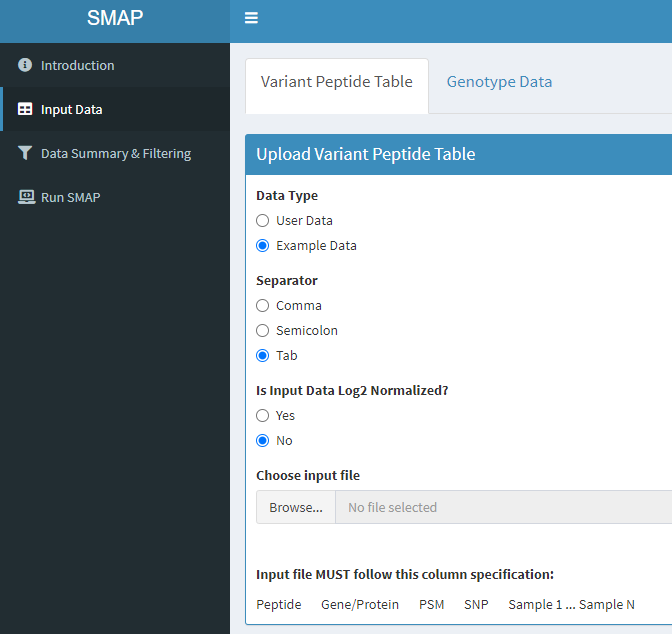
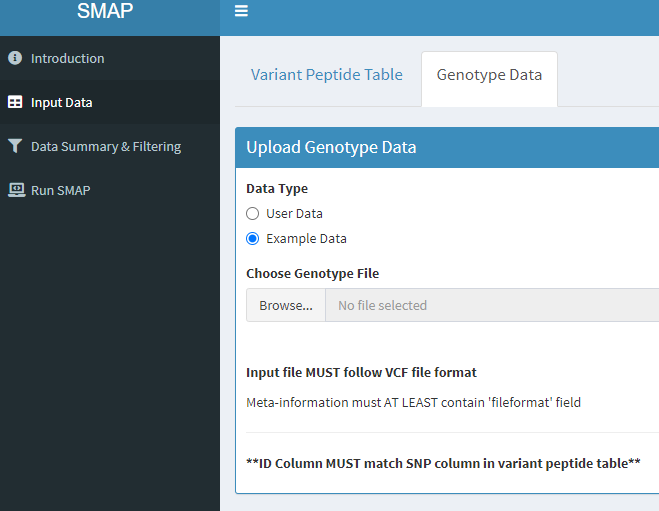
Once the website is loaded, the main page of SMAP is shown as below:

# 2.4.1 Introduction



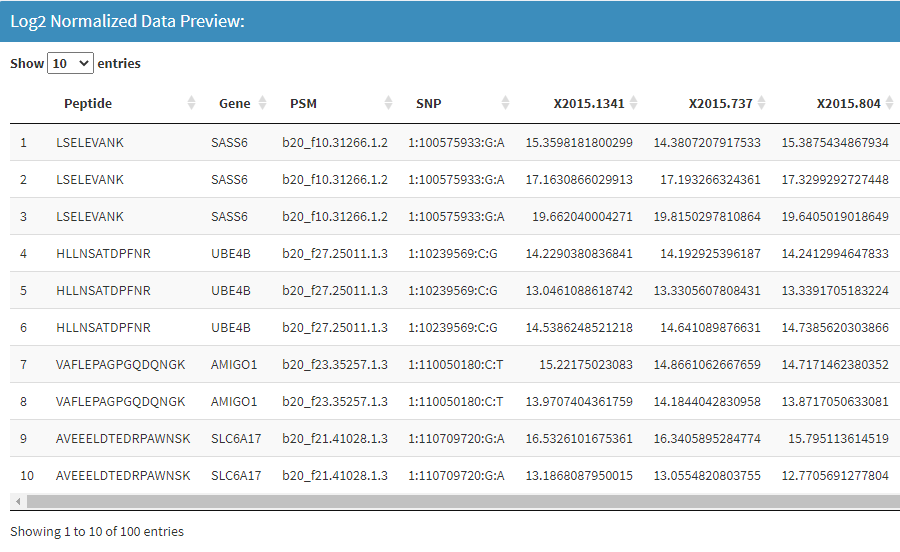
# 2.4.1 Input data

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.

1.  B. 

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

# Log2 transformed data (Preview)

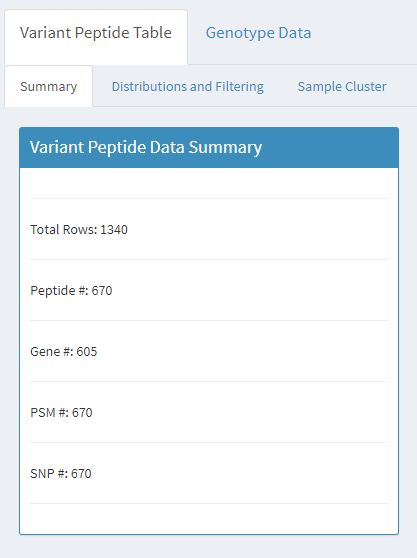
****

# 2.4.2 Data Summary & Filtering

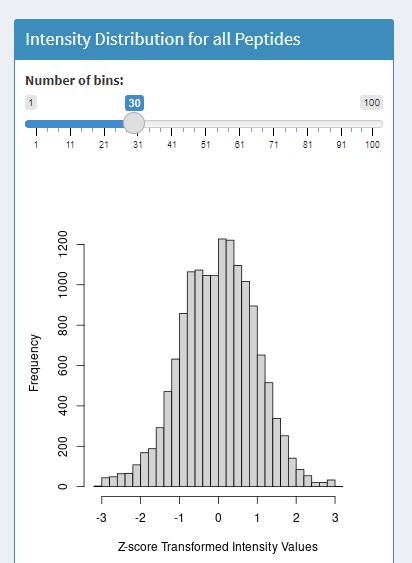
# 2.4.2.1 summary of variant peptide data

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)



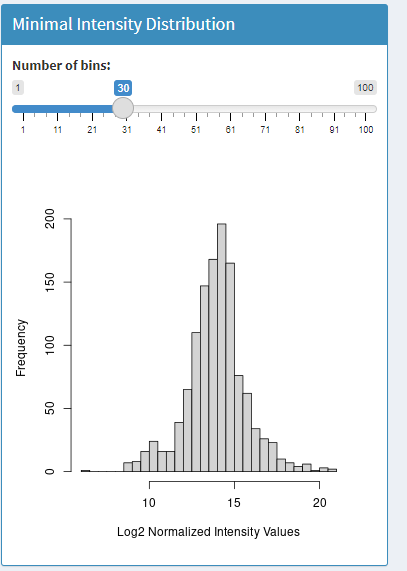
# Intensity distribution for all peptides (Preview)



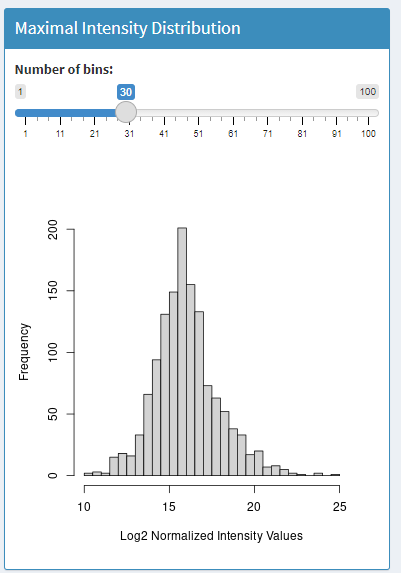
# 2.4.3.3 Distributions and filtering

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).

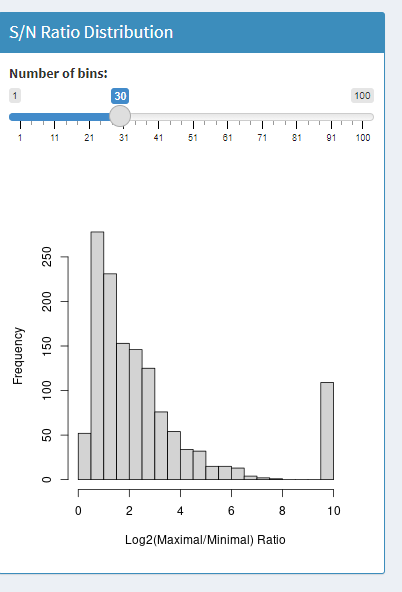
# Minimal intensity distribution (Preview)



# Maximal intensity distribution (Preview)



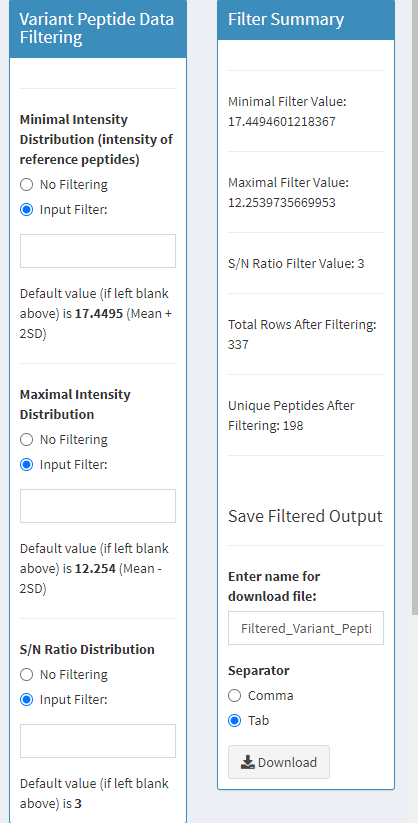
# S/N ratio distribution (Preview)



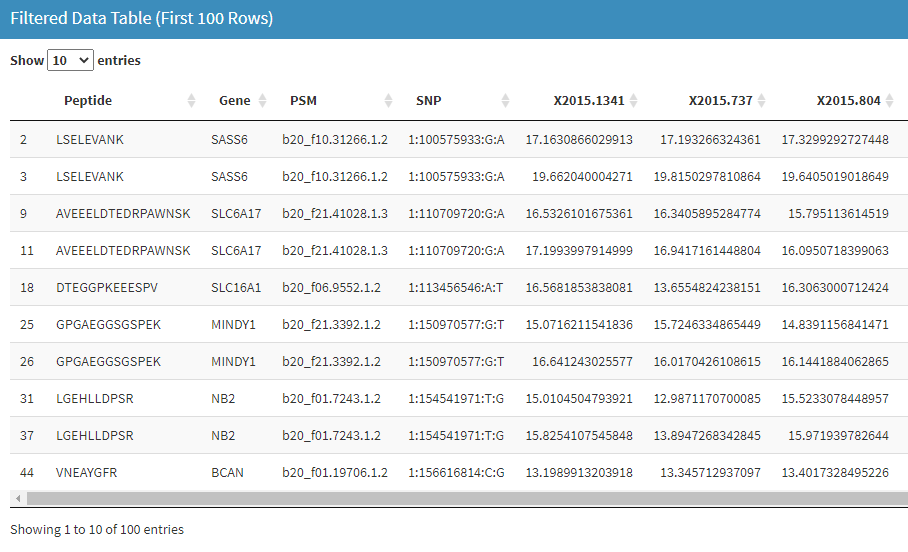
# Variant peptide data filtering

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)



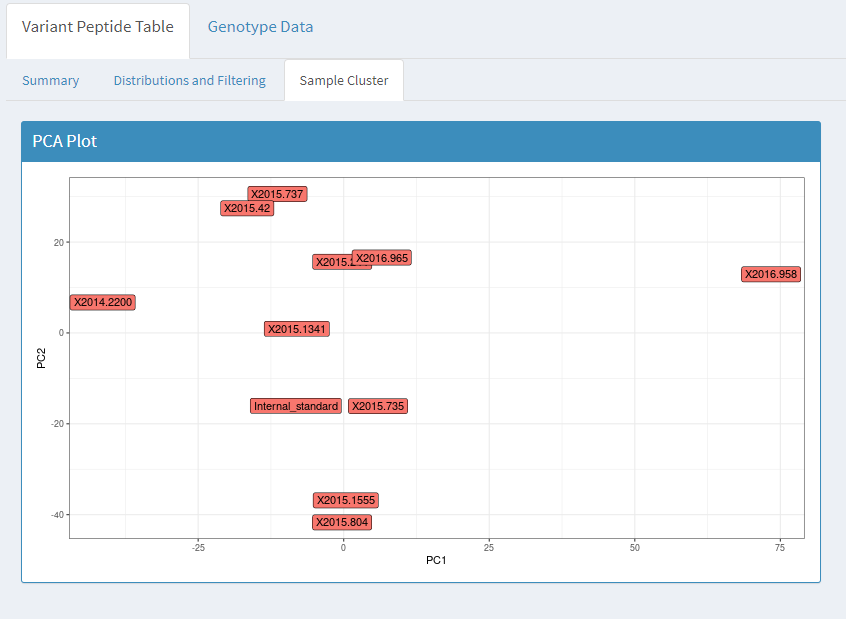
# Filtered data table (Preview)



# 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.

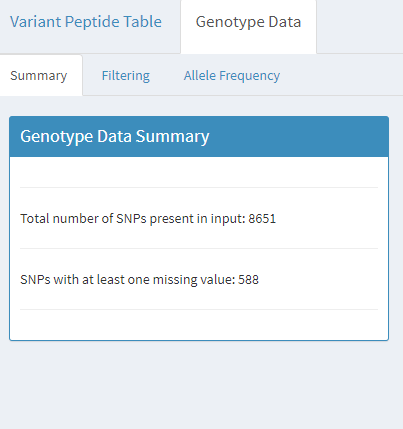
# PCA plot (Preview)



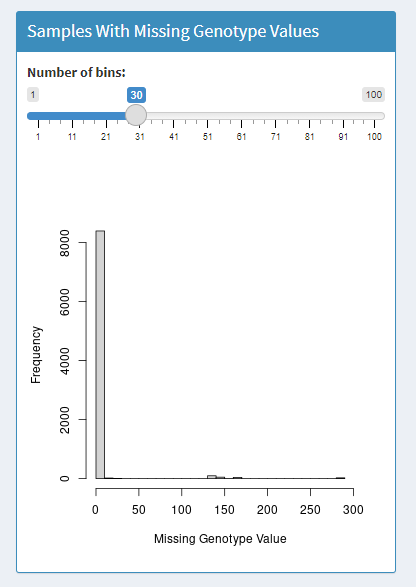
### 2.4.4 summary of Genotype data

1. Select Data Summary & Filter tab on left.
2. Select Genotype Data then select Summary, the default parameter is 30. Then the window will explore a plot for distribution of allele frequency. The basic information will also be listed in the left, such as total number of SNP present in input, SNPs with at least one missing value; average MAF(minor allele frequency).

# Geno type data summary (Preview)



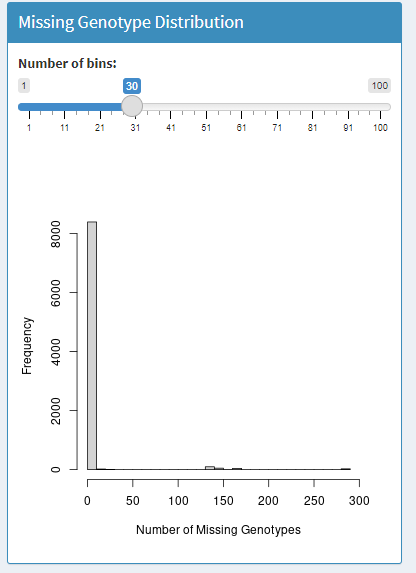
# Samples with missing genotype values (Preview)



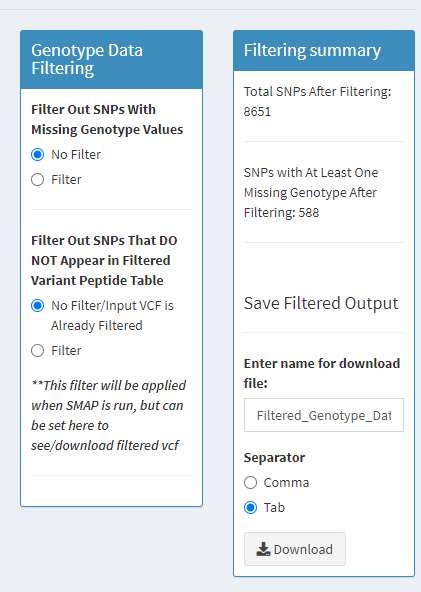
### 2.4.5 Filtering of Genotype data

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.

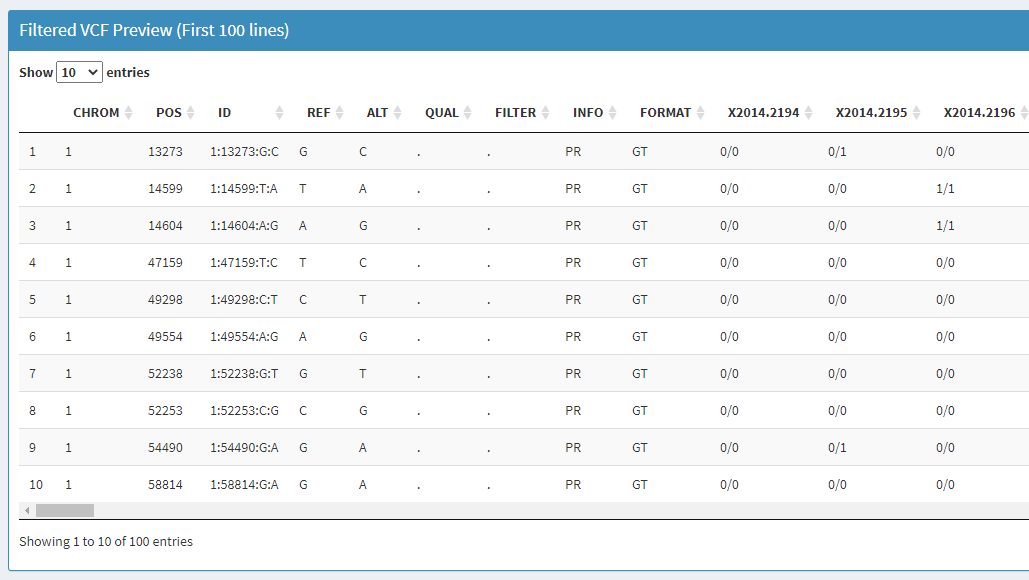
# Missing genotype distribution (Preview)



# Genotype filtering and summary (Preview)

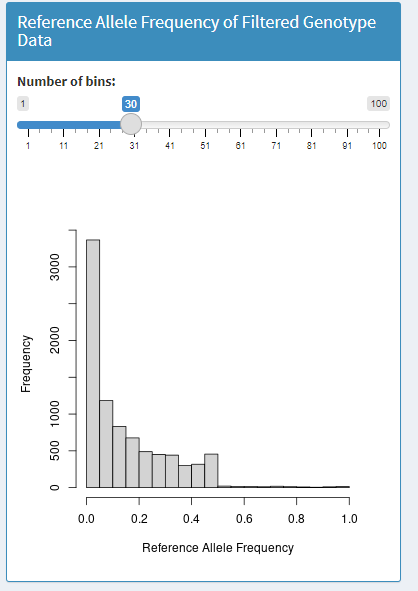


# Filtered VCF preview (Preview)

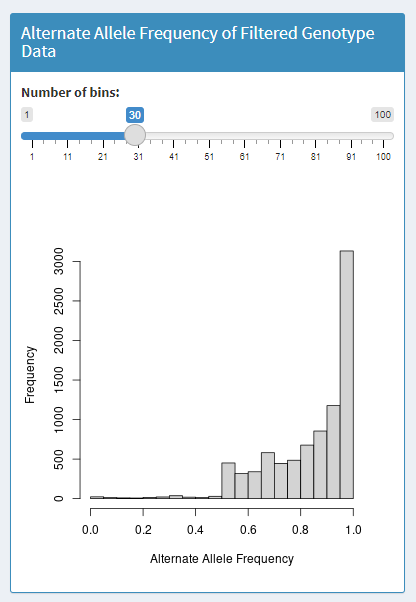


# 2.4.6 Allele frequency

# Reference allele frequency of filtered genotype data (Preview)

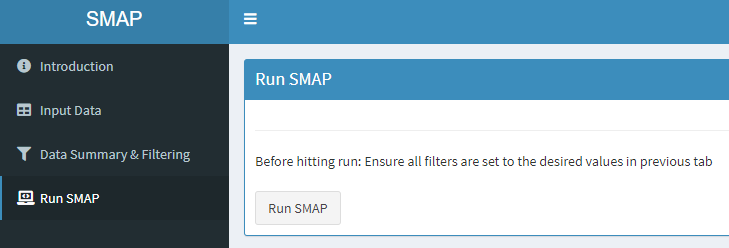


# Alternate allele frequency of filtered genotype data (Preview)

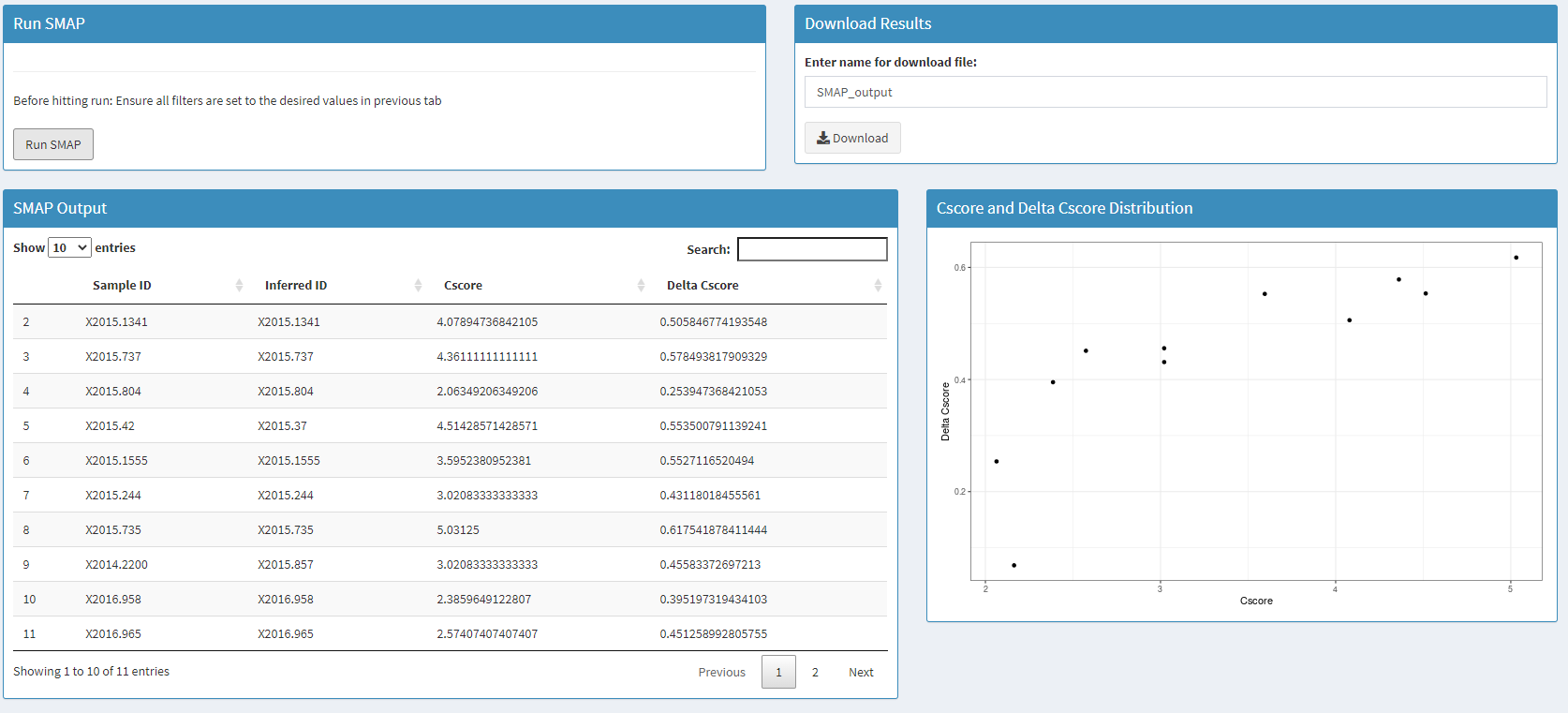


### 2.4.7 Run SMAP

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.



### 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&gt;. 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.