

本草基因组方法学研究*

□陈士林** 何 柳 刘明珠 钱 俊 闫海霞

(中国医学科学院 药用植物研究所 北京 100193)
(北京协和医学院)

程翔林 (中国生物技术发展中心 北京 100036)

钱忠直 (国家药典委员会 北京 100061)

摘 要:本草基因组计划(HerbGP)是针对具有典型代谢途径和重大经济价值的药用植物进行的全基因组测序和后基因组研究的系列计划。本文全面介绍物种基因组大小、染色体数目测定方法和第二代高通量测序方法,以及全基因组组装和基因组注释、基因组比较等生物信息学分析手段,并简要阐述重测序在药用植物全基因组研究中的应用方法。本草基因组计划的实施,将促进现代科学最新技术和方法在中药研究中的应用,使中药学研究位于生命科学的最前沿。

关键词:本草基因组计划 基因组大小 染色体数目 高通量测序 重测序

doi: 10.3969/j.issn.1674-3849.2010.03.001

2010年中国医学科学院药用植物研究所等提出本草基因组计划(Herb Genome Program),宣布将完成人参、丹参、灵芝、茯苓等多种本草的全基因组测序工作,这将为中药科学研究开拓新的基础平台,为开展本草功能基因组学、蛋白组学、代谢组学、遗传代谢工程和分子遗传育种等研究奠定基础,使选育高品质、高产量、抗胁迫的药用植物,建立现代中药有效成分生物工程体系成为可能。

本草基因组计划将采用第二代高通量测序技术对药用植物全基因组进行测序和研究,充分发挥不同测序平台的优势,实现不同平台数据的相互补充和校正,从而获得高质量、高精度的多种药用植物全基因组框架图,结合已经获得的人参、西洋参等药用

植物转录组分析结果^[1-4],开展药用植物后基因组学研究。以人参、丹参、灵芝、茯苓等为代表的本草基因组计划将为中医药现代化开辟新的视野与思路,提供新的研究与分析平台,对其他药用植物的研究具有很好的借鉴和示范作用。本草基因组计划的实施和完成将会对中药研究与发展产生深远影响。

一、中药植物及药用真菌染色体数目及基因组大小的确定

部分药用植物在形态上区别并不明显,为了确保测序物种的准确性,在开始正式测序以前,必须对本草进行分子水平的鉴定。多项研究发现 ITS2 是适合药用植物鉴定的分子标记^[5-8]。然后,必须对待测药用植物基因组大小进行估测。目前,国内外普遍采用流式细胞术和脉冲场电泳这两种基因组估测方法。

收稿日期: 2010-06-02

修回日期: 2010-06-12

* 国家自然科学基金面上项目(30970307):叶绿体全基因组条形码鉴定方法研究,负责人:陈士林。

** 联系人:陈士林,本刊学术副主编,研究员,主要研究方向:中药资源可持续利用及 DNA 条形码,E-mail:slchen@implad.ac.cn。

1. 流式细胞术

(1)简介。

流式细胞术(Flow Cytometry, FCM)是 20 世纪 70 年代发展起来的一种对细胞的物理或化学性质,如大小、内部结构、DNA、RNA、蛋白质、抗原等进行快速测量并可分类收集的高技术,它综合了激光技术、计算机技术、半导体技术、流体力学、细胞化学等多门学科^[9]。早在 20 世纪 70 年代,FCM 就已经被应用于测定基因组大小的研究中^[10~13]。

FCM 测定药用植物基因组大小的方法是将处于旺盛生长期的样品,如新鲜中药材的幼嫩叶片或指数生长期的药用真菌菌丝体制备成细胞核(直径 $\leq 20\mu\text{m}$ 生物颗粒)悬液,经特异性荧光染色后,以一定速度通过激光照射区,产生不同的荧光信号,经过一系列光电转换等电子设备后最终转化为可以被计算机处理和分析的数据文件。以已知基因组大小的物种为内参,根据其 G1 期光吸收峰的比值可以计算出样品的 C 值。

计算公式:样品 2C DNA 含量=样品 G1 吸收峰值/内参 G1 吸收峰值 $\times$ 内参 2C DNA 含量

FCM 测定药用植物倍性的原理是将样品与设定的内标物(内标物一般为已知染色体数目和倍性的同种物)比较,依据其光吸收峰的比值来计算倍性(见图 1)。

计算公式:样品的倍性=内标物的倍性 $\times$ 样品 G1 期的光吸收峰值/内标物 G1 光吸收峰值

(2)技术流程。

FCM 已广泛应用于动物、被子植物、微生物等的细胞 C 值及倍性的测定,其技术流程已非常成熟,一般可以分为以下几个具体步骤^[14]。

①材料的选择及样品的准备 选择药用植物,如,丹参幼嫩叶片或者处于对数生长期的药用真菌,如,灵芝的菌丝体。一般采用机械破碎法将样品制备成完整的细胞核悬液。

②内标的选定 对细胞基因组大小进行绝对定量和倍性确定时都需要选定内标,内标物的选择很重要,一般选用与样品同类的物种作为内标,如测定植物最好选用植物作为内标^[15],选择的内标物一般为与所测样

品 C 值大小接近的物种(但不能相同)。对内标物和样品应同时进行细胞核悬液的制备,这样才能保证数据的可靠性。

③细胞核悬液染色 选用碘化丙啶(PI)对药用植物细胞核 DNA 染色,其特点是灵敏且没有碱基偏好性,尤为适合测定细胞 C 值和染色体倍性^[16~18]。

④样品上机及结果分析 药用植物细胞核悬液染色完成之后即可上机分析,一般采用低速上样,这样可以获得更准确的数据,精确地测定细胞 C 值和倍性。选用内标法,即混合上样,测定物种的 C 值和倍性(见表 1)。

2. 脉冲场凝胶电泳技术测定药用真菌染色体数目及其基因组大小

(1)简介。

脉冲场凝胶电泳技术(Pulsed Field Gel Electrophoresis, PFGE)是 20 世纪 80 年代初发展起来的一种用来分离大片段(50 Kb~10 Mb)线性 DNA 分子的电泳技术。其原理是通过在两个方向上电场周期性交替变换,DNA 分子以蠕行的方式在琼脂糖凝胶的连续孔洞迁移,DNA 分子在交替变换方向的电场中做出反应所需的时间取决于它的大小,较小的 DNA 分子重新定向较快,因而在凝胶中移动的也快,于是不同大小的分子被成功分离^[16]。在对药用真菌等微生物的研究中 PFGE 应用十分广泛,早在 1983 年,

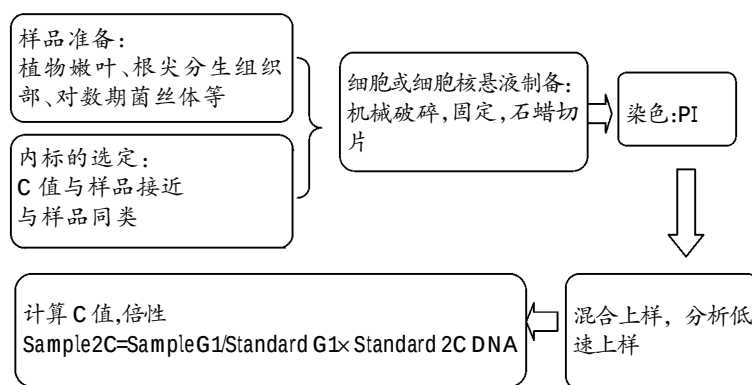


图 1 流式细胞术的技术流程图

表 1 部分物种流式细胞术测定结果与全基因组测序结果比较

物种名称	FCM 结果 (Mb)	倍性	全基因组测序得 大小(Mb)	参考文献
灵芝 <i>Ganoderma lucidum</i>	48.98 $\pm$ 0.60	2	48	未发表
紫芝 <i>Ganoderma sinensis</i>	53.75 $\pm$ 2.52	2	60	未发表

Schwartz 和 Cantor 就利用 PFGE 技术突破大分子 DNA 的壁垒,分离出大片段的 DNA 分子^[19-20]。

经过多年不断的发展和进步,PFGE 已有多种类型,箱位匀强电场技术 (Contour-Clamped Homogeneous Electric Field, CHEF) 是现在用于微生物电泳核型分析和 DNA 大小测定等领域最受欢迎和应用最为广泛的一种。PFGE 的应用十分广泛,其主要用于病原菌的核型分析,真菌的染色体分型和 DNA 大小测定,酵母人工染色体电泳分型, DNA 图谱制作等(见图 2)。

(2) 技术流程。

利用 PFGE 来进行真菌的染色体分型及 DNA 大小的估测,方法已非常成熟,相关研究文献也非常之多,2000 年, Ágnes Nagy 等对毛霉科霉菌 (Mucoraceous fungi) 进行 CHEF 电泳分型, Doolyi Kim 等对金针菇 (*Flammulina velutipes*) 进行染色体分型并估测其基因组大小,2008 年 Lisa Solieri 等利用 FCM 和 PFGE 技术同时对鲁氏接合酵母 (*Zygosaccharomyces rouxii*) 进行 DNA 大小的测定^[21-23]。

①样品的选择及准备 为了减少次生代谢产物的影响,最好选用对数生长期药用真菌菌丝体作为样品。首先对收集的药用真菌菌丝体进行等渗液清洗。根据不同真菌选用不同等渗液,一般为 KCl 及山梨酸等。因为需要对样品完整染色体进行电泳,所以实验中所有条件都应温和,不能对染色体的完整性造成破坏。

②原生质体的获得 每一种药用真菌细胞壁的结构和成分都有差异,为了从不同的药用真菌中获得完整原生质体,需要探索选择某种或多种酶对菌丝体进行酶解处理。

③制备胶栓及去膜 过滤药用真菌菌丝体得到的原生质体酶溶液,用少量等渗液洗涤并富集,将已融化的 1% 低熔点琼脂糖与原生质体等体积混合制成胶栓,根据所测样品基因组大小范围的不同,要求每片胶栓 (1~2μg) 包含酵母菌 (~12Mb) 为 6×10⁷ 个,大肠杆菌 (~5Mb) 为 2×10⁸,将制好的胶栓放入去膜的酶溶液中,过夜去膜,接着用苯甲基磺酰氟 (PMSF) 溶液将酶灭活。制备好的胶栓可以在 TE 溶液中 4℃ 长期保存^[24]。

④脉冲场凝胶电泳 根据样品的基因组大小设置电泳程序,以期达到最佳分离效果。如对样品没有任何基因组大小方面的了解,可以进行一个短时间如 24 小时电泳,观察分离效果之后,再设计更优电泳条件。

⑤凝胶染色观察及分析 电泳完成之后,对琼脂糖凝胶进行 EB 染色^[25],一般采用 1~2μg·mL⁻¹ 浓度 EB 溶液遮光染色 1/2 小时,或用 Gelred 染料进行染色亦可。再放入电泳缓冲液或水中脱色 30~60 分钟。然后用凝胶成像仪照相并用相关软件分析定量每一条带的大小。在对多种药用植物基因组大小和染色体倍性有清楚的了解以后,就可以进行全基因组测序的工作,如表 2 所示^[26-37]。

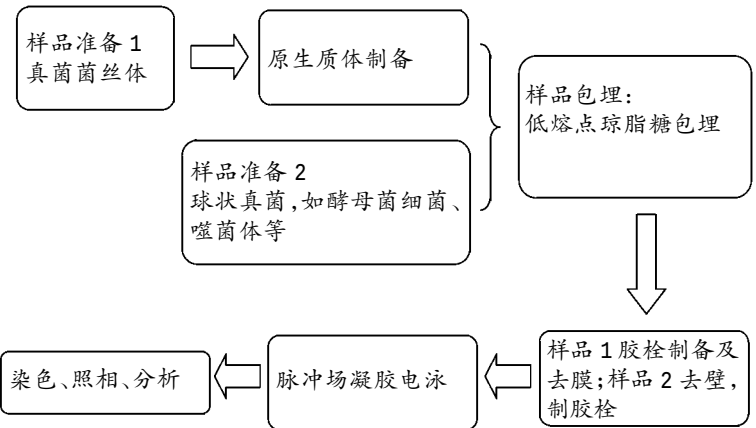


图 2 PFGE 的技术流程图

表 2 部分物种 PFGE 测定结果与全基因组测序结果比较^[26-37]

物种名称	染色体数目	PFGE 测定结果 (Mb)	全基因组测序结果 (Mb)	参考文献
巴斯德毕氏酵母 <i>Pichia pastoris</i>	4	9.7	9.43	26
柄孢霉 <i>Podospora anserina</i>	5	33.4	35.5~36	27
都柏林念珠菌 <i>Candida dubliniensis</i>	13	16~17	14.6	28,29
酿酒酵母 <i>Saccharomyces cerevisiae</i>	16	11.8	11.6	30
红球虫赤壳 <i>Nectria haematococca</i>	17	51	54.43	31
里氏木霉 <i>Trichoderma reesei</i>	7	33	34	32
产黄青霉 <i>Penicillium chrysogenum</i>	4	33.3~34.1	32.2	33,34
兔脑炎微孢子虫 <i>Nosema ceranae</i>	11	2.9	7.86	35,36
假丝酵母属 <i>Candida</i>	6-9	7.88~14.36	10.6~15.5	37

二、全基因组测序方法

1. Sanger 测序法

20 世纪 70 年代,英国生物化学家弗雷德·桑格(Frederick Sanger)发明了双脱氧链终止法测定 DNA 序列^[38]。20 世纪 80 年代,ABI 公司根据 Sanger 法,采用荧光代替同位素,通过计算机图像识别,研发了第一款测序仪 ABI 370。20 世纪 90 年代,ABI 公司将凝胶电泳改进为集束化的毛细管电泳,使 DNA 测序速度和质量得到进一步提高。

Sanger 法测序技术经过 30 年的不断发展和完善,现在测序长度已经可以达到 1000 bp,并且每个碱基的读取准确率高达 99.999%。由于其原始数据质量以及序列读长方面的优势,这一代测序仪目前仍在使用^[39]。

2. 高通量测序方法

随着基因组研究在各个领域的广泛应用,测序技术得到迅速发展,高通量自动化测序仪应运而生。Roche GS FLX、Illumina Solexa,以及 ABI SOLiD 构建了 3 个大高通量测序平台。新的高通量测序技术的应用使测序成本下降 100 倍以上,测序时间缩短 70%。本草基因组计划跨入第二代高通量测序时代。

(1) 样品 DNA 制备。

比起 Sanger 法测序,高通量测序的样品准备较为简单,只需要待测物种的基因组 DNA 即可。全基因组信息蕴藏在新鲜生物体的每个部位,为了保证获取足够的高纯度 DNA 样品,和后续测序实验中获得更多有用数据,选择样品部位时应注意以下几点:①可以选择黄化苗提取 DNA,避免测序得到较多的叶绿体基因组序列;②尽量选取幼嫩材料,避免 DNA 提取过程中混入较多 RNA,或者在 DNA 提取过程中加入适量 RNase A。

传统中药主要包括天然植物和药用真菌,都可以用 CTAB 法或 SDS 法进行样品 DNA 分离,最后选用合适的 DNA 吸附柱进行样品纯化,以获得高纯度的 DNA 样品。

(2) Roche 454 GS FLX。

2005 年,454 公司创始人 Jonathan Rothberg 博士,首次推出基于焦磷酸测序法的高通量基因组测序系统-Genome Sequencer 20 System(GS 20)。之后,454 与罗氏公司合并,推出性能更强的第二代测序系统-Genome Sequencer。

FLX System(GS FLX)。2008 年,Roche 公司再次将 GS FLX 升级成为 GS FLX Titanium。目前,GS FLX 系统在 10 小时的运行中可获得 100 多万个读长,一共获得 4~6 亿个碱基信息,准确率在 99% 以上。GS FLX 是采用边合成边测序的方法,核苷酸聚合过程中释放焦磷酸,焦磷酸在 ATP 硫酸化酶和荧光素酶的作用下,经过一系列化学反应产生光信号。将每一个碱基的聚合与荧光信号释放偶联起来,通过测定荧光信号的释放,就可以达到实时确定 DNA 序列。

(3) 合成法:Illumina GA。

2007 年,Illumina 公司与 Solexa 公司合并推出新一代高通量测序仪 Genome Analyzer,它采用边合成边测序(Sequencing By Synthesis)的方法,在每个循环过程里,荧光标记的核苷和聚合酶被加入到单分子阵列里。互补的核苷和核苷酸片段的第一个碱基配对,通过在反应中同时加入 4 种核苷标签,针对每种碱基特定波长的激光激发结合上的核苷的标记,这个标记会释放出荧光。荧光信号被 CCD 采集,CCD 快速扫描整个阵列检测特定的结合到每个片段上的碱基。通过上述的结合,检测可以重复几十个循环,这样就有可能决定核苷酸片段中的几十个碱基。

制备文库

454:将 gDNA 打断成 300~800bp 片段。
Solexa:将 gDNA 打断成 100~200bp 片段。
SOLiD:将 gDNA 打断成 100~200bp 片段。



片段 DNA 扩增

454:DNA 片段吸附在磁珠(20μm)上,进行乳化 PCR。
Solexa:DNA 片段吸附在芯片上,通过桥 PCR 扩增形成 DNA 簇。
SOLiD:DNA 片段吸附在磁珠(1μm)上,进行乳化 PCR。



测序反应

454:合成法(焦磷酸法),4 种碱基依次加入 PTP 板,如果发生碱基配对就释放 1 个焦磷酸,生成荧光信号,被 CCD 采集。
Solexa:合成法,反应中加入带有 4 种荧光染料的核苷,激光扫描每次聚合加入的核苷类型。
SOLiD:连接法,有效解决多聚核苷酸序列难读取的问题。

图 3 第二代高通量测序技术流程图

(4)连接法:ABI SOLiD。

20 世纪 80 年代,ABI 公司推出世界上第一台 DNA 测序仪 ABI 370,奠定了 ABI 公司在核酸分析领域的地位,但是随着 Roche GS FLX 和 Solexa 的出现使 ABI 公司在 DNA 测序领域的领先地位开始动摇,在迅速合并 Agencourt Personal Genomics 以后,2007 年底,ABI 推出 SOLiD (Supported Oligo Ligation Detetion)新一代测序平台。它以四色荧光标记寡核苷酸的连续连接合成为基础,取代了传统的聚合酶链式反应,可对单拷贝 DNA 片段进行大规模扩增和高通量并行测序。目前,SOLiD3 单次运行可以产生 50 GB 的序列(见表 3)。

三、De Novo 测序数据组装和基因组分析

1. Contig 的拼接和 Scaffolding 的拼接

针对于目前短读长的第二代测序数据,有专门的拼接软件和在线分析平台,如表 4 所示^[40~46]。当拼接的 Contig N50 长度达到 20Kb,Scafford N50 达到 300Kb,测序数据覆盖基因区域 98%左右,完成本草基因组框架图,就可以进行下一步基因组分析。

2. 基因组注释:发现新基因、蛋白质结构及功能预测和功能注释

完成全基因组测序仅仅是一个开始,得到精细图谱后最重要的就是对基因组信息挖掘。包括基因结构和功能预测,目前用得最多的软件和在线分析工具(见表 4)。

3. 基因组分析和分子进化:比较基因组学

药用植物种类繁多,中药植物中存在大量亲缘关系较近的物种,而且同种药材在不同地域生长的个体在药效上会有较大差异。这些差异可以通过对不同亲缘关系物种的基因组序列进行比较,得到生物系统发生、进化关系等信息来解释。

通过对同种群体内基因组比较,能发现不同个体基因组中存在的各种变异,而这种差异可能就是造成中药药效差异的遗传基础。目前对基因组分析得较多的就是单核苷酸多态性 (Single Nucleotide Polymorphism,

SNP)和拷贝数多态性(Copy Number Polymorphism, CNP)。利用比较基因组学可以以模式药用植物的基因组图谱为基础,通过基因组比较,来了解不同物种中相同基因的功能、表达机理和物种进化关系。

四、重测序

1. 重测序概述

所谓重测序,就是对已测物种的其他个体进行测序,利用已有的参照序列进行数据分析,主要包括全基因组重测序(Whole Genome Resequencing)和定向重测序(Targeted Resequencing)。

(1)全基因组重测序。

全基因组重测序是相对于 De Novo 测序而言,是在已有参考基因组序列情况下对已知物种的个体基因组进行测序,并在个体或群体水平进行差异性分析。

全基因组重测序方法与全基因组测序方法类似(图 4),在后期基因组组装时,由于参考序列提供了序列拼接所需的骨架,所以对于重测序来讲并不需要较长的测序读长^[47],而且测序覆盖度(8~12 倍)

表 3 第二代高通量测序仪比较

	测序原理	单机售价	获得 1 Mb 数据费用	单次获得数据量	单次运行时间	测序长度
454	合成法(焦磷酸法)	50 万美元	60 美元	600Mb	10 小时	400bp
Solexa	合成法	43 万美元	2 美元	20Gb	1.5 天~7 天	75bp
SOLiD	连接法	59 万美元	2 美元	50Gb	7 天	50bp

表 4 常用于基因组拼接和基因预测的生物信息学分析平台

名称	类别	网址	文献
SSAKE	Assembly	http://www.bcgsc.ca/platform/bioinfo/software/ssake	40
VCAKE	Assembly	http://sourceforge.net/projects/vcake	41
SHARCGS	Assembly	http://sharcgs.motgen.mpg.de	42
Edena	Assembly	http://www.genomic.ch/edena	43
Glimmer	Prediction	http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi	44, 45
Genscan	Prediction	http://genes.mit.edu/GENSCAN.html	46
NT	Annotation	http://www.ncbi.nlm.nih.gov	
NR	Annotation	http://www.ncbi.nlm.nih.gov	
SwissProt	Annotation	http://expasy.org/sprot/	
Kegg	Annotation	http://www.genome.jp/kegg/	
COG	Annotation	http://www.ncbi.nlm.nih.gov/COG/	
Interpro	Annotation	http://www.ebi.ac.uk/interpro/	
GO	Annotation	http://www.geneontology.org/	

的要求也比 De Novo 测序 (25~70 倍) 低得多^[48], 因此重测序得到的基因组质量取决于参考基因组的质量。

通过将短序列片段定位到参考基因组以及利用测序个体序列与参考基因组序列进行比对, 我们可以在全基因组范围内检测包括单核苷酸多态性 (SNPs)、短的插入缺失 (Indels)、结构 (大片段插入或缺失、复制、倒位、易位等) 和拷贝数等一系列不同类型的序列变异, 并根据已知信息对其注释。但片段序列能否正确定位取决于参考基因组的复杂性、测序读长、测序错误率以及物种个体本身的多样性等一系列因素^[49], 采用 Paired-End (或 Mate-Paired) 和不同插入片段长度相结合进行测序、改进短读长比对软件 (表 5) 的算法是解决上述问题的方法^[50]。

(2) 定向重测序。

定向重测序是对基因组中感兴趣的区域 (例如全部外显子、特定的某个基因区域) 进行测序分析。靶向富集基因组区域常用的策略有 PCR、分子反转探针 (Molecular Inversion Probes) 或杂交捕获 (Hybrid Capture) 等 (各个方法的比较、操作等具体内容请参考引文^[51])。

相对于全基因组重测序, 定向重测序针对性更高, 相同的经费可以获得更多个体的目标基因组信息且其测序覆盖度更深、数据的准确性更高, 更适合大样本量的测序研究。但由于其集中在目标基因组区域, 只能发现此区域范围内序列的各种变异信息。

2. 重测序在中药植物全基因组研究中的应用

第二代高通量测序平台的成功应用使得药用植物大规模重测序研究成为可能。2008 年, Ossowski 等^[52]运用 Solexa 对重要模式植物拟南芥的参考品系 (Reference Natural Strains) Col-0 和另两个品系 (Divergent

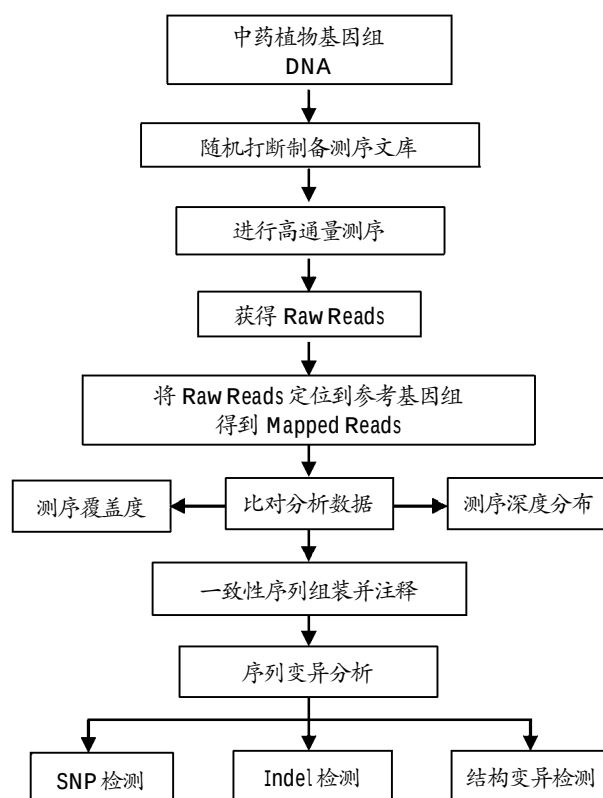


图 4 药用植物全基因组重测序实验及数据分析路线

表 5 短序列比对软件^[50]

	SNPs	indels	Mapped assembly	URL
Bowtie	✓			http://bowtie-bio.sourceforge.net
DIAL	✓	✓		http://www.bx.psu.edu/miller_lab/
ELAND				http://www.illumina.com
Exonerate		✓		http://www.ebi.ac.uk/~guy/exonerate/
GMAP				http://www.gene.com/share/gmap
GenomeMapper				http://1001genomes.org/downloads/genomemapper.html
MOSAIC	✓	✓	✓	http://bioinformatics.bc.edu/marthlab/Mosaik
MAQ	✓	✓	✓	http://maq.sourceforge.net
MUMer	✓	✓		http://mummer.sourceforge.net/
Novocraft	✓	✓		http://www.novocraft.com/
RMAP				http://rulai.cshl.edu/rmap/
SeqMap		✓		http://biogibbs.stanford.edu/~jiangh/SeqMap/
SHRIMP	✓	✓		http://compbio.cs.toronto.edu/shrimp/
SHORE	✓	✓		http://1001genomes.org/downloads/shore.html
SOAP	✓	✓		http://soap.genomics.org.cn/
SSAHA2	✓	✓		http://www.sanger.ac.uk/Software

Strains)Bur-0、Tsu-1 进行重测序研究鉴定了 80 多万个 SNPs 及近 8 万个 1-3bp 的插入缺失位点,并发现参考基因组中存在 2000 多个潜在错误。Orsouw 等将改进的 AFLP 技术和罗氏 454 测序技术结合,对玉米基因组进行了重测序,发现使用 CRoPS™ 推测的 75% 以上的 SNP 位点能够用 SNPWave 技术验证,提供了一条针对含高度重复序列植物基因组进行多态性分析的技术路线^[53]。越来越多的研究^[54-56]证实运用第二代高通量测序技术的重测序研究大大加速了种群内全基因组研究的深化。

对于来自不同产地有着不同药效的中药材,对其展开重测序研究有着重要价值。首先,通过重测序来进行中药原植物的系统进化和分类学研究。基于全基因组序列的系统进化研究由于囊括研究中药物种的全部核苷酸信息,将具有无可比拟的全面性和精确性。通过个体序列间的比对,将解决目前在分类学上仍存在争议的物种的归属,并找到每一个品系的唯一识别序列用于物种鉴定。其次,中药植物重测序可以将基因组遗传变异信息由单一个体扩展到一个生物种群。SNPs、插入缺失等位点信息将开发成更多的分子遗传标记,有助于从基因型到表型的研究。而通过与参考基因组的比对定位分析,可以预测新的基因和研究数量性状位点,有利于中药植物性状改良和育种等研究。

五、展 望

本草基因组计划运用基因组学的原理、方法和生物信息学手段,探讨中药基因组的结构、功能及其进化规律,从染色体和基因组水平上阐明中药活性成分的合成代谢规律,寻找中药活性成分合成、运输、积累的相关功能基因,发现和克隆中药次生代谢物的关键酶基因,利用基因工程生产中药的有效成分,利用分子育种来改良和培育新的中药品种,保护中药的种质资源和基因资源,保障中药资源的可持续利用^[57]。

本草基因组计划的实施,将开辟一个中药研究和应用的全新领域,把握这次历史性机遇,将大大提高我国开发中药资源的能力,增强我国中药基础研究实力、提高我国中药研究的自主创新能力,对于加速中药现代化进程具有重大的战略性科学意义,促进中药研究和产业的快速发展。

参考文献

- 1 陈士林,孙永巧,宋经元,等. 西洋参 cDNA 文库构建及表达序列标签(EST)分析. 药学学报,2008, 43(6):657-663.
- 2 Sun C, Li Y, Wu Q, et al. De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics, 2010, 11.
- 3 Li Y, Luo HM, Sun C, et al. EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. BMC Genomics, 2010, 11.
- 4 李滢,孙超,罗红梅,等. 基于高通量测序 454 GS FLX 的丹参转录组学研究. 药学学报, 2010, 45(4):524-529.
- 5 陈士林,宋经元,姚辉,等. 药用植物 DNA 条形码鉴定策略及关键技术分析. 中国天然药物,2009, 7(5):322-327.
- 6 罗焜,陈士林,陈科力,等. 基于芸香科的植物通用 DNA 条形码研究. 中国科学-C,2010,40(4):342-351.
- 7 Pang XH, Song JY, Zhu YJ, et al. Applying plant DNA barcodes for Rosaceae species identification. Cladistics, 2010, 26.
- 8 Chen SL, Yao H, Han JP, et al. Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. PLoS ONE, 2010, 5(1):e8613.
- 9 宋平根,李素文.流式细胞术的原理和应用.北京:北京师范大学出版社,1992:1-88.
- 10 Galbraith DW, Harkins KR, Maddox JM, et al. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. Science, 1983, 220: 1049-1051.
- 11 Bennett MD and Smith JB. Nuclear DNA amounts in angiosperms. Phil. Trans. R. Soc. Lond. B., 1976. 274:227-274.
- 12 Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter, 1991,9:208-218.
- 13 Greilhuber J, Dolezel J, Lysak MA, et al. The origin, evolution and proposed stabilization of the terms 'Genome Size' and 'C-Value' to describe nuclear DNA contents. Annals of Botany, 2005,95:255-260.
- 14 Dolezel J, Greilhuber J, Suda J. Estimation of nuclear DNA content in plants using flow cytometry. Nature Protocols, 2007, 2(9):2233-2244.
- 15 Roux N, Toloza A, Radecki Z, et al. Rapid detection of aneuploidy in Musa using flow cytometry. Plant Cell Rep, 2003, 21:483-490.
- 16 Temsch EM, Greilhuber J, Hammett KRW, et al. Genome size in Dahlia Cav. (Asteraceae Coreopsiadeae). Plant Syst Evol, 2008, 276:157-166.
- 17 Carvalho CR, Clarindo WR, Praça MM, et al. Genome size, base composition and karyotype of Jatropha curcas L., an important biofuel plant. Plant Science, 2008, 174:613-617.
- 18 Zonneveld BJM and Duncan GD. Genome sizes of Eucomis L' Hér. (Hyacinthaceae) and a description of the new species Eucomis grimshawii G.D.Duncan & Zonneveld. Plant Syst Evol, 2010, 284:99-109.
- 19 J·萨姆布鲁克,D·W·拉塞尔.分子克隆实验指南,3版.黄培堂,译.北京:科学出版社,2002: 429-450.
- 20 Schwartz DC, Cantor CR. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell, 1984, 37:67-75.
- 21 Schwartz DC, Saffran W, Welsh J, et al. New techniques for purifying

- large DNAs and studying their properties and packaging. Cold Spring Harb. Symp. Quant. Biol., 1983, 47 (Part 1):189-195.
- 22 Ágnes N, Palágyi Z, Vastag M, et al. Electrophoretic karyotypes of some related *Mucor* species. *Antonie van Leeuwenhoek*, 2000, 78: 33-37.
 - 23 Kim D, Tamai Y, Azuma T, et al. Analysis of the electrophoretic karyotype of *Flammulina velutipes*. The Japan Wood Research Society, 2000, 46:466-469.
 - 24 Solieri L, Cassanelli S, Croce MA, et al. Genome size and ploidy level: New insights for elucidating relationships in *Zygosaccharomyces* species. *Fungal Genetics and Biology*, 2008, 45:1582-1590.
 - 25 Herschleb J, Ananiev G, Schwartz DC. Pulsed-field gel electrophoresis. *Nature*, 2007, 2(3):677-684.
 - 26 De Schutter K, Lin YC, Tiels P, et al. Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nature Biotechnology*, 2009, 27:561-566.
 - 27 Espagne E, Lespinet O, Malagnac F, et al. The genome sequence of the model ascomycete fungus *Podospora anserina*. *Genome Biology*, 2008, 9:R77.
 - 28 Magee BB, Sanchez MD, Saunders D, et al. Extensive chromosome rearrangements distinguish the karyotype of the hypovirulent species *Candida dubliniensis* from the virulent *Candida albicans*. *Fungal Genet Biol*, 2008, 45: 338-350.
 - 29 Jackson AP, Gamble JA, Yeomans T, et al. Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. *Genome Res*, 2009, 12:2231-44.
 - 30 Argueso JL, Carazzolle MF, Mieczkowski PA, et al. Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. *Genome Res*, 2009, 19:2258-2270.
 - 31 Coleman JJ, Rounsley SD, Rodriguez-Carres M, et al. The Genome of *Nectria haematococca*: Contribution of Supernumerary Chromosomes to Gene Expansion. *PLoS Genet*, 2009, 5(8):e1000618.
 - 32 Martinez D, Berka RM, Henrissat B, et al. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nature Biotechnology*, 2008, 26:553-560.
 - 33 Fierro F, Gutiérrez S, Díez B, et al. Resolution of four large chromosomes in enicillin-producing filamentous fungi: the penicillin gene cluster is located on chromosome II (9.6 Mb) in *Penicillium notatum* and chromosome I (10.4 Mb) in *Penicillium chrysogenum*. *Mol Gen Genet*, 1993, 241:573-578.
 - 34 van den Berg MA, Albarg R, Albermann K, et al. Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. *Nature Biotechnology*, 2008, 26:1161-1168.
 - 35 Biderre C, Mathis A, Deplazes P, et al. Molecular karyotype diversity in the microsporidian *Encephalitozoon cuniculi*. *Parasitology*, 1999, 118: 439-445.
 - 36 Cornman RS, Chen YP, Schatz MC, et al. Genomic Analyses of the Microsporidian *Nosema ceranae*, an Emergent Pathogen of Honey Bees. *PLoS Pathog*, 2009, 5(6):e1000466.
 - 37 Butler G, Rasmussen MD, Lin MF, et al. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, 2009, 459: 657-662.
 - 38 Sanger F. Sequences, sequences, and sequences. *Annu Rev Biochem*, 1988, 57:1-28.
 - 39 Ji H, Shendure J. Next-generation DNA sequencing. *Nature Biotechnology*, 2008, 26:1135-45.
 - 40 Warren RL, Sutton GG, Jones SJ, et al. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 2007, 23:500-501.
 - 41 Jeck WR, Reinhardt JA, Baltrus DA, et al. Extending assembly of short DNA sequences to handle error. *Bioinformatics*, 2007, 23:2942-2944.
 - 42 Dohm JC, Lottaz C, Borodina T, et al. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res*, 2007, 17:1697-1706.
 - 43 Hernandez D, Francoil P, Farinelli L, et al. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res*, 2008, 18: 802-809.
 - 44 Salzberg SL, Delcher AL, Kasif S, et al. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*, 1998, 26:544-548.
 - 45 Delcher AL, Harmon D, Kasif S, et al. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*, 1999, 27:4636-4641.
 - 46 Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 1997, 268:78-94.
 - 47 Bentley DR. Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, 2006, 16:545-552.
 - 48 Schuster SC. Next-generation sequencing transforms today's biology. *NATURE METHODS*, 2008, 5(1):16-18.
 - 49 Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 2008, 18:1851-1858.
 - 50 Turner DJ, Keane TM, Sudbery I, et al. Next-generation sequencing of vertebrate experimental organisms. *Mamm Genome*, 2009, 20:327-338.
 - 51 Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. *Nature Methods*. 2010,7(2):111-118.
 - 52 Ossowski S, Schneeberger K, Clark RM, et al. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Research*, 2008, 18: 2024-2033.
 - 53 van Orsouw NJ, Hogers RCJ, Janssen A, et al. Complexity Reduction of Polymorphic Sequences (CROPSTM): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. *PLoS ONE*, 2007, 11: e1172.
 - 54 Hillier LW, Marth GT, Quinlan AR, et al. Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods*, 2008, 5 (2):183-188.
 - 55 Rubin CJ, Zody MC, Eriksson J, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, 2010, 464:587-591.
 - 56 Xia QY, Guo YR, Zhang Z, et al. Complete Resequencing of 40 Genomes Reveals Domestication Events and Genes in Silkworm (*Bombyx*). *Science*, 2009, 326:433-436.
 - 57 陈士林, 孙永珍, 徐江, 等. 本草基因组计划研究策略. *药学报*, 2010, 45(7):807-812.
- 323 [World Science and Technology/Modernization of Traditional Chinese Medicine and Materia Medica]

The Methodology for Herb Genome Program

Chen Shilin¹, He Liu¹, Liu Mingzhu¹, Qian Jun¹, Yan Haixia¹, Cheng Xianglin², Qian Zhongzhi³

(1. Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, China;

2. China National Center for Biotechnology Development, Beijing 100036, China;

3. Chinese Pharmacopoeia Commission, Beijing 100061, China)

Abstract: The Herb Genome Program (HerbGP) is introduced here based on next-generation sequencing techniques including Roche/454 GS FLX, Illumina/Solexa and ABI/SOLiD. The methodology for the HerbGP is illustrated in detail in terms of genome size measurement, whole-genome sequencing, genome assembly, genome annotation and re-sequencing. The HerbGP will improve the application of new technology and method in Chinese medicinal research, making the traditional Chinese medicine a frontier in modern life sciences.

Keywords: Herb Genome Program, Genome size, Next-generation sequencing, Whole-genome, Re-sequence

(责任编辑:王 瑀, 责任译审:张立崑)