Science Discovery
2017; 5(4): 257-261
http://www.sciencepublishinggroup.com/j/sd
doi: 10.11648/j.sd.20170504.13
ISSN: 2331-0642 (Print); ISSN: 2331-0650 (Online)

The Application of Next Generation Sequencing in Forensic DNA Database

Zhao Zhao*, Wang Le, Li Sheng, Wang Tong, Liu Bing

Institute of Forensic Science of Ministry of Public Security, Beijing, China

Email address:

zhaozhao@cifs.gov.cn (Zhao Zhao), wangle_02@163.com (Wang Le), lisheng@cifs.gov.cn (Li Sheng), wangtong@cifs.gov.cn (Wang Tong), liubing@cifs.gov.cn (Liu Bing)

*Corresponding author

To cite this article:

Zhao Zhao, Wang Le, Li Sheng, Wang Tong, Liu Bing. The Application of Next Generation Sequencing in Forensic DNA Database. *Science Discovery*. Vol. 5, No. 4, 2017, pp. 257-261. doi: 10.11648/j.sd.20170504.13

Received: April 5, 2017; **Accepted**: May 12 2017; **Published**: May 20, 2017

Abstract: With the developing of standardized forensic DNA profiling and information technology, the forensic DNA database has achieved a big success. By managing、searching and sharing through DNA database, the molecular biology and genetics play a crucial role in individual identification, in the meantime, the DNA database also promote the developing of large-scale forensic DNA testing. In recent years, the arising of second generation sequencing(SGS) has substantially changed the way of genetics research, and varies of platform、kit and system has been developed continuously, many of which has been applied in forensic laboratories. So for,the China National DNA Database has expanded to be the largest one in the world, therefore, the SGS needs to incorporated with the DNA database rather than instead of it. So, the compatibility of SGS-based data with massive PCR-CE data in the database is now becoming a critical problem.in this article, the China National DNA Database is introduced, the advantages of SGS and its possible role in the forensic DNA profiling and databasing are elaborated, and the consistency in data storage and the compatibility in comparing of the SGS-based data within the present DNA database are also discussed.

Keywords: Forensic DNA Profiling, DNA Database, Second Generation Sequencing, Data Compatibility

二代测序技术在法庭科学DNA数据库中的应用探讨

赵钊*，王乐，李盛，王彤，刘冰

公安部物证鉴定中心，北京，中国

邮箱

zhaozhao@cifs.gov.cn（赵钊），wangle_02@163.com（王乐），lisheng@cifs.gov.cn（李盛），wangtong@cifs.gov.cn（王彤），liubing@cifs.gov.cn（刘冰）

摘要：随着标准化的法医DNA分型技术及计算机技术的不断成熟和完善，DNA数据库得到了极大的发展。通过DNA数据库进行海量数据的存储、检索和共享，既能够发挥生物遗传学在个体识别上的作用，同时又反过来推动了法医DNA规模化检验技术的发展。近年来，二代测序技术兴起，对遗传学的研究方式产生了巨大影响，各种平台、试剂与技术方法不断涌现，并逐渐进入法庭科学领域。目前，中国公安机关DNA数据库作为世界第一大DNA数据库，已臻成熟，二代测序技术要深入发展，需要借助于DNA数据库推广应用，两者的兼容性问题凸显。本文介绍了中国公安机关DNA数据库的现状，综述了二代测序的技术优势及其在法医DNA检验与DNA数据库上的应用，并探讨了二代测序数据与中国DNA数据库的数据存储一致性和数据比对兼容性方案。

关键词：法医DNA分型，DNA数据库，二代测序，数据兼容

1．引言

生物技术的快速发展催生出各类生物信息学数据库，几乎覆盖了生命科学的各个研究领域，序列、结构、分型等多种类型的数据呈现指数级增长，日渐庞大的数据资源也为生物学领域科研进展提供了统计来源和分析基础[1-4]。其中广为应用的DNA数据库实现了DNA信息数字化组织、存储、管理和检索，成为主流生物信息数据库之一。随着各项法医DNA分析技术的成熟，信息化建设的不断跟进，法庭科学DNA数据库已成为刑事技术中最为高效精准的打击犯罪、维护公共安全的手段。法庭科学DNA数据库发展迅速，呈现出样本分布广，基础数据量大，数据类型不断增多，附属信息全面等优势。从其发展历程来看，法庭科学DNA数据库的应用与生物DNA技术的发展紧密相随，相辅相成。那么随着二代测序技术和平台日趋完善，SNP、mRNA、全基因组测序等众多新兴生物遗传技术的发展，DNA数据库下一步的发展也应该逐步结合新技术，从而提供更好的数据支撑与检索应用效果。

2．中国公安机关DNA数据库

出于社会安全和打击犯罪考虑，在国家层面的大力支持下，法庭科学DNA数据库现已成为主流DNA数据库中涉及人员样本最广，也最为公众熟知的数据库[5-6]。美国、英国、新西兰、日本等国家均建有法庭科学DNA数据库，其中美国CODIS系统在法庭科学DNA数据库建设领域影响深远，应用范围广，目前已有1200余万数据[7-9]。自90年代中期中国公安部提出“统一规划、统一标准、分步实施、滚动发展”的DNA数据库建设原则开始[10-12]，从科学研究到实际应用，从区域性建设到全国性部署，中国的法庭科学DNA数据库已经容纳了5300万以上的短串联重复序列（short tandem repeat,STR）数据，是目前世界最大的DNA数据库，仅2016年新增数据一千万余份，数据基础及活跃程度均令人瞩目。随着检测试剂耗材的费用下降及各地应用逐步规模化，同时为了降低海量数据比对的随机匹配概率，数据库中的数据质量也逐渐提高，主流新入库数据，尤其是建库数据，基因座数量已发展到15个及以上。公安部最新版本DNA数据库升级后，稀有位点、混合分型均可以入库应用。同时，Y-STR家系排查在各地公安机关也开展应用，Y-STR技术在“白银市连环杀人案”、“诸暨市多起珠宝抢劫杀人案”等一些多年未破的积案中的关键作用也促进了部分地区Y-STR、线粒体DNA等数据的入库。

DNA数据库已经成为中国刑事技术领域最重要的技术支撑，除了在精确打击犯罪方面不可替代的作用外，逐渐在社会管理、交通民事、灾难处置等多方面体现了重要价值。随着数据质量的提高和数据种类的增多，DNA数据库的应用模式也需要进一步拓展，不仅仅局限于简单的STR比中，综合多种生物信息进行主动排查研判也是数据库的发展趋势之一。

3．二代测序技术与DNA数据库

二代测序（Second Generation Sequencing, SGS，也称新一代测序）技术凭借其通量高、成本低、速度快等特点，迅速在生命科学的各个领域广泛应用，自2007年起发表的相关科研论文数量逐年递增，科研投入逐年增加。近几年尤其是2014年以来，二代测序也逐步进入了法医个体识别领域，并迅速成为国际法医领域的一个热门方向，该技术在法医学上的实际应用也急速增加，并不断有大量的国际期刊和法医专业会议对其进行报道[13]。国外法医学专家预测二代测序技术将代替现有的基于毛细管电泳的DNA检测成为未来主流DNA检测技术[14]。在中国已有多家司法鉴定机构开展了二代测序的研究应用，其中公安部物证鉴定中心已经开展了常染色体STR分型、线粒体全基因组等方面的二代测序平台的研究和应用。鉴于二代测序技术在法医DNA检测领域的深入推广和应用，DNA检测数据的形式也发生了根本改变，不再是单一的STR数据，信息量越来越大，详尽的生物信息对嫌疑人的刻画和案件的侦破能够提供更加强有力的支持。同时由于生物数据格式更为复杂，对于DNA数据库系统的存储和比对都提出了新的要求。

3.1．DNA数据库结合二代测序的优势

二代测序技术与片段分析技术比较具有明显的技术优势。第一，二代测序技术从碱基序列差异水平上获得了个体的遗传信息，能得到更为精细化的分析结果，充分提高STR基因座的个体识别率。目前，中国公安机关DNA数据库数据总量逾5300万，且数据量还在不断增长，而对于5000万数据，如使用11个共有基因座，随机匹配数量将接近200个[8]。尽管近两年来样本分型入库基因座数量多数提高到15个以上，全库数据中人员样本低于12个基因座的仍占比1.90%，物证样本低于12个基因座的占比4.31%，详见表1。这部分数据在不断攀升的数据总量基础上随机匹配概率偏高，识别率不高。尤其是新版国家库取消物证分型入库基因座限制后，低位点物证分型也将进入国家库。

表1 数据库中不同基因座数量的样本占比。

样本	基因座个数分类				
	<9	9~11	12~13	14~15	>15
人员	0.04%	1.86%	1.85%	39.33%	56.92%
物证	0.26%	4.05%	2.55%	82.22%	10.92%

二代测序数据的序列化结果，能以更少的基因座有效区分不同个体，可以有效提高基因座不足数据的识别率。第二，二代测序技术只需要少量DNA样品（一般1ng）就可以同时获得常染色体STR、Y-STR、X-STR、线粒体全基因组DNA等现有的所有检测数据，对于特殊案件中微量检材或降解样品的分析处理有明显优势。第三，应用二代测序可以区分混合检材中的个体。目前，DNA数据库中混合分型仅能入库，没有有效拆分比对功能，如果结合二代测序技术进行有效辨别，再利用数据库比对可以更好地解决疑难案件。第四，二代测序数据有助于提高亲缘比对与

家系推断效率。当前DNA数据库中的有效亲缘比中率偏低，不足1%。由于区分度不足，导致大量配偶-子女与人员样本的比中信息，一对配偶-子女动辄比中成千上万个体，跟本无法从基因信息上进行有效判定。另外，基于中国特殊国情建设的“全国公安机关查找被拐卖/失踪儿童DNA数据库”也亟需有效的亲缘判定技术，尤其是针对日益累积的单亲样本，很难从现存的常染色体STR比对上有所突破。引入二代测序结果以及更多的遗传标记，如 Y-STR、线粒体全基因组DNA等均可以有效缩小排查范围。特别是目前在中国，法庭科学DNA实验室普遍对线粒体DNA技术应用较少，而二代测序技术可以提升线粒体DNA的识别率，从而提高线粒体DNA技术的利用率，充分发挥线粒体DNA在微量检材或亲源认定等案件调查中的独特优势。

3.2．二代测序技术在DNA数据库中的应用模式

尽管二代测序技术有诸多技术优势，但是毕竟是近年于法庭科学领域兴起，囿于技术方法在法医DNA实验室开展不足，检测费用仍达不到规模化应用水平等，短期内并不能替代主流的片段分析方法。同时，已经建成的海量DNA数据库显然不能弃之不用，重建势必造成巨大的浪费。因此，在已有DNA数据库基础上逐步引入二代测序数据，寻求二者的兼容性才是可行之道。

目前在二代测序领域的研究主要是集中在测序技术本身的发展以及相关试剂、平台的验证等方面，在数据比对方面都鲜有报道，在二代测序技术与DNA信息化的结果方面探索较少。而法庭科学领域开展二代测序技术往往是希望应用于实战，那么实战应用离不开数据共享和比对。生物技术的发展为进一步应用提供了新的思路和技术保证。紧跟DNA检验技术发展的趋势，为下一代DNA数据库打下基础也是法庭科学DNA信息化应用的发展方向。当前，中国各级公安机关在现有DNA数据库上的投入逐年增加，即使传统测序技术逐渐为下一代测序技术替代，当前的DNA数据库也必然保留沿用，因此解决兼容问题，形成综合性的生物信息数据库才是科学合理的应用模式。这就需要首先从二代测序技术上与传统片段分型方法进行技术接轨，解决序列多态性一定程度上转化为长度多态性的问题，其次从信息技术方面，解决二代测序数据在现有DNA数据库中综合存储和兼容比对的问题。

4．二代测序数据与DNA数据库的兼容初探

目前，法医遗传学相关二代测序研究较为集中的领域有STR分析、SNP、线粒体全基因组等。鉴于法庭科学DNA数据库尚未引入SNP，本文只探讨与STR和线粒体DNA数据的兼容。其中线粒体全基因组测序结果与目前法庭科学实验室通常采用的高变区结果，从数据角度看只是碱基位置的长度范围变化。因此二代测序线粒体全基因组数据与目前DNA数据库中线粒体DNA数据从计算机系统处理上完全兼容，即以位置-值的格式存储比对即可。

如果说线粒体DNA\SNP在法庭科学DNA技术实战应用中只是重要补充，那么STR分型是个体识别中最为核心的技术，因此现有DNA数据库引入二代测序技术的突破在STR分型上的兼容。从二代测序技术上来看，将其引入常染色体STR核心基因座的研究已经初见成效。已有多个检测平台如Roche/454、Thermo Fishier 的Ion PGMTM、Illumina 的MiSeq FGxTM等都可以进行核心基因座的STR分型[15-17]。在法庭科学领域，部分法医DNA实验室渐次投入相关科研和验证工作，并已有利用二代测序进行案件检材的STR分型[18-19]。可预见的是，二代测序STR分型方法会日趋成熟，越来越多的进入法医DNA检验中，二代测序数据也会逐渐增多，这些数据只能散乱于各地，不能有效保存和利用。而相关样本的结果比较也只能以人眼逐一核对，在数据量日益增多的情况下，势必影响工作效率，就二代测序数据本身也需要有数据库可以进行管理、比对和数据共享。而二代测序在法庭科学领域的有效应用更需要纳入现有DNA数据库系统中，进而发挥其个体识别作用。

解决与DNA数据库兼容问题，其核心是存储与比对。以下从数据结构与比对策略两方面对常染色体STR分型的二代测序结果和原有片段分析数据的兼容做初步的技术性探讨。

4.1．数据存储

当前正在部署上线的新版国家公安机关DNA数据库中常染色体STR的数据结构采用JSON结构，即由“名称：值”键值对表示对象，由“,”分割数据，由“{}”保存对象。数据结构主体为{key1:value1,key2:value2,…}，形成可以任意扩展的键值集合，值也可以为任意类型。在DNA数据库中STR分型存储结构如下所示：

{"AMELOGENIN":{"X/Y"},
"D8S1179":{"14/16"},
"D21S11":{"28/33.2"},
"CSF1PO":{"10"},
"D3S1358":{"16/17"},
"TH01":{"6/9"},
"D13S317":{"8/11"},
"D16S539":{"11/12"},
"D2S1338":{"18/23"},
"D19S433":{"13"},
"D7S820":{"8/11"},
"VWA": {"14/17"},
"TPOX": {"8/11"}
}

该结构避免了旧版数据库中以一个字段存一个位点的局限性，存储长度灵活可变，具有很好的扩展性和兼容性。二代测序数据完全可以沿用JSON结构存储，只需在原有等位基因键值对的基础上扩展成“等位基因：值”与“序列：值”的集合{allele:value,seq:value}即可，可以很好的解决二者存储结构一致性的问题，示例如下：

{"AMELOGENIN
":{"allele":"X/Y","seq":"[G]1/[A]1"},
"D8S1179":{"allele":"14/16","seq":"[TCTA]1[TCTG]1[TCTA]12/[TCTA]2[TCTG]1[TCTA]13 "},
"D21S11":{"allele":"28/33.2","seq":"[TCTA]4[TCTG]6[TCTA]3[TA]1[TCTA]3[TCA]1[TCTA]2[TCCA]1[TA]1[TCTA]10/[TCTA]5[TCTG]6[TCTA]3[TA]1[TCTA]3[TCA]1[TCTA]2[TCCA]1[TA]1[TCTA]13[TA]1 [TCTA]1"},

"CSF1PO":{"allele":"10","seq":"[AGAT]10"},
"D3S1358":{"allele":"16/17","seq":"[TCTA]1[TCTG]2[TCTA]13/[TCTA]1[TCTG]3 [TCTA]13"},
"TH01":{"allele":"6/9","seq":"[TCAT]6 /[TCAT]9"},
"D13S317":{"allele":"8/11","seq":"[TATC]8/[TATC]11"},
"D16S539":{"allele":"11/12","seq":"[GATA]11/[GATA]12"},
"D2S1338":{"allele":"18/23","seq":"[TGCC]6[TTCC]12/[TGCC]7[TTCC]13 [GTCC]1 [TTCC]2"},
"D19S433":{"allele":"13","seq":"[AAGG]1[TAGG]1[AATT]11"},
"D7S820":{"allele":"8/11","seq":"[GATA]8 /[GATA]11"},
"VWA":{"allele":"14/17","seq":"[TCTA]1[TCTG]1[TCTA]1[TCTG]4[TCTA]3[TCCA]1[TCTA]3 /[TCTA]1[TCTG][TCTA]12"},
"TPOX":{"allele":"8/11","seq":"[AATG]8 /[AATG]11"}
}

4.2．数据比对

在数据结构一致的基础上，需要进一步扩展样本基因信息的比对算法。二代测序结果包括数字型的等位基因值和字母型的基因序列两部分信息，从比对上同样可以兼容原有STR分型数据。首先，二代测序数据与DNA数据库中片段分析STR数据比对，在算法上与两个片段分析STR数据比对完全相同，仅需提取等位基因键值对逐一比对即可。其次，两个二代测序数据之间比对方法如下：

步骤1：进行等位基因键值对的比对，如果两个样本的等位基因值容差数超过DNA数据库中设置的容差上限，则跳出；否则，再进入步骤2进行基因序列的比对。

步骤2：

（1）获取样本1和样本2等位基因值完全相等的基因座，记为集合L。

（2）从L中取一个基因座Li，以D8S1179为例：

样本1："D8S1179":{"allele":"14/16","seq":"[TCTA]1[TCTG]1[TCTA]12/[TCTA]2[TCTG]1[TCTA]13 "}

样本2："D8S1179":{"allele":"14/16","seq":"[TCTA]2[TCTG]1[TCTA]11/[TCTA]2[TCTG]1[TCTA]13 "}

（3）获取基因座Li的所有等位基因值，记为A，如"14，16"；

i．从A中取出一个等位基因值，如14，记其对应的样本1的基因序列为seq1，记其对应的样本2的基因序列为seq2，即：

seq1: [TCTA]1[TCTG]1[TCTA]12

seq2: [TCTA]2[TCTG]1[TCTA]11

ii．将seq1和seq2基因序列全部展开，进行序列化比对：

如序列完全相同则跳出本次循环，否则记录不同碱基的位置与值。

注：在序列展开时针对缺失、插入的情况，可插入设定字符如“-”以保证碱基序列的可比次序，如seq3: TCTATCTATCT-TCTATCTA……TCTA

iii．将初始比对结果再次合并，提高最终比对结果展示的可读性，最终可标示如下：

样本1: "D8S1179" 14 [TCTA]1*[TCTG]1[TCTA]1*[TCTA]11

样本2: "D8S1179" 14 [TCTA]1*[TCTA]1[TCTG]1*[TCTA]11

5．结论与展望

二代测序从技术原理上形成了重大的突破，也使法医遗传学进入了一个新的发展阶段，大大丰富了传统法医DNA技术手段，从多方面提高了法医DNA个体识别率，并且能够在特殊案件中挖掘更多生物信息主动锁定嫌疑人。从二代测序技术优势及其发展趋势来看，将逐步进入法庭科学领域扩展应用。在当前信息化时代，从10几年来法庭科学DNA检验技术与DNA数据库的发展历程与现状可以看出，未来二代测序技术如果在法医DNA领域常规化推广应用，必然需要通过DNA数据库进行数据处理和比对共享。本文以中国目前部署上线的国家公安机关DNA数据库系统为基础，从技术上初步探讨了二代测序STR分型（以部分基因座为例）、线粒体全基因组数据与DNA数据库的兼容方案。当然，DNA数据库引入二代测序数据具备可行性，并不意味着二代测序技术能够即刻常规应用于法庭科学DNA检验入库。第一，从法医学应用角度，二代测序的性价比较低，检测成本并不能被广泛接受。第二，能否直接被DNA数据库采用，首要因素是其技术解决方案必须成熟完整[7]。然而各种二代测序平台都不可避免地存在一定比例的错误率，并且针对STR分型的法庭科学数据分析方法尚不成熟。最后，DNA数据库是全国DNA实验室数据共享的平台，检测结果数据需要在实验室间通行使用，二代测序涉及的命名规则、结果解读必须统一。因此，鉴于法庭科学DNA数据库在打击犯罪、公共安全、灾难事故处置等各个领域的重要性，在相关新技术的应用上需要科学规划、深入探讨、谨慎验证。在解决好数据库存储一致性和兼容比对的基础上，在不影响已有数据库数据应用的前提下，如利用备份数据或者小范围数据，逐步在部分应用场景引入二代测序数据，充分利用新技术方法的优势，在特殊案件中进行综合研判，可能成为两者结合应用的良好开端。

致谢

本文为公安部科技强警基础工作专项项目《法医DNA二代测序数据批量比对关键技术研究》(2016GABJC18)的阶段性成果之一。

参考文献

[1] Cantor C R and Lim H A. Electrophoresis, Supercomputing and the Humangenomes [M]. Investigative Genetics, 2013, 4: 22.

[2] 张晓东，张传富等。生物信息学数据库研究进展［J］. 生物信息学，2006年，4（3）：143-145。

[3] 孙磊，胡学龙等。生物医学大数据处理的云计算解决方案［J］. 电子测量与仪器学报，2014年，28(11)：1190-1197。

[4] 张晓辉，高晓玲。基于生物信息学数据库的研究与应用［J］. 中国科技信息，2010年，22：44-45。

[5] 姜先华。中国DNA数据库建设应用技术现状及发展趋势［J］. 中国法医学杂志，2011年，26（5）：383-386。

[6] Richard Hindmarsh,Barbara Prainsack.Global Governance of Forensic DNA Profiling and Databasing[M]. UK,2010:1-11.

[7] 刘冰. 现阶段我国DNA数据库发展的几个关键问题［J］. 刑事技术，2015年，40（4）：318-323。

[8] 葛建业，严江伟等. 关于法庭科学DNA数据库若干问题的探讨［J］. 中国法医学杂志，2011，26(3):252-255。

[9] 焦文慧，宋辉. 英美国家犯罪DNA数据库建设及应用［J］. 上海公安高等专科学校学报，2013，23(2)：86-91。

[10] 胡兰，陈松，张国臣. 国家法庭科学DNA数据库建设势在必行［J］. 刑事技术，2003(6):3-5。

[11] 焦章平，唐晖，刘雅诚等. 建立法医DNA 数据库的初步探讨［J］. 中国法医学杂志，2003,18(1):58-59。

[12] 侯一平，王保捷，丛斌，等. 中国法医学会物证专业委员会法医 DNA 分析的若干建议[J］. 中国法医学杂志，2006,21(5):257-259。

[13] 王乐，叶健，白雪等。二代测序技术及其在法医遗传学中的应用，刑事技术，2015，40(5):353-358。

[14] Lutz Roewer. DNA fingerprinting in forensics: past, present, future [J]. Investigative Genetics, 2013,4:22.

[15] Van Neste C, Van Nieuwerburgh F, Van Hoofstat D, et al. Forensic STR analysis using massive parallel sequencing [J]. Forensic Sci Int Genet, 2012, 6: 810: 818.

[16] Bornman DM, Hester ME,Schuetter JM, et al Short-read, high-throughput sequencing technology for STR genotyping [J]. Biotech Rapid Dispatches, 2012: 1-6.

[17] Fordyce SL, Mogensen HS, Borsting C, et al. Second-generation sequencing of forensic STRs using the ion torrent HID STR 10-plex and the Ion PGM [J]. Forensic Sci Int Genet, 2015, 14: 132-140.

[18] Scheible M,Loreile O, et al. Short tandem repeat typing on the 454 platform:strategies and considerations for targeted sequencing of common forensic markers [J]. Forensic Sci Int Genet, 2014, 12: 107-119.

[19] Sarah L.Fordyce, Helle Smidt, et al. Second-generation sequencing of forensic STRs using the Ion Torrent TM HID STR 10-plex and the Ion PGM TM [J]. Forensic Sci Int Genet, 2015, 14: 132-140.