

亲缘关系预测系统

王鑫¹, 李悦², 文豪¹, 范虹^{1*}, 刘京^{2*}

(1. 陕西师范大学 计算机科学学院, 中国陕西 西安 710119; 2. 公安部物证鉴定中心 北京市现场物证检验工程技术研究中心 现场物证溯源技术国家工程实验室, 中国北京 100038)

摘要:报道了一个基于全基因组数据的亲缘关系预测系统(Kinship Prediction System Version 1.0, KPS v1.0), 旨在给定一个自定义群体中所有个体的全基因组数据文件, 通过亲缘关系预测算法, 计算出群体中所有存在的亲缘关系对, 并将结果可视化为家谱图, 进而搜寻某个未知个体与多个已知个体的复杂亲缘关系, 为物证鉴定提供强有力的线索。系统不仅将以往复杂的预测流程变得简单化、自动化, 而且将预测结果进行了界面化, 并和用户建立了更好的交互性, 大大地提升了运算效率, 在人类遗传学、法医遗传学等领域有重要的应用价值。

关键词: 法医遗传学; 亲缘关系预测; 亲缘关系预测系统(KPS v1.0); 生物信息学

中图分类号: D919.2, Q3-3

文献标识码: A

文章编号: 1007-7847(2020)03-0229-05

Kinship Prediction System

WANG Xin¹, LI Yue², WEN Hao¹, FAN Hong^{1*}, LIU Jing^{2*}

(1. College of Computer Science, Shaanxi Normal University, Xi'an 710119, Shaanxi, China; 2. National Engineering Laboratory for Forensic Science, Beijing Engineering Research Center of Crime Scene Evidence Examination, Institute of Forensic Science, Beijing 100038, China)

Abstract: The Kinship Prediction System Version 1.0 (KPS v1.0) based on genome-wide data aims to give the genome-wide data file of all individuals in a custom population. It uses the kinship prediction algorithm to calculate all existing kinship pairs in the population, makes the results visualized as a family tree, and then searches for the kinship between unknown individual and multiple known individuals, so as to provide strong clues for material evidence identification. KPS v1.0 not only simplifies and automates the previous complex prediction processes, but also interfaces the prediction results, establishing better interactivity with users, and greatly improving the computing efficiency. It has important practical values in the field of human genetics and forensic genetics.

Key words: forensic genetics; kinship prediction; Kinship Prediction System Version 1.0 (KPS v1.0); bioinformatics

(Life Science Research, 2020, 24(3): 229~233)

目前, 我国亲缘关系鉴定大多使用常染色体短串联重复(short tandem repeat, STR)遗传标记位点^[1-3]。在刑事侦查领域, 通过对 Y-STR 进行检验, 可以开展家系排查和辅助父系亲缘鉴定, 从而为

案件侦查提供线索^[4]。遗憾的是, Y-STR 只能针对父系亲属, 无法对母系亲属进行排查^[5], 而近年来半同胞、舅甥、姑侄等复杂亲缘关系鉴定诉求不断增加, 现有的鉴定方法已远远无法满足物证鉴定

收稿日期: 2019-12-15; 修回日期: 2020-04-12

基金项目: 国家重点研发计划资助项目(2017YFC0803501); 国家科技资源共享服务平台计划资助项目(YCZYPT[2017]01-3); 陕西省重点研发计划项目(2018SF-251); 公安部 2019 年度“双十攻关”项目(2019SSJH0601)

作者简介: 王鑫(1996—), 男, 陕西铜川人, 硕士研究生; *通信作者: 范虹(1969—), 女, 宁夏平罗人, 博士, 陕西师范大学教授, 主要从事模式识别与智能信息处理研究, Tel: 029-85310161, E-mail: fanhong@snnu.edu.cn; 刘京(1989—), 男, 山东沂水人, 硕士, 法医, E-mail: biojing@yeah.net。

的需要。随着全基因组测序技术的发展,使用全基因组单核苷酸多态性(single nucleotide polymorphism, SNP)数据进行亲缘关系预测的技术应运而生并快速发展^[6-7],解决了当下遇到的这一难题。该技术旨在通过亲缘关系预测算法,基于全基因组 SNP 测序数据计算某个群体中所有存在的亲缘关系对。

早在 2011 年这些亲缘关系预测算法就已经有报道^[8],且这些软件如 Plink/KING 均为国外免费开源软件。Plink 是一款用于基因组关联分析的开源软件,具有基因组数据格式转换、基础基因组信息统计、LD (linked dimorphisms)计算、共祖片段 IBD (identical by descent)距离计算等功能^[9],在亲缘关系预测中主要使用该软件进行数据归一化处理,为后续算法软件提供原始输入文件。KING 是一款使用全基因组 SNP 数据精确推断任何一对个体亲缘关系等级的软件,可计算高达 4 级的亲缘关系,支持接受 Plink 转换后的二进制格式文件^[10]。

然而,这一技术的原有预测流程十分繁琐,需要对全基因组数据进行预处理,并使用 Plink 软件将文件转化为二进制,再使用 KING 软件计算出亲缘关系等级;如果需要知道其实际亲属关系,还要按照亲缘关系等级判断方法逆向还原。鉴于此,本文提出并开发了一套具有极强交互性的分析系统——亲缘关系预测系统(Kinship Prediction System Version 1.0, KPS v1.0),该系统的处理流程界面化、自动化、简单化,可以自动预处理全基因组数据,并将最终结果自动绘制成家谱图。

KPS v1.0 极大地简化了亲缘关系预测的流程,不需要操作人员熟练掌握数据处理脚本语言和生物信息学软件,显著降低了操作难度,提高了分析效率,为物证鉴定工作提供了更好的服务,在人类遗传学、法医遗传学等领域有重要的应用价值。

1 KPS v1.0 简介

1.1 系统开发平台和运行环境

软件开发平台为 Ubuntu 18.04 操作系统,Python 语言 v3.6 版本,运行环境为安装了浏览器的 Windows 7 及其以上操作系统。

1.2 系统设计思路

KPS v1.0 亲缘关系模块主要涉及前端界面和后台计算两大部分。后台计算主要是基于 Python 语言将 4 个功能进行模块化串联设计并编程,包括输入数据格式预处理、基于 Plink 软件的数据二进制转换、基于 KING 软件的亲缘关系等级计算、亲缘关系等级与实际亲属关系的转换等。前端界面包括了批次名称录入、数据录入、数据类型选择、检测平台选择、亲缘关系结果查看等功能。具体结构如图 1 所示。后端部分的模块化串联设计可以降低程序的耦合度^[11],简化程序设计、调试和维护等操作,前端界面易于用户使用和直观查看结果。

1.3 系统实现

1.3.1 亲缘关系预测方法

亲缘关系预测是通过计算全基因组 SNP 数据的共祖片段大小判断亲缘关系的远近^[12]。KPS v1.0 系统主要基于 KING 软件计算亲缘关系等级,并

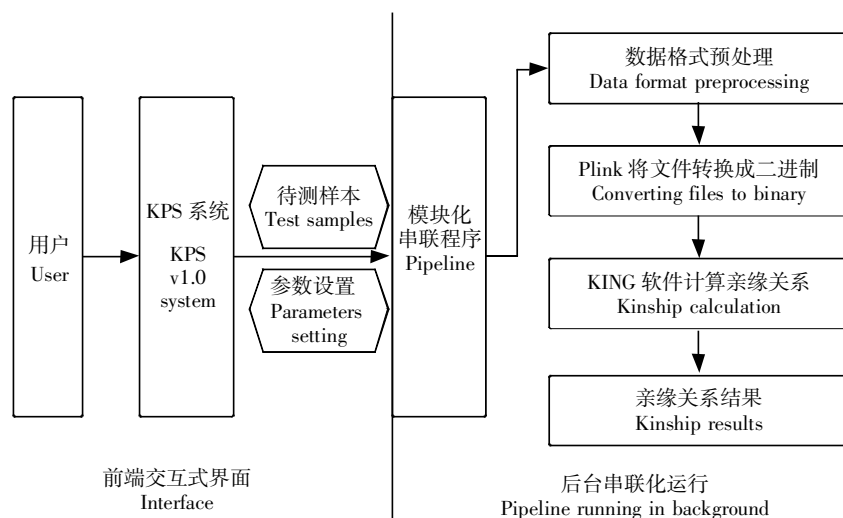


图 1 亲缘关系模块流程图

Fig.1 The kinship module flowchart

根据亲缘关系等级图推导其实际亲属关系。本系统中亲缘关系分为 4 级, 亲缘关系等级图如图 2 所示。亲缘关系等级判断方法: 父母子女之间的亲缘关系为一级, 两两个体亲缘关系等级为每个个体到共祖父母亲缘关系的加和^[13]。

1.3.2 亲缘关系等级计算

当系统输入完成后, 前端会将输入数据传递给后台, 后台只需新建一个 Celery^[14]任务, 并将其放置到 Redis^[15]任务缓冲队列末尾, 此时多核 CPU 会同时从队列中读取任务并执行, 用户提交完任务即可随时在任务列表中查看任务状态。任务处理示意图如图 3 所示。

每一个 Celery 任务都会定义一个亲缘关系计算函数, 并保存该任务的输入数据。亲缘关系计算函数首先根据批次数据类型不同调用不同

的接口, 并基于 pandas 库^[16], 将批次文件与相应的检测平台索引文件转化为 MAP 和 PED 格式^[17]文件, 接着自动调用后台脚本(集成了 Plink 和 KING 软件), 基于数据预处理结果计算得到本批次亲缘关系等级对, 并将其写入服务器指定文件。

1.3.3 系统输入与输出

系统输入需要用户填写新建任务的批次名, 选择批次文件的数据类型和检测平台, 并上传批次压缩包文件。其中检测平台有两种, 对应两种 SNP 索引文件, 每个文件保存多个 SNP 位点信息。数据类型也有两种: 个体数据和原始数据。原始数据为个体基于某种检测平台测序直接得到的碱基对序列文本; 个体数据为个体基于某种监测平台测序得到的基因型表格文件, 直观展示了个体在某个 SNP 位点处的基因型。上传的批次压缩

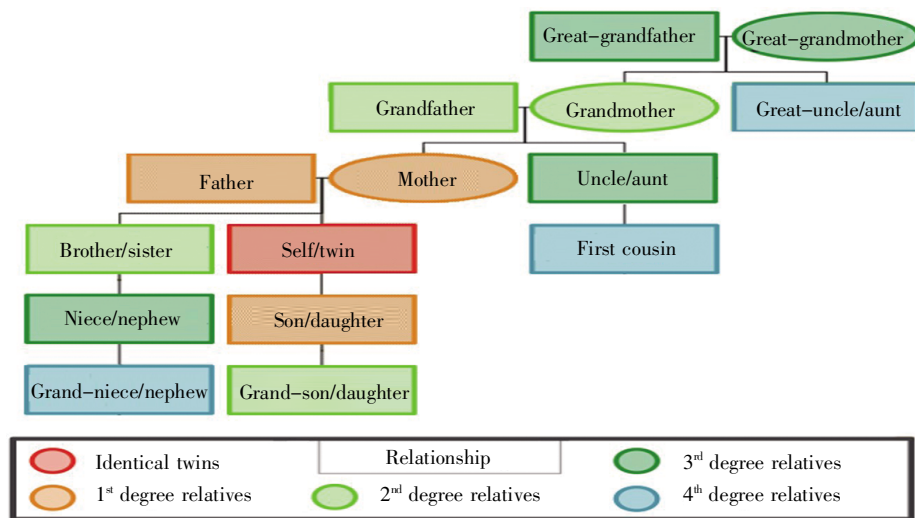


图 2 亲缘关系等级图

Fig.2 The kinship degree tree

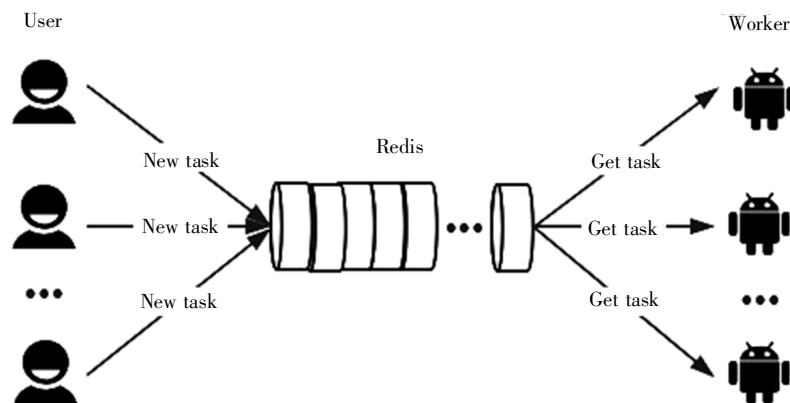


图 3 任务处理示意图

Fig.3 Task processing diagram

包文件包含多个个体的测序数据,所有测序数据格式保持一致。

任务计算成功后,用户可在页面查看亲缘关系对列表,列表中显示批次中所有存在的亲缘关系对及其对应的亲缘关系等级,点击某一对亲缘关系,系统会自动基于亲缘关系等级图计算这对样本可能存在的实际亲属关系,并将结果绘制为家谱图供用户直观查看。

1.4 系统特点

KPS v1.0 包含以下 5 个方面的特点: 1) 支持导入群体原始基因型数据; 2) 自动识别输入批次内样本个数, 若样本个数小于 2, 则运行失败; 3) 若计算时发生异常, 如用户新建任务时选错数据类型或检测平台, 任务直接会显示运行失败; 4) 可一次性提交多个批次进行计算, 自动识别电脑 CPU 核数, 将任务自动分配到单个 CPU 计算, 提升运算效率; 5) 可基于亲缘关系等级计算结果自动绘制家谱图供用户直观查看。

2 系统应用

2.1 样本来源

测试样本共计 37 份, 来自实验室 4 名志愿者家系, 涵盖 1~4 级亲缘关系。所有样本提供者均签署知情同意书且实际亲缘关系均基于样本来源者自述。

2.2 系统准确性评估

将 4 名志愿者家族的实际亲缘关系分别绘制家谱图, 与系统预测结果进行对比, 比较结果如图 4 所示。系统对 A、B、C、D 4 个志愿者家族的 163 对($C_{11}^2 + C_{12}^2 + C_7^2 + C_7^2 = 163$)亲缘关系进行了计算, 预测准确率分别为 92.73%、93.94%、100%、90.48%, 平均准确率达到 94.29%, 其中一级亲缘关系和四级亲缘关系的预测准确率均为 100%。虽然将 8 对二级亲缘关系预测为三级亲缘关系, 2 对三级亲缘关系预测为四级亲缘关系, 但对总预测的

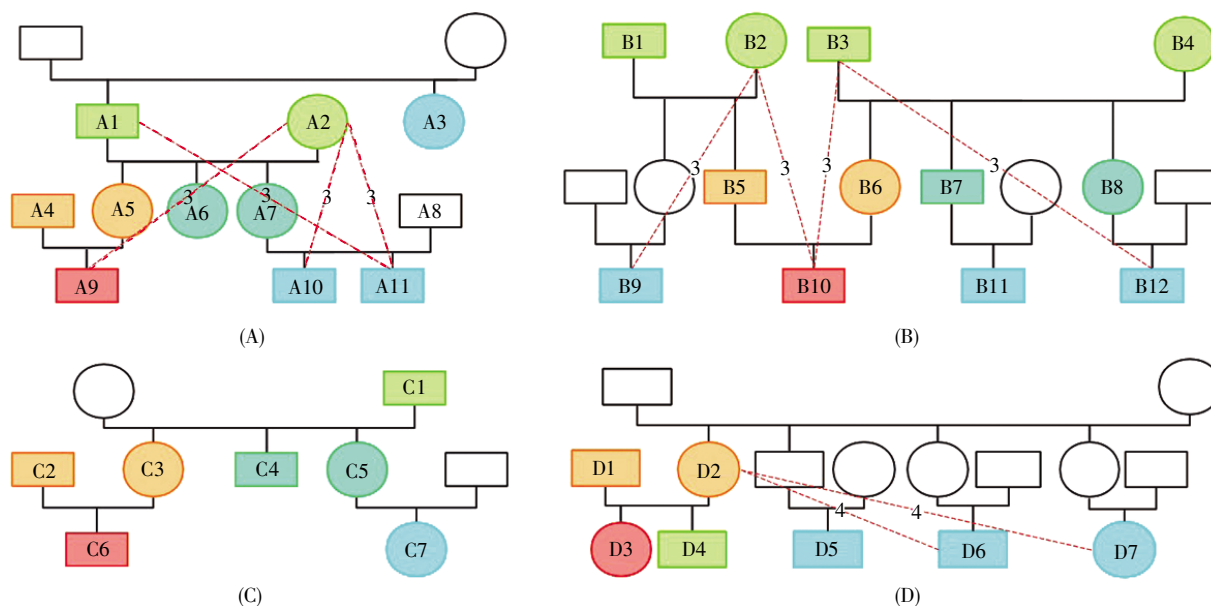


图 4 实际亲缘关系与预测亲缘关系对比图

(A) 志愿者 A; (B) 志愿者 B; (C) 志愿者 C; (D) 志愿者 D。图中方框和圆圈分别代表家族中的男性和女性, 检测的个体已为其编号, 未检测的个体未编号。不同颜色代表不同的亲缘关系等级(参考图 2 的亲缘关系等级图), 红色标注的个体代表本人, 未用颜色标注的个体为未检测的或与本人无实际亲缘关系的个体。虚线表示实际亲缘关系与系统预测亲缘关系不相符的亲缘关系对, 其上面的数字代表系统预测的亲缘关系等级。

Fig.4 The comparison between actual kinship and predicted kinship

(A) Volunteer A; (B) Volunteer B; (C) Volunteer C; (D) Volunteer D. The boxes and circles represent the males and females, respectively. The individuals with numbers are tested, and the ones without numbers are not tested. Different colors represent different kinship degrees which are consistent with Fig.2. The individuals in red mark represent themselves, and the unmarked ones are undetected or have no actual kinship with the red-marked individual. The dotted line indicates the kinship pair whose actual kinship doesn't match the kinship predicted by the system, and the number above it represents the kinship degree predicted by the system.

表1 系统与人工预测所需时间的比较
Table 1 Time comparison between system and manual prediction

Batch No.	Sample size	Kinship pairs	KPS v1.0 time	Manual time (about)
1	50	29	1 min 21 s	5 h
2	100	67	2 min 57 s	11 h
3	200	126	5 min 18 s	21 h
4	400	233	9 min 49 s	40 h
5	700	401	17 min 33 s	69 h
6	1 000	682	24 min 12 s	117 h

163 对亲缘关系而言, 错误率仅为 6.13%, 表明亲缘关系预测系统具有可观的准确性。

2.3 系统效率评估

在不同样本数下, 对系统的计算时间和原始人工预测方法所需时间进行了统计, 结果见表 1。从预测所需时间来看, 系统将预测效率提升了 200 多倍。同时由表中数据可知, 系统运行时间与一个批次内的样本数和亲缘关系对数正相关, 并且更易受批次内样本数的影响, 因为系统需要计算所有样本的两两亲缘关系, 最终输出有亲缘关系的样本对, 但人工预测所用时间主要受批次内亲缘关系对数的影响, 因为大多数时间主要用在最终亲缘关系结果的统计与家谱图的绘制。总之, 系统的开发节省了大量人力物力资源, 极大提高了亲缘关系预测的效率, 为物证鉴定工作提供了更好的工具。

3 讨论

KPS v1.0 是一个完整的 Web 系统, 需要部署在服务器上, 用户使用只需通过浏览器访问即可。系统支持多用户使用、人脸识别登录、个人信息管理、用户管理、任务管理及亲缘关系查看等功能。同时, 系统开发完毕已打包为完整的 Docker 镜像, 只需在服务器上安装 Docker 容器即可实现一键部署^[18], 解决了以往项目部署难的问题。目前, 系统在测试完成后就部署在线上投入使用, 各个功能都得到了物证鉴定部门工作人员的认可, 为物证鉴定工作带来了方便, 同时也极大地提高了亲缘关系预测效率, 实现了系统的开发意义。

本文仅着重介绍了系统亲缘关系的预测模块。需要指出的是, KPS v1.0 后台是基于 Plink 和 KING 软件计算的, 系统最终计算得到的亲缘关系对与 KING 软件计算结果一致, 目的是将以前人工预处理数据的流程和人工绘制家谱图的流程交由计算机来完成, 这样不仅节省了人力物力资源, 而且大大地提升了分析效率。同时, 以往每台

计算机每次仅执行一个批次计算, 一直要等到计算完毕才可进行下一批次的计算, 这样大大浪费了多核 CPU 的资源, 而 KPS v1.0 使用异步任务处理机制妥善解决了这一难题, 保证每台计算机可以并行处理多个批次计算, 达到 CPU 最高利用率, 用户提交完批次无需等待, 不仅可随时查看批次计算状态, 还可直接计算下一批次, 这样使计算效率得到了质的提升, 并且也方便了用户使用。

值得注意的是, 目前国内对亲缘关系这一方向的研究甚少^[13], 而 KPS v1.0 的亲缘关系计算软件也大都来源于国外, 在部分国内人群中的应用可能具有一定的局限性; 同时, KPS v1.0 系统只是针对本项目组前期采集的四大志愿者家族的 37 份测试样本进行了测试, 结果虽然很可观, 但无法直接用来推断实际案件, 直接确定嫌疑人, 只能作为参考, 辅助物证鉴定工作。

总之, KPS v1.0 使用简单, 界面直观, 显著降低了使用门槛, 而且将多种流程实现了程序自动化, 帮助物证鉴定工作人员提高了计算效率, 在人类遗传学、法医遗传学中为更多研究人员/分析人员提供了方便。

参考文献(References):

- [1] 刘聪, 肖超, 黄钰洁, 等. 新型连锁遗传标记的法医学研究进展[J]. 中国法医学杂志(LIU Cong, XIAO Chao, HUANG Yujie, et al. Advances in studies on the new type of linkage genetic marker[J]. Chinese Journal of Forensic Medicine), 2018, 33 (3): 269-272.
- [2] 侯一平. 法医物证学[M]. 第3版. 北京: 人民卫生出版社(HOU Yi-ping. Forensic Genetics[M]. 3rd ed. Beijing: People's Medical Publishing House), 2009: 2-35.
- [3] 柳燕, 赵珍敏, 李莉, 等. 多系统联合分析应用于复杂亲缘关系鉴定[J]. 中国司法鉴定(LIU Yan, ZHAO Zhen-min, LI Li, et al. Complex kinship analysis by multi-genotyping system and IBS score[J]. Chinese Journal of Forensic Sciences), 2014, 2: 61-66.
- [4] 张敏. 浅析 Y-STR 家系排查在实例侦查中的应用[J]. 法制博览(ZHANG Min. The application of Y-STR family check in case investigation[J]. Legality Vision), 2017, 9: 159-160.

(下转第 258 页)

- [57] BYUN M Y, CUI L H, KIM W T. Suppression of *OsKu80* results in defects in developmental growth and increased telomere length in rice (*Oryza sativa* L.)[J]. Biochemical and Biophysical Research Communications, 2015, 468(4): 857–862.
- [58] XIE X, SHIPPEN D E. DDM1 guards against telomere truncation in *Arabidopsis*[J]. Plant Cell Reports, 2018, 37(3): 501–513.
- [59] BYUN M Y, CUI L H, LEE H, *et al.* Telomere association of *Oryza sativa* telomere repeat-binding factor like 1 and its roles in telomere maintenance and development in rice, *Oryza sativa* L.[J]. BMB Reports, 2018, 51(11): 578–583.
- [60] WATSON J M, RIHA K. Telomeres, aging, and plants: from weeds to Methuselah – a mini-review[J]. Gerontology, 2011, 57(2): 129–136.
- [61] ZOU Y, SFEIR A, GRYAZNOV S M, *et al.* Does a sentinel or a subset of short telomeres determine replicative senescence?[J]. Molecular Biology of the Cell, 2004, 15(8): 3709–3718.
- [62] WU X, AMOS C I, ZHU Y, *et al.* Telomere dysfunction: a potential cancer predisposition factor[J]. Journal of the National Cancer Institute, 2003, 95(16): 1211–1218.
- [63] SALDANHA S N, ANDREWS L G, TOLLEFSBOL T O. Assessment of telomere length and factors that contribute to its stability[J]. European Journal of Biochemistry, 2003, 270(3): 389–403.
- [64] D'ADDA DI FAGAGNA F. Living on a break: cellular senescence as a DNA-damage response[J]. Nature Reviews Cancer, 2008, 8(7): 512–522.
- [65] LIM P O, KIM H J, NAM H G. Leaf senescence[J]. Annual Review of Plant Biology, 2007, 58: 115–136.
- [66] KIM M K, KIM W T. Telomere structure, function, and maintenance in plants[J]. Journal of Plant Biology, 2018, 61(3): 131–136.
- [67] CHAN S W L, HENDERSON I R, JACOBSEN S E. Gardening the genome: DNA methylation in *Arabidopsis thaliana*[J]. Nature Reviews Genetics, 2005, 6(5): 351–360.
- [68] XIAO W, CUSTARD K D, BROWN R C, *et al.* DNA methylation is critical for *Arabidopsis* embryogenesis and seed viability[J]. The Plant Cell, 2006, 18(4): 805–814.
- [69] OGROCKÁ A, POLANSKÁ P, MAJEROVÁ E, *et al.* Compromised telomere maintenance in hypomethylated *Arabidopsis thaliana* plants[J]. Nucleic Acids Research, 2014, 42(5): 2919–2931.

(上接第 233 页)

- [5] 朱传红, 史绍杏, 王海生, 等. Y-STR 家系排查法的应用原则及注意事项[J]. 中国法医学杂志(ZHU Chuan-hong, SHI Shao-xing, WANG Hai-sheng, *et al.* Application principles and precautions of Y-STR family investigation method[J]. Chinese Journal of Forensic Medicine), 2007, 22(6): 431–432.
- [6] LEE H, CHEN L. Inference of kinship using spatial distributions of SNPs for genome-wide association studies[J]. BioMed Central Genomics, 2016, 17: 372.
- [7] CHEUNG E Y Y, PHILLIPS C, EDUARDOFF M, *et al.* Performance of ancestry-informative SNP and microha-plotype markers[J]. Forensic Science International: Genetics, 2019, 43: 102141.
- [8] 赵书民, 张素华, 阙庭志, 等. 两个个体间常用亲缘关系指数的统一算法[J]. 法医学杂志(ZHAO Shu-min, ZHANG Su-hua, QUE Ting-zhi, *et al.* Establishment of universal algorithms for commonly used kinship indices between two individuals[J]. Journal of Forensic Medicine), 2011, 5: 330–333.
- [9] 李瑞, 张引娣, 刘奋进. 全基因组关联分析与 PLINK 软件的应用[J]. 中国数字医学(LI Rui, ZHANG Yin-di, LIU Fen-jin. The application of genome-wide association analysis and PLINK software[J]. China Digital Medicine), 2017, 12(3): 44–45, 78.
- [10] MANICHAIKUL A, MYCHALECKYJ J C, RICH S S, *et al.* Robust relationship inference in genome-wide association studies[J]. Bioinformatics, 2010, 26(22): 2867–2873.
- [11] KERNIGHAN B W, PIKE R. 程序设计实践[M]. 裘宗燕译. 北京: 机械工业出版社(KERNIGHAN B W, PIKE R. The Practice of Programming[M]. Translated by QIU Zong-yan. Beijing: China Machine Press), 2007: 55–62.
- [12] BETTINGER B T. The Family Tree Guide to DNA Testing and Genetic Genealogy[M]. New York: Oxford University Press, 2016: 10.6.43–10.6.48.
- [13] 刘京, 季安全, 李彩霞, 等. 法医系谱分析研究进展[J]. 刑事技术(LIU Jing, JI An-quan, LI Cai-xia, *et al.* Research progress of forensic genealogy analysis[J]. Forensic Science and Technology), 2019, 44(3): 189–194.
- [14] 李一风. 基于 Celery 和 Django 的分布式自动化测试系统设计[J]. 信息技术(LI Yi-feng. Distributed automated testing system based on Celery and Django[J]. Information Technology), 2019, 5: 97–100.
- [15] 黄健宏. Redis 设计与实现[M]. 北京: 机械工业出版社(HUANG Jian-hong. The Design and Implementation of Redis[M]. Beijing: China Machine Press), 2014: 55–62.
- [16] HEYDT M. Learning Pandas[M]. 3rd ed. Boston: Prentice Hall, 2017: 11.21.263–11.21.271.
- [17] RENTERIA M E, CORTES A, MEDLAND S E. Using PLINK for genome-wide association studies (GWAS) and data analysis[J]. Methods in Molecular Biology, 2013, 1019: 193–213.
- [18] MOUAT A. Docker 开发指南[M]. 黄彦邦译. 北京: 人民邮电出版社(MOUAT A. Using Docker[M]. Translated by HUANG Yan-bang. Beijing: Posts & Telecom Press), 2017: 155–185.