史东杰,胡金有,朱　华,等. 锦鲤基因组数据分析及体色相关基因的筛选[J]. 江苏农业科学,2019,47(16):52 – 56.
doi:10.15889/j.issn.1002 – 1302.2019.16.011

锦鲤基因组数据分析及体色相关基因的筛选

史东杰[1,2,3], 胡金有[4], 朱　华[1,2,3], 张　欣[1,2,3], 李荣妮[1,2,3], 孙砚胜[1,2,3]
[1.北京市水产科学研究所暨国家淡水渔业工程技术研究中心,北京 100068; 2.农业部都市农业(北方)重点实验室,北京 100097;
3.渔业生物技术北京市重点实验室,北京 100097; 4.中国农业大学,北京 100083]

摘要:为了获得红白锦鲤的基因组信息,筛选与其肤色相关的基因,采用 Illumina 高通量测序技术对红白锦鲤皮肤组织的基因组进行测序,获得 127.23 Gb clean data,Q20 碱基比例在 95.59% 及以上,Q30 碱基比例在 90.81% 及以上,GC 含量为 37.32% ~ 42.38%,测序错误率为 0.07。与鲤鱼基因组序列进行比对的结果显示,比对效率为 96.35%。研究共鉴定了 1 048 576 个 SNPs(单核苷酸多态性),其中 3.12 百万 ~ 5.40 百万个 SNPs 位于短 reads 比对不到的区域,其中变异位点位于外显子区域的有 579 778 个 SNPs。SNP 位点分布于锦鲤的 50 条染色体上,不包含 scaffold(染色体骨架)。经 ANNOVAR 软件进行功能注释,纯合类型的 SNPs 数量是 574 310 个,杂合类型的 SNPs 数量是 474 265 个。SNPs 位于基因间的数量最多,SNPs 位于基因内的外显子区域的多态性最高。通过对 8 个重要候选基因注释的理解,发现微管蛋白 LOC109046532、LOC109049213 这 2 个基因与色素颗粒运输有关。其中基因 LOC109046532 含有突变,而另 1 个基因 LOC109049213 则不含有任何突变。8 个候选基因都含有外显子 SNP 位点,但是没有发现终止密码子突变。

关键词:基因组重测序;锦鲤;体色基因;候选基因
中图分类号: S917　　**文献标志码**: A　　**文章编号**:1002 – 1302(2019)16 – 0052 – 04

全基因组重测序是对已知参考基因组序列的物种进行不同个体间的基因组测序,并在此基础上对个体或群体进行差异性分析[1]。近年来,随着测序技术的发展,人们已经在众多水产动物中开展了全基因组测序,目前,鲤鱼(*Cyprinus carpio*)[2]、大黄鱼(*Larimichthys crocea*)[3]、半滑舌鳎(*Cynoglossus semilaevis*)[4]、大西洋鲑(*Salmo salar*)[5]、鲇鱼(*Ictalurus punctatus*)[6]、凡纳滨对虾(*Litopenaeus vannamei*)[7]和牡蛎(*Ostrea gigas* Thunberg)[8]等的基因组计划已经完成。2011 年,由中国水产科学研究院和中国科学院北京基因组研究所共同实施的"鲤鱼基因组计划"成功完成了鲤鱼的全基因组测序,并绘制了鲤鱼基因组框架图谱、基因组物理图谱和高密度连锁图谱,进而利用各方面的资源和数据实现了鲤鱼基因组的基因识别定位和精确的功能注释等。全基因组序列海量数据的获得,为水产基因组辅助育种研究、优良品种的快速培育提供了重要基础。

锦鲤(*Cyprinus carpio* L.)是经济合作与发展组织(OECD)规定的 5 种试验生物之一,也是我国主养的观赏鱼类。该鱼隶属于鲤形目(Cypriniformes)鲤科(*Cyprinidae*)鲤属(*Cyprinus*)。锦鲤以其雄健的身躯、绚丽的色彩、华丽的斑纹、潇洒的泳姿、温顺的习性而享誉世界,被人们称为"水中活宝石"。该鱼经过几百年的自然分化、基因突变、人工选育,形成了体色艳丽、斑纹丰富、鳞片迥异等十三大品系 100 余个品种,是目前鲤科鱼类种质资源和基因组资源最丰富的鱼类。本研究通过对锦鲤进行基因组重测序,与鲤鱼进行参考基因组比对,以期找到大量单核苷酸多态性位点(SNP)、拷贝数变异(copy number variation,简称 CNV)、插入缺失(insertion/deletion,简称 InDel)、结构变异(structure variation,简称 SV)等变异信息,分析锦鲤与鲤鱼的遗传多样性,同时研究锦鲤是否有与驯化选择相关的差异位点,并在测序的基础上,筛选出与肤色相关的候选基因。本研究不仅对锦鲤基因组辅助育种研究、体色斑纹定向培育提供了重要基础,而且对鲤科鱼类的基础研究具有重大意义。

1　材料与方法

1.1　试验材料

试验用红白锦鲤来自观赏鱼产业技术体系北京市创新团队通州综合试验站。从生长状态良好的健康红白锦鲤成鱼上取适量皮肤组织样品(设 3 个生物学重复),采用 TIANamp Genomic DNA Kit(血液/细胞/组织基因组 DNA 提取试剂盒)进行 DNA 提取,并通过琼脂糖凝胶电泳、NanoDrop 检测和 Qubit 定量进行 DNA 样本的检测。取样前,采用 MS – 222(Sigma,USA)使试验鱼麻醉后安乐死,并根据我国在科学技术方面应用的法律法规人性化对待试验动物。

1.2　试验方法

基因组 DNA 利用 Covaris 破碎仪随机打断成长度为 350 bp 的片段,经末端修复和加 A 尾后,片段两端分别连接

收稿日期:2018 – 05 – 03
基金项目:北京市财政局、北京市农业农村局观赏鱼产业技术体系北京市创新团队建设专项(编号:BAIC03);北京市农林科学院项目(编号:KJCX20170101)。
作者简介:史东杰(1985—),女,北京人,硕士,副研究员,主要从事观赏鱼繁育及养殖技术的研究工作。E – mail:sdj19850104@163.com。
通信作者:朱　华,博士,研究员,主要从事水产繁殖、养殖以及水产养殖环境水质调控方面的研究与推广工作。E – mail:Zhuhua@bjfishery.com。

接头制备 DNA 文库。文库构建完成后，先使用 Qubit 3.0 进行初步定量，随后使用 Qseq 100 对文库的 insert size（插入片段大小）进行检测，insert size 符合预期后，使用 Q－PCR 方法对文库的有效浓度（2 nmol/L）进行准确定量，以保证文库的质量。库检合格后，根据文库的有效浓度及数据产出需求，进行 Illumina HiSeq X Ten PE150 测序。PE150（Pairend 150 bp）指高通量双端测序，每端各测 150 bp。在构建的小片段文库中，insert DNA，即插入片段是高通量测序直接测序的单位。双端测序是对插入片段的两端进行测序的方法，由于插入片段的长度分布已知，双端测序时不仅可以知道片段两端的序列，也能知道这两段序列之间的长度，从而便于后续组装和比对。

对测序获得的 reads 数据进行质量过滤得到 clean reads，用于后续生物信息学的分析。将 clean reads 与参考基因组进行比对，基于比对结果，使用 samtools[9] 进行去重复（mark duplicates），使用 GATK[10] 进行局部重比对（local realignment）、碱基质量值校正（base recalibration）等处理，再使用 GATK 进行单核苷酸多态性（single nucleotide polymorphism，简称 SNP）的小片段插入缺失（small InDel）的检测、过滤，并得到最终的 SNP 和 small InDel 的位点集。通过 BreakDancer[11] 可以得到结构变异（structure variation，简称 SV）数据集，其中一般以插入（insertion，简称 INS）和缺失（deletion，简称 DEL）为主。对 SNP、InDel、SV、CNV 的检测结果进行注释，从而实现 DNA 水平差异基因挖掘和差异基因功能注释等。

1.3　数据处理与分析

将下机数据进行过滤，得到 clean data，将其与指定的参考基因组进行序列比对，得到 mapped data，进行插入片段长度检验、随机性检验等文库质量评估；进行可变剪接分析、新基因发掘和基因结构优化等结构水平分析；根据基因在样品中的表达量进行差异表达分析、差异表达基因功能注释和功能富集等表达水平分析，从而筛选出与体色相关的功能基因。

2　结果与分析

2.1　红白锦鲤基因组重测序数据质量评估

共完成 3 个样品的基因组重测序分析，通过高通量测序法获得 127.23 Gb clean data，Q20 碱基的百分比在 95.59% 及以上，Q30 碱基的百分比在 90.81% 及以上，GC 含量为 37.32% ~42.38%，测序错误率为 0.07%。

2.2　红白锦鲤基因组与参考基因组的比对

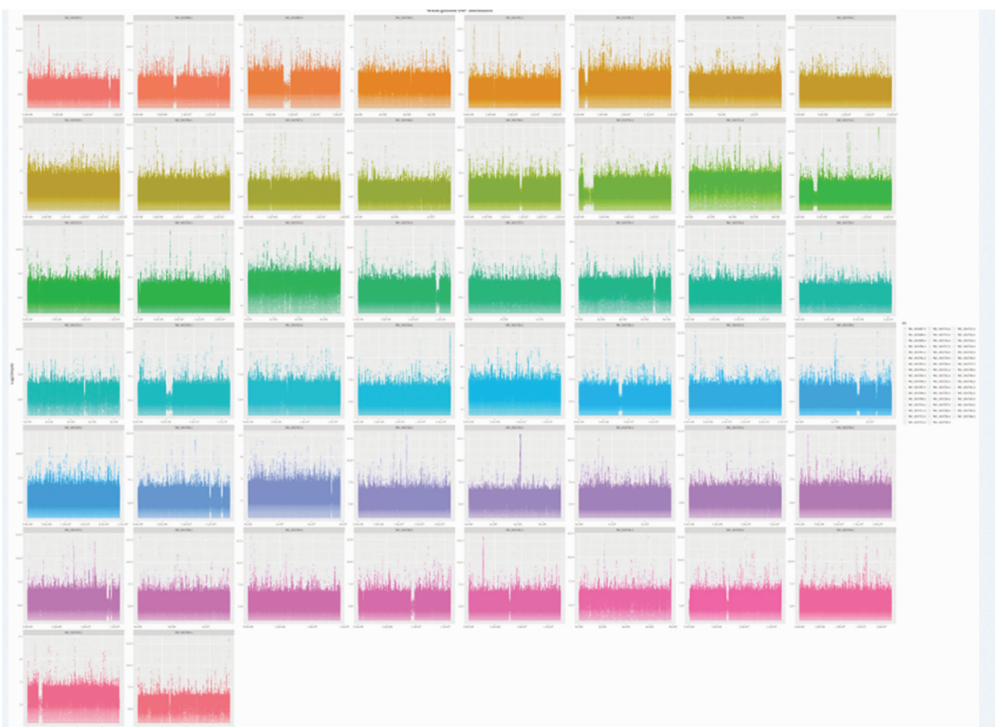
将红白锦鲤皮肤样品的 clean reads 与指定的参考基因组进行序列比对（网址为 ftp://ftp.ncbi.nlm.nih.gov/genomes），比对软件选择 BWA，结果显示，比对效率为 96.35%（表 1）。

表 1　与参考基因组比对的结果

样品编号	比对条带数（条）	比对碱基数（bp）	对比效率（%）	平均测序深度
S1－r2	695 275 942	104 291 391 300	96.35	58.72

2.3　红白锦鲤基因组的 SNP 检测及注释

由图 1、表 2 可知，利用重测序变异检测方法得到的结果显示，以鲤鱼基因组为参考，过滤掉测序深度在 10X 以下的位点，共鉴定了 1 048 576 个 SNPs，其中 312 万 ~540 万个 SNPs 位于短 reads 比对不到的区域，变异位点位于外显子区域的有 579 778 个 SNPs。SNP 位点分布于锦鲤的 50 条染色



横坐标为每条染色体的长度(SNP 所在位置的坐标)，纵坐标为每个 SNP 测序深度的 $\log_2$ 值，不同颜色的小图代表不同的染色体

图1　锦鲤全基因组单核苷酸突变染色体分布

表 2　SNP 注释结果统计

所在区域	SNPs 数量（个）
基因间区	535 520
基因内（无转录本信息）	415 166
基因上游与下游之间	2 260
基因上游区域	24 284
基因下游区域	23 761
基因的 5′UTR 内（UTR_5_PRIME）	5 385
基因的 3′UTR 内（UTR_3_PRIME）	11 934
外显子区域	29 061
剪切受体突变	249
外显子与剪切受体之间	21
5′UTR 与 3′UTR 之间（UTR5:UTR3）	10
非编码 RNA 的外显子区域（ncRNA – exonic）	39
非编码 RNA 的内含子区域（ncRNA – intronic）	885

注：样品为皮肤。

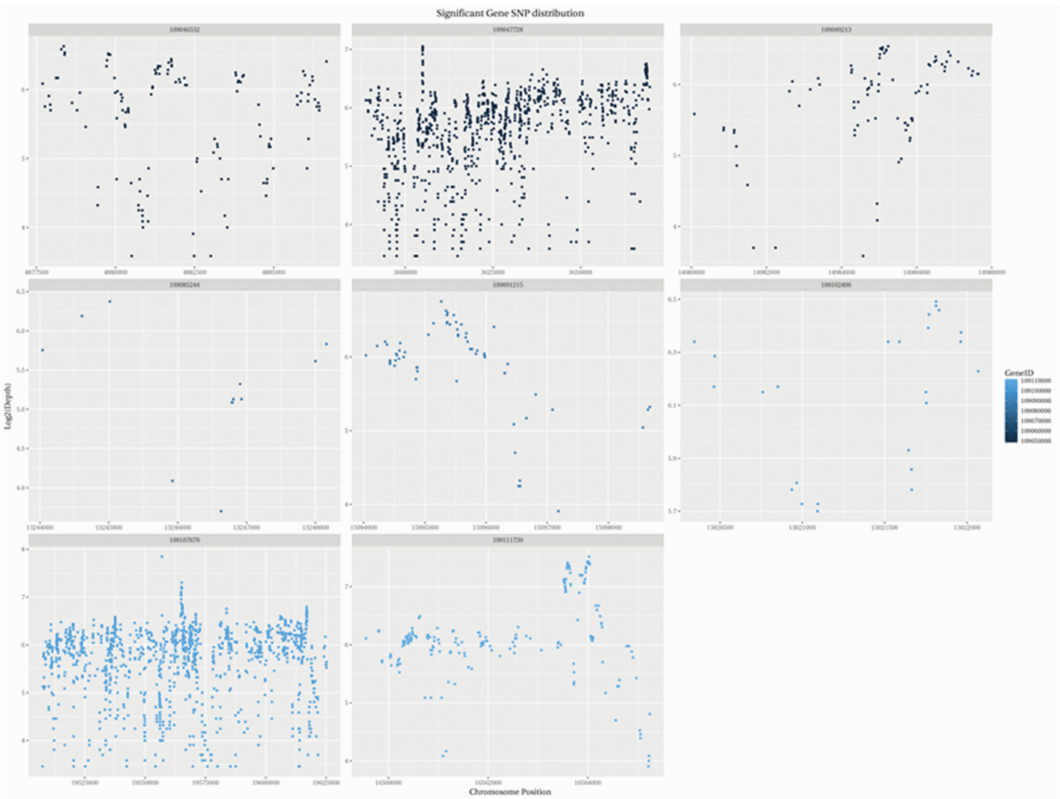
体上，不包含 scaffold（染色体骨架）。用 ANNOVAR 软件进行功能注释，结果显示，纯合类型的 SNPs 数量为 574 310 个，杂合类型的 SNPs 数量为 474 265 个。SNPs 位于基因间的数量最多，SNPs 位于基因内外显子区域的多态性最高，由此可以看出，与鲤鱼相比，红白锦鲤的变异位点很多，且分布在染色体的各个位置。

2.4　红白锦鲤肤色相关基因注释及 SNP 分析

通过对 8 个重要候选基因注释的理解，发现微管蛋白的 2 个基因 LOC109046532、LOC109049213 与色素颗粒运输有关。其中基因 LOC109046532 含有突变，而另 1 个基因 LOC109049213 则不含有任何突变。8 个候选基因都含有外显子 SNP 位点，但是没有发现终止密码子突变，详见图 2、表 3、表 4。

3　讨论

在全基因组测序过程中，基因组 DNA 的提取和检测是关



横坐标为每个基因的长度(SNP 所在位置的坐标)，纵坐标为每个 SNP 测序深度的 $\log_2$ 值，不同颜色的小图代表不同的基因

图2　候选基因 SNP 分析结果分布

键。通常情况下，DNA 的检测主要是通过 NanoDrop 检测 DNA 纯度（$D_{260\ nm}/D_{280\ nm}$ 值），用 Qubit 对 DNA 浓度进行精确定量[12]。其中 $D_{260\ nm}/D_{280\ nm}$ 值在 1.8 ~ 2.0 之间，DNA 浓度≥20 ng/μL，总量为 1 μg 以上的 DNA 样品被用来建库。在本试验中，红白锦鲤皮肤样品 DNA 的 Q20 碱基百分比在 95.59% 及以上，Q30 碱基百分比在 90.81% 及以上，GC 含量为 37.32% ~42.38%，测序错误率为 0.07%，可见样品质量满足建库测序要求，且总量满足 2 次或者 2 次以上的建库需要。对测序获得的 reads 数据进行质量过滤得到 clean reads，用于后续生物信息学的分析。将 clean reads 与参考基因组进行比对，基于比对结果，使用 samtools[1] 进行去重复（mark duplicates），用 GATK[2] 进行局部重比对、碱基质量值校正等处理，再使用 GATK 进行单核苷酸多态性的小片段插入缺失（small INDEL）的检测、过滤，并得到最终的 SNP 和 Small INDEL 的位点集。通过 BreakDancer[3] 可得到结构变异（structure variation，简称 SV）数据集，其中一般以插入和缺失为主。并对 SNP 的检测结果进行注释，实现 DNA 水平差异基因的挖掘和筛选等。利用基因组比对软件 BWA[1]，将过滤

表 3　8 个候选基因外显子 SNP 数量统计

序号	基因编号	外显子 SNP 数量（个）
1	109046532	26
2	109085244	2
3	109107670	2
4	109091215	4
5	109049213	0
6	109102406	3
7	109047728	62
8	109111739	97

后的 clean reads 比对到参考基因组上，统计比对结果。对于重测序分析而言，比对率以及覆盖度指标能反映样本、建库及测序以及参考序列等的质量。在本试验中，将 clean reads 与鲤鱼参考基因组序列进行比对，结果显示，mapping 率达到 96.3%，说明测序样本与鲤鱼参考基因组的相似度很高。

SNP 检测主要使用 GATK 软件工具包[2]。根据 clean reads 在参考基因组的定位结果，使用 SAMtools[3] 进行去重复（mark duplicates），使用 GATK 进行局部重比对、碱基质量值校正等预处理，以保证检测得到的 SNP 的准确性，再使用 GATK 进行单核苷酸多态性的检测、过滤，并得到最终的 SNP 位点集。SNP 是通过 ANNOVAR 软件进行注释的。SNP 分布

表 4　候选基因 LOC109046532 基因突变分析

染色体登录号	基因登录号	突变类型	SNP 类型	SNP 位置（bp）
NC_031739.1	109046532	同义突变（synonymous）	单核苷酸位点变异（SNV）	4 877 888
NC_031739.1	109046532	synonymous	SNV	4 877 891
NC_031739.1	109046532	synonymous	SNV	4 877 933
NC_031739.1	109046532	synonymous	SNV	4 877 945
NC_031739.1	109046532	synonymous	SNV	4 878 119
NC_031739.1	109046532	synonymous	SNV	4 878 176
NC_031739.1	109046532	synonymous	SNV	4 878 284
NC_031739.1	109046532	synonymous	SNV	4 878 359
NC_031739.1	109046532	synonymous	SNV	4 878 386
NC_031739.1	109046532	synonymous	SNV	4 878 392
NC_031739.1	109046532	synonymous	SNV	4 878 623
NC_031739.1	109046532	synonymous	SNV	4 878 870
NC_031739.1	109046532	synonymous	SNV	4 884 521
NC_031739.1	109046532	synonymous	SNV	4 884 548
NC_031739.1	109046532	synonymous	SNV	4 884 578
NC_031739.1	109046532	非同义突变（nonsynonymous）	SNV	4 884 679
NC_031739.1	109046532	synonymous	SNV	4 884 761
NC_031739.1	109046532	synonymous	SNV	4 884 773
NC_031739.1	109046532	synonymous	SNV	4 884 788
NC_031739.1	109046532	synonymous	SNV	4 884 794
NC_031739.1	109046532	synonymous	SNV	4 884 848
NC_031739.1	109046532	synonymous	SNV	4 884 854
NC_031739.1	109046532	synonymous	SNV	4 884 893
NC_031739.1	109046532	synonymous	SNV	4 884 914
NC_031739.1	109046532	nonsynonymous	SNV	4 884 985
NC_031739.1	109046532	synonymous	SNV	4 885 208

图通过 R 语言 ggplot2 包进行绘制展示。在本试验中，将锦鲤测序数据比对到参考基因组上，以分析 SNP 位点的分布情况，为了使 SNP 连续显示，过滤去除了测序深度在 10X 以下的位点，共鉴定了 1 048 576 个 SNPs，其中 3.12 百万 ~5.40 百万个 SNPs 位于短 reads 比对不到的区域，其中变异位点位于外显子区域的有 579 778 个 SNPs。此外，统计结果显示，SNPs 位于基因间的数量最多，SNPs 位于基因内的外显子区域的多态性最高，由此可以看出，与鲤鱼相比，红白锦鲤的变异位点很多，且分布在染色体的各个位置。此外，没有发现外显子 SNP 位点含有终止密码子突变，因此 SNP 位点并没有影响基因的正常编码和表达。可是就目前发现的 SNP 位点而言，由于鲤鱼基因组缺乏相应的 SNP 功能注释信息，无法看出 SNP 位点会对相应基因功能带来何种变化，可能需要进行进一步的功能验证试验。

鱼类细胞形态变化、定向运动、胞内物质（如色素颗粒）与“器官”的移迁（有丝分裂、减数分裂中的染色体极向移动）都与微管蛋白的聚合与解聚相关[13]。微管是由微管蛋白亚基组装而成的，每个微管蛋白亚基都是由 2 个非常相似的球状蛋白（α－微管蛋白和 β－微管蛋白）结合而成的异二聚体，这种 α－β 二聚体是微管组装的基本结构单位[14]。鱼类体色的重要调控机制之一是通过微管蛋白对色素颗粒的靶向运输[15]。在本试验中，1、5 号基因为微管蛋白基因，与色素颗粒运输有关。鲤鱼基因组 *gff* 的基因信息全部是由美国国立生物技术信息中心（NCBI）网站上 Gnomon 预测软件进行预测的结果，因此该基因组并没有完整、真实的数据来进行支撑。因此，由 BLAST 得到的这 8 个候选基因的名称都是以其在染色体上的位置进行命名的，至于其功能也是由预测软件进行功能注释的。

匡　琛,朱晓义,张　亮,等. 含多酶切位点融合黄色荧光蛋白的植物通用载体的构建与应用[J]. 江苏农业科学,2019,47(16):56 – 62.
doi:10.15889/j.issn.1002 – 1302.2019.16.012

含多酶切位点融合黄色荧光蛋白的植物通用载体的构建与应用

匡　琛[1,2], 朱晓义[1], 张　亮[1], 范世航[1], 华　玮[1]
(1. 中国农业科学院油料作物研究所/农业农村部油料作物生物学与遗传育种重点实验室,湖北武汉 430062;
2. 中国农业科学院研究生院,北京 100081)

摘要:以植物表达载体 pCAMBIA3301 为基本骨架,以黄色荧光蛋白(yellow fluorescent protein,简称 YFP)为标签蛋白构建可融合目标蛋白的表达载体,并包含可用于外源基因插入的单一识别位点的核酸酶酶切位点(*Spe* Ⅰ、*Xba* Ⅰ、*Sma* Ⅰ、*Bam*H Ⅰ)。为了验证载体的实用性,将构建完成的载体转化到感受态 GV3101 农杆菌上,进行菌落 PCR 鉴定,再分别瞬时转化烟草下表皮和稳定转化拟南芥。激光共聚焦显微镜观察结果显示,在阳性转基因植株上均观察到荧光,在阴性对照上没有观察到荧光,表明 YFP 标签蛋白在转基因受体细胞中能够正常表达。pCAMBIA3301::YFP 载体的成功构建为植物蛋白亚细胞定位及过表达转基因植株等相关领域的研究提供了稳定可靠的通用型载体资源。

关键词:多酶切位点;黄色荧光蛋白;通用载体;激光共聚焦;植物基因表征工具
中图分类号: S188　　**文献标志码:** A　　**文章编号:**1002 – 1302(2019)16 – 0056 – 07

植物表达载体是高等植物基因功能研究中不可或缺的工具,在后基因组时代具有广泛的应用[1]。植物表达载体可以用来研究启动子元件、蛋白质互作、亚细胞定位以及组成型表达。植物双元表达载体的发展和不同农杆菌菌株的选育,使得农杆菌介导的植物转化系统得到广泛应用[2]。早期的植物双元表达载体主要是由复制起始位点、2 个 T – DNA 边界重复序列之间的部分、大肠杆菌筛选标记、目的基因与启动子元件及植物筛选标记组成的。植物表达载体的构建方法有传统的酶切连接法[3]和 T – DNA 插入法[4]。随着现代生物技术的发展,出现了多种新的技术,如 Gateway 法[5]、不依赖于

收稿日期:2018 – 04 – 04
基金项目:湖北省自然科学基金(编号:2015CFB348);国家“863”计划(编号:2013AA102602)。
作者简介:匡　琛(1994—),男,湖南益阳人,硕士研究生,研究方向为功能基因组学与蛋白质组学。E – mail: junyuankuang@icloud.com。
通信作者:华　玮,博士,研究员,主要从事油菜功能基因组学研究。E – mail:huawei@oilcrops.cn。

参考文献:

[1]Altshuler D, Pollara V J, Cowles C R, et al. An SNP map of the human genome generated by reduced representation shotgun sequencing[J]. Nature,2000,407(6803):513 – 516.
[2]水　科. 鲤鱼全基因组序列图谱绘制完成[N]. 中国渔业报,2014 – 10 – 13(A03).
[3]陈小明,李佳凯,王志勇,等. 基于简化基因组测序的大黄鱼耐高温性状全基因组关联分析[J]. 水生生物学报,2017,41(4):735 – 740.
[4]刘　峰. 半滑舌鳎经济性状的遗传评估及基因组选择初步研究[D]. 上海:上海海洋大学,2015:37 – 40.
[5]Davidson W S, Koop B F. ICSASG international collaboration. Sequencing the Atlantic salmon (*Salmo salar*) genome the old fashioned way[R]. Plant & Animal Genomes XIX Conference,2011, San Diego,CA,USA:33 – 41.
[6]Liu J. Strategies for efficient assembly and annotation of the catfish whole genome sequence[R]. Plant & Animal Genomes XIX Conference,2011,San Diego,CA,USA:49 – 53.
[7]张晓军. 中国甲壳动物学会第十一届年会暨学术研讨会论文摘要集[C]//中国海洋湖沼学会甲壳动物学分会,中国动物学会甲壳动物学分会,2011:18 – 19.
[8]Zhang G F, Guo X M, Li L, et al. The oyster genome project: an update[C]// Ninth International Marine Biotechnology Conference. Qingdao,China,2010:371 – 379.
[9]Li H, Durbin R. Fast and accurate short read alignment with Burrows – Wheeler transform[J]. Bioinformatics, 2009, 25(14): 1754 – 1760.
[10]McKenna A,Hanna M,Banks E,et al. The genome analysis toolkit:a MapReduce framework for analyzing next – generation DNA sequencing data[J]. Genome Research,2010,20(9):1297 – 1303.
[11]Wang K,Li M,Hakonarson H. ANNOVAR:functional annotation of genetic variants from high – throughput sequencing data[J]. Nucleic Acids Research,2010,38(16):e164.
[12]莫惠栋,顾世梁. 基因组长度的估计方法[J]. 科学通报,2000,45(13):1414 – 1419.
[13]尹云厚. 中药复方制剂对缺氧大鼠微管蛋白和驱动蛋白表达影响的研究[D]. 长春:中国人民解放军军需大学,2003:156 – 158.
[14]Hirokawa N, Takemura R. Kinesin superfamily proteins and their various functions and dynamics[J]. Experimental Cell Research, 2004,301(1):50 – 59.
[15]薛继鹏. 三聚氰胺、氧化鱼油和脂肪对瓦氏黄颡鱼生长和体色的影响[D]. 青岛:中国海洋大学,2011:125 – 128.