PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2019,46(1):81~88
www.pibb.ac.cn

基于毛干蛋白质组的族群推断技术的建立与验证*

丰 蕾[1)] 江 丽[1)] 李姗飞[1,3)] 张 建[1)] 刘海渤[2)] 季安全[1)]
叶 健[1)] 王桂强[1)] 李彩霞[1)**]

([1)] 公安部物证鉴定中心，现场物证溯源技术国家工程实验室，法医遗传学公安部重点实验室，北京100038；
[2)] 新疆生产建设兵团公安局，乌鲁木齐 830002；[3)] 山西医科大学，太原 030001）

摘要 毛干是一种案件现场常见的生物物证，由于核DNA含量极少且高度降解，难以采用现有的短串联重复序列(short tandem repeat，STR)检验方法进行个人识别鉴定，目前仅使用线粒体DNA检验进行母系亲缘关系的判定，利用率非常低.毛干中蛋白质非常稳定，而且具有遗传多态性，表现为基因组中的非同义单核苷酸多态性(non-synonymous single nucleotide polymorphisms，nsSNPs)，转录翻译后形成蛋白质序列中的单氨基酸多态性(single amino acid polymorphisms，SAPs).充分利用毛干蛋白质中蕴含的遗传信息，为案件提供线索和证据，是实际公安业务的迫切需求，具有重要的应用价值.本文选取了104份中国汉族的毛干样本进行蛋白质组的检测，共获得了703个SAP位点，位于460个蛋白质上，共推导出552个nsSNP位点.进一步筛选在所有样本中检出率超过15%的位点，获得了88个nsSNP位点，使用毛干样本对应的口腔拭子DNA对88个nsSNP位点进行一代测序验证.为评估发现的nsSNP位点对于人群的区分能力，以千人数据库(1 000 Genome Project)为参考数据库，采用聚类分析和群体匹配概率等方法对检测的19份毛干样本进行人群来源推断.结果显示，通过检测毛干蛋白质组中的nsSNP可以实现东亚、欧洲、非洲三大洲际人群的区分.

关键词 毛干蛋白质组，单氨基酸多态性，非同义单核苷酸多态性，祖先来源推断
中图分类号 Q-3 **DOI**：10.16476/j.pibb.2018.0179

随着法医DNA检验技术的发展和进步，常见的血液/斑、唾液/斑、精液/斑、脱落细胞、带毛囊的毛发、骨骼等都能获得短串联重复序列（Short tandem repeat，STR）分型.然而，毛干由角质化细胞组成，细胞核DNA含量非常低而且降解严重，虽然也有报道采用低扩增体系、增加循环次数和多次平行扩增的方法可获得部分STR分型[1]，但是由于其准确性和稳定性差而未能在案件检验中应用.目前在实际案件中，通过检验毛干样本中线粒体DNA高变区的碱基差异，来进行母系亲缘关系的判定，存在识别率低、异质性、只能排除不能认定等缺点，限制了其在法医检验鉴定中的应用.

与毛干中的核DNA相比，蛋白质更加稳定，并且可以长期保持稳定[2].与基因组DNA类似，在不同的个体中，蛋白质氨基酸序列存在一定的差异，称作单氨基酸多态性（single amino acid polymorphism，SAP），是由编码基因上的非同义单核苷酸多态性（non-synonymous single nucleotide polymorphism，nsSNP）通过转录翻译后形成.目前蛋白质组学研究的首选平台是液质联用的串联质谱法鉴定.首先经胰酶消化，形成的肽段

*国家自然科学基金(81801877)，2017国家重点研发计划(2017YFC0803501)，基本科研业务费(2017JB025)和2018首都科技领军人才培养工程(Z18110006318006)资助项目.
**通讯联系人.
收稿日期：2018-06-28，接受日期：2018-12-05

先进入液相色谱进行分离，再进行质谱检测，通过与数据库比对分析鉴定出特异性多肽序列. 其中包含SAP的特异性多肽被称为遗传多样性多肽（genetically variant peptide，GVP）[3].

基因组中SNP作为新的法医遗传标记，研究报道了一些在洲际人群推断体系，可应用于法医人群推断. 比如Frudakis等[4]建立的27 SNP体系可实现非洲、东亚和欧洲三大人群推断，Kidd等[5-6]建立的55个SNP组合可以实现七个洲际人群的区分（非洲、欧洲、西南亚、南亚、东亚、大洋洲、美洲）. nsSNP作为一种与蛋白质序列相关的SNP，在人群中分布具有差异性. 一项美国的外显子测序计划（exome sequencing proiect， ESP）包含约2 203名非裔美国人和4 300名欧裔美国人，发现nsSNP在欧美人群中具有较好的差异性[7]. 2016年美国劳伦斯利福摩尔国家实验室法医科学中心首次将毛干蛋白质组中获得的nsSNP用于法医人群推断研究，证实该技术方法可行有效[8]. 在该研究中，作者整合了大量SNP数据，构建了SAP参考蛋白质数据库，通过质谱检测毛干蛋白质组获得89种包含SAP的特异性多肽序列，涵盖了53个SAP位点，进而反推出了32个nsSNP位点，以千人基因组人群数据为基础计算群体匹配概率和人群似然比，可以实现非洲与欧洲人群的区分.

本文选取了104个中国汉族毛干样本进行了蛋白质组的提取、检测方法研究，利用检测获得的nsSNP进行了人群推断研究.

1　材料与方法

1.1　毛干蛋白质组提取

生物样本来源于国家科技资源共享服务平台计划项目（YCZYPT［2017］01-3），共计104名中国汉族无关个体的毛干样本以及对应的口腔擦拭物样本. 样本均获得志愿者的知情同意，符合公安部物证鉴定伦理委员会的伦理学标准. 毛干样本均切去毛发的头尾，以保证不含毛囊和发尾，每份毛干长2 cm（单根长度不足2 cm时则使用两根同源毛干）. 使用10%甲醇、水各清洗2次，每次1~2 h，之后取出清洗后的毛干切碎至约1~2 mm. 切碎后的毛干每份分别加入100 μl蛋白质处理液（1 mol/L尿素、50 mmol/L NH_4HCO_3、0.1 mol/L DTT、7 mg/L胰酶），于37℃金属浴中振荡反应16 h. 吸取酶解液至新EP管中，进行ZipTip除盐，抽干后加入上样缓冲液复溶后进样，液质检测.

1.2　蛋白质组质谱检测与SAP位点定位

对液质检测. raw文件使用Proteome Discoverer1.4软件进行蛋白质定性鉴定，选择FDR≤1%的多肽作为高度可信蛋白质鉴定的过滤参数. 多肽中氨基酸多态性（SAP）位点定位方法如下：SAP参考蛋白质数据库为文献中建立的数据库（RefSeq protein variant database）[8]，该数据库既包括突变前的蛋白质序列，也包括突变后的蛋白质序列. 将上一步筛选得到的特异性多肽与参考蛋白质序列进行比对，筛选出其中存在氨基酸多态性位点的多肽，并对多态性位点在参考蛋白质序列的位置进行定位.

使用KOBAS（KEGG orthology based annotation system）系统[9]进行基因功能分类（gene ontology简称GO）分析. GO分析是基因功能国际标准分类体系，通过GO分析按照细胞组分（cellular component）、分子功能（molecular function）、生物学过程（biological process）对基因进行分类.

1.3　反推nsSNP统计学分析

根据SAP所在的蛋白质名称以及位置，将SAP与千人基因组数据库关联，找到SAP与SNP的对应关系. 根据SAP所在的蛋白质名称，与千人基因组中SNP所在的蛋白质名称进行对应关系的查找，通过比对分析进一步确定nsSNP的分型.

1.4　nsSNP一代测序验证

对挑选的10份口腔拭子，采用MagAttract DNA Mini M48（Qiagen）试剂盒提取基因组DNA，使用Primer Premier 5.0软件设计引物，一代测序方法检测相应的nsSNP的分型.

1.5　人群推断分析

使用SNPAnalyzer 2.0软件[10]分析连锁不平衡分析，R^2<0.2. 以千人基因组数据库为参考数据库，采用聚类分析以及群体匹配概率等分析方法进行人群推断分析. 利用R v3.2.3软件（https：//www.r-project.org/）进行主成分分析. 用STRUCTURE v2.3.4软件[11]进行聚类分析，分析各人群的遗传结构. 使用Distruct v1.1绘制人群聚类结果图，并统计个体祖先成分. 用族群推断软件Forensic Intelligence v1.0计算群体匹配概率[12].

2　结　果

2.1　毛干蛋白质组nsSNP检出情况

对全部104名中国汉族无关个体的毛干样本进行质谱检测，不同的样本检出的多肽数量不同，范围为304~1 509个多肽（均值为937±262个），其中包含SAP的特异性多肽为44~137个（均值为96±20个）.在104份样本检出的所有包含SAP的特异性多肽取并集，共获得772个包含SAP的特异性多肽和703个SAP位点，位于460个蛋白质上.GO分析显示，检出的蛋白质功能分布广泛，在细胞组分（cellular component）、分子功能（molecular function）、生物学过程（biological process）三大分类中均有分布，涉及到细胞功能、代谢、应急反应、信号转导等方面（附件表S1）.进一步通过与千人基因组数据库对比确定nsSNP的分型，其中140个蛋白质未找到对应的nsSNP位点，共获得320个蛋白质上的552个nsSNP位点.

2.2　一代测序验证毛干蛋白质组nsSNP

蛋白质中检测出的SAP位点反推出的nsSNP，可以通过基因组DNA测序验证其准确性.因此，使用了毛干样本对应的口腔拭子DNA，采用一代测序检测相应的SNP位点分型.在毛干样本检测中，共获得552个nsSNP位点，由于每个位点在104份样本中检出率不同，因此对nsSNP位点进一步筛选，选择检出率超过15%的nsSNP位点，共获得88个nsSNP位点，在10份毛干样本对应的口腔拭子进行一代测序验证.

通过比较质谱反推获得的nsSNP分型和测序检测到的真实nsSNP分型，质谱和一代测序结果一致定为真阳性（true positive，TP），质谱检测与一代测序不一致定为假阳性（false positive，FP），通过TP/（TP+FP）的比值计算准确性，10个样本平均准确性为95.88%.质谱未检测到而一代测序检测出分型定为假阴性（false negative，FN），通过TP/（TP+FN）的比值计算检出率，10个样本平均检出率为77.19%（图1）.

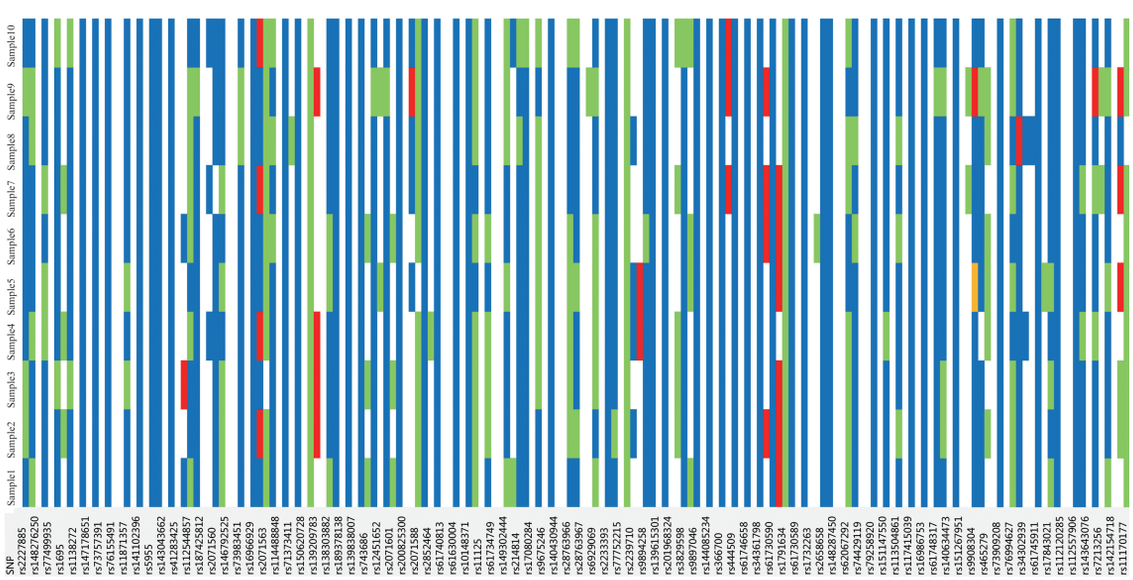

Fig. 1　Validation results of Sanger sequencing

Totally，buccal swabs of 10 individuals were sequenced to obtain the genotypes of 88 SNP. True positive（TP）represents that the mass spectrometry and Sanger sequencing results are consistent and are shown with blue color. False positive（FP）represents that the mass spectrometry and Sanger sequencing results are inconsistent and are shown with red color. False negative（FN）represents mass spectrometry failed to genotype but Sanger sequencing succeeded in genotyping and are shown in green color. True negative（TN）represents that both mass spectrometry and Sanger sequencing failed to genotype and are shown in white color. Orange represents Sanger sequencing failed to genotype

2.3　基于千人基因组数据库对毛干nsSNP人群推断效力评估

通过千人基因组数据库对比，毛干蛋白质组共推导出552个nsSNP位点. 其中6个位点（rs146291703、rs10274334、rs57670668、rs143643076、rs6580873、rs2229512）为三等位基因. 剩余546个nsSNP位点经连锁不平衡检验，删掉20个位点（R^2>0.2）. 剩余526个nsSNP位点（附表S1），基于千人基因组中非洲、东亚、欧洲三大人群数据（n=1 668）对该组位点进行评估.

目前关于人群祖先来源推断的常用是STRUCTURE分析，是基于一组SNP位点的群体样本基因型数据进行聚类分析. 首先假定有K个群体（K由用户指定可能的范围，最终根据结果确定最优值），程序模拟在K个群体的情况下使用贝叶斯算法和“有放回的重抽样方法”来推断群体结构和个体祖先成分. 每个个体按照概率被分配到一个群体或多个群体里. 使用的千人基因组中非洲、东亚、欧洲三大人群数据（n=1 668）作为参考数据库，STRUCTURE聚类分析结果显示，K=3时，非洲、东亚、欧洲三大人群为单一祖先成分为主（图2）. 每个群体在图中都会有一个单独颜色表示. 柱状图是由数据库中所有个体组成，每个个体颜色组成比例是根据STRUCTURE祖先成分聚类分析预测得到的. 同一人群中的个体虽然在画图时一起展示，但STRUCTURE软件聚类分析时是独立预测成分的. 同时，通过留一法对参考数据库进行自检，计算随机人群匹配概率和似然比（以似然比大于100判断其最可能祖先来源），仅有一份非洲样本（NA20314）被错判为欧洲样本，其余均正确. 因此，获得的526个nsSNP位点可有效区分非洲、东亚、欧洲三大人群.

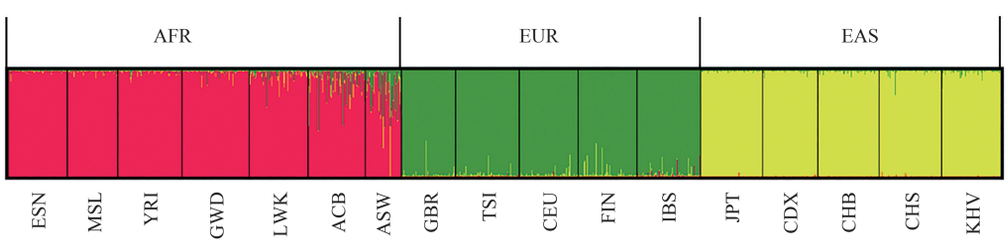

Fig 2　STURCTURE analysis results（K=3）

2.4　19份测试样本祖先来源分析

如前所述，检出率大于15%的nsSNP位点为88个，去掉连锁位点（R^2>0.2）和三等位基因位点后共获得81个nsSNP位点. 以千人基因组数据库中81个位点的分型数据为参考数据库，19个中国汉族人群为测试样本，计算随机人群匹配概率和似然比（以似然比大于100判断其最可能祖先来源），进行祖先来源的推断（表1）. 19个样本随机匹配概率最高的人群均为东亚，其中16个样本被正确预测为东亚人群，另外3个样本随机匹配概率最高的为东亚人群，与排第二的欧洲人群之间似然比小于100，因此不能判定这3个样本为东亚人群还是欧洲人群. 通过STRUCTURE（K=3）计算19个测试样本的祖先成分，结果显示东亚祖先信息成分范围为99.2%~84.2%，表明所有测试样本均以东亚祖先成分为主（表2）.

Table 1　The matching probability （MP） result of 19 test individuals

Samples	Population	MP value	LR value
CHH60	EAS	3.39E-06	1
	EUR	1.43E-07	2.37E+01
	AFR	1.99E-15	1.70E+09
CHH86	EAS	3.99E-09	1
	AFR	2.68E-28	1.49E+19
	EUR	1.14E-34	3.49E+25
CHH88	EAS	6.28E-08	1
	AFR	1.27E-15	4.93E+07
	EUR	2.69E-33	2.34E+25
CHH98	EAS	6.41E-06	1
	AFR	8.67E-12	7.39E+05
	EUR	8.86E-19	7.23E+12
CHH100	EAS	1.19E-09	1
	AFR	3.08E-13	3.86E+03
	EUR	6.57E-38	1.81E+28
CHH108	EAS	6.23E-07	1
	EUR	1.04E-10	5.97E+03
	AFR	2.80E-17	2.23E+10
CHH116	EAS	1.71E-09	1
	AFR	2.46E-21	6.95E+11
	EUR	9.41E-39	1.82E+29
CHH119	EAS	2.90E-08	1
	EUR	7.49E-10	3.86E+01
	AFR	7.24E-29	4.00E+20
CHH122	EAS	5.42E-12	1
	AFR	1.21E-33	4.48E+21
	EUR	4.39E-40	1.23E+28
CHH123	EAS	1.24E-10	1
	EUR	1.55E-18	8.00E+07
	AFR	7.23E-33	1.72E+22
CHH126	EAS	7.74E-07	1
	EUR	1.42E-09	5.44E+02
	AFR	9.31E-17	8.31E+09
CHH136	EAS	6.82E-06	1
	EUR	3.50E-09	1.95E+03
	AFR	3.91E-14	1.74E+08
CHH141	EAS	0.00490855	1
	EUR	0.00139902	3.51E+00
	AFR	5.21E-09	9.41E+05
CHH142	EAS	0.00048982	1
	EUR	9.97E-07	4.91E+02
	AFR	6.63E-14	7.39E+09
CHH146	EAS	3.02E-11	1
	AFR	9.22E-33	3.27E+21
	EUR	1.35E-39	2.23E+28
CHH149	EAS	2.31E-08	1
	EUR	5.95E-12	3.88E+03
	AFR	3.67E-14	6.29E+05
CHH152	EAS	1.04E-05	1
	EUR	3.43E-09	3.03E+03
	AFR	1.19E-13	8.73E+07
CHH153	EAS	6.78E-06	1
	EUR	1.93E-09	3.52E+03
	AFR	1.71E-14	3.97E+08
CHH162	EAS	8.09E-12	1
	EUR	1.06E-18	7.63E+06
	AFR	5.07E-19	1.60E+07

Table 2　Ancestry component result of 19 test individuals

Samples	Ancestry components		
	AFR	EAS	EUR
CHH60	0.0030	0.9431	0.0539
CHH86	0.0020	0.9910	0.0070
CHH88	0.0030	0.9860	0.0100
CHH98	0.0580	0.9280	0.0140
CHH100	0.0779	0.9161	0.0060
CHH108	0.0320	0.9590	0.0090
CHH116	0.0030	0.9920	0.0050
CHH119	0.0020	0.9610	0.0360
CHH122	0.0110	0.9830	0.0060
CHH123	0.0100	0.9830	0.0070
CHH126	0.0080	0.9810	0.0110
CHH136	0.0160	0.9740	0.0100
CHH141	0.0030	0.8460	0.1510
CHH142	0.0030	0.9830	0.0130
CHH146	0.0100	0.9830	0.0070
CHH149	0.0200	0.9640	0.0160
CHH152	0.0140	0.9740	0.0110
CHH153	0.0140	0.9760	0.0100
CHH162	0.1430	0.8520	0.0050

3　讨　　论

本文在中国国内首次建立了一种基于毛干蛋白质组进行法医人群推断的方法，对获得的81个nsSNP位点，采用聚类分析、群体匹配概率等方法进行人群来源推断，使用19份汉族毛干样本评估该组位点人群来源推断能力.结果显示，通过检测

毛干蛋白质组，可以初步实现东亚、欧洲、非洲三大洲际人群的区分.

毛干是案件现场的常见生物物证，目前主要用于同位素分析、毒品、药物等小分子的检验，从而推断个体的药物和饮食摄入情况[13-14]. 本文则是研究毛干中的生物大分子——蛋白质中蕴含的遗传信息（SAP），大分子与小分子检测相比，难度更大，但是蕴含的遗传信息更丰富. 本文检出的nsSNP不包含现有文献报道的关于种族、色素、秃发等相关的SNP位点[6，12，15-19].

毛干蛋白质组用于法医学族群推断检验中，最重要的问题就是有效性和准确性，本文以及Parker等[8]的研究初步证明该方法的有效性. 与Parker等[8]的方法相比，本文中nsSNP的检出率提高至77.19%， Parker等的方法需要1 mg毛干样本，本文改进了蛋白质组的提取方法，仅用2 cm的毛干进行质谱检测和分析. 在推断的准确性方面，19份汉族测试样本中16份均获得准确的结果，有3份样本无法与欧洲人群区分，主要原因在于获得的nsSNP位点数目不足，导致欧洲和东亚人群的区分度下降. 下一步可以通过使用性能更好的质谱仪，优化现有生物信息学分析方法，获得更高的检出率和灵敏度.

关于SAP检验的准确性，通过一代测序验证可以发现，准确性高达95.88%，表明SAP与反推的nsSNP具有非常高的一致性，但是仍然存在一定的假阳性，需进一步地研究. 三联体密码子具有简并性，除了甲硫氨酸和色氨酸外，每一个氨基酸都至少有两个密码子. 氨基酸的简并性会影响SAP到nsSNP推导的准确性. 本文使用的SAP参考数据库为Parker等[8]构建的数据库，包含了最小等位基因频率MAF>0.5%的多个nsSNP，通过质谱检出的特异性多肽，与该参考数据库比对，从而确定SAP位点分型，进一步通过千人数据库中的SNP分型信息，推导出nsSNP，在千人数据库中对应的这些nsSNP中，三等位基因的nsSNP在分析中被删除，而且没有发现简并密码子分型位点的存在. 除了上述遗传学因素以外，蛋白质存在各种修饰，比如在特定的位点存在磷酸化、糖基化等修饰，这些修饰会引起分子质量的变化而导致质谱检验结果的假阳性. 这些遗传和修饰因素都可能导致蛋白质质谱检测假阳性的出现. 而且本文使用参考数据库是文献里的数据库，只列出了包含SAP分型的特异性多肽序列信息，并不包含nsSNP分型，因此有必要基于东亚人群建立SAP参考数据库和对应的nsSNP数据库. 对不同中国人群的区分，由于中国人群内部遗传差异较三大洲际人群小，需要筛选更多的遗传位点，本课题组下一步将针对东亚南北方人群区分、高原适应人群的区分进行研究，目前相关人群的毛干样本已经收集.

总之，基于毛干蛋白质组的人群推断方法目前仍处于起步研究阶段，在法庭科学族群推断以及个体识别等方面有良好的应用前景，有望成为法医DNA分型技术之后的又一突破.

附件 表S1见本文网络版附录（http：//www.pibb.ac.cn）

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(05): 3-7
Tu Z, Chen S, Li WS, *et al*. Forensic Science And Technology, 2011, (05): 3-7

[2] Wadsworth C, Buckley M. Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone. Rapid Commun Mass Spectrom, 2014, **28**(6): 605-615

[3] Sheynkman G M, Shortreed M R, Frey B L, *et al*. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. J Proteome Res, 2014, **13**(1): 228-240

[4] Frudakis T, Venkateswarlu K, Thomas M J, *et al*. A classifier for the SNP -based inference of ancestry. J Forensic Sci, 2003, **48**(4): 771-782

[5] Kidd K K, Speed W C, Pakstis A J, *et al*. Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet, 2014, **10**: 23-32

[6] Pakstis A J, Haigh E, Cherni L, *et al*. 52 additional reference population samples for the 55 AISNP panel. Forensic Sci Int Genet, 2015, **19**: 269-271

[7] Tennessen J A, Bigham A W, O'Connor T D, *et al*. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science, 2012, **337**(6090): 64-69

[8] 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

[9] Wu J, Mao X, Cai T, *et al*. KOBAS server: a web-based platform for automated annotation and pathway identification. Nucleic Acids Res, 2006, 34(Web Server issue)**:** W720-724

[10] Yoo J, Lee Y, Kim Y, *et al*. SNPAnalyzer 2.0: a web-based

integrated workbench for linkage disequilibrium analysis and association analysis. BMC Bioinformatics, 2008, **9**: 290

[11] Pritchard J K, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics, 2000, **155**(2): 945-959

[12] Wei Y L, Wei L, Zhao L, *et al*. A single-tube 27-plex SNP assay for estimating individual ancestry and admixture from three continents. Int J Legal Med, 2016, **130**(1): 27-37

[13] 杨小林, 席焕久, 温有锋, 等 . 人类毛发的形态学研究及法医学意义 . 解剖科学进展, 2011, **17**(1):86-89
Yang X L, Xi H J, Wen Y F, *et al*. Progress of Anatomical Sciences, 2011, **17**(1):86-89

[14] 刘丹, 李上勋, 周鑫, 等 . 毛发分析的临床及法医学应用研究 . 中外医疗, 2010, **29**(18):178-179
Liu D, Li S X, Zhou X, *et al*. China Foreign Medical Treatment, 2010, **29**(18):178-179

[15] Medland S E, Nyholt D R, Painter J N, *et al*. Common variants in the trichohyalin gene are associated with straight hair in Europeans. Am J Hum Genet, 2009, **85**(5): 750-755

[16] Medland S E, Zhu G, Martin N G. Estimating the heritability of hair curliness in twins of European ancestry. Twin Res Hum Genet, 2009, **12**(5): 514-518

[17] Liu F, Hamer M A, Heilmann S, *et al*. Prediction of male-pattern baldness from genotypes. Eur J Hum Genet, 2016, **24**(6): 895-902

[18] Wu S, Tan J, Yang Y, *et al*. Genome-wide scans reveal variants at EDAR predominantly affecting hair straightness in Han Chinese and Uyghur populations. Hum Genet, 2016, **135**(11): 1279-1286

[19] Walsh S, Liu F, Wollstein A, *et al*. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet, 2013, **7**(1): 98-115

Development and Validation of Protein-based Forensic Ancestry Inference Method Using Hair Shaft Proteome*

Feng Lei[1)], Jiang Li[1)], Li Shan-Fei[1,3)], Zhang Jian[1)], Liu Hai-Bo[2)], Ji An-Quan[1)], Ye Jian[1)], Wang Gui-Qiang[1)], Li Cai-Xia[1)**]

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

[2)]*Xinjiang Production and Construction Corps public security bureau, Urumqi* 830002, China;

[3)]*Shanxi Medical University, Shanxi Taiyuan* 030001, China)

Abstract Hair shaft is one kind of important biological evidence and is widely collected in the crime scene. However, it is very difficult to obtain full STR profiles as the nuclear DNA of hair shaft has degraded seriously. Mitochondrial DNA examination is the routine method for hair shaft, but it can not be used to identify individuals. Therefore, it is important to explore more genetic information of hair shaft in order to offer more clues for crime investigation. Protein is chemically more robust than DNA and can persist for a longer period. Proteins also contains genetic variation in the form of single amino acid polymorphisms (SAPs), which can be used to infer the status of non-synonymous single nucleotide polymorphisms (nsSNPs) for forensic ancestry analysis. Here, we used mass spectrometry-based shotgun proteomics to characterize hair shaft proteins in 104 Chinese Han subjects. A total of 552 nsSNPs were imputed from 703 SAPs in 460 proteins. 88 nsSNPs were selected according to a call rate of 15%, then validated using Sanger sequencing. Finally, clustering analysis and population random match probability calculation were performed to infer the ancestry of 19 Han Chinese individuals, and individual genotypes of 1 000 genome populations were employed as reference data. The results demonstrated that nsSNPs imputed from hair shaft can be used to distinguish East Asian, European and African population.

Key words hair shaft proteome, single amino acid polymorphisms, non-synonymous single nucleotide polymorphism, ancestry inference

DOI：10.16476/j.pibb.2018.0179

---

*This work was supported by grants from The National Natural Science Foundation of China（81801877）, National Key R&D Program of China（2017YFC0803501）, Basic Research Project Grant（2017JB025）and Beijing Leading Talent Program（Z181100006318006）.

**Corresponding author

Received：June 28，2018 Accepted：December 5，2018