PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
XXXX,XX(XX):1~10
www.pibb.ac.cn

法医蛋白质组SAP分型自动化分析软件*

胡 峰[1)**] 王梦娇[3)**] 吴佳蕾[4)] 丁冬升[2)] 杨志远[2)] 季安全[2)] 丰 蕾[2)] 叶 健[1)***]

([1)] 中国人民公安大学研究生院，北京 100038；[2)] 公安部鉴定中心，现场物证溯源技术国家工程实验室，北京 100038；
[3)] 河南大学基础医学院，开封 475000；[4)] 江阴市公安局，江阴 214431）

摘要 **目的** 生物物证蛋白质组蕴含着丰富的遗传信息，即蛋白质序列的单氨基酸多态性（SAP）。然而，SAP分析工具的缺乏，严重制约了SAP在公安实战中的应用。本研究目的是为了满足法医样本蛋白质组SAP数据分析的应用需求。**方法** 分为3个模块设计SAP分型自动化分析软件。模块B内置了东亚人群常见非同义单核苷酸多态性（nsSNP）信息，与输入外显子组新增的nsSNP一起构建SAP蛋白质序列数据库。模块A利用模块B构建的SAP蛋白质序列数据库，调用预装的pFind或Maxquant搜索引擎分析质谱数据。模块C输出参考型与突变型SAP分型结果，反向推导对应的nsSNP分型（称为imputed nsSNP），并与输入的外显子组nsSNP分型结果比较，生成比对报告。使用2名中国个体的外显子组nsSNP数据、每人2根毛干蛋白质组数据，分别使用pFind与Maxquant搜索引擎对软件进行测试。使用文献中1个欧洲人、1个非洲人，每人3根毛干蛋白质组数据，勾选pFind搜索引擎测试软件，并与文献算法结果进行比较。**结果** 该软件以蛋白质组质谱数据与外显子组测序nsSNP结果为输入文件，输出SAP结果报告。测试结果显示，使用两种搜索引擎均可得到SAP结果，并且发现Maxquant得到的SAP数量略少于pFind的结果。使用文献数据测试结果显示，在文献方法完全匹配（即imputede nsSNP与外显子组nsSNP分型完全一致）的SAP位点中，SAPTyper得到了部分SAP结果，且分型一致。**结论** 针对东亚人群开发了一种自动化SAP分析算法，并形成软件SAPTyper。该软件为法医蛋白质组SAP进行个体识别与表型推断等方面研究与应用提供了一个便捷、高效的分析工具，具有良好的应用前景。

关键词 法医蛋白质组，SAP分型算法，单氨基酸多态性
中图分类号 Q3 **DOI**：10.3724/j.pibb.2025.0067 **CSTR**：32369.14.pibb.20250067

生物物证广泛存在于各类犯罪现场，在案件调查中发挥至关重要的作用。脱落毛发、陈旧骨骼、微量脱落细胞等生物物证，由于核DNA高度降解或含量极低无法获得个体遗传信息[1]，难以通过常规短串联重复序列（short tandem repeat，STR）分型技术获得有效个体遗传信息。然而，此类样本中往往保留相对完整的蛋白质组分，其中蕴含着丰富的遗传信息，表现为蛋白序列中的单氨基酸多态性（single amino acid polymorphism，SAP），与基因组中的非同义单核苷酸多态性（nonsynonymous single nucleotide polymorphism，nsSNP）之间存在直接的对应关系。通过质谱技术检测SAP位点分型，可反向推导对应的nsSNP分型（imputed nsSNP），而实现个体识别、族群推断及表型特征预测[2-3]等应用。这一策略为无法开展常规DNA分析的生物物证提供了新的技术路径，拓展了蛋白质组在法医学中的应用前景。

全面识别生物物证蛋白质组中的SAP位点分型，是实现蛋白质组遗传信息解读的关键环节。“鸟枪法”蛋白质组学技术（shotgun proteomics）能够对样本中所有蛋白质组分进行分析。该技术流程包括：首先使用胰蛋白酶将样品中的蛋白质酶切为短肽段；随后通过高分辨率液相色谱系统进行分离；最后进入质谱仪采用数据依赖性采集（data-dependent acquisition，DDA）模式进行检测，获取肽段的一级母离子和二级碎片离子谱图信息。谱图

* 国家重点研发计划课题（2022YFC3341003）和现场物证溯源技术国家工程实验室开放课题（2021NELKFKT04）资助项目。
** 并列第一作者。
*** 通讯联系人。
Tel：010-83752707，E-mail：yejian77@126.com
收稿日期：2025-02-12，接受日期：2025-06-30

的解析及蛋白质的鉴定依赖于专业的数据库搜索软件完成。谱图解析通常基于序列数据库搜索策略：将实验获取的质谱图与人类参考蛋白质数据库（如UniProtKB/Swiss-Prot）经理论酶切与碎裂模拟所生成的理论谱图进行匹配，并通过打分算法确定最可能的肽段序列。然而，该策略的局限性在于，只能识别数据库中已有的蛋白质序列。

目前，针对SAP位点的分析算法可分为以下几种。一种思路是将SAP变异按照一种氨基酸的修饰处理。一项研究[4]使用完全开放式搜索的策略，但结果表明此方法会带来极高的假阳性率，甚至高达96.7%，因此，该策略不适用于SAP鉴定。另外一种更为可行的策略是构建包含SAP信息的定制蛋白质序列数据库。该数据库除人类参考蛋白质序列之外，还需要含有SAP变异的蛋白序列数据库。SAP变异来源于基因组nsSNP信息经过生物信息学注释生成。在法医遗传学领域，Parker实验室[5]最早采用该策略分析人类毛干中SAP，选取欧洲或非洲人群中最小等位基因频率≥0.5%的nsSNP位点。此外，也Raj等[6]开发了一个蛋白质组学中变异肽质量综合评估工具PgxSAVy，将强度、变异信息等特征加入含有SAP位点的肽段，即遗传变异肽段（genetic variant peptide，GVP）的打分体系，从而对GVP的鉴定结果进行综合质量控制，为SAP的准确识别提供了量化依据。

然而，当前仍缺少专门面向东亚人群的蛋白质组SAP分析软件，难以从数万个的质谱原始谱图中高效获得SAP信息。这一问题在法医样本队列研究中尤为突出，已成为制约法医蛋白质组SAP应用研究的关键技术瓶颈。本文围绕法医蛋白质组SAP分析的应用需求，结合东亚人群遗传特征，开发了一套SAP自动化鉴定分析软件SAPTyper。该软件不仅可基于质谱DDA原始数据自动识别SAP，还能将识别结果与样本对应的外显子组测序所得到的nsSNP进行比对，从而对SAP的准确性进行系统评估与验证。与Parker实验室面向欧美人群开发的分析策略相比，SAPTyper通过集成构建东亚人群个性化SAP数据库与搜索结果解析的自动化流程，显著提升了我国法医样本队列与实际案件样本SAP分析的效率与适用性。

1 实验方法

1.1 软件总体设计

软件总体设计主要由3个核心模块构成（图1）：搜索引擎自动化调用模块（A）、SAP序列数据库构建模块（B）、以及搜索结果的SAP筛查与注释模块（C）。首先，用户输入的质谱数据文件作为分析起点，通过模块A自动调用搜索引擎，完成多肽序列的识别。模块B整合来自东亚人群公共数据库的nsSNP信息以及研究样本中通过外显子组测序新增nsSNP位点，生成个性化SAP蛋白质序列数据库，并作为模块A中搜索引擎分析所需的参考数据库。随后，模块C对模块A获得的多肽序列进行SAP位点的识别、分型和注释，自动判定多肽是否覆盖SAP位点，归类SAP分型结果，进而推导出对应的imputed nsSNP。还进一步自动提取该nsSNP在输入外显子组数据中的真实基因型信息，完成蛋白层与DNA层之间的比对。

1.2 模块设计与实现

代码的主体程序以Perl为主写成。分为3个模块。

模块A：搜索引擎的自动化调用。本模块集成了当前广泛使用的两个免费开源搜索引擎：国外的MaxQuant[7]和国内的pFind[8]，二者均支持多种质谱数据格式。用户也可使用MSRefine软件[9]对数据格式转换后再导入。首次运行前，需预设MaxQuant或pFind的参数配置文件，用户可根据实际需求进行自定义设置。该模块的输出为搜索引擎识别到的多肽和蛋白质信息。

模块B：SAP序列数据库的构建。SAP序列数据库分为固定与可变两部分。固定部分为东亚人群常见SAP序列数据库。具体生成过程如下：提取ExAC数据库（http：//exac.broadinstitute.org）中东亚人群最小等位基因频率（minor allele frequency，MAF）大于等于0.1%的SNP信息，仅选择二等位基因SNP位点。通过AnnoVar（2019Oct24）[10]形成nsSNP和对应SAP。以hg19基因组为标准，与标准序列一致的nsSNP对应的SAP定义为参考型（ref）型，与标准序列不同的nsSNP对应的SAP定义为突变型（mut）型。每个ref型SAP生成1条标准蛋白质序列。每个mut型SAP生成1条变异蛋白质序列，从sv0开始编号。与此同时，将构建一张SAP与nsSNP的对应关系表，包含SNP和SAP的位置信息、分型信息、所属基因及其在东亚人群中的频率等注释内容。SAPTyper内置的东亚人群常见SAP数据库包含88 856条参考蛋白质序列和247 390条变异蛋白质序列。可变部分根据输入的外显子组测序结果，仅选择新增的nsSNP位

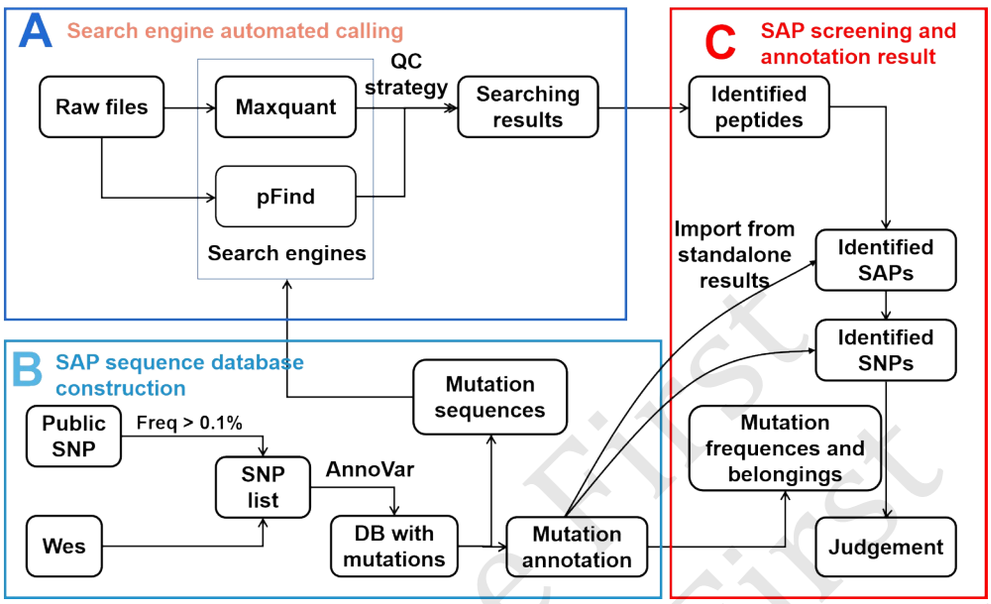


Fig. 1　Overall framework of the SAPTyper software

The workflow of SAPTyper consists of three main modules： module A is automated search engine invocation； module B is construction of the SAP sequence database； module C is SAP identification and annotation.

点，采用类似的方法动态生成含有SAP变异蛋白质序列，加入到数据库中。

模块C：搜索结果的SAP筛查与注释。该模块以模块A得到的多肽与蛋白信息为输入文件，对SAP位点进行筛查与功能注释。具体筛查主要是SAP基因分型为参考型或变异型，以及是否匹配上基因组的唯一位点。流程如下：肽段是否覆盖SAP位点（参考型或变异型均可）；匹配参考型还是变异型；氨基酸的变异是否是可替代（如氨基酸变异和氨基酸修饰的质量差类似，无法判断是因为变异还是因为修饰导致的差异）；匹配的变异如果出现在多个蛋白中，是否源于基因组上的同一个位点；由此总结出每个肽段分属的类型，共为6类（图2）。根据B模块中SAP与nsSNP的对应注释表，将识别出的SAP分型注释为nsSNP（imputed nsSNP）基因分型。之后与输入的外显子组基因分型比对，最终生成SAP结果报告。

1.3　测试数据与分析方法

选择实验室前期数据[11]进行测试。包含2个个体（M1、M2），每人2根毛干（A、B）的质谱检测结果，以及M1与M2的血液外显子组测序结果。采用SAPTyper软件分析，分别调用pFind v3.2.1与MaxQuant v2.6.7.0软件进行数据库搜索，采用反库控制结果的假阳性率（false discovery rate，FDR）。搜库设置为胰酶酶切，最多允许3个漏切，前体离子允许质量偏差为±10 ppm，碎片离子允许质量偏差为±20 ppm，*FDR*≤1%。半胱氨酸氨基甲基化修饰（carbamidomethyl［C］）为固定修饰，蛋白质N端乙酰化（acetyl［ProteinN-term］）和甲硫氨酸氧化（oxidation［M］）修饰为可变修饰。pFind中Open Search不勾选。对测试数据的SAP鉴定结果进行统计分析。

选取了Parker实验室2020年文献[12]的质谱数据，来源于1个欧洲人（E1）与1个非洲人（A2），每个人检测3根毛干样本。采用SAPTyper软件分析，调用pFind v3.2.1软件进行数据库搜索，参数设置同上。统计分析鉴定到的SAP结果，并与文献中SAP结果与部分外显子组测序结果比较。

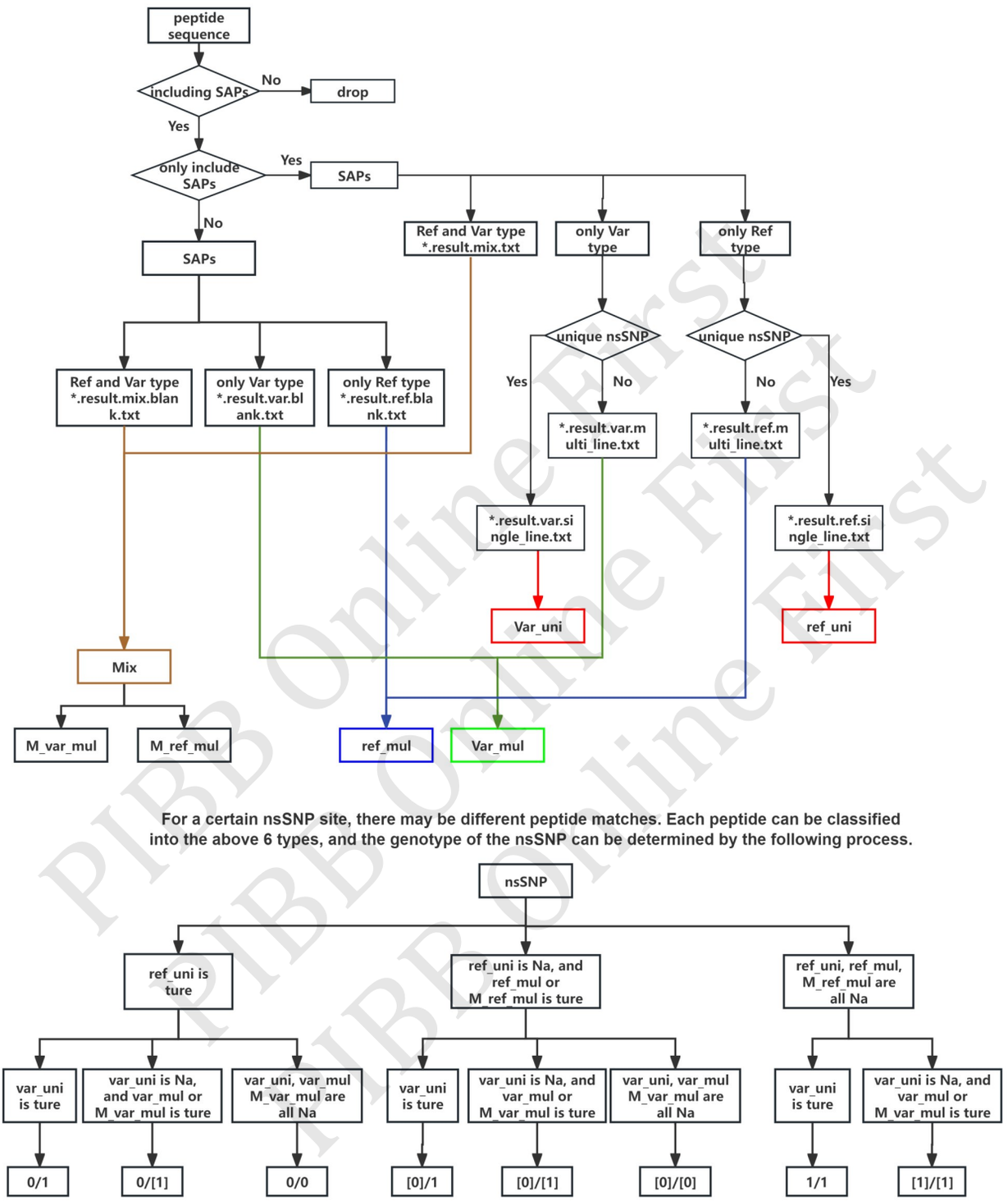


Fig.2　Logic diagram of SAP screening and genotype annotation for search results

2　结 果

2.1　软件运行

对测试数据（4个Raw文件，2个全外vcf文件）在Windows 10的台式机上分析（配置：CPU-Intel Xeon E3-1231；内存：8 GB）。硬盘占用主要是由于基因组和蛋白质组数据。内存和CPU的占用，以及运行时间主要是蛋白质组搜索引擎、以及

AnnoVar程序的需求。占用时间模块A为62 min，模块B和模块C为2 min。因此，该软件主要消耗的计算资源是由于搜索引擎占用的，其它模块占用资源相对很少。

2.2　测试样本SAP分型结果

该算法输出SAP与imputed nsSNP分型结果。具体计算过程以rs62623375位点的分型判定为例说明。rs62623375位于hg19基因组17号染色体39183304位置，为*KRTAP1-5*基因上一个nsSNP，该位点在东亚人群中存在2种分型，参考型为胞嘧啶脱氧核苷酸（C），变异型为胸腺嘧啶脱氧核苷酸（T）。这两种分型导致KRTAP1-5蛋白质序列第35位分别对应于半胱氨酸（Cys或C）和酪氨酸（Trp或Y）。*KRTAP1-5*基因对应于1个转录本，即ENST00000361883（表1）。该转录本上存在8个nsSNP位点，rs62623375位点排在第5个，因此变异型分型对应的蛋白质序列命名为ENST00000361883.sv4。在pFind搜库结果中，M1和M2样本都得到了特异性肽段为“CGYPSFSISGTCGSSCCQPSCCETSC**C**QPR”（图3a，b），匹配ENST00000361883、以及sv0、sv1、sv2、sv3、sv5、sv6、sv7共计8种蛋白质，仅未匹配sv4；另外一条特异性肽段为“CGYPSFSISGTCGSSCCQPSCCETSC**Y**QPR”（图3c，d）则仅匹配ENST00000361883.sv4。根据图2中判定逻辑，肽段A对应于SAP为参考型即Cys，肽段B对应于SAP变异型即Trp。即M1和M2两个样本在该SAP（ENST00000361883/C35Y）均检出CY分型，推导出nsSNP为CT分型。M1与M2样本rs62623375位点在全外显子组测序结果中也得到了CT的基因分型，与SAP推导的基因分型一致。

Table 1　rs6262375 site information

Gene	SNP					SAP			
	rsID	Chr：position（hg19）	Ref	Alt	Freq of Alt	Transcript	Position	Ref	Alt
KRTAP1-5	rs62623375	17：39183304	C	T	0.168 9	ENST00000361883	35	Cys（C）	Tyr（Y）

根据图2中判定逻辑，SAP基因分型分为杂合型0/1，参考纯合型0/0与变异纯合型1/1。若SAP仅检出参考型，则判定基因型为0/0；同样若仅检出变异型，则判定基因型为1/1。

在测试样本M1与M2中，使用pFind搜索引擎，0/0型检出数量最多，为103到131个；0/1型最少，为7到11个；1/1型检出数量为16到19个（表2）。这些位点大部分来自于构建的公共数据库，仅有2到3个来自于M2个体的外显子组测序结果。使用Maxquant搜索引擎略少于pFind的结果，0/0型检出数量为93到115个；0/1型为4到8个；1/1型为16到21个（表2）。

为了进一步分析软件的适用性，使用该软件测试了文章[13]的质谱数据，共计10个样本的6馏份DDA质谱结果，对结果进行整体统计分析。结果与文章中采用分步分析的结果一致。

2.3　测试样本与nsSNP比对结果

如果多个蛋白质组数据来自于同一个个体，如M1的A、B两个数据结果，会合并为单一个体的SAP与imputed nsSNP结果，再与外显子nsSNP比对。该软件自动生成imputed nsSNP与外显子组nsSNP分型结果比对报告。匹配不一致的结果情况包括半匹配（half-match），即nsSNP为杂合而imputed nsSNP仅检测到其中一种分型；以及不匹配（mismatch），即imputed nsSNP检出了nsSNP不存在的分型。

在测试样本中，使用pFind搜索引擎，合并M1和M2的A、B两次结果。M1个体共检测到171个SAP位点，158个在外显子组中被检测到，其中111个为完全匹配，37个半匹配，10个错误匹配。M2个体共检测到210个SAP位点，194个在外显子组中被检测到，其中152个为完全匹配，35个半匹配，7个错误匹配（表3）。Maxquant搜索引擎SAP检出数量略低于pFind。M1个体139个SAP位点在外显子组中被检测到，其中87个为完全匹配，45个半匹配，7个错误匹配。M2个体154个SAP位点在外显子组中被检测到，其中108个为完全匹配，41个半匹配，5个错误匹配（表3）。

2.4　其他人群样本质谱数据测试

为测试SAPTyper软件对于其他人群样本的适用性，使用Parker实验室文献[12]中1个欧洲人（E1）与1个非洲人（A2）的3次生物学重复质谱数据进行测试。E1样本检出SAP位点为157个到167个，其中0/0型为126到134个；0/1型为15到

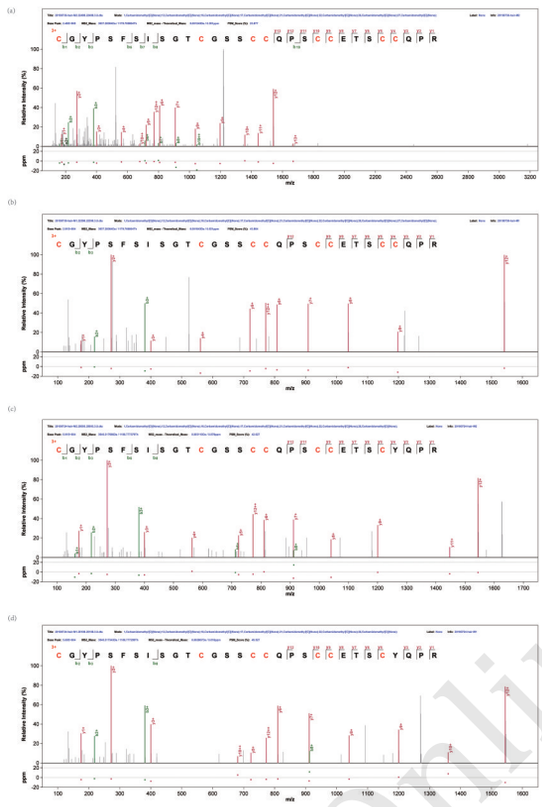


Fig. 3　Fragment mass spectrogram of peptides including SAP using pFind search engine

（a，b）The reference peptides containing Cys for M2 and M1.（c，d）The variant peptides containing Trp for M2 and M1.

Table 2　Statistics of SAP typing results in test samples

Search engine	Sample	SAP genotype			
		0/0	0/1	1/1	Total
pFind	M1-A	103	7	19	129
	M1-B	105	8	16	129
	M2-A	131	11	18	160
	M2-B	118	8	17	143
Maxquant	M1-A	93	4	19	116
	M1-B	94	4	20	118
	M2-A	115	8	16	139
	M2-B	107	6	21	137

23 个；1/1 型为 16 到 18 个（表 4）。A2 样本检出 SAP 位点为 120 个到 148 个，其中 0/0 型检出为 91 到 110 个；0/1 型为 15 到 20 个；1/1 型为 14 到 19 个（表4）。

将上述个体检出 SAP 对应的 imputed nsSNP 分型与文献中的外显子测序结果比对。由于文献中未给全部外显子组检测结果，为 237 个 nsSNP 位点的分型结果。因此，与外显子组匹配的位点数量为

Table 3　Statistics of imputed nsSNP from SAP comparing with nsSNP genotyping from exome sequencing in test samples

Search engine	Sample	Match#			Half-match #	Mismatch#	Total#
		0/0	0/1	1/1			
pFind	M1-A	78	4	8	26	7	123
	M1-B	70	3	9	30	6	118
	M1	97	4	10	37	10	158
	M2-A	98	6	7	29	6	146
	M2-B	94	5	7	28	4	138
	M2	134	8	10	35	7	194
Maxquant	M1-A	65	2	8	31	6	112
	M1-B	56	2	11	35	6	110
	M1	73	2	12	45	7	139
	M2-A	78	6	4	34	5	127
	M2-B	68	4	6	38	6	122
	M2	95	6	7	41	5	154

Three types of comparison results between SAP and nsSNP： match/half match/mismatch. Match means that SAP and nsSNP are completely consistent. Half match means that nsSNP is heterozygous and SAP only detects one of the types. Mismatch means that SAP detects a type that does not exist in the nsSNP.

22 到 33 个，更多的是文献未列出的 nsSNP 位点，为 89 到 136 个（表 5）。文献中只列出了两个个体包含所有检出 SAP 对应的外显子组测序结果，因此，SAPTyper 检出了更多的 SAP 位点。但是由于没有外显子组结果，无法判断这些 SAP 的准确性。

根据文献中 SAP 鉴定结果，分为完全匹配与半匹配（表 6）。在文献方法完全匹配的位点中，SAPTyper 检出的部分位点也包含在内（表 6）。对于半匹配位点，即 nsSNP 为杂合分型，SAPTyper 与文献方法都存在漏检的风险。

Table 4　Statistics of SAP typing results of European and African samples with pFind

Sample	SAP imputed nsSNP genotype			
	0/0	0/1	1/1	Total
E1-1	126	15	16	157
E1-2	131	18	18	167
E1-3	134	23	18	175
A2-1	95	20	17	132
A2-2	110	19	19	148
A2-3	91	15	14	120

Table 5　Statistics of SAP typing results of European and African samples from the literature with SAPTyper （pFind） compared with the nsSNP genotype

Sample	Match#			Half-match #	Mismatch#	Total#	Not detected*			
	0/0	0/1	1/1				0/0	0/1	1/1	Total
E1-1	19	3	7	7	1	37	102	10	8	120
E1-2	21	4	8	4	1	38	106	13	10	129
E1-3	23	7	9	4	2	39	111	16	9	136
A2-1	12	5	5	5	3	31	79	12	10	101
A2-2	13	5	7	7	4	37	91	11	9	111
A2-3	12	6	6	4	2	31	75	7	7	89

#It is the number of imputed nsSNP that is listed in the article’s exome sequencing result. *It is the number of nsSNP imputed from SAP but not listed the DNA sequence in the article （PMID：32505640）.

Table 6　Parker’s result of imputed nsSNPs compared with nsSNPs from exome sequencing

Sample	Matched		Half-matched	Total
	0/0 and 1/1	0/1		
E1-1	59 （22）*	3 （2）	10	72
E1-2	60 （22）	2 （1）	11	73
E1-3	62 （20）	3 （2）	9	74
A2-1	47 （12）	7 （4）	13	67
A2-2	42 （13）	6 （3）	13	61
A2-3	37 （12）	5 （3）	13	50

*The first number is Parker’s result and that with （） is the number of SAP also detected by SAPTyper.

3　讨论

本软件实现了从毛干蛋白质组DDA质谱数据到SAP结果输出的自动化分析，不仅显著减少人为操作带来的干扰，而且大幅提升了数据处理效率。SAPTyper为法医蛋白SAP应用研究提供了一个便捷有效的工具，可应用于分析如毛发等案件现场的生物物证，具有良好的应用前景。本软件具有以下几个特点：一是适合东亚人群。选取东亚人群常见nsSNP位点（*MAF*≥0.1%），覆盖了东亚人群常见SAP变异，确保分析结果的群体适应性和准确性。二是同时识别参考型与突变型SAP，全面反映个体蛋白质组中的遗传多态性信息。三是SAP序列数据库具有拓展可扩展性。当输入的外显子测序结果中有新增的nsSNP位点时候，软件将自动生成相应的SAP变异蛋白序列并添加至搜库数据库中，用于后续分析。四是满足了法医毛干蛋白SAP科研的需要。目前本软件已在毛干样本中进行了验证。从理论上讲，该软件也可应用于其他类型的样本分析，如皮屑、骨骼碎片或体液斑等，但仍需开展更多样本类型的实证研究以进一步确认其适用性。五是考虑实际案件应用场景。针对应用需求，本软件支持整合案件现场毛干的多次蛋白质组数据结果，并与多名嫌疑人的外显子测序数据进行逐一自动比对并输出结果报告，为案件侦破提供辅助支持。

从SAP的检出数量来看，pFind的检出数量普遍多于MaxQuant，这与两者所采用的搜索算法存在差异有关。若取两种搜索引擎SAP检出结果的并集，可以在一定程度上提升SAP的总体检出数量，但同时也会导致假阳性率的上升。相反，若取两者检出结果的交集，尽管可能会降低SAP的检出数量，但能够有效提高SAP识别结果的可信度与准确性。因此，在实际应用中，应根据具体研究目的权衡敏感性与特异性，选择合适的结果整合策略。

关于本软件在其他主要人群中的适用性，本文对文献[12]中来自欧洲人群与非洲人群的各3份毛干质谱数据进行了测试与验证。将SAPTyper的分析结果与文献中报道的SAP检测结果进行比对发现，部分由文献方法识别但SAPTyper未检出的imputed nsSNP位点，其在东亚人群中的MAF低于0.1%，因此未被纳入本软件构建的东亚人群常见nsSNP数据库。本软件由于数据库构建是基于东亚人群的遗传变异信息，目前无法有效识别欧洲或非洲人群中具有代表性的特有常见变异位点，但仍可识别这些人群与东亚人群共有的SAP位点。因此，SAPTyper目前主要适用于东亚人群的蛋白质组SAP分析。未来可通过引入非洲、欧洲等人群的群体特异性nsSNP数据，构建相应的SAP蛋白序列数据库，并在经过验证后，实现对多种人群的适用拓展。此外，对于杂合型（heterozygous）nsSNP位点，SAPTyper与文献方法均存在一定的漏检现象。可能原因在于DDA数据采集策略本身具有一定的随机性，质谱系统在选择母离子进行碎裂时可能偏向于某一等位基因表达较强的肽段，从而导致另一种分型未被检测，影响SAP位点的全面识别。

本软件当前未对同一检测肽段中存在双突变的情况进行处理，因此无法识别此类SAP类型。由于质谱分析中获取的肽段长度一般为6~40个氨基酸，理论上单个肽段同时包含两个SAP的概率较

低，因此该部分突变的未检出对整体分析结果的影响有限。Parker实验室构建的SAP数据库在一个蛋白质序列中整合了所有变异，理论上更有利于识别双突变。然而，至今尚无研究报告明确识别出同一肽段上存在双SAP的情况。因此可以推测，目前尚未纳入双突变序列对主流分析结论影响较小。然而，双突变肽段的检出率受到多个因素的影响，不仅与数据库构建策略密切相关（如SNP位点的选择数量及其分布位置），还受到质谱检测深度、样本的遗传背景等因素的共同制约。随着质谱检测精度和样本数量的提升，未来出现双突变甚至多突变肽段的概率将逐步增高，这类序列的识别将为个体识别提供更加丰富的遗传信息。双突变的蛋白质序列类似于基因组中的微单倍型（micro-haplotype），相比单一SAP可提供更高的信息含量。未来的研究中可考虑两种策略以提升对双/多突变的识别能力。个体特异性策略，即结合个体的外显子测序结果，在数据库中加入其真实存在的双突变或多突变SAP序列，提升个性化识别精度。穷举策略，即基于常见SAP组合构建包含多突变的候选蛋白质序列库，配合谱图匹配分析，系统探索可能存在的多突变肽段。

最后，针对法医实际应用场景的复杂性，最好可以进行常见污染物蛋白噪音信号的控制。当前，尽管会对毛干进行清洗，仍不能保证完全去除环境中的污染物。未来，可以通过模拟案件场景，如比较直接剪取头发与案件中提取毛干样本间的蛋白质组差异，采用单样或者混样建库的方式，需逐步建立法医场景特异性污染物数据库（如常见载体材料蛋白、地区性微生物蛋白组），在数据分析阶段动态设定污染信号阈值，实现对常见污染物蛋白噪音信号的控制。

4 结论

针对东亚人群开发了一种自动化SAP分析算法，并形成软件SAPTyper。本软件为法医蛋白质组SAP研究与应用提供了一个便捷、高效的分析工具，在基于SAP进行个体识别与表型推断等方面具有重要的应用前景。

参考文献

[1] 涂政, 陈松, 李万水, 等. 脱落毛发及毛干DNA的STR分型研究. 刑事技术, 2011, **36**(5): 3-7
Tu Z, Chen S, Li W S, *et al*. Forensic Sci Technol, 2011, **36**(5): 3-7

[2] 丰蕾, 江丽, 李姗飞, 等. 基于毛干蛋白质组的族群推断技术的建立与验证. 生物化学与生物物理进展, 2019, **46**(1): 81-88
Feng L, Jiang L, Li S F, *et al*. Prog Biochem Biophys, 2019, **46**(1): 81-88

[3] Parker G J, McKiernan H E, Legg K M, *et al*. Forensic proteomics. Forensic Sci Int Genet, 2021, **54**: 102529

[4] Salz R, Bouwmeester R, Gabriels R, *et al*. Personalized proteome: comparing proteogenomics and open variant search approaches for single amino acid variant detection. J Proteome Res, 2021, **20**(6): 3353-3364

[5] Parker G J, Leppert T, Anex D S, *et al*. Demonstration of protein-based human identification using the hair shaft proteome. PLoS One, 2016, **11**(9): e0160653

[6] Raj A, Aggarwal S, Singh P, *et al*. PgxSAVy: a tool for comprehensive evaluation of variant peptide quality in proteogenomics – catching the (un)usual suspects. Comput Struct Biotechnol J, 2024, 23: 711-722

[7] Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc, 2016, **11**(12): 2301-2319

[8] Chi H, Liu C, Yang H, *et al*. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. Nat Biotechnol, 2018, **36**(11):1059-1061

[9] Tang M, Huang P, Wu L, *et al*. Comprehensive evaluation and optimization of the data-dependent LC-MS/MS workflow for deep proteome profiling. Anal Chem, 2023, **95**(20): 7897-7905

[10] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 2010, **38**(16): e164

[11] 吴佳蕾, 季安全, 丁冬升, 等. 东亚人群毛干蛋白中单氨基酸多态性检测方法建立与个体识别应用. 生物化学与生物物理进展, 2022, **49**(9): 1774-1784
Wu J L, Ji A Q, Ding D S, *et al*. Prog Biochem Biophys, 2022, **49**(9): 1774-1784

[12] Goecker Z C, Salemi M R, Karim N, *et al*. Optimal processing for proteomic genotyping of single human hairs. Forensic Sci Int Genet, 2020, **47**: 102314

[13] Wu J, Liu J, Ji A, *et al*. Deep coverage proteome analysis of hair shaft for forensic individual identification. Forensic Sci Int Genet, 2022, **60**: 102742

Development of an Analytical Software for Forensic Proteomic SAP Typing*

HU Feng[1)**], WANG Meng-Jiao[3)**], WU Jia-Lei[4)], DING Dong-Sheng[2)], YANG Zhi-Yuan[2)], JI An-Quan[2)], FENG Lei[2)], YE Jian[1)***]

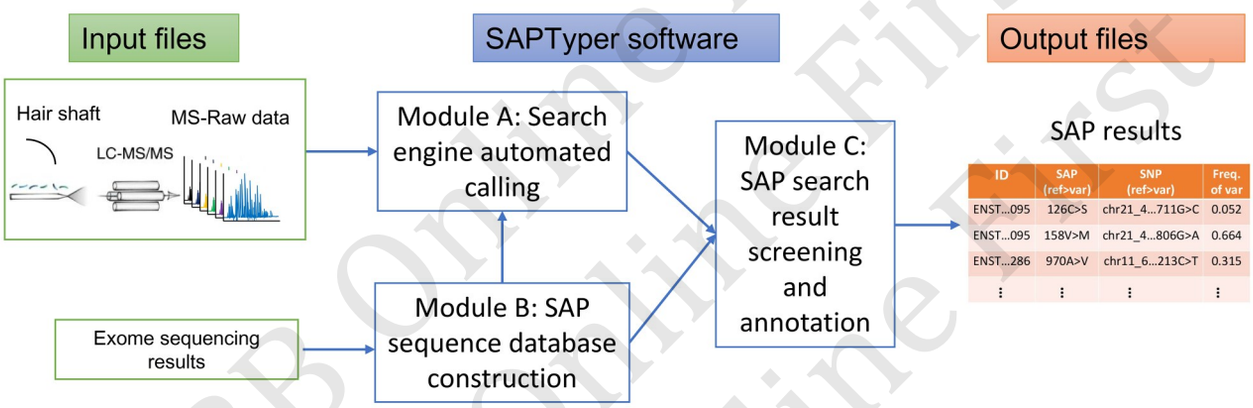
([1)]*Graduate School, People's Public Security University of China, Beijing* 100038, China;

[2)]*National Engineering Laboratory for Forensic Science, Institute of Forensic Science, Ministry of Public Security, Beijing* 100038, China;

[3)]*School of Basic Medical Sciences, Henan University, Kaifeng* 475004, China;

[4)]*Public Security Bureau of Jiangyin, Jiangyin* 214431, China)

Graphical abstract



Abstract **Objective** The proteome of biological evidence contains rich genetic information, namely single amino acid polymorphisms (SAPs) in protein sequences. However, due to the lack of efficient and convenient analysis tools, the application of SAP in public security still faces many challenges. This paper aims to meet the application requirements of SAP analysis for forensic biological evidence's proteome data. **Methods** The software is divided into three modules. First, based on a built-in database of common non-synonymous single nucleotide polymorphisms (nsSNPs) and SAPs in East Asian populations, the software integrates and annotates newly identified exonic nsSNPs as SAPs, thereby constructing a customized SAP protein sequence database. It then utilizes a pre-installed search engine—either pFind or MaxQuant—to perform analysis and output SAP typing results, identifying both reference and variant types, along with their corresponding imputed nsSNPs. Finally, SAPTyper compares the proteome-based typing results with the individual's exome-derived nsSNP profile and outputs the comparison report. **Results** SAPTyper accepts proteomic DDA mass spectrometry raw data (DDA acquisition mode) and exome sequencing results of nsSNPs as input and outputs the report of SAPs result. The pFind and Maxquant search engines were used to test the proteome data of 2 hair shafts of 2 individuals, and both obtained SAP results. It was found that the results of the Maxquant search engine were slightly less than those of pFind. This result shows that SAPTyper can achieve SAP fingding function. Moreover, the pFind search engine was used to test the proteome data of 3 hair shafts from 1 European person and 1 African person in the literature. Among the sites fully matched by the literature method, sites detected by SAPTyper are also included; for semi-matching sites, that is, nsSNPs are heterozygous, both literature method and SAPTyper method had the risk of missing detection for one type of the allele. Comparing the analysis results of SAPTyper with the SAP test

results reported in the literature, it was found that some imputed nsSNP sites identified by the literature method but not detected by SAPTyper had a MAF of less than 0.1% in East Asian populations, and therefore they were not included in the common nsSNP database of East Asian populations constructed by this software. Since the database construction of this software is based on the genetic variation information of East Asian populations, it is currently unable to effectively identify representative unique common variation sites in European or African populations, but it can still identify SAP sites shared by these populations and East Asian populations. **Conclusion**　An automated SAP analysis algorithm was developed for East Asian populations, and the software named SAPTyper was developed. This software provides a convenient and efficient analysis tool for the research and application of forensic proteomic SAP and has important application prospects in individual identification and phenotypic inference based on SAP.

Key words　forensic proteome, SAP typing algorithm, single amino acid polymorphisms
DOI：10.3724/j.pibb.2025.0067　　**CSTR**：32369.14.pibb.20250067

* This work was supported by grants from National Key R&D Program of China (2022YFC3341003) and Open Project of National Engineering Laboratory for Forensic Science (2021NELKFKT04).

** These authors contributed equally to this work.

*** Corresponding author.
Tel：86-10-83752707，E-mail：yejian77@126.com

Received：February 12，2025　Accepted：June 30，2025