华大基因
BGI

科技服务方案手册

农学篇

Agronomy Research

基因科技造福人类

Omics for all

深圳华大基因股份有限公司

华大科技

电话：400-706-6615

邮箱：info@genomics.cn

网址：www.bgitechsolutions.com

地址：深圳市盐田区洪安三街21号（518083）

此书献给

“

广大前沿的科研工作者

希望您能从本书中寻找到

新的科研思路

”

CONTENTS
目 录

01

02

03

BGI

引 言

动植物研究

动植物育种的历史源远流长。从数千年前神农尝百草开始，采用存优汰劣的留种技术，选育出了对于中华文明几千年繁衍至关重要的五谷(稻、黍、稷、麦、菽)和六畜(马、牛、羊、鸡、犬、豕)，以及丰富的蔬菜(白菜、萝卜等)和水果(桃、杏、李、梨、桔等)品种资源。

最初的育种基本上是依靠育种家的经验以适应环境或最佳产量等表型为依据进行选择。在孟德尔的遗传学分离规律和自由组合规律被发现和重视后，育种家们才逐渐了解关注的表型背后的遗传逻辑。在传统育种中，育种家们通过系统选择、杂交、理化诱变等方法培育人类需要的动植物新品种。但是，传统育种也有其局限性，主要是通过表型进行选择，往往我们知其然不知其所以然，并且表型容易受到环境的影响；另外有一些表型需要培育很久才能体现或者某些性状无法鉴定（如公牛的产奶性状）；再者，传统育种技术只能适用于那些能够发生有性杂交的物种，而难以突破种与种之间的遗传屏障，所以导致传统育种周期长、效率低等缺点。

近20年来，随着分子生物学、分子遗传学、基因组学的迅速发展，通过基因型选择辅助进行动植物育种得到广泛应用，分子育种应运而生。基因型从本质上决定了性状能够发展的极限，是生物体在适当环境条件下获得某种表型的内因。而表型则是基因型和环境条件共同作用的结果。分子育种过程中挖掘基因或找到与性状相关基因紧密连锁的分子标记是十分关键的，当然性状的表现还会受到表观调控的影响。现在，高通量测序技术在分子育种的方方面面都发挥着强大的推动作用，不仅能够帮助快速检测目标性状关联基因，同时挖掘基因、蛋白表达的调控机制。

中心法则是动植物生命最基本的遗传规律，它揭示了遗传信息在DNA、RNA和蛋白质间的传递方式，以及它们彼此间的相互作用。中心法则任何阶段的变化都会影响表型的展现。华大基因拥有多种平台，可以在DNA水平、RNA水平、表观遗传学等各个水平对动植物各种感兴趣表型性状进行全方位的研究，并结合质谱技术开展蛋白质组水平的研究，利用贯穿组学分析深度解析动植物界的科学问题，检测与人类息息相关的农艺性状相关基因、研究动植物进化、抗病、抗逆、生殖发育等生理机制，为育种挖掘多样性的遗传资源，为提高动植物育种进程奠定坚实的理论基础。

基于多平台的动植物基因组 *de novo*研究方案

001

研究背景

动植物*de novo*测序即动植物从头测序，指不需要任何参考序列信息即可对某个物种进行测序，用生物信息学分析方法进行拼接、组装，从而获得该物种的基因组序列图谱。利用全基因组从头测序技术，可以获得动植物的全基因组序列，带动这个物种下游一系列研究的开展，推动该物种的研究。全基因组序列图谱完成后，可以构建该物种的基因组数据库，为该物种的后基因组学研究搭建一个高效的平台，为后续的基因挖掘、功能验证提供DNA序列信息。

测序技术的不断进步和大数据处理能力的提升让我们对于参考序列的精准度提出了更高的要求。随着PacBio、Nanopore测序技术的商业化，越来越多已经用illumina测序技术的物种又开始重新回炉组装就足以说明这个趋势。为了让组装结果更趋于完整和准确，很多辅助组装技术也应运而生，比如Hi-C，这些辅助技术和长读长测序技术（PacBio/Nanopore）联合，让组装结果更加完整和准确，直逼染色体水平。

目前动植物基因组*de novo* 产品涉及的平台有PacBio、Nanopore、Hi-C、Genomics 以及DNBSEQ、illumina测序平台，每个平台都有自己的特点和使用范围，各平台的详细介绍见文后附录部分。

方案设计

2.1 拟解决的关键科学问题

本研究借助PacBio /Nanopore长读长测序为主，辅助以Hi-C技术，通过多种技术或平台进行结合从而获得连续性最完善的高质量的基因组参考序列。同时采用新技术手段也可以为以后发表文章增加亮点。

2.2方案

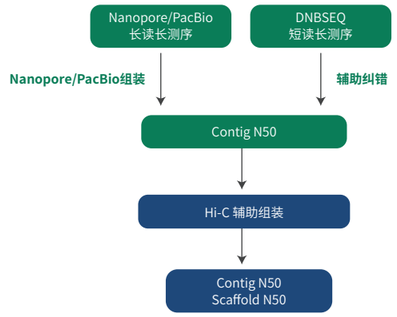


图1 多组学组装策略

2.3测序策略

表1 多平台组装测序策略

分析内容	测序平台	产品类型	测序深度	测序平台
组装	PacBio/Nanopore	*De novo*测序	100 X	PacBio/Nanopore
	Hi-C	Hi-C辅助定位	100 X	Hi-C
	DNBSEQ	300-400bp文库PE150用于纠错	50-100X	HiSeq/BGISEQ平台
注释	全长转录组	辅助注释	10-20G	PacBio

2.4送样要求

送样请遵循以下原则：

1) 所有测序样本DNA取自同一个个体;

2) 如果不能满足也要尽量保证取样的植物遗传背景尽量一致，遗传差异尽可能小。

表2 不同平台样本选择

测序平台	文库类型	样品类型	样品浓度及纯度	样本量
PacBio	20Kb/30KbDNA文库	完整且无污染的DNA	20ng/ul	20ug/文库
Hi-C	Hi-C文库	甲醛固定交联好的样本	——	——
HiSeq/BGISEQ平台	300-400bp 文库	gDNA	20ng/ul	3ug

2.5项目执行周期

从样本合格开始计算周期，周期6-8个月。

组装步骤

组装步骤	详细分析内容
组装	1) 数据纠错; 2) 组装; 3) 组装结果长读长纠错; 4) 组装结果短读长纠错; 5) BUSCO 评价;
Hi-C数据辅助组装	1) 数据测试; 2) Hi-C分析; 3) 手工矫正，获得染色体; 4) 近缘物种比较，染色体定名。

研究案例

4.1 多平台联合组装经典案例

表3 多平台组装案例

发表时间	杂志名称	物种名称	组装策略	组装指标
2019.12	Nature	睡莲	PacBio 122X +Hi-C	ContigN50=2.1Mb
2019.09	Nature Genetics	菠萝	PacBio+HIC+illumina	Contig N50 = 427Kb
2019.07	Nat Plants	香蕉B	PacBio 113X+Hi-C 138X+illumina166X	Contig N50=1.83Mb; Scaffold N50=5Mb;
2018.08	GigaScience	单叶省藤和黄藤[2]	单叶省藤: Illumina209X+PacBio 40X+Hi-C75X; 黄藤: illumina 216X+PacBio 49X+Hi-C96X	单叶省藤: Contig N50=99Kb; Scaffold N50=160Mb; 黄藤: Contig N50=90Kb; Scaffold N50=119Mb
2017.04	Nature	大麦[7]	BAC 4.5T+ Hi-C 50X + BioNano 57X	Contig N50=79Kb; Scaffold N50=1.9Mb
2017.03	Nature Genetics	山羊[8]	PacBio 69X +BioNano 98X+Hi-C 8X +Illumina 23X	Contig N50=18.7Mb; Scaffold N50=87.3Mb

注:标蓝色文章华大有参与。

案例一：睡莲基因组和早期开花植物进化

The water lily genome and the early evolution of flowering plants

发表期刊: *Nature*

发表日期: 2019年12月

主要研究团队: 福建农林大学, 南京农业大学、比利时根特大学和美国宾州州立大学、华大基因

测序策略: PacBio +Hi-C

研究结果: 1.本项目通过122X的PacBio测序数据，组装得到蓝星睡莲的高质量基因组，组装基因组大小为409Mb,组装指标为Contig N50=2.1Mb。通过系统发育组分析显示无油樟是最早分化出的被子植物类群,其次是睡莲。

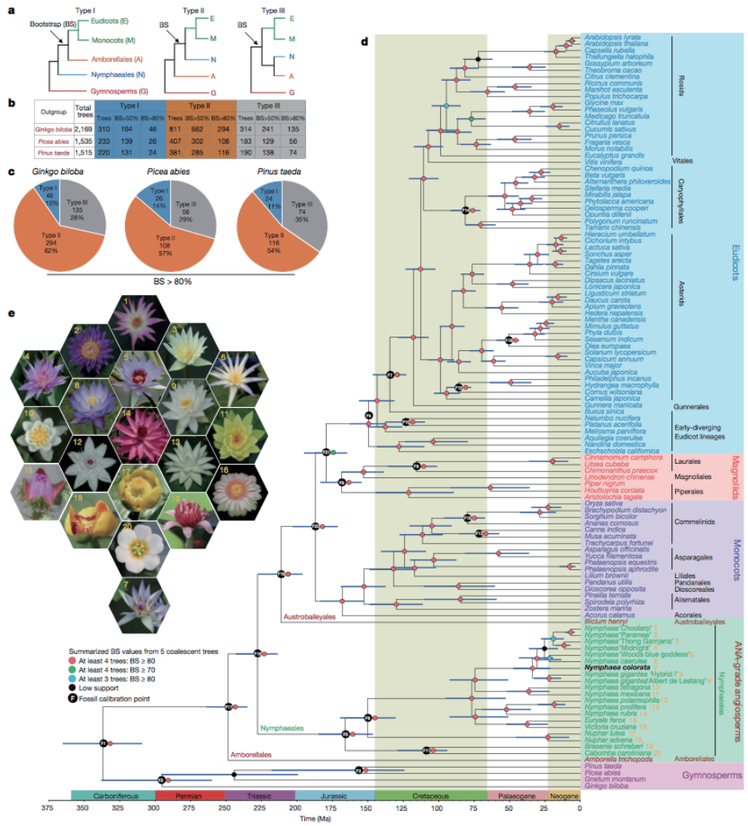


Fig. 1 | Phylogenomic relationships of angiosperms. a, Three different evolutionary relationships among major clades of angiosperms. **b**, Number of LCN gene trees with different bootstrap support (BS) values based on nucleotide sequences from six eudicots, six monocots, *N. colorata*, *Amborella* and three different gymnosperms. **c**, Comparison of gene trees supporting the three evolutionary relationships using each gymnosperm in turn as the outgroup. The percentage was calculated by dividing the number of type I, II or III trees (BS > 80%) by the total number of trees. **d**, Summary phylogeny and timescale of 115 plant species. Blue bars at nodes represent 95% credibility intervals of the estimated dates. **e**, The flowers of the 20 sampled water lilies in Nymphaeales used in **d**.

图1 睡莲基因组进化关系

2.通过基因组和转录组分析，睡莲科祖先发生了一次基因组加倍事件，这个事件可能在睡莲目祖先时期发生的。

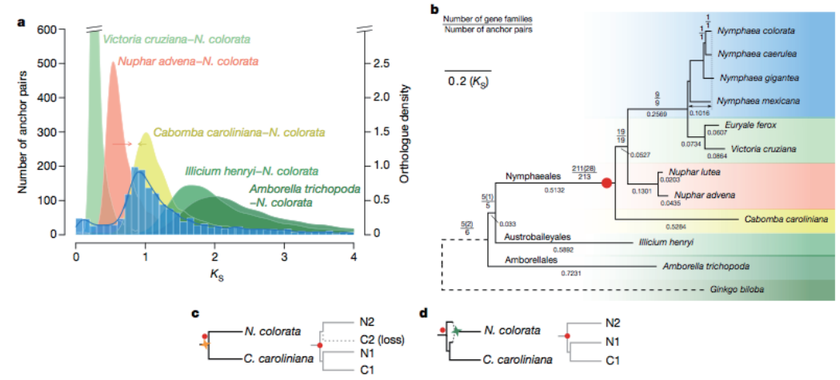


Fig. 2 | A Nymphaealean WGD shared by Nymphaeaceae and possibly Cabombaceae. a, K_S age distributions for paralogues found in collinear regions (anchor pairs) of *N. colorata* and for orthologues between *N. colorata* and selected Nymphaealean and angiosperm species. Red and yellow arrows indicate under- and overestimations of the *N. colorata*–*Nuphar advena* and *N. colorata*–*C. caroliniana* divergence, respectively. **b**, WGD phylogenomic analysis. Numbers in parentheses are the number of gene families with retained *C. caroliniana* duplicates supporting the duplication events. Numbers below branches show branch lengths in K_S units. The double-arrowed line denotes total K_S from the pointed node to *N. colorata*. We used *G. biloba* (dashed branch) as an outgroup. The red dot denotes the branch on which most of the anchor pairs in *N. colorata* coalesced. All mapped duplication events have BS ≥ 80% in the gene trees. **c**, Left, the scenario of a WGD (yellow four-pointed star) before the divergence between Nymphaeaceae and Cabombaceae. Right, a possible gene tree under this scenario, with loss of one duplicate in *C. caroliniana*. Two red dots show where the anchor pair of *N. colorata* would coalesce. **d**, Left, scenario of a WGD in the stem lineage of Nymphaeaceae involving an allotetraploid (green four-pointed star) that formed between two ancestral parents after the divergence of the lineages leading to *N. colorata* and *C. caroliniana*, with one of the parents being more closely related to *C. caroliniana*. Right, a gene tree under such a scenario. Red dots are as in **c**.

图2 睡莲基因组全基因组复制事件

3.确定了睡莲ABCE 模型，揭示早期被子植物花发育特征。

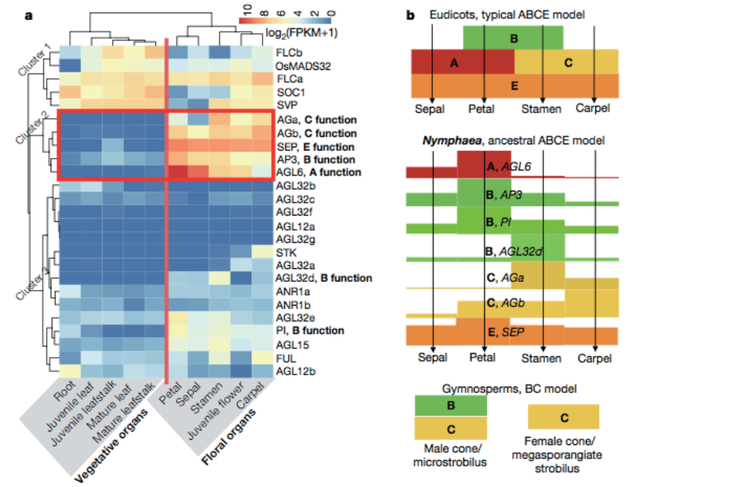


Fig. 3 | MADS-box genes in *N. colorata* and proposed floral ABCE model in early angiosperms. a, Gene expression patterns of MIKC^c from various organs of *N. colorata*. Three clusters of genes were classified according to the expression of type II MADS-box genes. The organ types (vegetative organs and floral organs) were matched to the expression patterns of type II MADS-box genes. Expression values were scaled by $\log_2$(FPKM + 1), in which FPKM is fragments per kilobase of exon per million mapped reads. **b**, The flowering ABCE model in *N. colorata* that specifies floral organs is proposed based on the gene expression values (bar heights) from **a**.

图3 睡莲花发育模型

4.本通过研究通过比较自然变异的白色花瓣蓝星睡莲与蓝色花瓣的转录组，发现两个重要基因可能编码蓝色花瓣合成途径关键酶。睡莲的花香有11种成分主要是萜类和脂肪酸等，其中合成倍半萜基因与单双子叶中已知基因不一样。

附录各平台技术介绍

6.1 PacBio 平台

PacBio 平台是Pacific Bioscience公司开发的长读长测序技术。目前上市的测序仪有PacBio RSII和Sequel，动植物基因组主要采用Sequel进行测序 。

（1）PacBio测序原理介绍

PacBio 平台的测序原理为单分子实时测序。一张芯片即一个反应管（SMRTCell：单分子实时反应管）中有许多圆形纳米小孔，即ZMW（零模波导孔），外径100多纳米，比检测激光波长小（数百纳米），激光从底部打上去后不能穿透小孔进入上方溶液区，能量被限制在一个小范围里，正好足够覆盖需要检测的部分，使得信号仅来自这个小反应区域，孔外过多游离核苷酸单体依然留在黑暗中，将背景降到最低。单个ZMW底部有一个结合了模板DNA的聚合酶，这个DNA聚合酶是实现超长读长的关键之一，读长主要跟酶的活性保持有关，主要是激光会对它造成损伤。当加入测序反应试剂后，经过Watson配对后不同的碱基加入，4色荧光标记4种碱基，会发出不同光，根据光的波长与峰值可判断进入的碱基类型。PacBio RSII一个SMRTCell中有15万个ZMW，Sequel 仪器一个SMRTCell 有100万个ZMW，每个孔中有一个单分子DNA链在高速合成，如众星闪烁。原始检测数据的结果，每合成一个碱基即显示为一个脉冲峰，每分钟>100个碱基的速度，配上高分辨率的光学检测系统，就能实时进行检测。

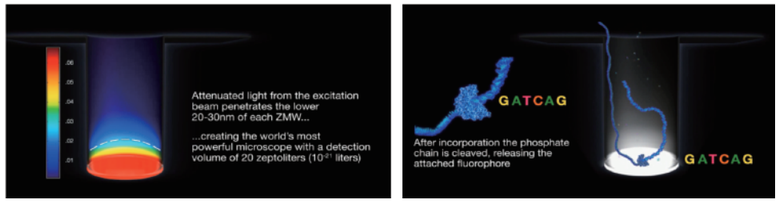


图5 PacBio 测序原理图

（2）PacBio建库流程

1）建库DNA制备：高质量的全基因组长片段DNA；

2）片段筛选：Bluepippin筛选15K以上的片段；

3）接头连接：先把片段粘末端变成平端，两端分别连接环状单链，单链两端分别与双链正负链连接上，得到一个类似哑铃（“套马环”）的结构，称为SMRT Bell。

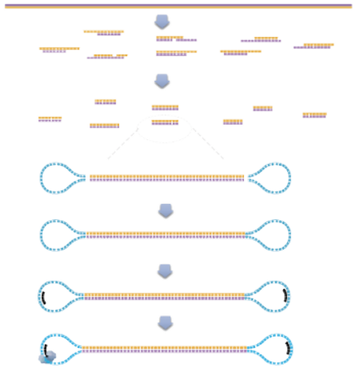


图6 Seqeul DNA文库建库流程图

6.2 Nanopore 平台

（1）Nanopore 平台测序原理

Oxford Nanopore Technologies公司的纳米孔测序技术是基于电信号的单分子实时测序技术，可以直接读取DNA/ RNA分子双链。在测序过程中DNA/RNA双链首先与马达蛋白酶连接并与镶嵌在生物膜上的纳米孔蛋白相结合并解螺旋，蛋白酶马达通过纳米孔控制DNA/RNA双链的移动，位移过程中电流波动的信号被捕获，通过算法计算转换成可识别的碱基序列，完成实时测序。

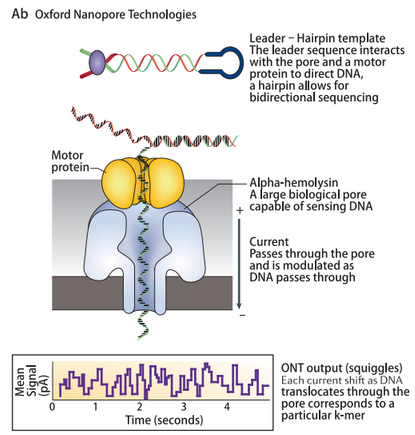


图7 Nanopore 测序原理示意图

（2）Nanopore 平台建库流程

在获得合格的高质量DNA后，采用Oxford Nanopore Technologies 公司的建库Kit进行测序文库构建，具体步骤如下：

1）BluePippin全自动核酸回收系统筛选大片段DNA；

2）文库构建（SQK-LSK109连接试剂盒）；

a）每个文库取BluePippin筛选后约1ug的DNA进行损伤修复和末端修复；

b）磁珠法纯化回收修复的DNA，Qubit测定浓度并计算回收率；

c）用试剂盒自带的测序接头进行DNA与接头的连接，使每个DNA片段均带有可被识别的特异接头序列；

d）磁珠法纯化连接后的样品DNA，Qubit测定浓度并计算回收率；

3）文库上机测序。

6.3 Hi-C技术介绍

（1）Hi-C技术原理介绍

美国麻省大学医学院分子遗传学家Job Dekker认为基因组空间结构对基因调控极其关键，“染色体的所有其他作用也涉及3D结构”。大约1999年，Dekker开创了一种新的技术，叫做染色体构象捕捉（C3），染色体构象捕捉技术基于染色体间彼此接近区域的物理交联，对这些区域进行测序后可以确定哪些区域发生了交联。2009年，Dekker和同事利用该技术的高通量版本——Hi-C技术，发现人类基因组似乎采用了“不规则球体（fractal globule）”的方式来确保染色体在扭曲时不出现打结的问题，并利用该技术得到人类正常淋巴细胞基因组的三维结构图。

Hi-C技术源于染色体构象捕获（Chromosome Conformation Capture，3C）技术，利用高通量测序技术，结合生物信息分析方法，研究全基因组范围内整个染色质DNA在空间位置上的关系，获得高分辨率的染色质三维结构信息。Hi-C技术不仅可以研究染色体片段之间的相互作用，建立基因组折叠模型，还可以应用于基因组组装、单体型图谱构建、辅助宏基因组组装等，并可以与RNA-seq、ChIP-seq等数据进行联合分析，从基因调控网络和表观遗传网络来阐述生物体性状形成的相关机制。

(2) Hi-C技术流程

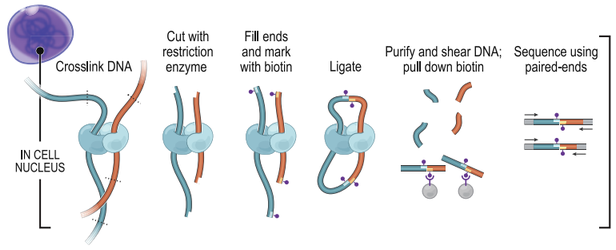


图9 Hi-C建库流程图

Hi-C技术流程如下：

1) 细胞甲醛交联；

2) Hi-C文库构建；

3) 测序及数据质控；

4) Hi-C数据质控比对统计；

5) 辅助基因组组装。

基于*De novo*测序的动物进化研究方案

研究背景

随着测序技术的不断进步和基因组学分析方法的飞速发展，将基因组数据与进化生物学相结合，在基因组水平研究生物进化机制正成为趋势，于是就产生了进化基因组学（Evolutional Genomics）。

从目前的研究来看，进化基因组学的研究主要在两个方面上进行。第一个是系统进化研究，即物种间的进化研究，从基因组水平理解生物学功能和生命现象；第二个是物种自身的进化研究，是依据基因组数据研究基因组自身的进化过程和规律。

1) 系统进化研究

随着人类和一系列模式物种全基因组测序的完成，如何读懂基因组，即阐明基因和基因网络的生物学功能，成为后基因组时代的重大课题。强调历史和比较的进化生物学在这方面也扮演了十分重要的角色，即通过生物信息学手段比较不同物种的基因组或表达组，发现新的功能基因或结构，理解功能的遗传学基础。

利用基因组数据，进化基因组学还可以探讨功能变化和进化的机制。生物在家养条件下变异的遗传学机制是生物界长期以来悬而未决的问题。基于性状和少数基因的观察，一直未能告诉我们究竟是哪些有关的基因导致家养品系和野生品系间的不同。随着高通量测序技术的发展，越来越多的组学数据包括家养生物的基因组序列、相关的转录组学数据的获得，我们有机会从基因组和转录组层面阐明人工选择的基因组学和遗传学机制。类似的研究也适用于野生群体的生态和功能适应研究。

构建生物系统发育的进化树也是进化生物学中的一个传统课题。但我们离重建整个生命之树的目标还很遥远。现在人们提出了雄心勃勃的计划，希望利用整个基因组的数据来重构不同物种的系统发育，并提出了系统发育基因组学（Phylogenomics）这一概念。随着基因组测序技术的不断改进和成本的不断降低，相信人们将有希望彻底解决生物界的系统发育问题。

2）基因组自身的进化研究

在基因组水平，生物的进化体现为基因组的进化。如今繁杂多样的生物界其遗传基础其实就是基因组的多样性。理解基因组演化的历史，破解其规律意味着将达尔文以来的进化生物学推到前所未有的高度和深度，最终有可能实现遗传、发育和进化的统一。

在基因组进化的研究中，基因组的结构是我们审视的重要对象。每一个基因组测序工作完成后，研究人员都要看看基因组的大小、基因分布的型式、重复序列和转座子的多少、与其它生物的差异等等。对进化生物学家来讲，最感兴趣的就是新基因和新结构的产生，因为它们是生物进化的标志。

我们已经知道，不同生物在基因组大小及基因数目上存在巨大的差异。如一种细菌基因组大小为$1.7X10^7$bp，仅含1,700多个基因，而人的基因组大小为$3.0X10^9$bp，基因数目约为3万多个，两者基因数目相差数十倍。从横向上看，正如我们已在果蝇中所观察到的那样，即使分化时间很短的近缘物种间，基因的种类和数目也不尽相同。目前人们已发现了大量的系谱特异的基因，即所谓的独有基因（orphan gene）。这说明生物进化的过程伴随着基因组的大小及基因数目的不断变化。由此引出一个根本性的生物学问题：这些新基因是如何产生的?经过初步的研究，我们发现基因重复、外显子重排、转座和逆转座、强烈正选择等在新基因发生过程中起着重要作用。

在基因组结构进化的探讨中，基因、特别是基因组重复（也叫基因组复制）也是人们关注的一个重要话题。现在认为，基因和基因组重复是生物进化和复杂性增加的重要原因。有科学家提出脊椎动物（vertebrates）的基因组可能经历了两轮重复。目前有许多研究者在进行基因组是否重复、重复以后如何进化等问题的研究，这是进化基因组学的一个热门的研究课题。

目前，研究物种进化的信息分析方法包括1）基因家族鉴定；2）物种进化树构建及系统发育分析；3）正选择分析；4）物种分歧时间估算；5）基因组共线性分析；6）全基因组复制分析；7）基因家族扩张和收缩分析。这是目前基于基因组序列研究物种进化常用的分析方法。

除了目前的信息分析手段，传统的研究进化的方法也可以作为辅助手段相互验证。

1）基于古生物学与化石纪录研究进化：古生物学是以生物化石为基础，研究生物亲缘关系的一种研究。只要是古代生物造成的痕迹，或是生物体本身，都可以称为化石。化石对于了解生物演化历程相当重要，因为它是较为直接的证据，且带有许多详细的资讯。较为常见的化石，通常源自骨骼或外壳等坚硬部位，并经由类似铸模的过程形成。坚硬的骨骼在动物死亡之后，会因为有机物的腐败，而产生一些漏洞。将骨骼掩埋的砂石或矿物，则会经由这些漏洞侵入骨骼内部，并将其填满。这种过程称为置换作用，属于型体的保留，而不是生物体本身的保留。也有一些化石是生物体本身，例如被冰冻的猛犸象、琥珀里的昆虫。此外，古代动物的脚印、或植物在地底下因温度与压力作用而碳化，都可称为化石[2]。

不同时代的生物化石，会出现在不同的地层中，如此便能够研究古生物之间，以及它们与现代生物之间的关系。“失落的环结”指演化过程可能出现过，却尚未被发现的物种。连接两个物种之间的化石，则称为“过渡化石”。例如可能位于鸟类与恐龙中间的始祖鸟（Archaeopteryx）化石，以及一种具有四肢的大型浅水鱼（Tiktaalik）可能是鱼类与两栖类的过渡化石。

2）基于生物地理学与物种分布研究进化：由于板块移动造成的大陆漂移（如南美洲与非洲），以及冰河时期前后造成的海平面高度变化（如白令海峡陆桥），改变了陆地间的相连性。一些相距遥远的地区，可在地下挖出许多相似的生物化石，而海洋或山脉的隔离作用，却使现有的物种具有相当大的差异。例如南美洲的猴、美洲豹、骆马，与非洲的猴、狮子、长颈鹿。此外，与世界上其它地方的胎盘动物相比较为原始的有袋类动物虽已大多灭绝，但澳大利亚大陆却依然保存着袋鼠、无尾熊等许多有袋类动物。

除不同大陆之间有这种现象，大陆与其邻近岛屿因曾在地理上相连，也能够找到相似但变异了的物种。例如中国台湾地区、中国大陆与日本的猕猴之间的差异。

3）基于形态比较研究进化：对脊椎动物五趾肢的比较，支持了脊椎动物具有共同祖先的理论。举例而言，虽然人类、猫、鲸鱼与蝙蝠的五趾肢在型态上有所差异，但是主要架构都很相似。这些“同源”的构造，适应了不同的功能，如抓握、行走、游泳与飞行。

此外有一些构造在功能上相似，但却具有不同的型态。例如蝙蝠、鸟类与昆虫的翅膀；昆虫与脊椎动物的腿；章鱼与脊椎动物的眼睛；鱼类、鲸鱼与龙虾的鳍等。这类“异源”的构造，适应了相同的功能，如飞行、行走、感光与游泳。

4）基于发育形态研究进化：在许多动物的发育初期都非常相似，在发育的过程中，这样的相似会逐渐减少，最后形成各物种的型态。举例而言，虽然各种成熟的脊椎动物差异很大，但是它们的胚胎型态在发育初期却非常相似，鳃裂仍然出现在已经没有鳃的爬虫类、鸟类与哺乳类胚胎中。

在胚胎重演论提出直到被推翻的期间，胚胎学对于演化机制的解释并没有太大的进展。但是演化发育生物学（Evolutionary developmental biology）研究，将分子生物学与发育生物学等学科结合，解释基因的改变对于动物形态的控制过程。同时也发现外表差异相当大的动物之间，也拥有相同的调控基因。以及相同的基因在不同的时间与空间，具有不同的作用。这些调控动物发育过程的基因，主要为一类Hox基因[2]。

在研究物种进化的过程中，要综合考虑多水平的进化证据，相互佐证，如此得出的结论才会比较准确、令人信服。

方案设计

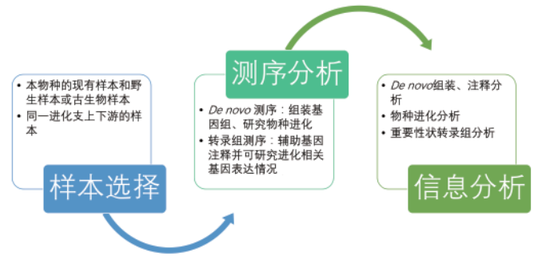


图1 方案设计图

2.1 样本建议

***De novo*选样：**

1）系统发育树根据需要选择常见重要分支节点上的已测物种；

2）本物种的祖先种或野生种或古生物样本。

转录组选样：

不同发育时期的组织样本，选择多时空下的样本以便获得更多的表达基因用于辅助基因组基因注释。

2.2 实验技术

包含但不限于短读长技术及长读长技术平台涉及的不同插入片段大小文库，根据基因组复杂度提供不同的测序策略。

2.3 测序参数

表1 测序策略推荐

分析内容	测序平台	产品类型	测序深度	测序平台
组装	PacBio/Nanopore	*De novo*测序	100 X	PacBio/Nanopore
	Hi-C	Hi-C辅助定位	100 X	Hi-C
	DNBSEQ平台	300-400bp文库PE150用于纠错	50-100X	HiSeq/BGISEQ平台
注释	全长转录组	辅助注释	10-20G	PacBio

注：上文提及的辅助组装手段可全选，也可根据物种特性及技术特点进行选择。

2.4 送样要求

送样请遵循以下原则：

1) 所有测序样本DNA取自同一个个体;

2) 如果不能满足也要尽量保证取样的植物遗传背景尽量一致，遗传差异尽可能小。

具体样本选择如下：

表2 不同平台样本选择

测序平台	文库类型	样品类型	样品浓度及纯度	样本量
PacBio	20Kb/30KbDNA文库	完整且无污染的DNA；	20ng/ul	20ug/文库
Hi-C	Hi-C文库	甲醛固定交联好的样本	——	——
DNBSEQ	300-400bp 文库	g DNA	20ng/ul	3ug

注：以上技术手段根据实际项目需要进行取舍；Hi-C部分我们会提供甲醛交联的流程。

2.5 分析结果

2.5.1 基因家族鉴定

用treefam的方法定义基因家族，基因家族是由来自一个祖先基因的一组基因组成。基因家族的鉴定，是进化分析很重要的一个方面。通过同源基因的聚类及基因家族的鉴定分析，可以得到单拷贝基因家族和多拷贝基因家族。这些家族在物种之间都是比较保守的，可用于物种间亲缘关系的分析。我们还可以得到物种特有的基因和家族，它们可能和物种的特异性表型有关。

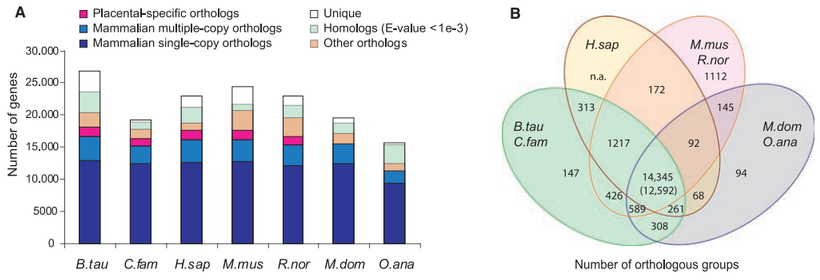


图2 不同物种间直系同源基因种类和数量

A图表示不同物种间直系同源基因的种类及数量；B图表示不同物种间直系同源基因的种类及数量韦恩图[3]

2.5.2 物种系统发育树构建

利用直系同源基因的四重简并位点构建系统发育树；每个分支长度代表中性进化速率；树枝上的数字代表dN/dS。而dN/dS可以反映出物种所受到的纯化选择压力的大小。

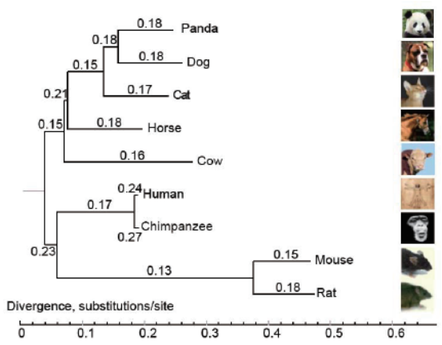


图3 系统发育树[4]

2.5.3 物种分歧时间估算

分化时间和替换速率的估算。分歧时间是在计算序列之间距离的基础上，参考化石分离时间，进行估算。

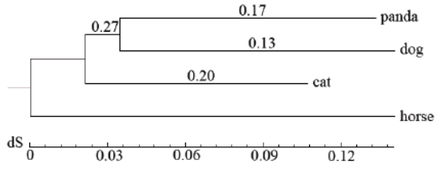


图4 物种分歧时间估算[4]

绿色的数字代表替换速率；蓝色的数字表示估算出来的分化年代，单位是百万年。人和狗的分化年代来自TimeTree database (http://www.timetree.org)，用来作为校正的时间。

2.5.4 基因组共线性分析

全基因组比对结果是比较基因组分析中的一个重要基础，它一般用于识别基因组中的功能元件。例如，通过基因组的多序列比对结果得到的多个远缘物种的同源序列，一般暗示着这些序列是保守的，具有一定的生物特性。

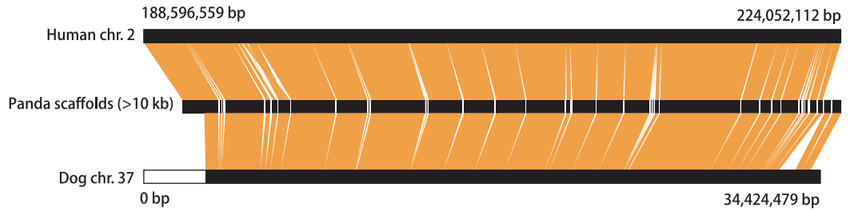


图5 基因组共线性分析[4]

人2号染色体与熊猫Scaffold(＞10Kb)及狗37号染色体进行共线性分析，其中狗37号染色体末端3M区域是完全比对不上。

2.5.5 全基因组复制分析

在植物中一般发生全基因组复制事件，全基因组区段重复分布，由circos软件生成。

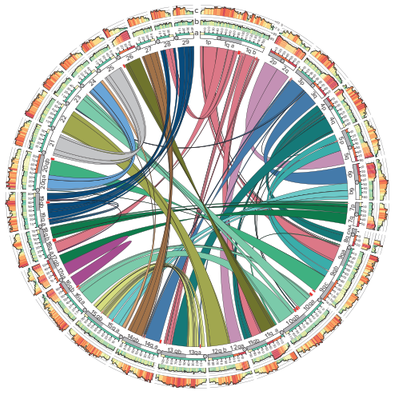


图6 全基因组区段分布图[5]

2.6 项目周期

从样本合格开始计算周期，周期6-8个月。

2.7 预期研究结果

1) 物种进化地位的确定；

2) 验证或修正已有的进化理论；

3) 特殊进化事件基因组层面上的发生机制。

2.8 辅助研究策略

群体重测序可以辅助研究物种内的进化信息，提供更多的进化依据。

2.9 后期验证手段

与化石及历史文字资料记载等信息进行相互验证以保证进化研究结果的准确性。

应用案例

3.1 案例一：44种反刍动物进化分析（华大参与）

标题：大规模全基因组测序揭示反刍动物演化背后的遗传机制

Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits

发表期刊：*Science*

发表日期：2019年6月

主要研究团队：华大生物多样性基因组学研究团队与西北工业大学生态与环境保护研究中心、昆明动物研究所及哥本哈根大学

研究结果：

1) 反刍动物虽然在生物学史上地位显赫，迄今却连科一级的分类都缺乏共识，其独特形状的遗传基础更无从知晓。本项目挑选了44个反刍动物代表物种开展了基因组比较分析，这些物种覆盖了反刍亚目的全部6个科和一半以上属。在该项目旗舰文章中，研究人员通过全基因组数据构建的全新反刍动物系统发育树，解决了关于该类群生命之树的长久争议。

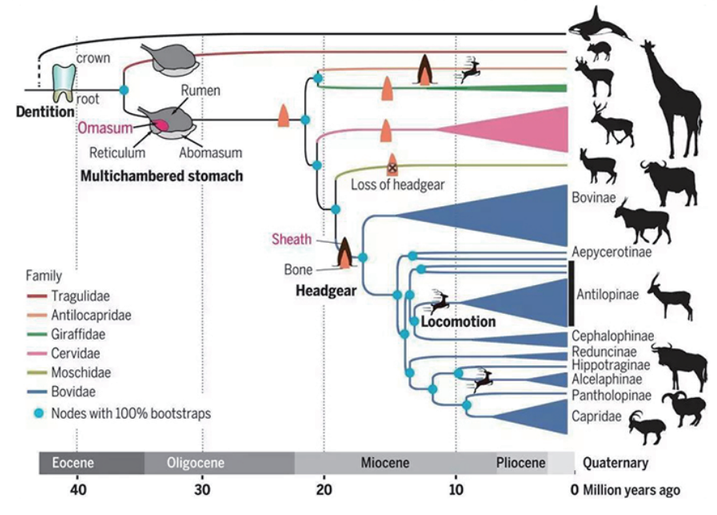


图7 反刍动物系统发生树及关键性状演化历程图

2) 尽管反刍动物在演化过程中是非常成功的物种类群，然而研究发现在最近10万年内反刍动物种群数量急剧下降。通过比较基因组学分析揭示了反刍动物特异的多胃室、体型大小、奔跑能力、独特牙齿形态、免疫和代谢等性状相关的基因演化过程。

3.2 案例二：现代鸟类的进化之谜（华大参与）[6]

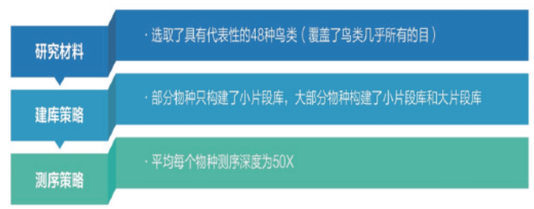


图7 鸟进化研究策略

本项目是由深圳华大基因和深圳国家基因库主导、世界上20多个国家，80多家机构的200多名科学家参与，历经4年完成的鸟类基因组系统演化史项目。

研究结果：

1）完成了48个鸟类物种的基因组测序、组装和全基因组比较分析，包括乌鸦、鸭、隼、鹦鹉、企鹅、朱鹮、啄木鸟、鹰等，囊括了现代鸟类的主要分支；根据所得的鸟类演化树（图8），发现测序物种中占现存鸟类95%物种的Neoaves（新鸟小纲）分成两大分支，分别独立演化出了各自的陆生鸟类和水生鸟类；在Passerea分支中，其陆生鸟类的共同祖先应该是位于生态位顶端的捕食者，而其中具有鸣唱学习能力的鸟类是独立多起源的；在Columbea中，发现鸽子和火烈鸟其实是姐妹分支。

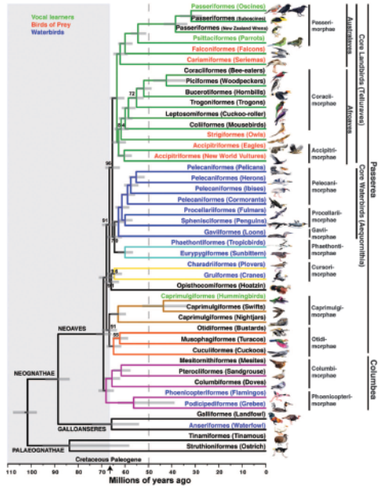


图8 鸟类系统发育树

2）通过全基因组数据的方法推断出的鸟类物种树发展史，与之前得到的结果有巨大差异。研究发现，单纯使用编码蛋白基因来构建演化树与真实物种树具有极大的差异，因此还需要利用非编码序列，包括基因间区。研究还发现，编码蛋白序列在一些具有相似生活史的物种之间存在有意思的趋同演化现象。

3.3 案例三：现代蛙类-高山倭蛙的进化之路（华大参与）[7]

图9 高山倭蛙研究策略

目前蛙类仅有热带爪蟾（Xenopustropicalis）一个物种的基因组被测定。爪蟾属于"古老蛙类"（Archaeobatrachia）的一种，而目前95%以上的蛙类属于"现代蛙类"（Neobatrachia），这在一定程度上限制了我们对两栖动物基因组特性的认识。

研究结果：

1）分歧时间估算：研究人员估算了高山倭蛙和热带爪蟾的分歧时间大概在2.66亿年前，比TimeTree项目中记录的时间早了四千万年。尽管两者分歧时间很久，但是两物种染色体间的重排特别少，说明蛙类基因组可能具有相对较慢的进化速率。

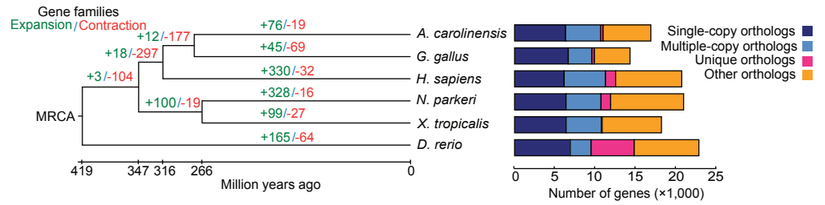


图10 基因家族的扩张与收缩

选取脊椎动物中代表物种构建基因家族动态进化树，发现非洲爪蟾和高山倭蛙的分化时间在2.66亿年前。

2）通过比较高山倭蛙和热带爪蟾二者的基因组，发现前者拥有更大的基因组，且这种基因组大小上的差异主要归因于两者基因组中转座元件的含量不同。高山倭蛙的转座元件主要以长末端重复序列（LTR）为主，而非洲爪蟾则是以DNA转座子（transposons）为主。相对于非洲爪蟾，高山倭蛙中的LTR具有更高的保守性。进一步分析了转座元件在基因组的分布关联情况，发现两物种的转座元件分布模式差异很大。

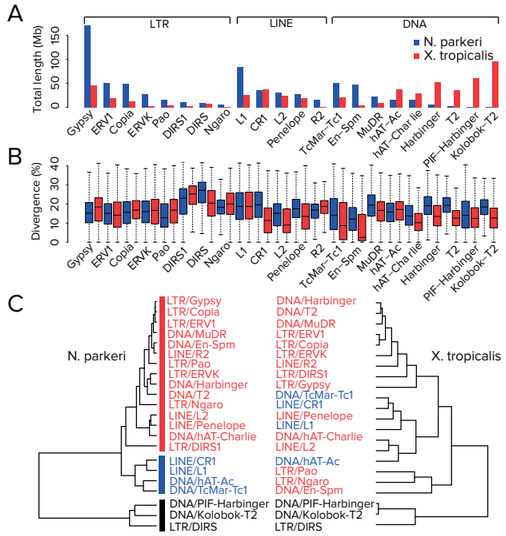


图11 非洲爪蟾和高山倭蛙转座子比较

A图：两个物种中不同转座子家族扩张速率比较，高山倭蛙的扩张速率高于非洲爪蟾；
B图：两个物种中不同基因家族聚类结果。

可能存在的风险

组装结果不好的情况下有些信息有可能丢失，导致最终结果与实际结果有偏差。

常见问题

1、研究物种进化取样需要注意什么？

答：研究不同的物种进化选择样本时需要注意以下几点：

1）根据研究目的选择该物种进化支 上下游的物种及进化支上重要节点的物种；

2）选取进化关系尚不明确或有争议的物种。

华大优势

测序平台丰富：Nanopore/PacBio/BioNano/Hi-C/10X Genomics平台均提供；

质控严格：从样本接收到数据交付都有严格的质量控制流程，保证数据准确性；

丰富的项目经验：已完成几百个*de novo*项目，在CNS等顶级刊物上发表文章超过100篇。

基于*de novo*测序的物种进化文章汇总

中文名	发表时间	刊物
12种果蝇基因组进化分析(10种新)	2007.11	Nature
48种鸟类（45种为新测）	2014.12	Science
10种蜂类(5种为新测）	2015.05	Science
66个水稻泛基因组	2018.01	Nature Genetics
44个反刍动物	2019.06	Science
21种企鹅（19种新测）	2019.09	GigaScience
12只犬类（10只野生澳洲野犬和2只新几内亚歌唱犬）	2020.02	Nature comminications

参考文献

[1] 王文, 宿兵. 进化基因组学简介[J]. 科学中国人, 2004 (5): 50-51.

[2] 恩斯特.迈尔.进化是什么[M].田洺译.上海科学技术出版社，2009.

[3] Elsik C G, Tellam R L, Worley K C. The genome sequence of taurine cattle: a window to ruminant biology and evolution [J]. Science, 2009, 324(5926): 522-528.

[4] Li R, Fan W, Tian G, et al. The sequence and de novo assembly of the giant panda genome[J]. Nature, 2010, 463(7279): 311-317.

[5] Lien S, Koop B F, Sandve S R, et al. The Atlantic salmon genome provides insights into rediploidization[J]. Nature, 2016.

[6] Jarvis E D, Mirarab S, Aberer A J, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds [J]. Science, 2014, 346(6215): 1320-1331.

[7] Sun Y B, Xiong Z J, Xiang X Y, et al. Whole-genome sequence of the Tibetan frog Nanorana parkeri and the comparative evolution of tetrapod genomes[J]. Proceedings of the National Academy of Sciences, 2015, 112(11): E1257-E1262.

[8] Luo Y J, Takeuchi T, Koyanagi R, et al. The Lingula genome provides insights into brachiopod evolution and the origin of phosphate biomineralization[J]. Nature communications, 2015

基于高通量测序的种内群体进化研究方案

研究背景

现在我们赖以生存的重要的粮食作物、家禽、家畜，以及其它的动植物都是由原始祖先经过自然选择、人工驯化等过程发展而来。在物种进化的过程中，伴随着表型更适应环境的变化、迎合人类的需求，其遗传物质也发生了很大的改变。利用高通量测序技术，科研工作者可以从DNA层面开发遗传标记，研究在进化过程中物种遗传结构的变化、哪些基因受到定向选择、进化过程中种群历史的变化、物种间基因交流情况等。研究群体进化问题可以揭示物种起源，同时开发重要经济性状和环境适应性相关的基因，为育种提供基因资源。

动植物的群体进化研究可以解决以下几个问题：

1）通过研究野生种和驯化种间的遗传差异，获得驯化动植物的驯化位点（选择性清除位点，selective sweep），结合相关区域的基因功能注释信息，挖掘驯化相关的基因。

2）对不同亚种的育成动植物进行群体研究，了解育成动植物的驯化条件、驯化过程及其进化动力、复杂性状变异及其分子机制等问题。

3）研究不同地理环境下野生动物的遗传特点，分析种群遗传结构及其演化历史。通过绘制该物种完整的种群演化历史，可以探寻其种群变迁及濒危的主要原因，为深入理解野生动物的种群演化规律，同时也为保护濒危动物提供基因组学的分析依据。

4）通过研究野生种、或化石、或群体多态性，研究该物种起源，同时通过评估各亚群遗传信息的来源，判断亚群间基因交流情况，研究物种进化路径。

过去的几十年中，人们借助考古学与分子遗传学等方法对动植物的起源与驯化问题开展了大量研究。在动物的驯化中，考古学通过对考古遗址的研究提供人类驯化的直接证据。例如，在出土于近东和欧洲东南部墓葬的陶罐内找到了人类7,000年前已饮用牛奶的证据[1]，在哈萨克斯坦的EneolithicBotai文化遗址中发现5,500年前的马骨骼有套过缰绳的痕迹[2]。分子遗传学通过对分子标记的分析来追溯驯化动植物的起源历史，阐明了许多考古学无法解决的问题。而两个学科的联合为解决驯化问题提供了更广阔的视角，例如运用分子遗传学方法分析考古发现的古DNA样本，已成功揭示了家犬和家猪的起源地与迁移模式[3,4]。而大豆、水稻等重要农作物则有考古学证据和分子生物学证据表明其起源于东亚[5,6]。

自达尔文以来，人工选择被认为是驯化动植物的主要进化动力，通过持续定向的高强度选择压，使满足人类需求与喜好的表型在短时间内固定[7]。受到人工选择的位点，通常会具有选择性清除（selective sweep）的信号，如核苷酸多样性显著下降、连锁不平衡（LD）值升高，与野生祖先种遗传距离（Fst）变大，或者该基因的频率分布谱有所改变等。

进化导致的性状改变被称为进化相关性状，从分子进化角度来看，这些表型的改变是对目标基因强烈定向选择的结果。寻找进化基因并探讨其进化的分子机制是动植物进化研究中的一个重要课题。寻找进化相关基因主要有两种策略：一是经典的从表型到基因的方法，检测进化性状相关的基因。如通过数量性状座位（quantitative trait loci, QTLs）定位，但这种方法不仅慢而且要做大量实验，因此找到的进化相关基因很少。二是运用群体遗传学方法寻找受选择的基因。由野生祖先尚存的群体（野生种）来代替家养物种的驯化祖先，通过对比野生群体和驯化群体的核苷酸多样性和连锁不平衡情况来初步判定受选择的目标基因。

当前测序技术飞速发展，基于测序的群体遗传学分析也为探索野生动物的种群变迁和保护提供了强有力的工具。通过对野生动物基因组学的研究，阐明其与种群衰退相关的分子机制，为建立濒危野生动物的种群复壮奠定理论基础。

方案设计

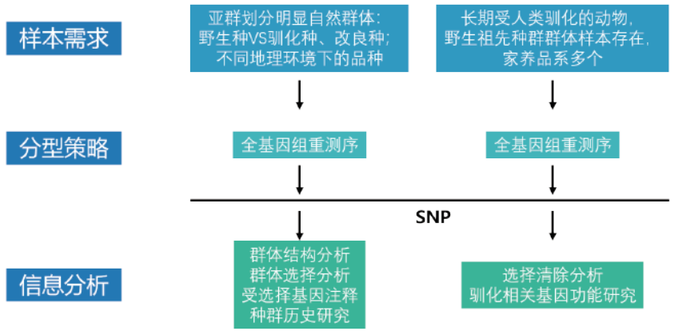


图1 种内群体进化研究方案设计

2.1 样本建议

1) 方案一：单个样本取样研究进化：要求亚群划分明显的自然群体，各亚群样本数不要过少（例如每个亚群样本数>30）。研究驯化问题时，选择的样本要包括野生种、驯化种；研究环境适应性时，选择的样本要包含不同地理来源，有明显环境特征差异的不同品种；研究物种起源时，要包含目前已知的祖先种，或化石，或可能起源地的野生种；研究种群历史时，最好有化石数据相互验证。研究改良问题时，样本包括地方种、改良种。

2) 方案二：混合样本研究驯化：长期受人类驯化的动物，野生祖先种群30个个体，家养品系每个品系30个个体，分别混合测序。

2.2 实验技术

采用全基因组重测序进行基因分型。

2.3 测序参数

采用全基因组重测序，每个个体推荐测序10X；混合样本测序，建议每个文库测序>20X。

2.4 分析结果

2.4.1 检测变异信息

利用测序结果，检测群体中存在的变异信息，对于全基因组重测序可以检测SNP、InDel、SV、CNV等，简化基因组测序主要是检测SNP信息。变异信息是进行其他信息分析的基础。

2.4.2 群体结构分析

通过构建群体的系统进化树（图2a）、主成分分析（图2b）和Structure分析（图2c），研究样本间的亲缘远近和样本间的进化关系。进化树是根据样本间亲缘关系的远近，把各样本安置在有分枝的树状的图表上，简明地表示生物的进化历程和亲缘关系。主成分分析（Principal Component Analysis，PCA），是将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。群体结构研究的过程中通过将测序品系和SNP位点构成二维矩阵数据，经过PCA分析，计算出几个主要的特征向量，并且将每一个品系在各特征向量上进行定位，也是研究群体品系间亲缘关系的方法之一。Structure分析则是假设若干个品系起源于K个截然不同（或差异较大）的祖先，分析每一个品系的遗传成分中，所具有的每一个假想祖先成分的比例。三种分析方法的结果可以相互验证。

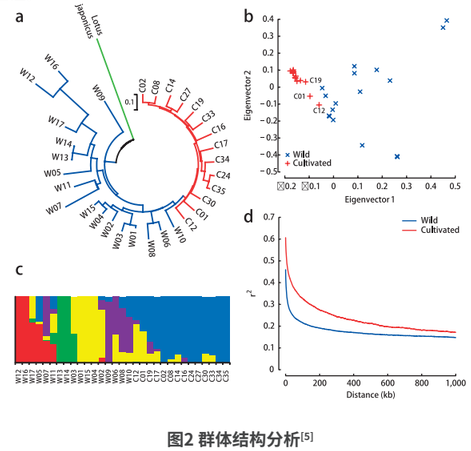


图2 群体结构分析[5]

a为进化树；b为PCA分析；c为Structure分析，不同颜色代表不同的假想祖先；d为连锁不平衡分析。

2.4.3 连锁不平衡分析

连锁不平衡（linkage disequilibrium，LD），指群体内不同座位等位基因之间的非随机关联，包括两个标记间或两个基因间或一个基因与一个标记座位间的非随机关联，可以用r^2计算两个标记间的连锁不平衡度。LD受重组、人工选择、群体类型等的影响，不同的物种LD变化情况不同，一般情况下我们会统计LD值衰减到一半的距离（图2d）。LD值会对信息分析需用的标记个数有指导意义，LD大的物种所需要的标记数量低。

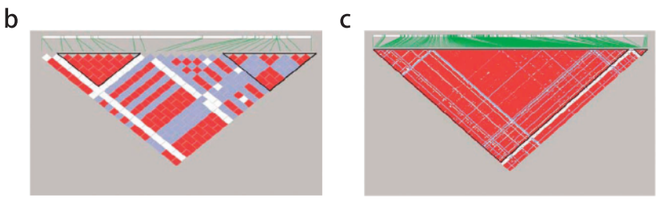


图3 LD blocks展示[5]

不同颜色代表不同的LD值，红色代表两个位点显著相关，连锁紧密。

2.4.4 群体多态性分析

群体多态性指的是同一群体中两种或两种以上变异类型并存的现象。群体多态性的量度：群体多态性参数θ=4Nμ，其中，N和μ分别是有效群体大小和每个位点的突变速率。常采用两种近似的方法来估计θ，分别是θw和θπ；θw计算方法：基于全部序列内分离位点个数，θπ计算方法：基于两两序列之间的平均距离。

表1 基于SNP的群体多态性统计结果展示[5]

Whole genome						
	Number of SNPs	$\theta_\pi(10^{-3})$	$\theta_W(10^{-3})$	Non-synonymous SNPs	Synonymous SNPs	Nonsyn/Syn
Wild soybean	5,924,662	2.966	2.307	106,716	78,701	1.36
Cultivated soybean	4,127,942	1.894	1.689	77,291	55,883	1.38
Genic regions						
	CDS			UTR		Intron		
	Number of SNPs	$\theta_\pi(10^{-3})$	$\theta_W(10^{-3})$	Number of SNPs	$\theta_\pi(10^{-3})$	$\theta_W(10^{-3})$	Number of SNPs	$\theta_\pi(10^{-3})$	$\theta_W(10^{-3})$
Wild soybean	185,145	1.063	0.829	74,476	1.768	1.415	621,432	2.002	1.582
Cultivated soybean	132,976	0.723	0.626	53,730	1.118	1.073	426,897	1.318	1.180

2.4.5选择分析

选择在物种的遗传变异形成过程中有巨大的贡献，其中搭便车效应会对种群水平的分化产生剧烈的影响，于较强的选择效应使得一个突变位点相邻DNA上的核苷酸之间的差异下降或消除（selective sweep）。通过分析大量的比较基因组学数据集和大量的SNP集，我们可以确定在野生种到驯化种、由地方种到改良种的过程中，以及在不同的环境情况下，哪些区域的多态性发生了巨大的改变，检测驯化或环境适应性相关的候选基因，而且选择清楚相关的候选基因与进化相关的性状也有关系。Fst是衡量群体间差异的量，Fst=(πBetween-πWithin)/πBetween，πBetween和πWithin分布代表亚群间和亚群内的不同样本两两差异平均值。选择性清除区域θπ值低，LD大，Fst值大。

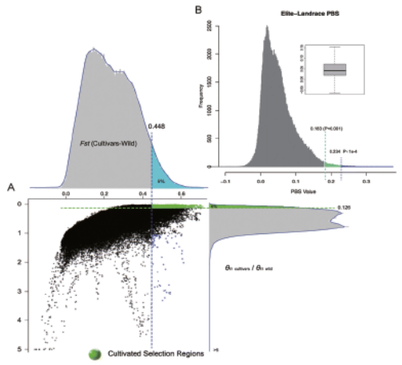


图4 选择分析结果示例[6]（绿色区域代表栽培种驯化过程中受选择区域）

2.4.6基因流分析

基因流（gene flow）又叫基因迁移或等位基因流动，指遗传信息从一个生物的群体传入另一个群体，导致不同种群之间基因交流的过程，可发生在同种或不同种的生物种群之间。基因流的产生至少需要两个条件：1)某亚群存在至少一个以上的居群或亚居群；2) 不同的居群或亚居群间有基因交流的机会。基因流的存在会消弱种群间的遗传差异。检验亚群间存在基因流的指标有基因频率、基因型频率、Fst、Nm等。

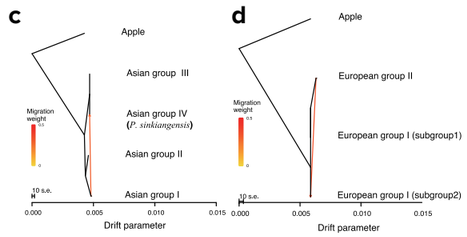


图5 基因流分析[7]

2.4.7种群历史研究

根据溯祖理论利用分析数学和统计学理论来回溯对应物种序列之间的变异过程。通过构建模型模拟不断进化的群体，从中抽取一定的样本，利用样本的突变情况来描述这个群体的遗传状况，群体的突变可以有重组、转换、迁移或是群体大小的历史变化。利用此研究可以分析物种的起源地，以及在历史发展过程中遗传信息的改变和群体大小的变化，并根据其他史料记载分析种群变化的原因。进行种群历史研究，最好有化石数据或史料信息进行佐证，否则很难说明模型的准确性。

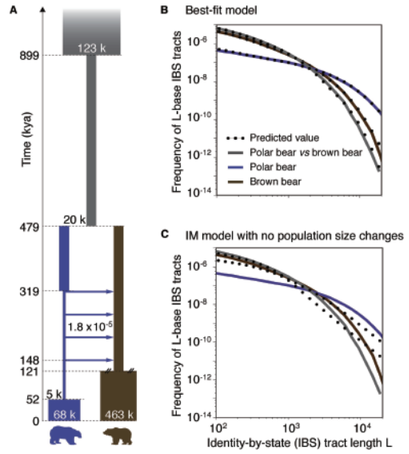


图6 群体历史变化分析[8]

2.5 项目周期

样品检测合格后，建库+测序+信息分析约60个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.6 预期结果

借助高通量测序平台，通过对所选样本的全基因组测序分型，检测变异信息，并基于SNP标记对群体进行相关的信息分析，了解样本间的亲缘关系远近，进化过程中发生的遗传和结构上的变化，挖掘与进化性状相关的基因位点，有助于人们更深层次的了解进化机理，同时可以了解该物种的遗传多样性，以利于育种资源的选择。

2.7 辅助研究策略

DNA测序获得的SNP标记，可以利用基因分型芯片验证其基因分型的准确性。

样本量大（如>200）的自然群体，还可以对驯化相关性状进行GWAS分析，定位区间与群体选择分析结果相互比较，看看是否位置有重叠。

可以通过RNA测序研究具有不同进化相关性状的样本，比较不同性状的样本间表达量不同的基因，通过比较DNA层面挖掘到的进化相关的候选基因和RNA表达层面的差异基因，寻找交集，提高数据挖掘的准确性。

2.8 后期验证手段

分析得到的候选基因，可以利用基因表达芯片或者是转录组测序等结果相互验证，找到与进化或进化性状相关的基因位点。

应用案例

3.1 案例一：高粱进化，一个项目七篇文章(华大参与)[9-15]

华大和昆士兰大学共同合作，利用44株高粱的重测序数据研究群体进化问题，从2013年到2017年间，在著名期刊发表了7篇文章。样本选择：44株高粱，其中17株是改良种，18株是地方种，还有2株驯化种以及7株野生种，另外还有同属的2个拟高粱（S. propinquum）。利用全基因组重测序技术获得了基因型数据，数据平均有22X的深度，基于全基因组水平。

表2 高粱项目发表7篇论文汇总

发表时间	发表期刊	研究方向	文章名	影响因子
2013.8	Nature Communications	利用“抗病基因的SNP”研究高粱进化	Whole-genome sequencing reveals untapped genetic potential in Africa’ s indigenous cereal crop sorghum[9]	11.47
2014.9	BMC Plant Biology	利用“抗病基因的SNP”研究高粱进化	The plasticity of NBS resistance genes in sorghum is driven by multiple evolutionary processes[10]	3.813
2016.1	Biotechnology for Biofuels	构建SNP数据库	SorGSD: a sorghum genome SNP Database[11]	6.044
2016.5	Plant Biotechnology Journal	利用“淀粉代谢途径相关基因”研究进化	Domestication and the storage starch biosynthesis pathway: Signatures of selection from a whole sorghum genome sequencing strategy[12]	5.752
2016.12	Frontiers in Plant Science	利用“氮代谢途径相关基因”研究进化	Whole Genome Sequencing Reveals Potential New Targets for Improving Nitrogen Uptake and Utilization in Sorghum bicolor[13]	4.495
2017.7	Frontiers in Plant Science	利用“高粱谷粒大小和重量基因”研究进化	Whole-Genome Analysis of Candidate genes Associated with Seed Size and Weight in Sorghum bicolor Reveals Signatures of Artificial Selection and Insights into Parallel Domestication in Cereal Crops[14]	4.495
2017.11	Molecular Breeding	高粱不同品系有关硝酸还原酶和谷氨酸合成酶的不同等位基因影响植物氮反应	The vegetative nitrogen response of sorghum lines containing different alleles for nitrate reductase and glutamate synthase[15]	2.246

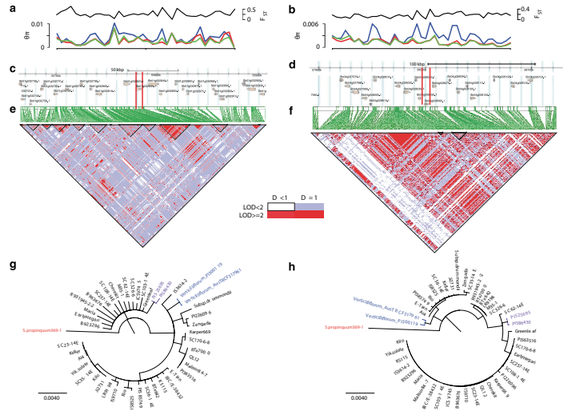


图7 高粱选择分析结果[9]

(a,b) θπ值(红色是野生种，蓝色是地方种，绿色是改良种)和FST值（黑色）；(c,d) 受选择区间的候选基因；(e,f) 受选择的候选基因LD blocks (红的代表强选择，白的代表弱选择)；(g,h) GS3 和SSIIb 的基因树。

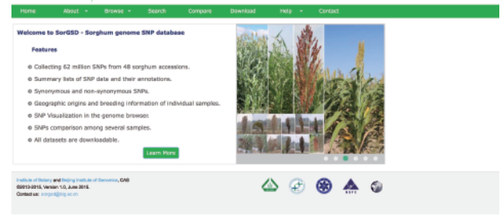


图8 高粱SNP数据库[11]

3.2 案例二：3K水稻重测序&泛基因组研究（华大参与）[16]

由中国农业科学院作物科学研究所牵头，联合IRRI、上海交大、华大基因、深圳农业基因组研究所、安徽农大等16家单位共同完成“3000份亚洲栽培稻基因组研究”，并于2018年4月发表在《Nature》上。研究针对水稻起源、分类和驯化规律进行了深入探讨，揭示了亚洲栽培稻的起源和群体基因组变异结构，剖析了水稻核心种质资源的基因组遗传多样性。

3000份水稻（来自全球89个国家和地区）代表了全球78万份水稻种质约95%多样性的核心种质。通过全基因组重测序，每个样本平均测序深度14X，利用重测序数据共检测到32M的高质量SNPs和InDels。对亚洲栽培稻群体的结构和分化进行了更为细致和准确的描述和划分，由传统的5个群体增加到9个，分别是东亚（中国）的籼稻、南亚的籼稻、东南亚的籼稻和现代籼稻品种等4个籼稻群体，东南亚的温带粳稻、热带粳稻、亚热带粳稻等3个粳稻群体，以及来自印度和孟加拉的Aus和香稻。研究首次揭示了亚洲栽培稻品种间存在的大量微细结构（>100bp）变异（SVs，包括易位、缺失、倒位和重复）。着重研究453个测序深度>20X的品系的SVs，利用SVs构建的进化树与SNP构建的进化树类似。大量的SVs可能是不同程度杂种不育和XI与GJ杂种衰退的遗传基础。同时构建了亚洲栽培稻的泛基因组，包括12,770个（62.1%）核心（core）基因家族和9,050个（37.9%）分散式(distributed)基因家族,发现了1.2万个全长新基因和数千个不完整的新基因。核心基因比较古老，大多数的新基因表现更年轻和长度偏短。

研究策略：最初测序3024份水稻样本，后来进行质控过滤掉14份，最终保留3010份水稻样本进行深度研究。3K RG测序数据比对到参考基因组日本晴Nipponbare上检测SNPs、indels。合并Nipponbare基因组序列和无冗余的新组装的基因组序列构建泛基因组。利用测序深度>20X，比对深度>15X的453个水稻材料进行SVs和PAVs分析。

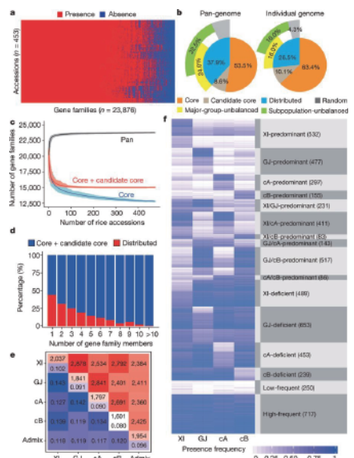


图9 水稻泛基因组

a、基因家族PAVs；b、泛基因组和一个单独的基因组的组成成份；c、基于500个随机筛选的水稻基因组模拟泛基因组和核心基因组；d、核心和分散式基因家族比例；e、两个品系间基因家族平均数量差异；f、5733主要群组不平衡基因家族特性。

3.3 案例三：梨地理起源与独立驯化（华大参与）[7]

梨属于蔷薇科，是世界性栽培的重要果树，其栽培历史可以追溯到3000多年前。梨的种质资源较为丰富，至今至少已经认证了22个种，全世界有超过5000个品系。这些品系从形态学、生理学上都有较大的差异，对不同的地理环境表现了广泛的适应性。同时由于其表现典型的自交不亲和性，品种资源间存在广泛的基因交流和遗传重组。为探明梨的起源、传播与驯化特征，研究人员对广泛来源的梨样品进行全基因组重测序，测序结果以“砀山酥梨”作为参考基因组（相比组装结果更好），检测了18M的SNP，进行了系统发育分析、群体结构分析、基因流分析和选择分析。研究结果揭示了梨起源于中国的西南部，经过亚欧大陆传播到中亚地区，最后到达亚洲西部和欧洲；发现了我国的新疆梨存在亚洲梨与西洋梨间的基因交流，遗传相似度分析进一步证明了新疆梨是来自栽培种间的杂交，该种间杂交种的形成与2000多年前丝绸之路的文化物资交流有关。

研究发现不同来源的梨品种资源总体分为亚洲梨和欧洲梨两大组。其中，欧洲梨遗传背景相对简单、遗传多样性低，可分为野生种与栽培种组群；而亚洲梨遗传组成复杂，表现出较高的遗传多态性，并且野生与栽培种形成多个亚组。亚洲梨与西洋梨的分化时间大约在6.6 -3.3百万年以前，群体驯化研究进一步揭示了两大组群受选择区域的显著差异，支持了东、西方梨的独立驯化事件，从遗传水平上解答了两个种群在果实品质等生物学性状上的显著差异。此外，研究还发现了梨的花柱S-RNase基因快速进化和平衡选择有利于保持自交不亲和性，从而促进了梨的异交和高度遗传多样性。

研究策略：对来自26个国家33个梨属品系的113个样品进行全基因组重测序，其中57个野生种梨和56个栽培种梨，平均测序深度11X。

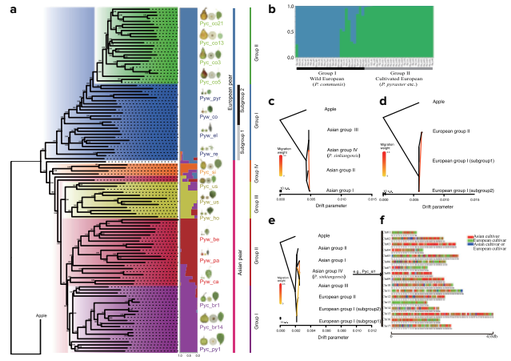


图10 进化树和基因流分析

a. 构建进化树，测序样本被分到两个亚群中。b. 群体结构分析将欧洲梨分到两个亚群。c. 亚洲梨基因流分析。d. 欧洲梨基因流分析。e. 亚洲梨和欧洲梨间基因流分析。f. 新疆梨IBD分析。

3.4 案例四：江豚基因组+环境适应性研究（华大参与）[17]

江豚，鼠海豚科江豚属，它们外形与其它海豚不同，没有背鳍且吻部平钝。江豚一般被分为两个种：广布种宽脊江豚（*N. phocaenoides*）和窄脊江豚（*N. asiaeoridentalis*）。长江江豚曾与东亚江豚（海江豚的一种）共同被认为是窄脊江豚的两个亚种，两者形态十分相似、难以区分。本研究通过构建长江江豚的基因组、获取来自长江和中国沿海不同水域的48只江豚的基因组数据进行比较分析，发现长江江豚与海洋江豚之间存在着显著而稳定的遗传分化，已形成独立的进化支系。窄脊江豚主要生活在温带沿岸海域及长江流域，均为深度不超过150米的浅水区。而宽脊江豚的生活水域要深得多，比如中国南海水深可达1212米，最深处超过5500米。利用复合似然率算法分别鉴定出218与144个宽脊、窄脊江豚群体中受选择的基因。宽脊江豚与窄脊江豚在不同环境的遗传适应中显示出独特的个体化特征（缺氧胁迫等）。江豚在海水和淡水两种生活环境中的类群大致分化于五千至十万年前。文章利用XP-EHH选择分析对窄脊江豚海洋-长江两个群体进行研究，发现了83个（长江）和187个（海洋）受选择基因，这些基因与肾脏发育、尿管发育等相关。

研究策略：通过短片段高深度106X测序，构建长江江豚的基因组、并对来自长江和中国沿海不同水域的48只江豚进行全基因组重测序，测序深度10-30X，测序结果进行群体分析。

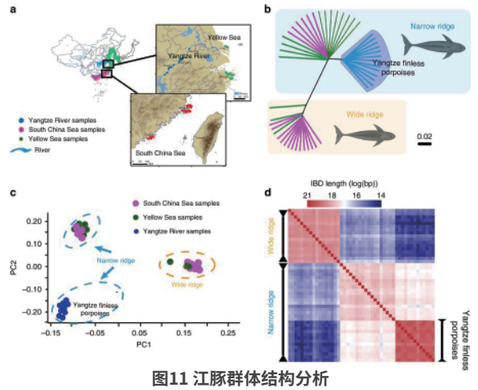


图11 江豚群体结构分析

a、江豚样本来源；b、48个江豚的进化树；c、48个江豚的PCA分析；窄脊江豚被显著的分为两个亚群。d、评估江豚间共享单倍型情况，热图不同颜色代表成对样本间IBD blocks的总长度。

3.5 案例五：大熊猫群体进化-种群历史演化（华大参与）[18]

科研人员对来自六大山系的34只野生大熊猫进行全基因组重测序，依据基因组信息确定这六个山系熊猫可划分为秦岭、岷山、邛崃山-大小相岭-凉山三个遗传系，通过构建大熊猫从起源到如今的演化历史，揭示出其间所经历的两次种群扩张、两次瓶颈和两次种群分化现象。此外，研究表明全球气候变化可能导致大熊猫的种群波动，但近期的人类活动则是导致熊猫种群分化和数量严重下降的主要因素。

研究策略：全基因组重测序，平均测序深度为4.7X。比较34个大熊猫的基因组序列，检测遗传变异并基于SNP标记进行群体结构和种群历史演化的分析。

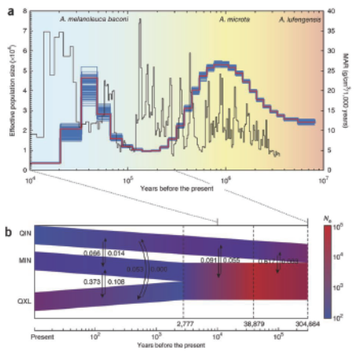


图12 大熊猫进化历史

3.6 案例六：银杏种群进化历史研究 [19]

对采自全球51个种群545棵银杏大树进行了全基因组重测序，分析发现在中国有3个银杏避难所：东部（浙江天目山为代表）、西南（贵州务川、重庆金佛山为代表）以及南部（广东南雄、广西兴安为代表），而现今分布在不同大洲的银杏多源自中国东部种群。进一步研究表明，生物气候变量影响银杏的地理分布、环境适应性及复原力。

研究策略：全球9个国家，51个种群，545棵银杏大树进行全基因组重测序，每个样本测序深度4-10X，平均测序深度6.1X。

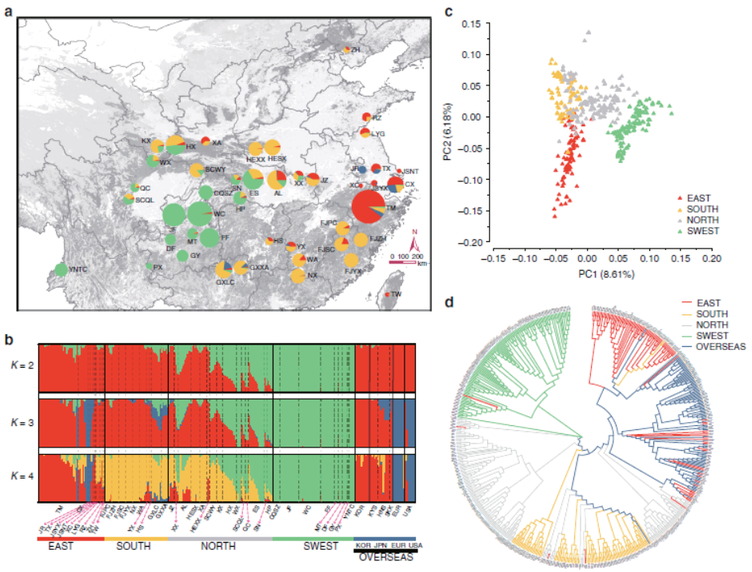


图13 银杏种群的进化关系及群体结构分析

a: 样品地理分布; b:群体结构分析;c:中国银杏样品PCA分析;d: 545棵银杏进化树分析

可能存在的风险

群体进化研究选择的样本要具有明显的亚群划分，例如研究驯化要有野生种、驯化种、改良种等；研究环境适应性，样本要来自明显不同的环境条件下；研究地理起源要包含研究物种可能的起源野生种和不同进化途径中的样本。没有明显的亚群分化是无法进行选择分析的。同时选择分析过程需要借助滑动窗口，所以此部分信息分析只能应用于有参考基因组物种。群体进化分析的样本起始量要>30个，并且各亚群的样本个数不要太少，否则代表性不强，选择分析获得的候选基因位点准确度不高。

文章小结

研究动植物群体进化的文章能发表到什么水平的杂志？我们筛选了一些有代表性的文章，从中可以看到一些规律。

1、只针对基因组序列进行研究的基因组项目，已经很难发表高分文章。一般会增加DNA重测序进行物种内进化研究，或增加RNA、蛋白或表观调控水平的研究内容，为整个文章增色。

2、某个物种基于全基因组重测序做第一篇群体进化的文章，样本量>30，早期文章影响因子基本都超过20，但现在一般情况下文章点数接近10。要注意的是，文章中要解决该物种一个独特的生物学问题，如果只是研究这个物种的一些群体结构的问题，将越来越难发表好文章。

3、样本量的要求越来越高，即使是相同的研究问题，取样范围也会影响研究问题的广度。如全球范围取样与全国范围取样，前者的发文点数能更好一些。

4、样本量大的情况下，群体同时做进化和GWAS分析，或者是用其他群体定位进化相关表型的QTL，或是从多层面研究进化问题，不仅包含DNA测序，同时有RNA、甲基化、蛋白互作等。同一个群体用不同研究方法分析，或是用不同的群体和技术手段研究同一个相关问题，都可以提升文章的层次。

5、做同一类型的研究，全基因组重测序分型能进行全基因组水平、不同功能元件的分析，优于简化基因组测序或是目标区域测序。

6、选择特殊样本，研究的问题点区别于其他常规的研究，有助于文章发表。如别人是研究驯化，我们研究反驯化、野生化。

7、充分利用研究物种已发表的测序数据，自己只用再测序少量的样本，高性价比研究自己关心的问题。

8、全基因组重测序主要用于种内进化。但是也可以利用保守序列相关的变异信息研究大尺度的进化事件，也可以基于全基因组比对的数据研究群体进化问题。但是如果样本间遗传距离比较大，全基因组水平的比对率会偏低，这样也会使得用于进行群体研究的标记数受限。

表3群体进化文章

发表日期	期刊	IF	物种	研究内容	样本选择	测序策略
2018.02	Nature	43.07	柑橘[20]	柑橘起源与进化	60个柑橘品系：58个不同地理来源的品系+2个外群，其中12个之前有测过序	WGS平均测序深度约60X，9.2-178X
2018.04	Nature Communications	11.88	梅花[21]	木本植物梅花中花性状的遗传结构	351个品系	WGS 平均19.3X
2018.04	Nature	43.07	水稻[16]	亚洲栽培稻的起源和群体基因组变异结构，遗传多样性	3000份水稻（来自全球89个国家和地区）代表了全球78万份水稻种质约95%多样性的核心种质	WGS平均14X
2018.04	Nature Communications	11.88	江豚[17]	鲸类对淡水环境的适应性进化机制	不同地理环境下的48个江豚	10-30X
2018.05	Nature Ecology & Evolution	10.97	牛[22]	牛属基因交流	大额牛、印度野牛、爪哇野牛、欧洲野牛和美洲野牛等	WGS 5-20X
2018.06	Genome Biology	14.03	梨[7]	亚洲梨和欧洲梨的的多样性与独立进化	来自26个国家属于33个梨属的113个品系，其中57个野生种梨和56个栽培种梨	WGS 平均11X
2018.07	Science	41.06	山羊[23]	古山羊驯化	来自肥沃新月周围东部、西部和南部地区的83只野生和驯化山羊遗骸进行线粒体测序，并对其中DNA保存较完整的51个古代样本进行全基因组重测序	线粒体测序71X，古样本WGS0.01-15X
2018.07	Science	41.06	狗[24]	美洲犬的进化史	时间跨度大约9000年，来自古代北美和西伯利亚的狗	71只狗线粒体基因组进行测序，对7只狗的核基因组进行测序
2018.07	Nature Ecology & Evolution	10.97	家蚕[25]	家蚕基因组驯化	137种代表性家蚕品系	WGS 13X
2018.12	Nature Communications	11.88	青稞[26]	青稞的起源与进化	69个青稞地方品种、35个青稞育成品种以及10个西藏半野生大麦进行全基因组重测序，结合已经发表的260份全球野生和地方品种的外显子测序数据，共437个大麦材料一起分析。	WGS 9.6X，结合原有外显子数据
2018.12	Nature Communications	11.88	桃[27]	桃起源进化	58个桃栽培种和近缘种：包括44个新测序品种和14个以上研究的品系	新测序WGS 41.25-72.18X,原有样本>10X
2019.03	Nature Communications	11.88	油菜[28]	甘蓝型油菜起源和进化机制，整合全基因组关联研究、选择信号和转录组分析	来自21个国家的588份有代表性的甘蓝型油菜材料	WGS平均5X
2019.03	Nature Communications	11.88	葡萄[29]	葡萄属多样性及种群历史进展	广泛地理分布现存472葡萄属品系，包括的60个葡萄属种中的48个	WGS平均15.5X
2019.04	Nature Genetics	25.46	鹰嘴豆[30]	鹰嘴豆基因组多态性与迁移路线	45个国家的429个鹰嘴豆品种：地方种205个、优良栽培种163个、育种品系44个、未知生物学地位样本10个、野生种7个，新测序300，其他为原有数据	WGS平均10.22X

发表日期	期刊	IF	物种	研究内容	样本选择	测序策略
2019.07	Genome Biology	14.03	小麦[31]	小麦起源、进化和驯化历史	93个小麦及其近缘种材料：包括20个野生二粒小麦（wild emmer）、5个粗山羊草(Ae.tauschii)、5个硬粒小麦(durum)、29个六倍体农家种(landrace)和34个栽培种(variety)	WGS+24个转录组和90个外显子捕获测序数据
2019.09	Nature Communications	11.88	银杏[19]	种群进化历史、避难所、进化潜力	全球51个种群，545棵银杏大树	WGS 平均6.1X

常见问题

1、进行群体进化研究的样本要满足什么样的条件？

答：进行群体进化研究的样本要是自然样本；并且具有明显的亚群分化；同时亚群内的样本具有代表性；群体大小在30个以上。

2、群体进化研究测序深度推荐？

答：采用全基因组重测序，每个个体推荐测序>5X；混合样本测序，建议每个文库测序>20X。

3、两个具有表型差异的群体样本（非家系群体）可不可以采用驯化研究的分析思路检测与表型相关的基因位点？

答：如果两个具有表型差异的群体不是像驯化这样经过长期定向选择的样本，利用相似的方法检测表型相关的基因位点，在信息分析流程实现上是可以做的，但是结果的准确度会有问题。因为对于具有表型差异的两个群体间多态性差异显著的区域会很多，从中间再选择真正的与表型相关的位点，工作难度比较大。

华大优势

项目经验丰富：迄今动植物群体进化研究华大参与发表文章40+，包括《Nature》、《Nature Genetics》《Science》等顶级杂志；

根据客户需求可以提供方案设计并完成个性化分析内容；

测序平台多样，选择空间大，能满足不同需求；

质控严格：从样本接收到数据交付都有严格的质量控制流程，保证数据准确性；

提供不同类型产品服务，一站式完成您的需求。

参考文献

[1] Evershed R P, Payne S, et al. Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding [J]. Nature, 2008, 455(7212): 528-531.

[2]Outram A K, Stear N A, et al. The earliest horse harnessing and milking [J]. Science, 2009, 323(5919): 1332-1335.

[3]Leonard J A, Wayne R K, et al. Ancient DNA evidence for Old World origin of New World dogs [J]. Science, 2002, 298 (5598): 1613-1616.

[4]Larson G, Albarella U, et al. Ancient DNA, pig domestication, and the spread of the Neolithic into Europe [J]. Proceedings of the National Academy of Sciences, 2007, 104(39): 15276-15281.

[5]Lam H M, Xu X, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection[J]. Nature genetics, 2010, 42(12): 1053-1059.

[6]Zhao S, Zheng F, et al. Impacts of nucleotide fixation during soybean domestication and improvement [J]. BMC plant biology, 2015, 15(1): 81.

[7]Wu J, Wang Y, Xu J, et al. Diversification and independent domestication of Asian and European pears[J]. Genome biology, 2018, 19(1): 77.

[8]Liu S, Lorenzen E D, et al. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears [J]. Cell, 2014, 157(4): 785-794.

[9]Mace E S, Tai S, Gilding E K, et al. Whole-genome sequencing reveals untapped genetic potential in Africa’s indigenous cereal crop sorghum[J]. Nature communications, 2013, 4.

[10]Mace E, Tai S, Innes D, et al. The plasticity of NBS resistance genes in sorghum is driven by multiple evolutionary processes[J]. Bmc Plant Biology, 2014, 14:253(1):221-230.

[11]Luo H, Zhao W, Wang Y, et al. SorGSD: a sorghum genome SNP database[J]. Biotechnology for biofuels, 2016, 9(1): 1.

[12]Campbell B C, Gilding E K, Mace E S, et al. Domestication and the storage starch biosynthesis pathway: Signatures of selection from a whole sorghum genome sequencing strategy[J]. Plant Biotechnology Journal, 2016.

[13]Massel K, Campbell B C, Mace E S, et al. Whole Genome Sequencing Reveals Potential New Targets for Improving Nitrogen Uptake and Utilization in Sorghum bicolor[J]. Frontiers in Plant Science, 2016, 7.

[14]Tao Y, Mace E S, Tai S, et al. Whole-genome analysis of candidate genes associated with seed size and weight in sorghum bicolor reveals signatures of artificial selection and insights into parallel domestication in cereal crops[J]. Frontiers in plant science, 2017, 8: 1237.

[15]Diatloff E, Mace E S, Jordan D R, et al. The vegetative nitrogen response of sorghum lines containing different alleles for nitrate reductase and glutamate synthase[J]. Molecular Breeding, 2017, 37(11): 138.

[16]Wang W, Mauleon R, Hu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice[J]. Nature, 2018, 557(7703): 43.

[17]Zhou X, Guang X, Sun D, et al. Population genomics of finless porpoises reveal an incipient cetacean species adapted to freshwater[J]. Nature communications, 2018, 9(1): 1276.

[18]Zhao S, Zheng P, et al. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation [J]. Nature genetics, 2013, 45(1): 67-71.

[19]Zhao YP, Fan G, Yin PP, et al. Resequencing 545 ginkgo genomes across the world reveals the evolutionary history of the living fossil. Nat Commun. 2019;10(1):4201.

[20]Wu G A, Terol J, Ibanez V, et al. Genomics of the origin and evolution of Citrus[J]. Nature, 2018.

[21]Zhang Q, Zhang H, Sun L, et al. The genetic architecture of floral traits in the woody plant Prunus mume[J]. Nature communications, 2018, 9(1): 1702.

[22]Wu D D, Ding X D, Wang S, et al. Pervasive introgression facilitated domestication and adaptation in the Bos species complex[J]. Nature ecology & evolution, 2018.

[23]Daly K G, Delser P M, Mullin V E, et al. Ancient goat genomes reveal mosaic domestication in the Fertile Crescent[J]. Science, 2018, 361(6397): 85-88.

[24]Leathlobhair M N, Perri A R, Irving-Pease E K, et al. The evolutionary history of dogs in the Americas[J]. Science, 2018, 361(6397): 81-85.

[25]Xiang H, Liu X, Li M, et al. The evolutionary road from wild moth to domestic silkworm[J]. Nature ecology & evolution, 2018: 1.

[26]Zeng, X., Guo, Y., Xu, Q. et al. Origin and evolution of qingke barley in Tibet. Nat Commun 9, 5433 (2018).

[27]Yu, Y., Fu, J., Xu, Y. et al. Genome re-sequencing reveals the evolutionary history of peach fruit edibility. Nat Commun 9, 5404 (2018).

[28]Lu, K., Wei, L., Li, X. et al. Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. Nat Commun 10, 1154 (2019).

[29]Liang, Z., Duan, S., Sheng, J. et al. Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. Nat Commun 10, 1190 (2019).

[30]Varshney R K, Thudi M, Roorkiwal M, et al. Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits[J]. Nature genetics, 2019, 51(5): 857.

[31]Cheng H, Liu J, Wen J, et al. Frequent intra-and inter-species introgression shapes the landscape of genetic variation in bread wheat[J]. Genome biology, 2019, 20(1): 136.

全基因组关联分析定位QTL研究方案

034

研究背景

随着分子育种研究的深入，越来越多的证据表明，物种的大部分表型性状（如株高、株型、生长速度、产量）属于数量性状，掌握数量性状的表型和基因型是育种工作的基础。家系连锁作图（Family-based linkage mapping，FBL mapping）与自然群体关联作图（association mapping）是现今解析研究动植物数量性状基因型的主要方法。前者通过双亲杂交，建立作图群体，进行高密度分子连锁图谱的绘制，对作图群体进行各种性状高精确度的表型鉴定，再进行连锁分析，将相应性状的QTL 定位在特定的遗传连锁区段内。双亲遗传差异、分离群体大小及高密度分子标记连锁图谱是决定QTL 定位精确性的基础。由于在特定的两个亲本间一些位点不发生分离与重组，该方法所获结果常常有一定的局限性，而且所能检测出的等位变异只限于双亲所有的两个。

关联作图（association mapping），又称连锁不平衡作图（linkage disequilibrium mapping，LD mapping）或关联分析（association analysis），曾广泛应用于人类遗传学研究中[1]。该方法以自然群体为研究对象，以长期重组后保留下来的基因（位点）间连锁不平衡（LD）为基础，将目标性状表型的多样性与基因（或标记位点）的多态性结合起来分析，可直接鉴定出与表型变异密切相关且具有特定功能的基因位点或标记位点[2]。与传统的QTL作图技术相比，关联分析具有明显的3个特点：1）不需要专门构建作图群体，自然群体或种质资源都可作为研究材料；2）广泛的遗传材料可同时考察多个性状大多数QTL的关联位点及其等位变异，不受传统的FBL的“两亲本范围”的限制；3）自然群体经历了许多轮重组后，LD衰减存在于很短的距离内，保证了定位的更高精确性[3]。

全基因组关联分析（Genome wide association study，GWAS）是应用基因组中数以百万计的SNP为分子遗传标记，进行全基因组水平上的对照分析或相关性分析，通过比较发现影响复杂性状的基因变异的一种新策略。已广泛应用于人类复杂疾病研究，近年来，这种方法在农业动植物重要经济性状主效基因的筛查和鉴定中得到了应用。但是GWAS对于动植物的研究可能将更加具有优势，这主要是由于动植物可以随意构建群体以及随意重组。

动植物重要经济性状GWAS分析方法的原理是，借助于SNP分子遗传标记，进行总体关联分析，在全基因组范围内选择遗传变异进行基因分型，比较异常和对照组之间每个遗传变异及其频率的差异，统计分析每个变异与目标性状之间的关联性大小，选出最相关的遗传变异进行验证，并根据验证结果最终确认其与目标性状之间的相关性。例如对拟南芥开花期、抗病性的研究[4]，在玉米开花期上的研究[5-7]，对水稻抽穗期和产量的研究[8]，在狗毛色上的研究[9]以及对牛产奶量研究[10]等。

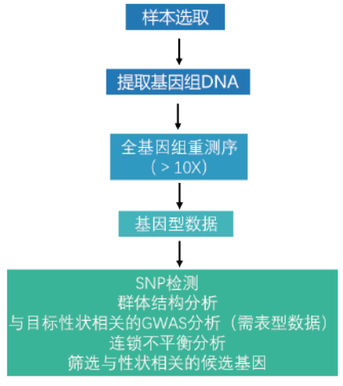


图1 GWAS分析方案设计

2.1 样本建议

有参考基因组物种的自然群体，群体大小推荐>300个；严格控制群体结构即样本间不要有明显的亚群分化；构建多亲本衍生群体（如NAM群体），可打破群体结构的影响。

2.2 实验技术

采用全基因重测序进行基因分型。

2.3 测序参数

采用全基因组重测序，每个个体推荐测序10X。

注：对于大基因组或者无参考基因组物种也可以采用转录组测序检测SNP进行GWAS研究，但是转录水平的SNP与DNA层面的肯定会有所差异，所以一般情况下不推荐。

2.4 分析结果

2.4.1检测变异信息

利用测序结果，检测群体中存在的变异信息，对于全基因组重测序可以检测SNP、InDel、SV、CNV等。变异信息是进行其他信息分析的基础。

2.4.2群体结构分析

通过构建群体的系统进化树、主成分分析和Structure分析，研究样本间的亲缘远近和样本间的分类关系。不同群体分析结果可以相互验证。

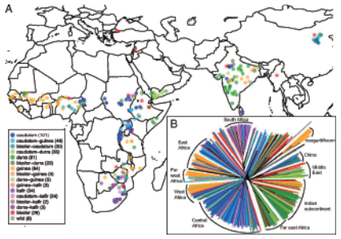


图2 样本选取区域及进化树分析[11]。图2B进化树是根据样本间亲缘关系的远近，把各样本安置在有分枝的树状的图表上，简明地表示生物的进化历程和亲缘关系。

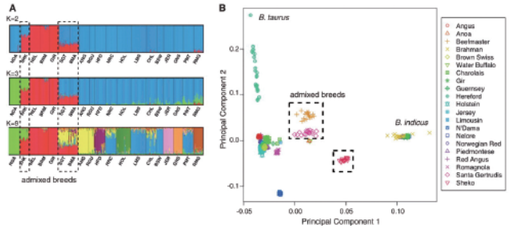


图3 群体结构分析[12]

Structure分析（图3A），Structure分析是假设若干个品系起源于K个截然不同（或差异较大）的祖先，分析每一个品系的遗传成分中，所具有的每一个假想祖先成分的比例。主成分分析(图3B,主成分分析（Principal Component Analysis，PCA），是将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。群体结构研究的过程中通过将测序品系和SNP位点构成二维矩阵数据，经过PCA分析，计算出几个主要的特征向量，并且将每一个品系在各特征向量上进行定位，也是研究群体品系间亲缘关系的方法之一。

2.4.3连锁不平衡分析

连锁不平衡（linkage disequilibrium，LD），指群体内不同座位等位基因之间的非随机关联，包括两个标记间或两个基因间或一个基因与一个标记座位间的非随机关联。当两个基因座位A（等位基因A和a）和B（等位基因B和b）位于同一条染色体或是连锁群上，则认为他们在遗传上是连锁的。位点间的连锁程度用重组率r来衡量。重组率表示在一次减数分裂的过程中，两个连锁座位之间发生交换的概率。所谓连锁平衡指的是配子基因型的频率等于等位基因频率的乘积，在随机交配群体中，习惯上把配子基因型的实际频率与平衡时频率的偏差程度作为连锁不平衡度，用D表示，如：$D_{AB}=f_{AB}-f_Af_B$。连锁的程度决定了连锁不平衡的大小，即连锁越紧密，连锁不平衡度越高。

LD受重组、人工选择、群体类型等的影响，不同的物种LD变化情况不同，一般情况下我们会统计LD值衰减到一半的距离（图4c）。连锁不平衡度会对信息分析需用的标记个数有指导意义，连锁不平衡度大的物种所需要的标记数量低。

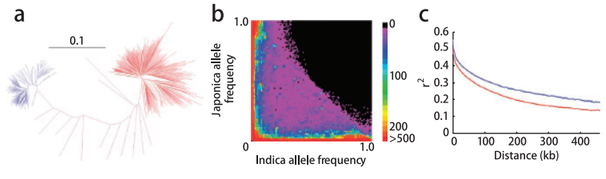


图4 群体样本特性展示[13]

2.4.4HapMap图谱构建

HapMap（haplotype map）是某一基因组中常见遗传多态位点的目录，它是建立存储某一物种常见SNP变异以及LD值等相关信息的Database。同一物种个体的遗传序列极为相似，在群体水平上，常见的SNPs（MAF>5%）数量不到基因组大小的0.1%。而这些相邻的常见SNP，在群体中常处于关联状态。因此，可以从常见SNP中挑选出更具代表性的标签SNP（Tag SNP），来简化SNP集的数据量。仅仅利用这些相对数据量较少的标签SNP集合所包含的基因型信息，就可以代表整个基因组的大部分遗传信息。

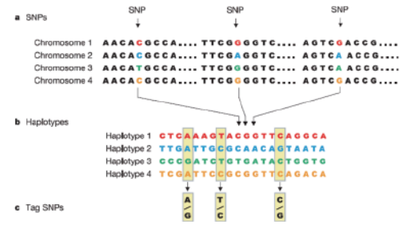


图5 HapMap示意图[14]

2.4.5全基因关联分析（GWAS）

利用分布于全基因组水平的分子标记（例如SNP）通过一定的模型（一般线性模型或混合线性模型等）与表型进行关联分析，检测目标性状基因位点。但是由于连锁的存在，往往我们检测到的标记并不是直接决定目标性状的变异，如果进行基因克隆时还是要在一定的定位区间内完成。Manhattan plot（图6左）和QQ plot（图6右）是查看GWAS定位结果和计算模型合理性的标配图，几乎每个GWAS文章中都有。Manhattan plot横坐标是表示位置，纵坐标表示-lgP，在纵坐标上超过一定阈值的点被认为和表型关联。我们看Manhhatan图时，认为孤点是不可信的，山峰状的比较可信。QQ图的意义在于基因型和性状无关联的情况下，各个标记P-value的观察值和期望值是相等的（红线），但是由于出现了基因型和性状有关联的情况，P-value往往会偏离y=x这条线。

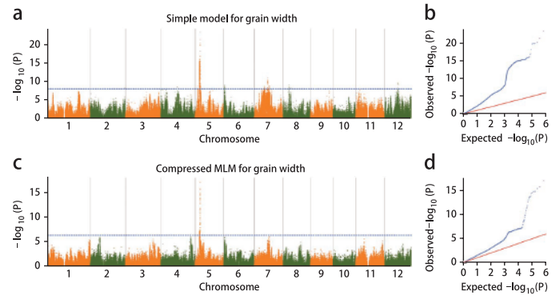


图6 GWAS结果示例[13]

2.4.6目标性状候选基因位点功能注释

根据各种全基因组关联分析方法定位到目标性状候选基因区间后，比对回参考基因组，可以检测到此区间的基因信息，通过与过往研究比较可以得知定位的区间与哪些以前定位到的基因位置重合或靠近。同时利用现有的基因功能数据库，对候选基因进行功能注释和聚类分析，更深层次的挖掘目标性状的分子机制。

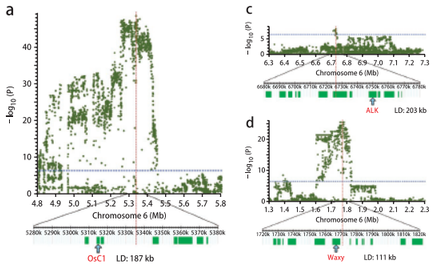


图7 目标性状候选基因展示[13]

2.5 项目周期

样品检测合格后，建库+测序+标准信息分析约60个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.6 预期结果

借助高通量测序平台，通过对所选样本的全基因组测序分型，检测变异信息，并基于SNP标记进行群体相关的信息分析，了解样本间的群体结构信息，群体的遗传多样性信息，挖掘与目标性状相关的基因位点，有助于人们检测到目标性状的功能基因位点，并对目标性状相关的基因进行功能注释，深层次地了解目标性状的分子机制。

2.7 辅助研究策略

可以通过RNA测序研究在目标性状上存在显著差异的样本，比较性状不同的样本间差异表达的基因，通过比较DNA层面挖掘到的目标性状候选基因和RNA表达层面的差异基因，寻找交集，提高数据挖掘的准确性。或者是检测目标性状候选基因的表达情况，研究基因对性状的调控机制。

2.8 后期验证手段

分析得到的候选基因，可以利用转基因、基因敲除、基因沉默（RNAi）等方式验证基因功能。

应用案例

3.1 案例一：水稻GWAS，从定位到基因克隆 [15]

亚洲栽培稻划分为两个亚种：*indica*（籼稻）和*japonica*（粳稻）。粳稻在中国南部最早由一小撮野生稻进化而来，而籼稻衍生自粳稻栽培种和其他的 O. rufipogon生态型杂交。粳稻和籼稻无论是在形态上、遗传上都存在着区别。典型的籼稻粒长、温带粳稻粒短偏圆。但是典型的热带粳稻比温带粳稻籽粒大，这些形态差异的遗传基础现在还不知道。

本研究通过对381份粳稻材料（包含40份热带稻和341份温带粳稻）的粒长和千粒重开展全基因组关联分析（GWAS），定位到一个既控制粒长又决定千粒重的关键数量性状位点QTL-GLW7，并进一步整合基因组变异和全基因组基因表达谱信息，

以及水稻突变体材料的分析和转基因实验鉴定，最终成功克隆到控制水稻粒长和粒重的关键基因*GLW7*。研究发现，*GLW7*为编码一类高等植物特有的SPL转录因子基因，因此被命名为*OsSPL13*。该基因在穗发育及颖壳发育时期特异性表达，但是*GLW7*在大粒型水稻材料中的mRNA和蛋白表达量均显著高于小粒型品种材料。转基因实验表明，该基因不仅可以使水稻籽粒增大，同时还能显著增加穗长、一级枝梗和二级枝梗数目以及每穗粒数，最终增加水稻产量。研究发现并证明*GLW7*主要是通过增加细胞的大小而使籽粒的体积变大。进一步群体遗传学分析发现，在水稻的遗传改良过程中，大粒基因*GLW7*是从籼稻通过遗传漂移渗透到热带粳稻以及少量的温带粳稻中，从而改良了粳稻的千粒重和产量。

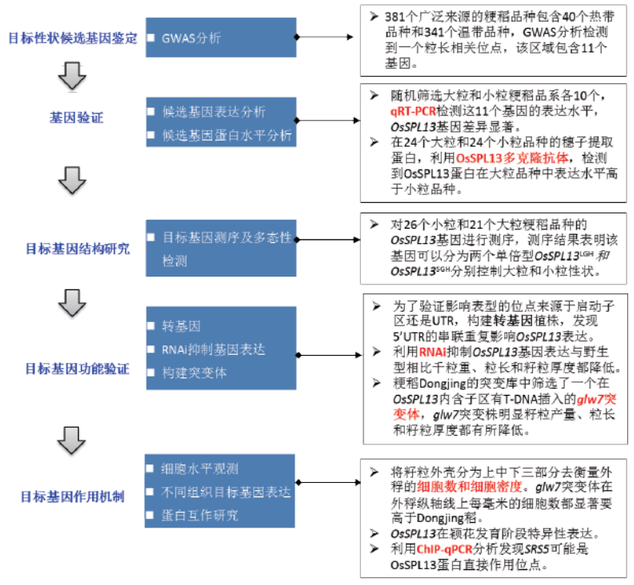


图8 案例一研究思路

3.2 案例二：玉米地方种杂交F_1群体研究开花期适应性 [16]

本研究选择来自35个美洲国家的4471个玉米地方种，这些样本根据生长环境的地势高低划分为3类，在这些样本中筛选部分亲本构建F_1群体。利用GBS对4471个品系和3552个F_1样本进行基因分型，并利用以下两个研究思路研究玉米开花性状。第一、玉米地方种是适应于当地环境的，研究人员利用环境信息作为性状研究玉米适应性；第二、利用田间控制下的样本通过FOAM（F-one association mapping）方法研究开花期性状。FOAM首先在不同的群体中选择样本，然后不同样本间杂交获得F_1群体，利用F_1群体样本进行GWAS分析（图9）。

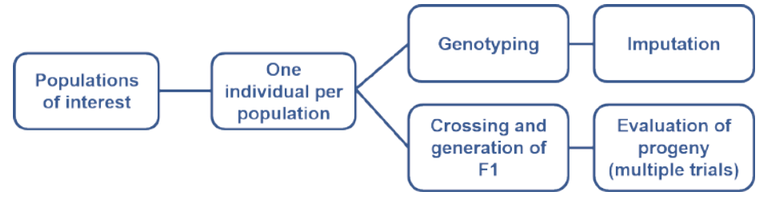


图9 实验设计。图中展示了研究的两个思路，上面一条线代表利用自然群体进行适应性研究；下面一条线代表利用从自然群体中筛选的代表性样本杂交构建F1群体进行GWAS分析的FOAM方法。

玉米开花期是很重要的适应性相关的性状，当前研究已知该性状是受微效多基因控制的。在许多的物种中，开花期相关的遗传结构主要是与纬度和海拔相关，所以本研究根据海拔高度对样本进行分类研究环境适应性问题。根据思路一研究发现染色体2、5、6、8的着丝粒区和染色体3着丝粒上游，以及从高原墨西哥类蜀黍中渗入玉米的13M片段 Inv4m与海拔高度关联。除了这些低重组区域，研究还发现有366个基因与海拔高度关联。根据思路二，研究人员构建F1群体，历时2年在墨西哥的13个地方进行了22次实验，每次实验涵盖不同的玉米品系。针对每次实验利用MLM方法进行开花期的GWAS分析，发现染色体3、5、6着丝粒区，Inv4m染色体3上的6M区域与之关联。部分结果与NAM群体QTL定位的结果重合。除了结构变异的影响，分别检测到881和883个基因与雌性和雄性开花相关（图10）。本研究定位检测到的开花期相关基因大部分是通过间接效应作用于开花期性状，一些与环境相互作用，如检测到关联的SNP中分别有61.4%和19%与海拔和纬度重合。研究人员还利用高密度的分子标记预测开花期，发现其也具有此方面的潜能。

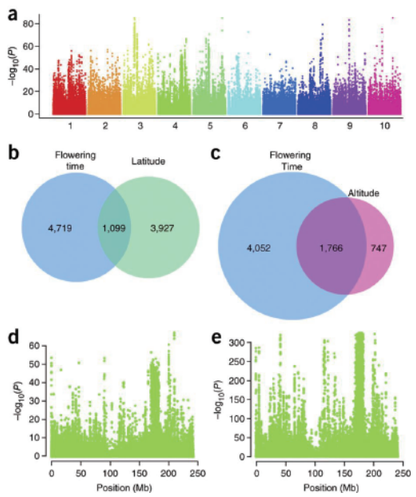


图10 开花期GWAS定位结果以及开花期和海拔高度相关SNP的overlap。(a) 雌蕊开花期manhattan plot；(b、c) 维恩图--开花期和海拔高度相关SNP的overlap；(d) chr4雌蕊开花期GWAS检测结果；(e) chr4海拔高度GWAS检测结果。

3.3 案例三：牛单基因和复杂性状定位[17]

牛是人类生活中重要的肉源、奶源和劳力，已经被驯化了一万年。如今牛的健康、营养成分指标是育种产业重要关注性状，但是这些性状大部分主要是由大量的微效基因控制，如果要进行基因组预测和遗传缺陷检测等技术需要依赖大群体基因组数据。本研究中测序了234头牛，通过比较基因组学和GWAS分析，定位到与胚胎死亡、骨骼畸形、产奶能力以及毛发卷曲相关的关键基因。利用US荷斯坦牛基因组数据与234个牛测序数据进行比较，在US荷斯坦牛Chr8的94.0–96.5 Mb区域中没有荷斯坦牛单倍型3（HH3）为纯合的个体，而在234个测序的牛中只检测到一个牛携带有杂合型HH3，并且有一个突变位点（g.95410507T>C）。对10个已知的HH3携带者进行PCR产物测序和在5,606荷斯坦牛中进一步验证了这个突变位点与胚胎期流产相关。

经常作为父本的Igale有1%的小牛呈现斗牛犬性状-颈短、头盖鼓起、舌头突出和口腔腭裂（沮丧表情），这种病会导致软骨发育异常，有致命危险。根据前人研究这种病不属于隐性遗传病，可能是显性突变导致的。两个染病小牛作为cases，234牛

测序数据作为control，最终检测到Chr5上 g.32475732G>A 的突变影响功能基因COL2A1，而COL2A1编码了II型胶原蛋白的α1链，据报道在人类中会引起骨骼疾病，包括ACGII 。

对3513头西门塔尔牛和2327头荷斯坦牛根据雌性后代泌乳早期牛奶中脂肪含量衡量父本性状进行关联分析，获得了6个未报道的QTL位点与该性状相关。其中在Chr14和Chr27上有两个位点位于DGAT1和 AGPAT6基因附近。前人研究已知DGAT1基因影响牛奶脂肪含量。AGPAT6基因上游转录起始位点有一段缺失。

研究策略：选择129头荷斯坦牛祖先种、43头西门塔尔牛祖先种、15头娟珊牛祖先种，4头经过低或高饲料转化率筛选的安格斯牛，共232头公牛和2头母牛进行全基因组重测序，平均测序深度8.3X。

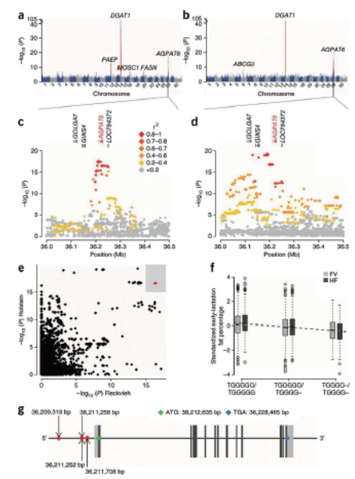


图11 牛奶脂肪含量关联分析结果。(a) 3513西门塔尔公牛关联分析 (b) 2,327荷斯坦公牛关联分析 。红色点代表P < 7.7×10^{-8}的变异. (c) 西门塔尔牛chr27上3,507关联位点 (d) 荷斯坦牛 chr27 上3,954 关联位点，不同颜色代表变异间的连锁不平衡 (r2) ，红色点对应基因 AGPAT6. (e) 比较西门塔尔牛和荷斯坦牛的P值，红色点代表四个变异位点显著性高 (P < 1×10^{-16})在所有的品种中，有一个indel at 36,211,252 bp. (f) 西门塔尔牛和荷斯坦牛缺失频率分别为0.25和 0.38，这个缺失在所有的品种中都与泌乳早期低牛奶脂肪含量相关. (g) *AGPAT6*基因结构。黑色代表外显子，浅色代表UTRs，缺失位于 *AGPAT6*上游1,383 bp 转录起始位点处 。

3.3 案例四：木豆驯化及相关农艺性状位点定位[18]

在亚洲、非洲和热带美洲的发展中国家中，木豆(Cajanus cajan)在食物和营养安全方面发挥着重要作用。在本研究中，研究人员选取292份木豆材料（包括117个育成品种、166个地方种、2个其他品种和7个野生种）进行全基因组重测序，测序深度5-12X。测序数据与参考基因组比较，检测了17.2 M的变异数据，包括SNP、InDel、SV。利用全基因组的变异数据分析发现，栽培木豆（育成品种和地方种）积累了更高的有害突变到无害突变的比例；同时发现从野生种到地方种，地方种到栽培种检测到的SV的长度都有变长；STRUCTURE分析结果发现大量的材料显示混合性，说明育成品种和地方种之间存在大量的遗传交流。FST值表明木豆的进化路线是从南亚到撒哈拉以南非洲地区，最后到南美洲和中美洲。为了检测由驯化和育种导致的选择消除，研究人员计算了多样性的减少（ROD），通过野生物种和地方品种的比较与地方品种和育成品种的比较，找到了2945个和1323个基因组区域，被确定为有更高ROD值。

对286份材料的重测序数使用SUPERGWAS的方法，针对2012-2013年和2013-2014年收集的8个农艺性状数据进行分析，估计标记-性状关联（MTA）。当P <0.05时共鉴定出241个MTAs。在第1年（2012-2013）检测到53个，在第2年（2013-2014）检测到90个，其余98个在合并的数据被检测到。大部分情况下，定位的区域都不是共享的，因此许多数量性状与环境适应有关。

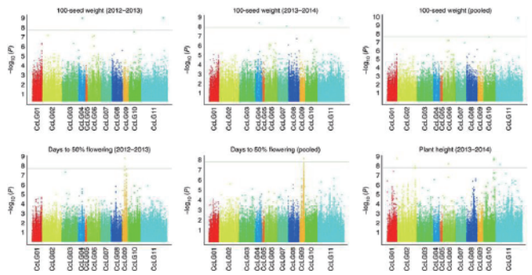


图12 木豆SUPER GWAS分析结果MTAs。
Y轴代表MTAs的P值，X轴代表连锁群

可能存在的风险

4.1 稀有变异和微效基因影响GWAS定位

GWAS分析可以检测目标性状常见的或候选的基因。但是一个性状可以由稀有的大效应变异控制，也可能是由许多常见的微效基因控制，这两类基因用GWAS研究存在难度[19,20]。因为GWAS定位功效决定于对应标记能够解释的表型变异大小（Figure 1a）[21]，而表型变异决定于等位基因效应大小的差异和它们在样本中出现的频率。微效基因达到某一检测功效相对需要更大的群体样本量。

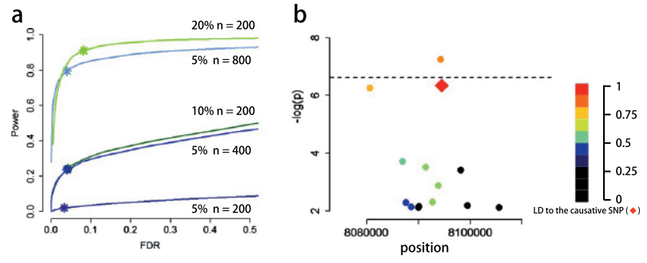


图13 群体大小和检测功效。
a、假设一个SNP能解释5%、10%和20%的表型变异，模拟计算不同的群体大小下的检测功效和FDR值；b、模拟causative SNP（红色方块）并不是检测结果最显著的[20]。

那么如何提高稀有变异或微效基因的检测功效呢？解决的方法包括提高样本数、只针对目标区域进行基因分析和研究、提高遗传多样性、降低遗传背景噪音等。但是提高样本量也不一定能完全解决稀有变异的问题，最好是用连续多个标记作为整体标记来进行研究，未来单倍型作为标记进行GWAS研究也许会成为趋势。对于稀有变异利用家系群体进行QTL定位可能效果会更好。

4.2 样本量大小影响GWAS定位

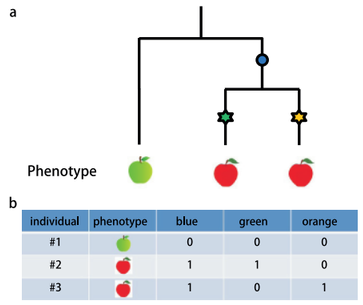
一些性状是由大效应的位点控制，进行GWAS研究需要的样本量比较低，即使低于100也能检测到有意义的位点[22]。对于复杂性状由若干的微效基因控制，那么样本量至少要达到几千[23,24]。从现在的动植物的发表文章来看，GWAS研究的样本数从100-5000不等。进行GWAS研究群体大效果更好。而且在群体大的情况下，LD相对会缩小，在标记密度足够的情况下，定位的区间小，有利于基因克隆。

那么如果不想定位效应很小或者频率很低的基因，我们可以用GWASpower/QT软件（http://www.mybiosoftware.com/gwaspowerqt-1-0-statistical-power-calculation-software-designed-gwas.html）辅助群体大小的选择，通过输入遗传力、标记个数等参数，计算达到预期检测功效需要的群体大小。根据我们的经验，一般情况下推荐群体大于300个，另外可以利用选择不同地理分布、表型差异大的品系，以最大化样本间的遗传变异，但是同时也可能引入遗传异质性。

4.3 遗传异质性影响GWAS定位

遗传异质性(genetic heterogeneity)是指某一种表型可以由不同的等位基因或者基因座突变所引起的现象。遗传异质性分为等位基因异质性和基因座异质性。

遗传异质性会降低变异检测的功效，因为它会减弱表型和任一变异的关联性，遗传异质性能够引起non-causative标记与表型关联性更强（图12,13）[20]。解决的方法之一是提高表型多样性高地区的样本量。



individual	phenotype	blue	green	orange
#1		0	0	0
#2		1	1	0
#3		1	0	1

图14 遗传异质性导致综合性关联。a、进化关系树，星星是近期发生的能引起表型变化(红果)的突变位点；b、早期的蓝色突变不会引起果皮颜色的变化但是和其关联[21]。

4.4 群体结构影响GWAS定位

群体结构指的是不同的亚群间同一等位基因频率差异显著。遗传结构不同的混合群体也会产生不平衡，如下所示：

表1 遗传结构不同的群体混合带来的连锁不平衡

群体	基因座A		基因座B		配子型			
	allele A	allele a	allele B	allele b	AB	Ab	aB	ab
群体1	0.4	0.6	0.2	0.8	0.08	0.32	0.12	0.48
群体2	0.2	0.8	0.6	0.4	0.12	0.08	0.48	0.32
1:1混合	0.3	0.7	0.4	0.6	0.1（≠0.3*0.4）	0.2（≠0.3*0.6）	0.3（≠0.7*0.4）	0.4（≠0.7*0.6）

关联分析是基于连锁不平衡来识别分子标记之间或候选基因与性状之间关系的方法。但是如果样本是来自不同遗传结构的亚群，混合群体也会计算到连锁不平衡，但是这样两个基因座位间的不平衡是来源于群体结构，对于GWAS定位目标性状相关基因来说属于假阳性。

现在的一些算法（如混合模型），引入群体结构和亲缘关系作为协变量帮助解决群体结构对GWAS定位结果的影响[25]，能有效的降低假阳性关联（图14）[21]。

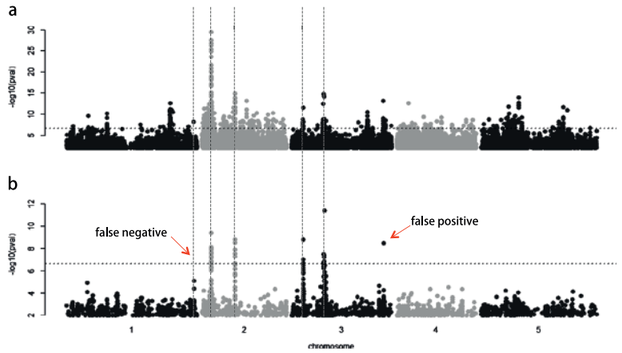


图15 GWAS结果manhattan plot，考虑群体结构以改善GWAS定位结果。五条竖虚线是模拟数据中预设的causative 位点，每个位点能解释最高10%的表型变异。a、一般线性模型结果；b、混合线性模型结果。前者假阳性较多，后者效果要好一些，但是同样存在一个假阳性和一个假阴性[21]。

群体结构对GWAS分析的结果影响大，虽然至今开发了若干的算法有助于消除群体结构的影响，但是有一些性状是和群体结构紧密连锁的，如植物的开花期[26]，如果控制了群体结构那么就降低了此类性状的检测功效。当然我们可以在选样的过程中，控制群体结构，例如同时对籼稻和粳稻进行了群体遗传特性进行了分析，但是由于粳稻和籼稻间存在显著差异，在进行GWAS分析时只针对籼稻进行研究[27]。利用多亲本衍生群体是一个不错的选择，康奈尔大学研究人员通过多个亲本和同一亲本杂交并不断自交构建了多亲本的NAM（Nested Association Mapping）群体，由于拥有统一的亲本作为遗传背景，打破了群体结构的影响[26]。多亲本衍生的群体克服了双亲作图群体中包含的变异信息过少，自然群体有群体结构影响的缺点。所以多亲本衍生群体结合连锁分析和关联分析的优点并克服二者的缺点，是QTL定位的上佳群体类型选择。

4.5 标记密度不足影响GWAS定位

对于大多数表型来说，基于PCR的分子标记如SSR和现有的SNP分型芯片中可能并不包含所有的causal变异，那么在进行GWAS分析的过程中可能就意味着标记密度不足，无法检测到causative位点。

但是因为有连锁的存在，如果每个LD block上有标记，那么即使标记的数量不是特别多也能够用于GWAS分析。不过随着测序技术的发展，样本全基因组数据的获得使得标记密度和标记类型将不再是问题。全基因组水平的SNP、InDel和CNV等都可以作为标记进行GWAS研究。

常见问题

1、GWAS分析的样本要满足什么样的条件？

答：进行GWAS分析样本量要大，推荐300个以上，进行GWAS分析可以选择自然群体、NAM群体，动物的全同胞或半同胞家系也可以，但要控制群体结构。进行GWAS分析最好有参考序列。

2、GWAS研究测序深度推荐？

答：采用全基因组重测序，每个个体推荐测序10X 。

3、没有参考基因组可以做GWAS吗？

答：没有参考基因组的物种做GWAS，分析流程是可以做的，但是分析的结果是点与表型间的关联。由于连锁效应，定位结果是个区间会比较可信，并且有了区间信息能用于后续的精细定位或基因克隆或相关基因的功能研究。如果只是点与表型的关联信息，首先有一些超过阈值的标记是假阳性位点，同时候选的基因克隆等相关研究都比较难做下去。所以如果是为了进行目标基因候选基因的检测的话，对于无参考基因组的物种不推荐采用GWAS。

华大优势

项目经验丰富：华大参与的动植物GWAS研究文章发表在《Nature Genetics》等顶级杂志上；

根据客户需求可以提供方案设计并完成个性化分析内容；

测序平台多样，选择空间大，能满足不同需求；

质控严格：从样本接收到数据交付都有严格的质量控制流程，保证数据准确性；

提供不同类型产品服务，一站式完成您的需求。

参考文献

[1] March R E. Gene mapping by linkage and association analysis [J]. Molecular biotechnology, 1999, 13(2): 113-122.

[2] Mackay I, Powell W. Methods for linkage disequilibrium mapping in crops [J]. Trends in plant science, 2007, 12(2): 57-63.

[3] Yu J, Buckler E S. Genetic association mapping and genome organization of maize [J]. Current Opinion in Biotechnology, 2006, 17(2): 155-160.

[4] Aranzana M J, Kim S, Zhao K, et al. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes[J]. PLoS Genet, 2005, 1(5): e60.

[5] McMullen M D, Kresovich S, Villeda H S, et al. Genetic properties of the maize nested association mapping population [J]. Science, 2009, 325(5941): 737-740.

[6] Gore M A, Chia J M, Elshire R J, et al. A first-generation haplotype map of maize[J]. Science, 2009, 326(5956): 1115-1117.

[7] Buckler E S, Holland J B, Bradbury P J, et al. The genetic architecture of maize flowering time[J]. Science, 2009, 325 (5941): 714-718.

[8] Huang X, Zhao Y, Wei X, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm[J]. Nature genetics, 2012, 44(1): 32-39.

[9] Karlsson E K, Baranowska I, Wade C M, et al. Efficient mapping of mendelian traits in dogs through genome-wide association[J]. Nature genetics, 2007, 39(11): 1321-1328.

[10] Guo J, Jorjani H, Carlborg Ö. A genome-wide association study using international breeding-evaluation data identifies major loci affecting production traits and stature in the Brown Swiss cattle breed[J]. BMC genetics, 2012, 13(1): 1.

[11] Morris G P, Ramu P, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum [J]. Proceedings of the National Academy of Sciences, 2013, 110(2):453-458.

[12] Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds[J]. Science, 2009, 324(5926): 528-532.

[13] Huang X, Wei X, et al. Genome-wide association studies of 14 agronomic traits in rice landraces [J]. Nature genetics, 2010, 42(11): 961-967.

[14] http://www.ts.cn/GB/channel5/34/200412/10/128491.html

[15] Si L, Chen J, Huang X, et al. OsSPL13 controls grain size in cultivated rice[J]. Nature genetics, 2016, 48(4): 447-456.

[16] Navarro J A R, Willcox M, Burgueño J, et al. A study of allelic diversity underlying flowering-time adaptation in maize landraces[J]. Nature genetics, 2017, 49(3): 476.

[17] Daetwyler H D, Capitan A, Pausch H, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle[J]. Nature genetics, 2014, 46(8): 858.

[18] Varshney R K, Saxena R K, Upadhyaya H D, et al. Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits[J]. Nature Genetics, 2017.

[19] Asimit J, Zeggini E. Rare variant association analysis methods for complex traits[J]. Annual review of genetics, 2010, 44: 293-308.

[20] Gibson G. Rare and common variants: twenty arguments[J]. Nature Reviews Genetics, 2012, 13(2): 135-145.

[21] Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review[J]. Plant methods, 2013, 9(1): 1.

[22] Atwell S, Huang Y S, Vilhjálmsson B J, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines[J]. Nature, 2010, 465(7298): 627-631.

[23] Manolio T A, Collins F S, Cox N J, et al. Finding the missing heritability of complex diseases[J]. Nature, 2009, 461(7265): 747-753.

[24] Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases[J]. Nature genetics, 2008, 40(6): 695-701.

[25] Kang H M, Zaitlen N A, Wade C M, et al. Efficient control of population structure in model organism association mapping[J]. Genetics, 2008, 178(3): 1709-1723.

[26] Buckler E S, Holland J B, Bradbury P J, et al. The genetic architecture of maize flowering time.[J]. Science, 2009, 325 (5941):714-.

[27] Huang X, Wei X, Sang T, et al. Genome-wide association studies of 14 agronomic traits in rice landraces[J]. Nature genetics, 2010, 42(11): 961-967.

家系群体定位QTL研究方案

研究背景

动植物研究领域，特别是与人类息息相关的农艺性状，大多属于数量性状。与质量性状相比较，数量性状的遗传研究要困难得多，因为质量性状的基因型可以通过表现型来辨别，而对于数量性状，基因型间的差异是量上的差异，基因型与表现型之间难以找到准确的对应关系。过去多认为数量性状由微效多基因控制，实际研究中发现，表现为连续、数量变化的性状，也可以是有一对或少数几对基因控制，但由于环境的修饰效应使得表型表现为数量化和连续性。遗传学的发展证实，大量数量性状是由一对或若干对主基因控制，并受微效多基因修饰。数量性状基因的DNA序列一般比较难以搞清楚，而数量遗传学把“从量上的差异研究基因差异”作为出发点。分子标记的出现使人们能够通过标记基因型与数量性状表现型之间的关联，定位数量性状基因（又称数量性状基因座位，简称QTL，quantitative trait loci）。

利用家系群体，通过连锁分析进行QTL定位。当两个基因位点A（等位基因用A和a表示）和B（等位基因用B和b表示）位于同一条染色体或连锁群上时，我们认为它们是连锁的，连锁的程度用位点间的重组率r去衡量。连锁在育种中的作用具有两面性，如果两个优良基因连锁在一起，我们希望连锁越紧密越好，两个基因能一起传递到下一代，这种连锁最好不要被打破；如果一个优良基因和一个不利基因连锁在一起，则我们希望能打破这种连锁。连锁不平衡常存在于两个连锁的基因位点间，QTL作图实际上就是利用连锁不平衡去发现与数量性状基因连锁的分子标记，而轮回选择这种育种方法则是希望通过群体间的随机交配打破有利基因和不利基因间的连锁，获得新的重组基因型。

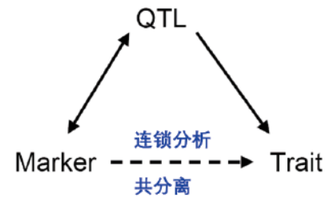


图1 QTL作图（QTL mapping），寻找QTL在染色体上的位置并估计其遗传效应的过程，利用标记与性状的连锁分析反映QTL与性状的关系。

连锁作图是一个高度控制下的试验，个体间杂交形成一个作图群体，且个体间关系已知。针对动植物群体，通常是用两个目标性状差异显著的亲本杂交，再通过自交或回交等方式构建作图群体（图2）。DH和RIL群体被认为所有的位点都是纯合的，自交不会再发生性状分离，所以属于永久性作图群体。对于F_1、F_2这类杂合位点多的群体来说，不能永久保存，被称为临时性作图群体。作物特别是异花授粉的作物构建作图群体相对比较容易，在林木和动物群体中，构建家系难度提高，对于无法自交的个体一般通过姊妹交来构建自交群体。实验者通过构建家系创造了一个封闭系统，利用少量的遗传标记来推测这些较少的重组位点的位置。利用杂交后代的基因型数据，实验者就可以推断两个重组位点间的染色体区段是否与表型有关。

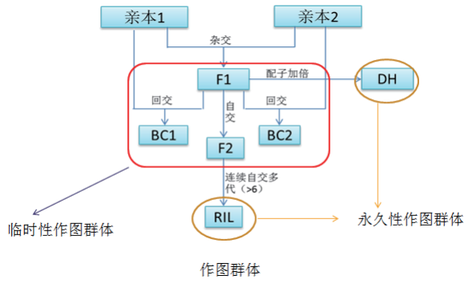


图2 家系群体类型

1961年，Thoday首次利用一对侧翼标记定位一个QTL[1]，随后QTL的研究得到飞速发展。作为构建遗传图谱必不可少的分子标记，从最早的RFLP一直到近年流行的SSR，标记的密度不断提高。当SSR等第二代标记的密度和均匀性已无法满足育种学家对于遗传图谱质量的需求时，第三代分子标记——SNP标记应运而生。

SNP是基因组中最常见的变异，而且有多种手段可以进行快速、高通量的基因分型。采用SNP标记制作的遗传图谱，是目前技术下质量最好、密度最高、最为均匀的图谱。目前进行SNP分型的方式包括全基因组重测序、简化基因组测序（RAD、GBS）、SNP分型芯片等。SNP构建遗传图谱同传统遗传图谱相比具有巨大的优势，国内外各科研机构对该技术的开发，始终保持着极为热情的态度。

方案设计

方案类型	内容	特点
Bin Map	双亲全基因组重测序（20X） 子代低深度重测序（~1X） ↓ 构建Bin Map ↓ Bins作为标记构建遗传图谱 —表型 ↓ QTL定位	适用于DH、RIL群体，子代低深度测序，连续的多个SNP（bins）判断染色体片段来源，以bins作为标记构建遗传图谱并进行QTL定位
全基因组重测序构建遗传图谱	双亲&子代全基因组重测序 （双亲：20X；子代：5X） ↓ 筛选SNP标记 ↓ 构建遗传图谱 —表型 ↓ QTL定位	针对有参考基因组物种，全基因组扫描SNP变异构建图谱，标记密度高

图3 遗传图谱构建方案

2.1 样本建议

F_1、F_2、BC、RIL、DH等动植物作图群体，推荐群体大小200个以上。

林木、鱼类高杂合样本构建的F_1群体，可以用于QTL作图。纯合样本杂交构建的F1群体，无表型分离不能用于QTL作图。

2.2 实验技术

采用全基因重测序,根据群体类型选择适宜的构图方案。

2.3 测序参数

采用全基因组重测序，亲本建议测序深度>20X，子代个体推荐测序1-10X（构建Bin Map子代测序深度可以比较低）

2.4 分析结果

2.4.1检测变异信息

利用测序结果，通过与参考序列比对，检测亲本间中存在的多态性SNP位点信息（图4）。

图4 有参测序比对检测SNP

2.4.2亲本间多态性SNP筛选

基于SNP检测结果，筛选亲本间多态性SNP。

对于F_2、BC、DH和RIL群体：筛选双亲间具有多态性的纯合位点（aa×bb型）；过滤掉亲本信息缺失的位点。

对于F_1群体：筛选双亲间具有多态性的杂合位点（lm×ll、nn×np、ab×cd、ef×eg、hk×hk型）；过滤掉亲本信息缺失的位点。

表1 不同群体遗传图谱构建标记类型

亲本基因型 父本×母本	标记类型解释	F_1群体	F_2/BC/ DH/RI群体	研究内容		
				F_1群体	F_2群体	DH/RI群体
aa×bb	双亲为不同的纯合位点		√		aa、ab、bb	aa、bb
lm×ll	第一个亲本为杂合位点， 第二个亲本为纯合位点	√		ll、lm		
nn×np	第一个亲本为纯合位点， 第二个亲本为杂合位点	√		nn、np		
ab×cd	两个亲本都是杂合位点， 4个allele	√		ac、ad、bc、bd		
ef×eg	两个亲本都是杂合位点， 3个allele	√		ee、eg、ef、fg		
hk×hk	两个亲本都是杂合位点， 2个allele	√		hh、hk、kk		

2.4.3子代基因分型

根据筛选获得的亲本标记类型对子代进行基因分型，得到的标记经过卡方检验（显著性水平α=0.01）去除偏分离标记（如F_2群体aa、ab、bb三种基因型出现的期望概率比为1:2:1，明显偏离此比例认为该标记偏分离）；过滤掉子代群体中缺失率>20%的标记位点；根据不同群体的子代分离基因型过滤掉异常标记型。最终得到的标记，整理为软件Joinmap4.1输入文件格式，用于遗传图谱的构建。

表2 不同群体子代基因型编码方式

群体类型	群体描述	子代基因分型代码
BC1	纯合亲本杂交F1与亲本之一回交1代	(a,h)或(h,b)
F2	F1自交构建的F2	(a,h,b)
RIx	F1自交了X代构建的重组自交系F2=RI2	(a,b)
DH	单倍体加倍构建的加倍单倍体	(a,b)
IMxFy	F2单株随机交配x-2代，然后自交y代，IM2F0=F2，IM2Fy=RIx（x=y+2）	(a,b)
BCpxFy	BC1再单株连续回交x-1代，然后自交y代，BCp1F0=BC1	(a,h,b)
CP	BC1再单株连续回交x-1代，然后自交y代，BCp1F0=BC1	lm×ll：(ll, lm);nn×np：(nn, np) ab×cd：(ac, ad, bc, bd);ef×eg：(ee, eg, ef, fg) hk×hk：(hh, hk, kk)

表3 子代群体分型结果示例（F_2群体）

标记名 / 称基因型	SNP1	SNP2	SNP3	……	SNP N
父本	a	a	a	a	a
母本	b	b	b	b	b
子代1	a	a	b	b	a
子代2	b	b	a		b
子代3	b	b	a		a
……					
子代M	b	b	b		b

2.4.4遗传图谱构建

根据基因分型结果，利用作图软件如Joinmap4.1进行图谱构建。根据各标记在染色体上的重组值(或交换率)，将染色体上的各标记之间的距离和顺序标出来，绘制成遗传图谱。标记间距离是为遗传距离（m），以M（摩）或cM（厘摩）表示，图距是交换率r的函数。遗传距离是个相对距离，不同作图函数选择影响图距，作图函数在作图软件中可选。作图函数中：Morgan 作图函数，1cM相当于交换率1%；Haldane作图函数和Kosambi作图函数，1cM对应交换率<1%。一般情况下，选择Haldane作图函数。

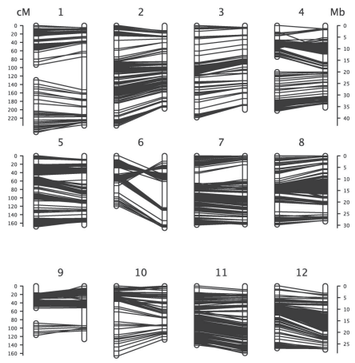


图5 遗传图谱示例[3]

不同物种1cM遗传距离对应的物理距离是不同的，即使是同一条染色体，不同区域1cM遗传距离对应的物理距离也不同。

表4 不同物种遗传距离与物理距离的对应关系[4]

物种	单倍体基因组大小（kb）	遗传图谱的长度（cM）	碱基对（kb）/cM
酵母（Yeast）	2.2×10^4	3700	6
Neurospora	4.2×10^4	500	80
Arabidopsis	7.0×10^4	500	140
Drosophila	2.0×10^5	290	700
西红柿（Tomato）	7.2×10^5	1400	510
人类（Human）	3.0×10^6	2710	1110
小麦（Wheat）	1.6×10^7	2575	6214
水稻（Rice ）	4.4×10^5	1575	279
玉米（Corn ）	3.0×10^6	1400	2140

2.4.5Bin Map构建

根据子代群体的基因型信息，将一定数量的连续SNP作为判断染色体重组事件的最小单位（recombination bin），判断子代每个bin来源于父母本的情况，得到每个子代的全基因组物理重组图谱，并构建Bin Map（图7）。Bin map中标记间的距离是真实的物理距离。如果要进行连锁分析，还需要将bin作为分子标记构建遗传图谱如图6。

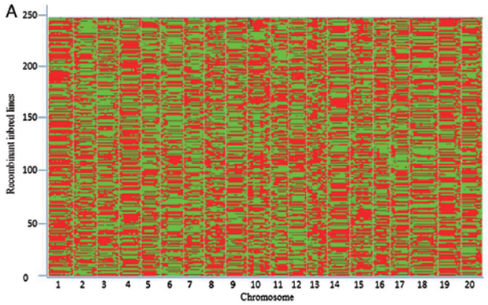


图6 Bin Map示意图[5]

2.4.6遗传图谱整合（需提供同一作图群体原图谱数据）

结合SNP和SSR/AFLP 等其他类型的标记构建的同一作图群体的遗传图谱。可以加密标记，用于精细定位。

2.4.7QTL定位

根据连锁分析可以检测连锁群上哪些区间影响目标性状，QTL定位信息包括定位区间，效应大小等信息。QTL定位软件包括MAPQTL、QTL IciMapping、QTL Cartographer等。根据QTL定位到目标性状候选基因区间后，比对回参考基因组，可以检测到此区间的基因信息，通过与过往研究比较可以得知定位的区间与哪些以前定位到的基因位置重合或靠近。同时利用现有的基因功能数据库，对候选基因进行功能注释和聚类分析，更深层次的挖掘目标性状的分子机制。

表5 QTL定位结果示例[6]

Trait		Year	Maximum	Minimum	Mean(±SD)	Distribution
Fruit weight (g)		2008	253.80	56.80	128.77 (±30.33)	non normal
		2009	205.60	50.00	104.55 (±23.67)	non normal
		2010	193.60	52.00	110.94 (±26.64)	non normal
		2011	170.00	33.00	81.30 (±30.36)	non normal
Firmness (kg cm^{-2})		2010	15.67	3.84	8.78 (±1.81)	non normal
		2011	12.16	5.49	8.71 (±1.39)	normal
acidity	total acid	2011	11.547	2.118	6.887 (±2.285)	normal
(mg g^{-1})	malic acid	2011	11.211	1.653	6.452 (±2.322)	normal
	citric acid	2011	0.200	0.016	0.081 (±0.039)	normal
	tartaric acid	2011	0.310	0.010	0.109 (±0.056)	non normal
	oxalic acid	2011	0.099	0.048	0.070 (±0.011)	normal
	acetic acid	2011	1.021	0.000	0.169 (±0.214)	non normal
	succinic acid	2011	0.027	0.000	0.006 (±0.004)	non normal
sugar content	total sugar	2011	127.175	79.564	104.737 (±10.811)	normal
(mg g^{-1})	fructose	2011	68.477	33.302	51.014 (±6.854)	normal
	glucose	2011	35.065	9.239	19.690 (±6.176)	non normal
	sucrose	2011	61.584	11.059	34.033 (±9.765)	normal

2.5 项目周期

样品检测合格后，建库+测序+标准信息分析：约60个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.6 预期结果

借助高通量分型平台，快速鉴定作图群体亲本和子代样本的基因型信息，并构建基于SNP标记的高密度连锁图谱，定位QTL。

应用案例

2.7 辅助研究策略

可以通过RNA测序研究QTL定位区间内的基因，在目标性状上存在显著差异的样本间是否表达量不同，亦可以检测对应蛋白的表达情况，进一步缩小范围，找到目标基因。

2.8 后期验证手段

分析得到的候选基因，可以利用转基因、基因敲除、基因沉默（RNAi）等方式验证基因功能。

3.1 案例一：谷子遗传图谱构建提升基因组组装及QTL定位（华大参与）[7]

2012年，谷子基因组序列被公布，其中Zhanggu和Yugu的基因组序列都被构建完成，但是组装的结果还有提升空间。本研究利用作图群体构建图谱辅助谷子基因组图谱质量的提升。Zhanggu和A2(雄性不育系)杂交构建RIL群体，Zhanggu绿叶、红毛、黄色花药和稀禾定抗性，A2黄叶、绿毛、棕色花药、对稀禾定敏感。全基因组重测序184个谷子的RIL子代样本，每个样本~2X。所有测序数据比对回Zhanggu基因组序列，检测到483414个SNP，构建了bin map，共构建了3437个bins。利用bins构建遗传图谱，图谱有9个连锁群总长1927.8 cM。利用图谱信息将Zhanggu基因组序列锚定到染色体上，将之前未定位的16Mb scaffolds锚定到基因组上，基因组序列提升到416M。A2和RIL测序数据比对回Yugu的基因组序列，构建bin map，基于bin map数据修正了Yugu参考基因组上的错误，并利用Zhanggu同源序列填补了3158个gaps。基于RIL群体对9个农艺性状进行基因定位，研究性状有单基因控制的质量性状-稀禾定抗性、叶色、毛色、花药颜色和穗硬度，和多基因控制的数量性状-株高、抽穗期、旗叶宽度和旗叶长度。定位到2个与株高相关的QTL，并且一个候选基因与已知赤霉素合成基因sd1有89%的一致性。3个QTL与抽穗期相关。

研究策略：RIL群体全基因重测序2X，构建bin map，通过构建Bin map/遗传图谱辅助基因组锚定到染色体水平，同时有助于修正参考基因组的错误，并能利用其他亚种的参考序列，通过同源序列比对弥补参考基因组的gap信息。

表6 Zhanggu和Yugu基因组二次编辑数据统计

Strain	Chromosome length (bp)	Gap length (bp)	Gap number	Gap ratio	Filled gap number
Zhanggu	399 854 594	26 817 695	31 942	6.7%	/
Zhanggu ^{2th}	415 979 272	28 962 873	34 452	7.0%	/
Yugu	401 300 876	4 616 102	6171	1.2%	/
Yugu ^{2th}	402 520 233	2 175 332	3297	0.5%	2874

3.2 案例二：花生遗传图谱构建及QTL定位（华大参与）[8]

早期叶斑病（ELS）、晚期叶斑病（LLS）及番茄斑萎病毒引起的疾病（TSWV）会导致严重的花生产量下降。本研究利用高通量测序手段，对花生的重组自交系进行测序，利用20个连锁群中的8869个SNP标记，构建高质量的遗传图谱。图谱长度3120 cM，有效位点平均间距1.45 cM。对以上三种疾病进行QTL定位：在染色体B05及B03定位到两个与ELS相关的QTL；在染色体A05和B03定位到两个与LLS相关的QTL；以及在染色体B09上定位1个与TSWV相关的QTL。

研究策略：母本Tifrunner表现出对三种疾病的抗性，父本GT‐C20则对三种疾病具有易感性，二者构建重组自交系RIL群体。利用母本、父本及91个RIL个体进行遗传图谱的构建，其中，母本、父本进行高深度测序（10-100X），子代测序深度为2-5X。

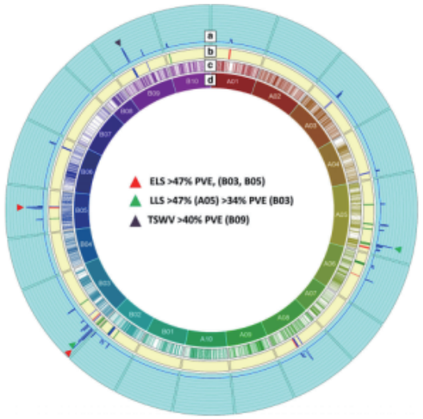


图7以SSR标记及SNP标记建立的遗传图谱及QTL定位情况

可能存在的风险

利用作图群体进行QTL定位的精度不仅受标记密度的影响，同时也受群体大小的影响。由于群体小，重组事件发生少，那么最终QTL定位的区间会比较大，为基因克隆带来难度。同时在作图群体中进行的QTL定位，检测的是存在于双亲中的基因信息，并且QTL的效应值也只适用于这个群体，也许在我们鉴定的群体中是主效QTL，换个群体效应值降低。

常见问题

1、连锁图谱构建适用于什么样的群体？

答：连锁图谱的构建适用于作图群体，它是由性状差异显著的亲本杂交衍生的群体。亲本选择的要求：要考虑亲本间的遗传多态性、目标性状差异、亲本的纯合度和杂交后代的可育性。构建分离群体类型，根据遗传稳定性可将分离群体分成两大类：暂时性分离群体如F_1、F_2、BC等，永久性分离群体如RIL、DH等。

2、加密标记才能精细定位QTL？

答：寻找QTL（数量性状基因座）在染色体上的位置并估计其遗传效应过程叫QTL作图或QTL定位。标记稀疏会导致双交换检测不到，如果标记加密不仅能降低漏掉双交换的概率同时能使连锁的QTL间有空白标记区间，有利于QTL定位。增加标记有增加检测功效的作用，但更有利于提高效应较小QTL的检测功效。但是增加标记也会使得假阳性QTL有增加的趋势。另外如果一个有限群体中多个连锁的标记之间没有发生重组，那么这些标记即使数量比较多，但是发挥的作用也只是一个标记的效果。所以在关注加密标记提高检测功效的同时，还要注意增加标记只有在大群体中才会产生更大的作用。

3、 作图群体大小如何确定？

答：遗传图谱的分辨率和精度，很大程度上取决于群体大小。群体越大，则作图精度越高，但群体太大，不仅增大实验工作量，而且增加费用。目前大部分已发表的分子标记连锁图谱所用的分离群体大小多为100-200个。但是在群体较小的情况下，由于重组事件有限，那么就会导致最终与性状进行定位分析的过程中，定位的区间范围会比较大。因为即使标记密度足够大，但是没有重组事件发生，那么连续多个SNP的作用相当于一个SNP。由于不同的分离群体中子代的基因型种类不同，例如F_2群体中存在更多种类的基因型，而为了保证每种基因型都有可能出现，就必须有较大的群体。所以在分子标记连锁图的构建方面，为了达到彼此相当的作图精度，所需的群体大小的顺序为F2>RIL>BC1和DH。

4、 遗传图谱都有哪些作用？

答：①QTL定位：家系群体构建遗传图，表型性状与标记进行连锁分析，确定与目标性状相关的基因位置、数量、效应值等。②辅助基因组组装：作图亲本之一是进行基因组组装的品种，通过筛选标记构建遗传图谱，确定了不同标记间的位置关系，辅助基因组组装获得的scaffold进一步延长，甚至定位到染色体水平。

5、提高QTL检测功效的途径有哪些？

答：①增大作图群体；②减小表型测定时的误差（即提高性状的遗传力）；③降低表型变异也可间接提高PVE，从而提高QTL的检测功效。遗传研究中近等基因系和染色体片断置换系都是通过这种途径提高遗传分析的可靠性。

6、缺失标记的影响有多大？

答：一个群体大小为n、缺失率为p的群体的作图功效与大小为n(1-p)、无缺失群体的作图功效大致相同。所以标记缺失相当于群体样本数降低。

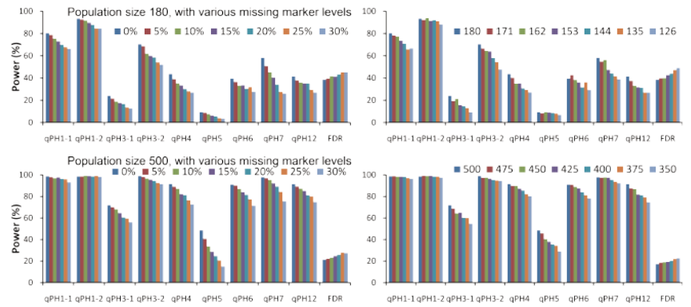


图8 不同水平标记缺失对检测功效的影响[9]

7、标记奇异分离对检测功效的影响？

答：如果奇异分离位点与QTL不连锁对定位结果无影响；如果奇异分离位点与QTL连锁，对定位功效有影响，但是当群体增大到500时，影响很小。

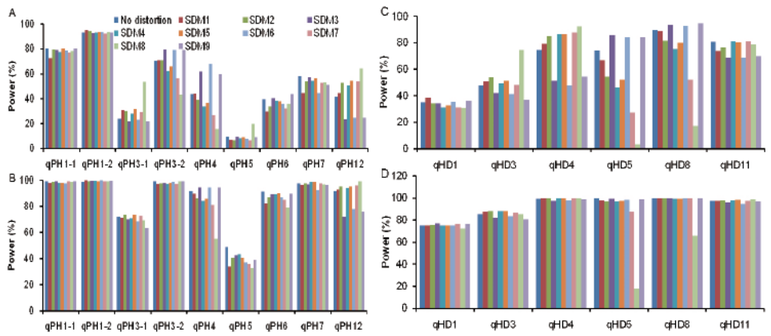


图9 奇异分离标记对检测功效的影响。A、C群体大小180，B、D群体大小500[9]

8、复合性状适合QTL定位吗?

答:复合性状是指由构成性状通过加减乘除计算出来的表型值,如千粒重、长宽比等。QTL定位推荐用构成性状,谨慎用复合性状。因为①复合性状遗传结构更复杂:复合性状受较多QTL控制、QTL具有更复杂的遗传效应和连锁关系;②复合性状遗传力变小:独立QTL,复合性状的遗传力与构成性状基本一致。连锁QTL,积和商性状的遗传力有所下降;③复合性状的检测功效有不同程度的降低,错误发现率升高;选择较大的作图群体,复合性状检测功效下降的程度减小,同时错误发现率也略有减小,但错误发现率仍然高于构成性状;④复合性状会有一些独有QTL:遗传机制尚不明确,模拟实验结果表明复合性状独有QTL或者是由构成性状中的微效QTL引起的,或者是"幻影"QTL。

9、 QTL作图群体中的表型数据要服从正态分布?

答:QTL定位表型不需要服从正态分布,但随机效应要求符合正态分布。

10、连锁分析与关联分析各自的局限性是什么?

答:连锁分析的局限性包括:需要构建作图群体;检测QTL的个数有限,只能检测到作图群体双亲间存在差异的QTL位点;QTL分辨率低,作图群体有限代重组,定位区间大;QTL有效性的限制性,换个群体效应值会发生改变。

关联分析的局限性包括:随机交配掩盖基因座间连锁关系;群体结构导致定位结果假阳性;关联分析中需要大量的分子标记。

华大优势

项目经验丰富:华大具有各种类型群体进行遗传图谱构建经验;

根据客户需求可以提供方案设计并完成个性化分析内容;

测序平台多样,选择空间大,能满足不同需求;

质控严格:从样本接收到数据交付都有严格的质量控制流程,保证数据准确性;

提供不同类型产品服务,可一站式完成您的需求。

参考文献

[1]Thoday J M. Location of polygenes [J]. Nature, 1961, 191: 368-370.

[2]Catchen J M, Amores A, Hohenlohe P, et al. Stacks: building and genotyping loci de novo from short-read sequences [J]. G3: Genes, Genomes, Genetics, 2011, 1(3): 171-182.

[3]Guo Y, Yuan H, Fang D, *et al.* An improved 2b-RAD approach (I2b-RAD) offering genotyping tested by a rice (Oryza sativa L.) F2 population [J]. BMC genomics, 2014, 15(1): 1.

[4]翟虎渠, 王建康.应用数量遗传学, 2017

[5]Xu X, Zeng L, Tao Y, et al. Pinpointing genes underlying the quantitative trait loci for root-knot nematode resistance in palaeopolyploid soybean by whole genome resequencing[J]. Proceedings of the National Academy of Sciences, 2013, 110(33): 13469-13474.

[6]Sun R, Chang Y, Yang F, *et al.* A dense SNP genetic map constructed using restriction site-associated DNA sequencing enables detection of QTLs controlling apple fruit quality [J]. BMC genomics, 2015, 16(1): 1.

[7]Ni X, Xia Q, Zhang H, *et al.* Updated foxtail millet genome assembly and gene mapping of nine key agronomic traits by resequencing a RIL population[J]. GigaScience, 2017, 6(2): 1.

[8]Agarwal G, Clevenger J, Pandey M K, et al. High‐density genetic map using whole‐genome resequencing for fine mapping and candidate gene discovery for disease resistance in peanut[J]. Plant biotechnology journal, 2018.

[9]李慧慧 , et al. 数量性状基因定位研究中若干常见问题的分析与解答. 作物学报, 2010, 36(6): 918—931

突变检测研究方案

研究背景

突变体是某个性状发生可遗传变异或某个基因发生突变的生物体材料。长期以来,育种学家们一直在尽力地发现和分离一些有价值的变异材料。早期的遗传学家和育种学家们主要是在自然界中筛选和寻找各种突变体。这种因自然界中各种条件的驱使,使得生物体内发生可遗传的DNA变异并进行稳定遗传的个体称为自发突变体。然而自发突变的频率比较小,并且突变的方向并不都是有利的,大多数并非人们感兴趣的突变体,同时这种小概率的突变还有一大部分在自然界选择过程中被摒弃。因此仅仅利用自然变异是远远不能满足人类利用和研究的需要。所以利用人工手段诱导遗传物质产生可以稳定遗传的变异是现代遗传学获取突变体的最有效手段之一。

自20世纪70年代以来,γ射线、快中子、MNU(N-甲基-亚硝基脲)、EMS(甲基磺酸乙酯)等理化诱变创建人工突变体在动植物遗传育种中开始广泛应用。此外,一些针对基因组内部基因进行修饰和改进的新技术如T-DNA插入、RNAi抑制基因表达、转座子标签的出现和发展,大大地加快了突变体的创建步伐。随着航天技术的发展,航空育种也获得了大量的突变体。突变体创建技术手段的改进大大的促进了植物基因功能的注释和作物遗传育种。

在测序技术大规模应用以前,突变体的鉴定方法有表型鉴定、细胞学鉴定、生理化学鉴定、分子标记鉴定以及TILLING和EcoTILLING鉴定,但这些传统方法都有局限性:前三种方法鉴定结果不够准确且不能精确到DNA水平;分子标记鉴定只能将突变基因定位在某一区域,要精确定位还需进行图位克隆,操作繁琐且耗时长久;TILLING和EcoTILLING鉴定只能对已知基因进行鉴定,应用范围比较小。

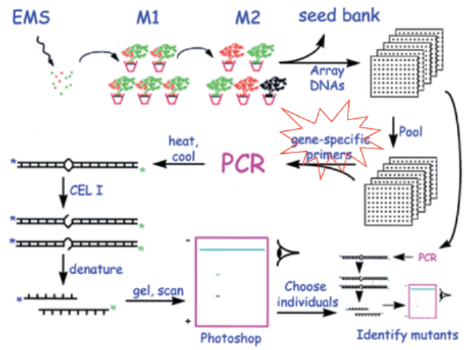


图1 植物中TILLING检测点突变位点示意图[1]

近几年来,应用新一代高通量测序技术寻找突变位点取得了成功并逐渐成为一种趋势。目前采用重测序技术定位点突变的技术已经在线虫[2-3]、果蝇[4]、酵母[5-6]、拟南芥[7-8]等模式生物中成功应用并逐渐延伸至其他物种。在研究过程中原来用于数量性状定位的选择基因型(selective genotyping)分析方法也得到了充分的应用。

选择基因型分析方法是在表型鉴定的基础上，只对群体中表型值最高和最低的双尾或单尾群体进行基因型鉴定，在假设控制某数量性状的QTL与标记连锁，即QTL的位置和标记的位置完全重合的情况下，选择会引起控制性状的基因频率发生变化，对于和性状无关联的基因，其频率选择前后将保持不变。混合分组分析法（Bulked Segregant Analysis，BSA）[9]及选择DNA池法（selective DNA pooling）[10]是一种简单的选择基因型分析方法，它是将目标性状在F_2或BC代中表型极端的高、低两组个体的DNA分别混合成两个DNA池，然后利用分子标记在两池中进行标记与性状间的共分离分析检测QTL。BSA分析在过去的研究过程中属于选择基因型中的一种极端情况，将两个表型极端的个体混合成两个DNA pool如抗池和感池，那么就要求抗池中只包含抗病等位基因，感池中只包含感病等位基因。但是测序技术的出现，在我们假设混池中每个个体被测到的概率相等的情况下，可以近似的利用测序reads计算等位基因频率，此时分析方法等同于选择基因型分析。

方案设计

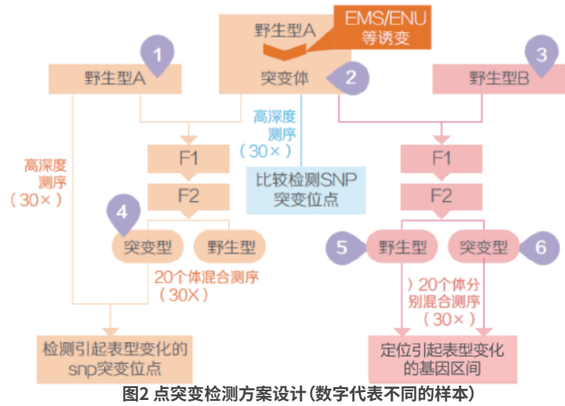


图2 点突变检测方案设计（数字代表不同的样本）

2.1 样本建议

样品需要是点突变个体或者是构建的家系群体，根据样本特性推荐不同的研究方案。以下研究方案更适合于隐形纯合突变位点。

1）方案1：没有作图群体，只有突变体和野生型：对野生型（样品1）和突变体（样品2）进行高深度的测序（30X），比较检测突变位点；

2）方案2：野生型（样品1）和突变体（样品2）构建了F_2群体（或RIL群体），选择野生型（样品1）和F_2突变株混合样品（20株，样品4）进行高深度的测序（30X），检测引起表型变化的突变位点；

3）方案3：无突变表型的其他品种和突变体构建了F_2群体（或RIL群体）：对F_2双亲（样本2、3）进行高深度的全基因组重测序（>20X），选择F_2中表型极端的样本各20株分别进行混合测序（样本5、6，30X）或是F_2中突变表型的样本20株混合测序（样本6，30X），利用双亲间的多态性SNP位点进行混合分组分析（BSA，Bulk Segregant Analysis）检测与表型相关的候选区间。

4）方案4：选择大样本量的F_2（或RIL）突变株混合样品（>50株，样品6）进行高深度的测序（30X），检测与表型关联的候选区间；

5）方案5：选择F_2/RIL群体中表型极端的样本各20株分别进行混合测序（样本5、6，30X），检测与表型相关的候选区间。

表1 各方案比较

方案	标记类型	标记区间	检测突变位点数	检测结果
方案1	发生突变的SNP	直接利用SNP，不进行窗口扫描	多	检测突变的位点，与表型无关的变异较多
方案2	发生突变的SNP	密度小	最少	检测与表型相关的点突变位点
方案3	不同品种间的SNP	密度大	较少	检测与表型相关的区间
方案4	等位基因频率与期望频率差异显著的位点	密度中等	中等	检测与表型相关的区间
方案5	两极端池中等位基因频率差异显著的位点	密度中等	中等	检测与表型相关的区间

2.2 实验技术

全基因组重测序。

2.3 测序参数

如方案设计所示，无论是个体或者混合样本，都建议采用全基因组重测序，每个文库建议30X以上，混合样本建议平均每个样品数据1X以上。

2.4 分析结果

2.4.1突变体与野生型亲本比对（方案1）

突变体和野生型进行高深度测序，检测两者间的差异SNP，在去除野生型亲本背景变异后，检测到所有突变位点，其中包含与表型相关的功能突变位点。以水稻为例，突变体直接与参考基因组比对能检测到上万个SNP，去除野生型亲本的背景变异后也要剩余几千个位点，这些位点大部分与表型关联度低。大量的标记位点被检测到，这对后期突变位点的验证带来很大的障碍。但是如果突变位点已经通过初定位到一个有限的区间（如几M），利用此方法进行检测能大大的降低最终突变位点数。

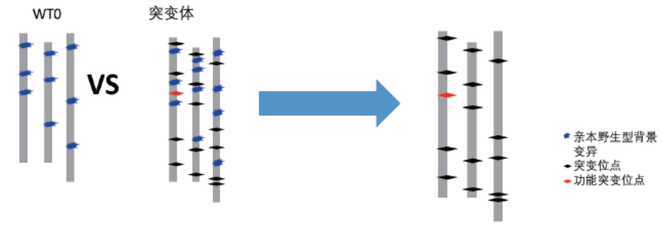


图3 野生型vs突变体

2.4.2突变体与野生型亲本构建群体检测突变位点（方案2）

突变体与野生型亲本杂交并构建F_2或BIL群体，通过分离重组，使得群体中的野生型个体背景不断替换为野生型亲本的基因型。对作图群体中的多个突变型个体（如20个）pooling测序，然后去除野生型亲本背景变异，那么检测到的突变位点数比方案一少的多，同时这些位点和表型关联。BIL群体多次回交，重组事件多，所以利用BIL群体检测到的位点较F_2群体更精细。

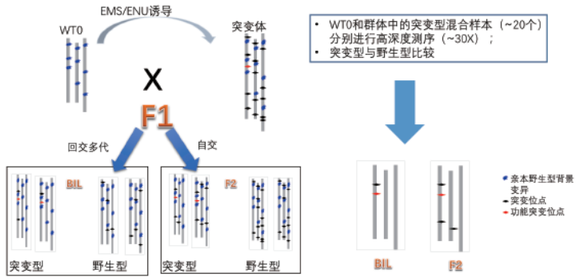


图4 检测示意图

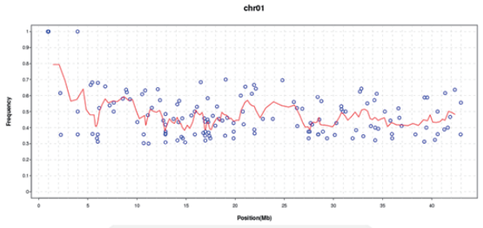


图5 结果示意图

2.4.3无突变表型的其他品种和突变体构建了家系群体（方案3）

检测突变体和非亲本野生型之间的纯合多态性SNP位点，在家系群体中两个极端表型池中检测这些多态性位点的等位基因频率（SNP-index，支持某一基因型reads深度除以该位点总的reads深度），频率差异显著的位点认为是与表型相关的候选位点。此分析方法不仅能进行点突变定位，同时能对数量性状的主效QTL进行定位。

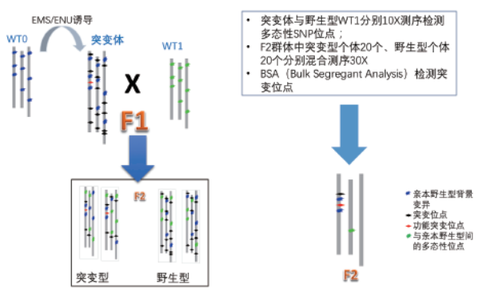


图6 检测示意图

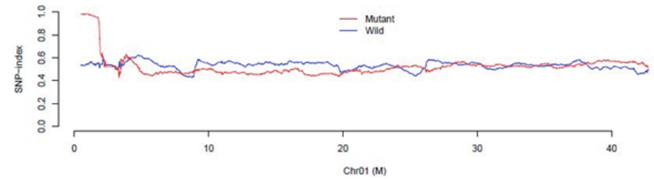


图7 结果示意图

2.4.4候选基因功能注释

对定位区域涵盖的基因进行GO功能分析和KEGG Pathway分析。GO数据库适用于各个物种，能对基因、蛋白质进行限定和描述。通过GO分析按照Cellular component、Molecular Function、Biological process对基因进行分类，找到该基因最初的功能。KEGG是有关Pathway的主要公共数据库，KEGG分析能让我们更清楚的知道候选基因参与的新陈代谢过程。

2.5 项目周期

样品检测合格后，建库+测序+标准信息分析：约50个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.6 预期结果

借助高通量测序平台，一次性扫描，准确、快速、高效的确定引起表型变化的点突变的区间范围。

2.7 辅助研究策略

通常情况下，候选区域会涵盖较多的候选基因，可以通过转录组测序研究这些基因在突变体和野生型间是否存在差异表达，同时结合已有基因功能数据库（KEGG、GO等）进行基因注释，辅助确定控制性状的真正候选基因。

2.8 后期验证手段

分析得到的候选基因，可以利用转基因、基因敲除、基因沉默（RNAi）等方式验证基因功能。

应用案例

3.1 案例一：MutMap检测水稻点突变[11]

水稻品种*Hitomebore*的突变体（叶子浅绿色）*Hit1917-pl1*和*Hit0813-pl2*与野生型杂交，构建了大于200个的F_2群体中，选择20个具突变表型子代混合测序。在突变体*Hit1917-pl1*和*Hit0813-pl2*分别检测到1,001和1,339个G→A/C→T的突变位点；*Hit1917-pl1*在10号染色体检测到一个包含7个SNP的集合，一个SNP位于基因OsCAO1的编码区，使得植株叶绿素含量降低；*Hit0813-pl2*在1号染色体检测到一个包含5个SNP的集合。

研究策略：突变体与野生型亲本杂交构建F_2群体，利用F_2中20个突变型混合样本进行全基因组重测序总测序量>5G，覆盖>12X，利用MutMap方法检测与表型相关的点突变。

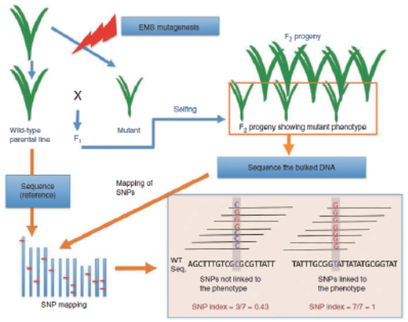


图8 MutMap定位水稻突变位点

3.2 案例二：MutMap+检测点突变[12]

用化学诱变剂EMS对水稻受精后幼年胚芽进行诱变，诱变后产生的M1（突变位点为杂合位点）。M1自交获得M2，目标突变性状是由一个隐性基因控制的，在M2代出现性状分离；选择杂合突变基因型的M2自交，获得M3；M3发生性状分离，选择M3子代中野生型和突变型各20-40株个体，混池测序。测序reads比对到亲本基因组上，计算SNP-index。M3突变混合池中目标突变位点SNP-index=1，同时通过比较突变池和野生池的SNP-index差值，计算△SNP-index，利用Fisher精确检验判断候选区间，检测突变所在的基因区域。

水稻HitomeboreEMS诱变后获得突变体Hit9188，突变体表现矮小和叶子颜色淡，并在3周后死亡，所以研究材料无法使用MutMap分析思路。M3突变表型和非突变表型植株1:3，突变位点为隐形纯合位点。选择M3代中40个突变表型和40个非突变表型植株分别混合测序利用MutMap+的分析方法在1号染色体上检测到Os01g0127300 基因与突变表型相关。并利用RNA干扰和转基因进行了验证。

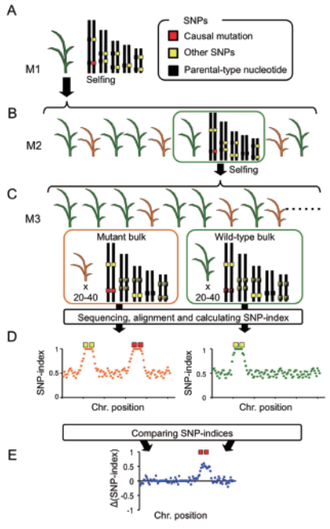


图9 MutMap+定位突变位点

3.3 案例三：MutMap-Gap检测点突变[13]

利用MutMap可以检测点突变，但是如果目标位点位于参考基因组没有组装上的gap区，或是参考基因组不具有的序列中，那么利用MutMap检测方法则不能有效的检测到目标突变位点(图10)。MutMap-Gap方法结合了MutMap和*de novo*组装，首先将所要研究的样本进行MutMap分析，通过计算SNP-index，分析SNP-index peak区，发现找不到跟突变性状相关的基因；那么突变位点很有可能就在品系特有基因区域内，将之前比对不上参考基因组的野生型亲本unmapped reads和MutMap分析中SNP-index peak区域的野生型亲本比对上的reads一起进行*de novo*组装，获得潜在的新基因，并以此为参考再计算SNP-index，检测目标突变位点。

利用MutMap-Gap的方法检测水稻稻瘟病抗性基因*Pii*，由于Nipponbare不包含*Pii*基因，如果要定位此基因，MutMap方法是无效的。Hitomebore含有*Pii*基因，采用MutMap-Gap的方法在Hitomebore突变体中定位到了染色体9上定位到*HIT7*基因（同*Pii*基因）。

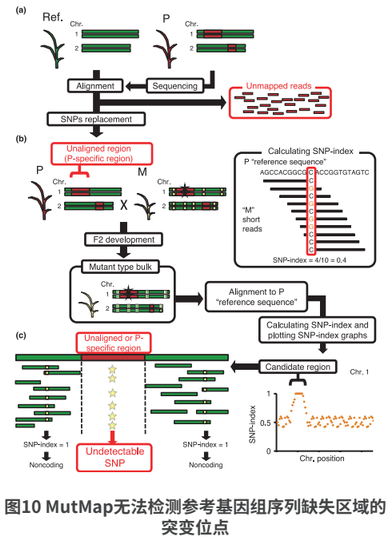


图10 MutMap无法检测参考基因组序列缺失区域的突变位点

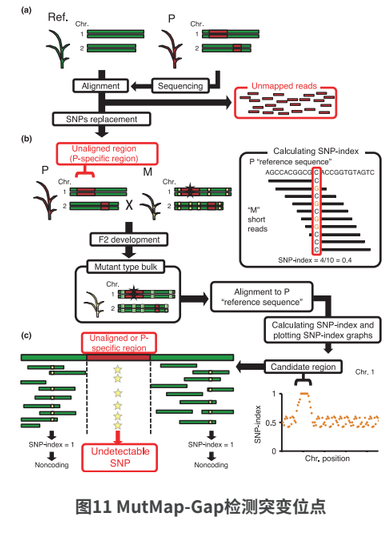


图11 MutMap-Gap检测突变位点

3.2 案例二：SHOREmap检测拟南芥点突变[14]

EMS诱导哥伦比亚（*Col-0*）拟南芥突变导致未知基因受损引起生长缓慢且叶子颜色变浅，该突变体与*Ler-1*杂交衍生F_2群体，混合提取500株有突变表型的水稻DNA进行全基因组重测序。将测序的reads比对回*Col-0*参考序列上，SHORE软件筛选纯和的SNP，与82,127个*Col-0/Ler-1*间高质量的SNP以及1,219处已知参考序列错误集输入SHOREmapINTERVAL软件中，以大小为3Mb作为滑动窗口对突变体的排列reads进行扫描，排除*Col-0/Ler-1*间高质量的SNP以及1,219处已知参考序列错误集的情况下，检测候选基因位点。对候选基因位点进行研究，联系基因注释信息分析突变位点发生在基因内还是基因间，检测到一个突变位点引起基因*AT4G35090*中丝氨酸转变为天冬酰胺。

研究策略：点突变亲本构建的F_2群体，混合提取500株有突变表型的水稻DNA进行全基因组重测序22X。利用SHOREmap检测影响表型变化的点突变。

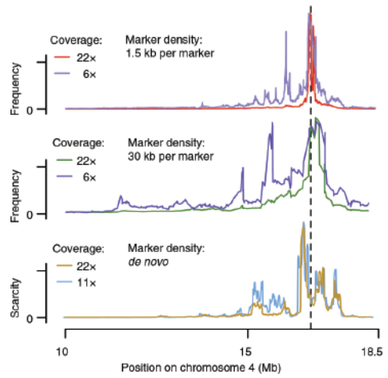


图12 SHOREmap检测点突变

3.3 案例三：线虫回交群体鉴定点突变位点[15]

携带有*a cog-1::gfp*基因的N2线虫PS3662经EMS诱变获得的缺失PDA的突变体*fp6*、*fp12*和*fp9*，*fp6*和*fp12*回交六代获得BC_6，*fp9*回交四代获得BC_4。每个回交群体中筛选3个突变型植株混合全基因组重测序。以野生型N2的基因组序列作为参考序列进行比对，去除有两个突变体共有的SNP位点，检测每个突变体在染色体上高密度突变的peak区；分别在*fp6*突变体Chr III上检测到一个4.29Mb；*fp9*突变体ChrX上检测到一个7.11Mb；*fq12*突变体的ChrX上检测到一个1.28Mb的peak区。进一步分析，在*fp6*、*fp9*和*fp12*分别发现在突变体的定位区间发现有6、10和4个候选突变位点与功能相关。

研究策略：多代回交群体，选择突变表型混合样本进行全基因组重测序，平均测序深度52.2-55.3X。与野生型亲本比对去除背景，去除突变体共有位点，多次过滤检测引起表型变化的突变位点。

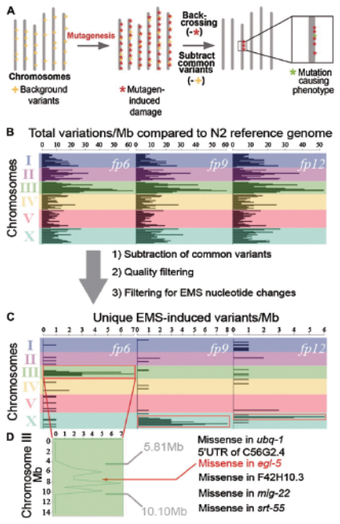


图13 回交后测序信息分析流程

可能存在的风险

进行点突变检测，由于物种和材料的区别，使得最终获得突变位点的数据量有所差别。如果构建了作图群体，能有效的减少最终检测到的SNP位点数，同时能与表型进行关联，便于后期验证。

常见问题

1、点突变检测的方案能用于定位数量性状吗？

答：上述中的方案3，利用双亲间的多态性位点进行选择基因型分析可以用于定位数量性状。

2、混池的样本数选择多少为好？

答：群体大小、标记密度和选择比例都会影响到选择基因型分析的检测结果。如下图所示，群体大小和标记密度对效应值大的QTL影响略小，对微效QTL影响很大。如果想要检测到微效的QTL，要求群体大，标记密。根据表型分布进行双尾选择，每尾选择10%-25%定位效果相对要好一些，如果群体比较大（如1000），每尾选择比例可以下降到5%。要求混池中表型鉴定准确。对于点突变检测，如果是隐形纯合的单基因突变，推荐混合样本在20个以上。

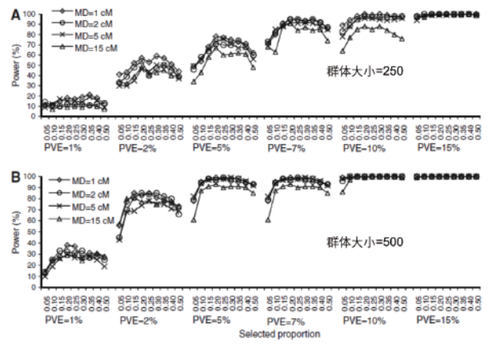


图14 检测功效与表型极端样本选择比例、标记密度关系[16]

华大优势

项目经验丰富：文案中介绍的BSA、MutMap和突变个体比较的点突变项目均有经验，交付结果客户满意；

根据客户需求可以提供个性化分析内容；

测序平台多样，选择空间大，能满足不同需求；

质控严格：从样本接收到数据交付都有严格的质量控制流程，保证数据准确性；

提供不同类型产品服务，可一站式完成您的需求。

参考文献

[1] Colbert T, Till B J, Tompa R, et al. High-throughput screening for induced point mutations [J]. Plant physiology, 2001, 126(2): 480-484.

[2] Sarin S, Prabhu S, O'Meara M M, et al. Caenorhabditis elegans mutant allele identification by whole-genome sequencing[J]. Nature methods, 2008, 5(10): 865.

[3] Doitsidou M, Poole R J, Sarin S, et al. C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy[J]. PloS one, 2010, 5(11): e15435.

[4] Blumenstiel J P, Noll A C, Griffiths J A, et al. Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing[J]. Genetics, 2009, 182(1): 25-32.

[5] Smith D R, Quinlan A R, Peckham H E, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies [J]. Genome research, 2008, 18(10): 1638-1642.

[6] Irvine D V, Goto D B, Vaughn M W, et al. Mapping epigenetic mutations in fission yeast using whole-genome next-generation sequencing[J]. Genome research, 2009, 19(6): 1077-1083.

[7] Schneeberger K, Ossowski S, Lanz C, et al. SHOREmap: simultaneous mapping and mutation identification by deep sequencing [J]. Nature Methods, 2009, 6(8): 550-551.

[8] Ashelford K, Eriksson M E, Allen C M, et al. Full genome re-sequencing reveals a novel circadian clock mutation in Arabidopsis [J]. Genome biology, 2011, 12(3): 1.

[9] Michelmore R W, Paran I, Kesseli R V. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations [J]. Proceedings of the National Academy of Sciences, 1991, 88(21): 9828-9832.

[10] Gallais A, Moreau L, Charcosset A. Detection of marker–QTL associations by studying change in marker frequencies with selection [J]. Theoretical and Applied Genetics, 2007, 114(4): 669-681.

[11] Abe A, Kosugi S, Yoshida K, et al. Genome sequencing reveals agronomically important loci in rice using MutMap [J]. Nature biotechnology, 2012, 30(2): 174-178.

[12] Fekih R, Takagi H, Tamiru M, et al. MutMap+: genetic mapping and mutant identification without crossing in rice[J]. PloS one, 2013, 8(7): e68529.

[13] Takagi H, Uemura A, Yaegashi H, et al. MutMap‐Gap: whole‐genome resequencing of mutant F 2 progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene Pii[J]. New Phytologist, 2013, 200(1): 276-283.

[14] Schneeberger K, Ossowski S, Lanz C, et al. SHOREmap: simultaneous mapping and mutation identification by deep sequencing [J]. Nature Methods, 2009, 6(8): 550-551.

[15] Zuryn S, Le Gras S, Jamet K, et al. A strategy for direct mapping and identification of mutations by whole-genome sequencing [J]. Genetics, 2010, 186(1): 427-430

[16] Sun Y, Wang J, Crouch J H, et al. Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement [J]. Molecular Breeding, 2010, 26(3): 493-511.

基于高通量测序的作图群体BSA分析定位QTL方案

研究背景

选择基因型（selective genotyping）分析是在表型鉴定的基础上，只对群体中表型值最高和最低的双尾或单尾群体进行基因型鉴定，在假设控制某数量性状的QTL与标记连锁，即QTL的位置和标记的位置完全重合的情况下，选择会引起控制性状的基因频率发生变化，对于和性状无关联的基因，其频率选择前后将保持不变。选择基因型分析就是利用这一特性，通过检测标记位点上等位基因频率的差异来检验这个标记附近是否存在控制性状的QTL。如图1所示，在假定标记有两个等位基因，并且在两个亲本间存在多态性的情况下，进行双尾选择基因型分析时，如果上尾与下尾间某标记的等位基因频率差异显著，则认为该标记与QTL连锁，否则该标记与影响目标性状的基因没有关系。

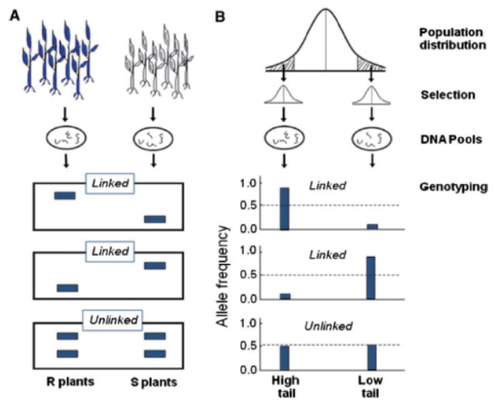


图1 双尾选择基因型分析[1]

Lebowitz等[2]和Zhang等[3]的检测方法中假设两尾群体中等位基因频率的变化是对称的，同时假设两尾具有相同的方差，那么选择基因型分析可直接利用t-统计量检验两尾群体中某标记的等位基因频率是否差异显著（双尾测验），如果显著则认为这个标记是一个QTL。

混合分组分析法（Bulked Segregant Analysis，BSA）[4]及选择DNA池法（selective DNA pooling）[5]是一种简单的选择基因型分析方法，它是将目标性状在F_2或BC子代中表型极端的高、低两组个体的DNA分别混合成两个DNA池，然后利用分子标记在两池中进行标记与性状间的共分离分析检测QTL。目前，运用BSA法已对各种不同作物的育性基因、抗性基因、生理基因及形态基因进行定位[6]。BSA分析在过去的研究过程中属于选择基因型中的一种极端情况，将两个表型极端的个体混合成两个DNA pool如抗池和感池，那么就要求抗池中只包含抗病等位基因，感池中只包含感病等位基因。但是测序技术的出现，在我们假设混池中每个个体被测到的概率相等的情况下，可以近似的利用测序reads计算等位基因频率，此时分析方法等同于选择基因型分析。

群体大小、标记密度和选择比例都会影响到选择基因型分析的检测结果。如图2所示，群体大小和标记密度对效应值大的QTL影响略小，对微效QTL影响很大。如果想要检测到微效的QTL，要求群体大，标记密。同时双尾选择，每尾选择10%-25%定位效果相对要好一些，但如果群体比较大（如1000），每尾选择比例可以下降到5%。

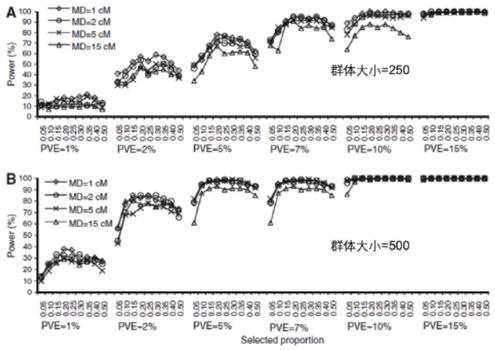


图2 群体大小、标记密度、选择比例对作图功效的影响[1]

横坐标代表不同的选择比例；PVE代表QTL效应值大小及能够解释表型变异的比例；纵坐标为检测功效，即定位多次能够检测到该QTL的占比；MD代表标记密度。

基于高通量测序的BSA分析，在亲本和子代的极端表型DNA混池中，实现了全基因组水平的SNP标记高密度扫描。BSA分型适合各种基因组特点、各种群体类型（F_2、BC、RILs、DH）和各种性状（质量和数量性状均可），是目前比较高效的基因初定位方法。目前已经有利用全基因组测序进行BSA分析定位QTL已经有多个成功案例[7-10]。BSA法大大减少了基因型鉴定的费用和耗时，但是对表型鉴定的准确性要求比较高。同时BSA也有一定的局限性，如一次只能针对一个性状进行分析；仅适用于质量性状或有明显主效基因的数量性状。

基于PCR扩增电泳跑胶基础的标记和高通量测序层面进行BSA分析的区别：

1）基于PCR扩增电泳跑胶基础上的标记，在混合池中如果AA这种基因型的DNA量比较多，其他基因型如Aa、aa的DNA量比较少，那么Aa、aa这些基因型很可能无法检测到。除非每个样本单独进行基因型鉴定，然后再统计某个极端池中某种基因型出现的频率。所以基于PCR扩增电泳跑胶基础的标记进行BSA分析，是比较双亲间存在多态性的标记在极端池中的有无。而高通量测序是比较敏感的基因分型方式，即使DNA含量较低，也能被检测出来，所以可以利用测序reads统计等位基因频率，但是对于杂合位点无法确定是样本混合造成的还是本来就是杂合位点。

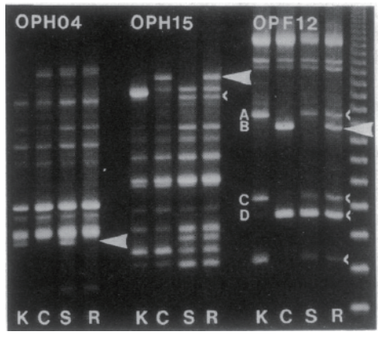


图3 在混合样本间利用RAPD标记检测结果
K和C分别代表亲本；S是易感病混合池；R是抗病混合池[4]。

2）如果表型鉴定不准确，特别是数量性状受环境影响大，很容易混入不同基因型的样本；混合样本测序，由于建库和测序的不可控因素，会出现数据的不均一情况，即每个样本测序的数据量是不同的，这样可能会放大某些基因型出现的频率。而混合测序基因频率在0-1变化，只是通过频率数据无法判断是低频基因型或者是测序错误亦或是比对错误导致的。另外基于高通量测序进行基因分型开发SNP标记，标记密度大，若错误的将其他基因型样本混入对结果的检测影响大，因为高密度的SNP噪音大。

3）有参考基因组的物种，可以直接测序进行BSA分析，无需构建遗传图谱。

方案设计

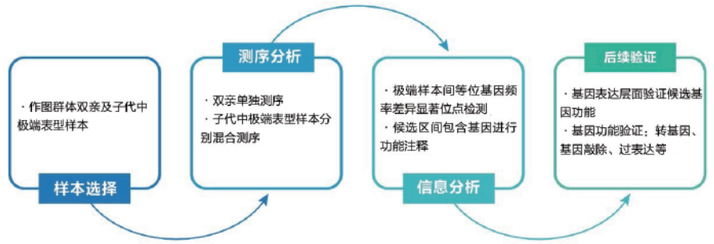


图4 BSA分析方案设计

2.1 样本建议

1）选择动植物作图群体如F_2、BC、DH、RIL、NIL等群体类型，不建议F_1群体进行基因高通量测序的BSA分析（F_1需要选择亲本之一为杂合，另一个亲本为纯合的标记进行QTL定位，在混合样本中计算等位基因频率时，杂合位点无法判断是混合导致的还是个体样本真实存在的杂合，同时混合样本测序等位基因频率变化范围为0-1，判断是否杂合出现误差的几率大于纯合型）；

2）建议选择双亲+子代极端表型混合样本进行测序，其中双亲可以各选择一个个体进行测序，子代极端表型样本两尾选择（如抗病表型，选择极抗和易感）20个分别进行DNA提取，然后等量混合进行建库测序，所以最终相当于构建4个文库进行测序。如果选择混合提取DNA样本，建议表型鉴定准确的极端表型样本数提高如50个以上（要注意是在子代样本足够多的情况下，否则混合样本数量多，可能引入非表型极端样本）。当然子代极端样本也可以单尾选择（如抗病表型，只选择极抗）进行单尾测验进行BSA分析，但是检测效果不如双尾选择；

3）极端样品选择的前提是表型鉴定准确，由于表型鉴定引入的偏差通过提高混合样本数可以有效降低，但是要保证群体足够大，能选择到适宜的大的样本（极端样本选择比例推荐10%-25%）。

2.2 实验技术

采用全基因组重测序，检测SNP，并利用reads数统计等位基因频率（SNP-index），通过计算表型极端池间或是与期望值间的差异程度，判断对应SNP是否与目标性状有关联。

2.3 测序参数

全基因组重测序，极端池样本≥30X；

2.4 分析结果

2.4.1 SNP检测及基因分型

有参考基因组物种，与参考基因组比对检测SNP等变异信息。检测双亲间存在多态性的SNP位点，并根据亲本基因型判断子代基因型来源。

2.4.2 SNP-index计算

对于多个样本pooling建库测序，我们假设每个样本被测到的概率相等，那么我们可以利用测序的reads来统计样本间的等位基因频率。我们利用SNP作为标记统计等位基因频率也叫SNP-index。将每一个标记位点支持亲本1的reads深度除以该位点总的reads深度，获得所有位点亲本1的reads的频率index（SNP-index）；同样也将每一个标记位点支持亲本2的reads深度除以该位点总的reads深度，获得所有位点亲本2 reads的频率index（SNP-index）。根据遗传连锁可以推测，如果该Marker与目标性状相关，则两个群体中的SNP-index都会显著偏离期望值如F2群体中为0.5（隐形纯合池中的SNP-index会接近于1），Δ(SNP-index)（极端池间SNP-index的差值）会显著偏离0。该Marker与目标性状越关联，Δ(SNP-index)偏离0的程度越大，即更接近1。根据SNP-index和Δ(SNP-index)可以检测与目标性状相关的候选QTL位点。

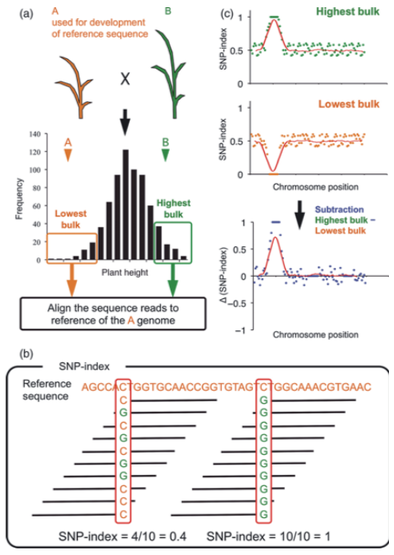


图5 表型极端池SNP-index示例[11]

2.4.3 定位区间包含的基因进行功能注释

有参考基因组物种，根据基因组的注释信息，可以分析获得定位区间包含哪些基因，并针对这些基因进行功能注释。如GO功能注释，GO分为分子功能、细胞组分和生物过程三大功能类。同时可以对这些基因参与的代谢途径进行注释。如Pathway分析，KEGG（Kyoto Encyclopedia of Genes and Genomes）是有关Pathway的主要公共数据库，该数据库整合了基因组、化学以及系统功能信息，特别是基因与细胞、生物体以及生态环境的系统性功能相关联。

2.5 项目周期

样品检测合格后，建库+测序+标准信息分析：约50个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.6 预期结果

借助高通量分型平台，快速、高效、低成本的检测目标性状相关的主效QTL，并针对候选区域包含的基因进行功能注释。

2.7 辅助研究策略

可以通过RNA测序研究QTL定位区间内的基因，在目标性状上存在显著差异的样本间是否表达量不同，亦可以检测对应蛋白的表达情况，进一步缩小范围，找到目标基因。

2.8 后期验证手段

分析得到的候选基因，可以利用转基因、基因敲除、基因沉默（RNAi）等方式验证基因功能。

应用案例

3.1 案例一：QTL-seq检测水稻抗稻瘟病和苗活力QTL[11]

水稻稻瘟病抗性品系*Nortai*和*Hitomebore*杂交衍生RIL群体，对241株RIL群体抗性鉴定，20个抗性和20个易感性植株分别混合测序>6.88x，*Nortai*和*Hitomebore*间检测到161,563个SNPs，利用QTL-seq检测这些SNP在混合样本中的SNP-index值，在6号染色体2.39-4.39Mb检测到混合样本间index值差异显著，与抗性相关（图6）。利用芯片分型进行连锁分析验证了此结果。同时针对水稻苗活力强的品系*Dunghan Shali*和*Hitomebore*杂交衍生F_2群体，531株F_2群体的种子吸水14天后测苗高，50个最高和50个最低植株分别混合测序>19x，利用QTL-seq（作者开发的分析软件）检测到两个SNP-index值差异显著的峰（图7），其中很可能包含了之前在RIL群体中利用连锁分析检测到的相关基因*OsGA20ox1*。利用QTL-seq进行BSA分析在F_2中只能检测主效QTL，在RIL、DH大群体中也能检测微效QTL。

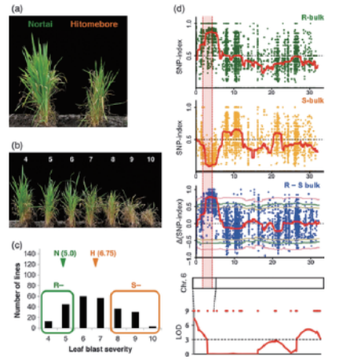


图6 水稻抗稻瘟病QTL-seq检测结果

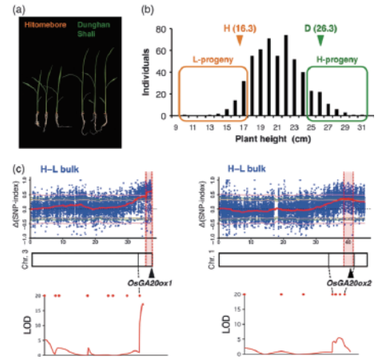


图7 水稻苗活力QTL-seq检测结果

3.2 案例二：BSA分析定位番茄灰叶斑病抗性基因*Sm*[12]

马铃薯灰叶斑病由匍柄霉引起，是栽培番茄主要的病害之一。在野生番茄*S. pimpinellifolium.*中检测到的*Sm*基因对该病有抗性、该基因被定位到11号染色体上。本研究利用重测序数据进行BSA分析定位*Sm*基因。本研究中在双亲间检测到50,968差异标记，其中46,941位于11号染色体。11号染色体上差异区间包含了37个基因，进一步利用F_2S-pool基因型，BSA分析缩小候选区间范围。最终有8个SNP的SNP index值接近0.33，候选区间缩短到0.26Mb，此区间包含37个基因。利用qRT-PCR检测双亲间候选基因的表达情况，发现Solyc11g011870.1.1 和Solyc11g011880.1.1与抗性相关，这两个基因在接种前低水平表达，在接种后5天表达水平快速增长。对抗性植株和易感植株这些基因的PCR产物进行测序，在抗性植株Solyc11g011880.1.1突变体中发现一个标记D5与抗性基因共分离。

研究策略：BSA测序策略：（P1：40.51×，P2：40.87 ×，F_2 R-pool: 52.17×，F_2S-pool: 55.89×），qRT-PCR取样：P1和P2在3-4真叶期接种，分别在0、3、5、8天后采取嫩叶提取RNA。

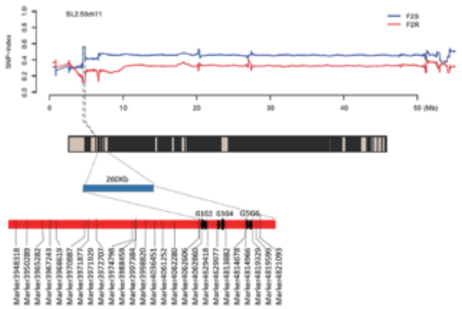


图8 BSA分析结果

可能存在的风险

由于表型鉴定的准确度、测序的均一性、测序错误和比对错误等问题，导致检测结果出现假阳性。因为表型鉴定不准确，引入错误的基因型信息；混合样本测序，各个样本的测序均一性不同，导致计算的等位基因频率与真实值间不同；抑或因为混合的样本少，那么错误的基因型信息和测序不均一性问题会被放大，而导致假阳性的出现；测序错误的存在使得低频基因型无法准确鉴别；另外由于参考基因组重复序列的存在，比对也可能出现错误而最终导致假阳性。

常见问题

1、简化基因组测序和全基因组重测序进行BSA分析的差异？

答：简化基因组是只针对酶切位点相关的片段进行测序，检测到的标记位点相对全基因组重测序稀疏。如果在作图群体比较大的情况下，重组事件比较多，简化基因组测序检测到的标记可能会漏掉某些交换信息。

2、高通量测序进行BSA分析是不是必须要测亲本呢？

答：对于作图群体进行QTL定位我们主要利用的是子代的重组信息，所以只有双亲间存在多态性的位点才是有意义的。如果有亲本基因型信息的情况下，我们可以首先筛选双亲间存在多态性的标记进行后续的BSA分析，不仅降低了计算量同时提高了准确度。如果不测双亲，那么就存在这样的可能：一个标记在混池中是存在多态性的，但是双亲间是没有多态性，例如双亲都是Aa的杂合位点，这样的标记会对我们的定位产生不良影响，提高假阳性。但是如果有测亲本的话，就可以将此类标记过滤掉。

华大优势

项目经验丰富：文案中介绍的 BSA项目无论是全基因组重测序或是简化基因组测序均有丰富的项目经验并发表文章，交付结果客户满意；

根据客户需求可以提供个性化分析内容；

测序平台多样，选择空间大，能满足不同需求；

质控严格：从样本接收到数据交付都有严格的质量控制流程，保证数据准确性；

提供不同类型产品服务，可一站式完成您的需求。

参考文献

[1] Sun Y, Wang J, Crouch J H, et al. Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement[J]. Molecular Breeding, 2010, 26(3): 493-511.

[2] Lebowitz R J, Soller M, Beckmann J S. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines[J]. Theoretical and Applied Genetics, 1987, 73(4): 556-562.

[3] Zhang L P, Lin G Y, Nino-Liu D, et al. Mapping QTLs conferring early blight (Alternaria solani) resistance in a Lycopersicon esculentum× L. hirsutum cross by selective genotyping[J]. Molecular Breeding, 2003, 12(1): 3-19.

[4] Michelmore R W, Paran I, Kesseli R V. Identification of makers linked disease-resistance genes by bulked segregant regions by using segregating populations [J]. Proceedings of the National Academy of Science, 1991, 88: 9828-9832.

[5] Gallais A, Moreau L, Charcosset A. Detection of marker–QTL associations by studying change in marker frequencies with selection[J]. Theoretical and Applied Genetics, 2007, 114(4): 669-681.

[6] 张小明, 李春寿. 混合分组分析法在作物基因定位上的研究进展 (综述)[J]. 上海农业学报, 2002, 18(3): 24-27.

[7] Ehrenreich I M, Torabi N, Jia Y, et al. Dissection of genetically complex traits with extremely large pools of yeast segregants[J]. Nature, 2010, 464(7291): 1039-1042.

[8] Parts L, Cubillos F A, Warringer J, et al. Revealing the genetic structure of a trait by sequencing a population under selection[J]. Genome research, 2011, 21(7): 1131-1138.

[9] Swinnen S, Schaerlaekens K, Pais T, et al. Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis[J]. Genome research, 2012, 22(5): 975-984.

[10] Swinnen S, Thevelein J M, Nevoigt E. Genetic mapping of quantitative phenotypic traits in Saccharomyces cerevisiae [J]. FEMS yeast research, 2012, 12(2): 215-227.

[11] Takagi H, Abe A, Yoshida K, et al. QTL‐seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations[J]. The Plant Journal, 2013, 74(1): 174-183.

[12] Yang H, Zhao T, Jiang J, et al. Mapping and screening of the tomato Stemphylium lycopersici resistance gene, Sm, based on bulked segregant analysis in combination with genome resequencing[J]. BMC plant biology, 2017, 17(1): 266.

动植物不同器官发育基因挖掘研究方案

074

研究背景

纵观生物的发育过程，总结起来就是基因的选择性的表达，即发育前后的细胞拥有相同的基因组的信息，但是在分化的过程中选择了特异的基因进行表达。生物器官的发育具有严格的时间控制和空间的次序性，是遗传信息按照特定的时间和空间表达的结果，是生物体基因型与内外环境因子相互作用，并逐步转化为表型的过程。目前，人们对真核生物的基因组的结构和基因表达的调控已有一定的了解。近十年来，转录组测序技术，在动植物器官发育研究领域发挥了重要的作用，是目前生命科学领域的研究重点之一。

方案设计

图1 RNA测序研究发育机制方案设计

2.1 样本建议

1) 针对动植物的不同器官的样本，建议选择目标发育过程中的多个关键的器官，例如根、茎、叶、花、果实等；

2) 至少2个生物学重复，3个以上的生物学重复更好；

3) 对于无参考序列的物种，需要对所有样本的转录组测序结果进行拼接，从而得到参考序列，然后作为基因表达定量的参考序列；对于有参考序列的物种，以基因组序列为基因表达定量的参考序列。

4)转录组测序和RNA-Seq，植物总RNA≥1μg，样品浓度40-1000ng/μL，RIN≥6.0，28S/18S≥1.0；动物总RNA≥1μg，样品浓度40 -1000 ng/μL，RIN≥7.0，28S/18S≥1.0，昆虫无RIN和28S/18S要求。

2.2 实验技术

采用转录组测序或者RNA-Seq测序技术，通过样本间的比较及筛选，寻找差异表达的基因，并对差异表达的基因进行GO和Pathway的富集分析等。

2.3 测序参数

建议转录组项目每个样本6Gb，RNA-Seq项目每个样本20Mb Reads。

2.4 分析结果

2.4.1 Unigene功能注释

对于无参考序列的物种，需要对多个组织（混合样本或是分别测序的样本）的转录组测序下机数据进行组装，然后对组装得到的Unigene进行七大功能数据库注释（NR、NT、GO、KOG、Pfam、KEGG 和 SwissPprot），并且用维恩图来展示注释结果。

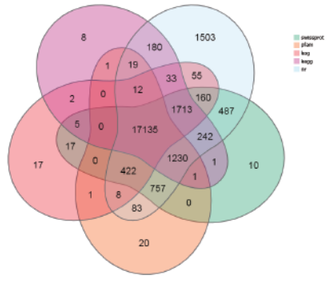


图2 NR、KOG、KEGG、Pfam、以及 SwissProt功能注释维恩图

2.4.2 差异表达基因分析

根据各个样品基因表达水平数据，我们可以检测样品（或者样品组）之间的差异表达基因。对于设置生物学重复的实验，我们可以采用DEGseq、DEseq2、EBseq、NOIseq进行组间样品基因差异表达分析，从而比较处理组与对照组，得到上下调基因个数。对于无生物学重复样品，则采用PossionDis进行基因差异表达分析。差异表达基因数量、比较组之间共有和特有的差异基因、及差异表达基因层次聚类都可以用不同形式的图片进行展示。

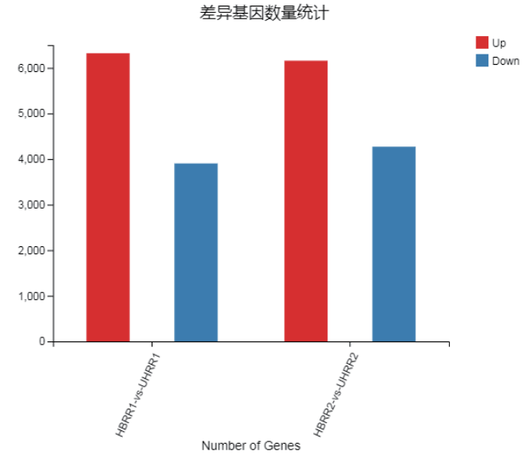


图3 差异表达基因数量统计图

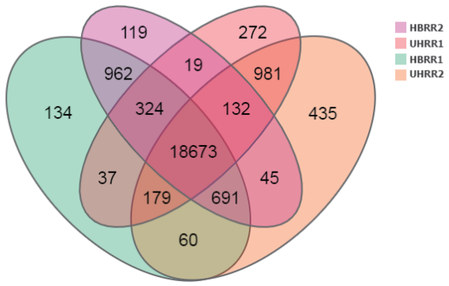


图4 样本特异表达及共有表达基因Venn图

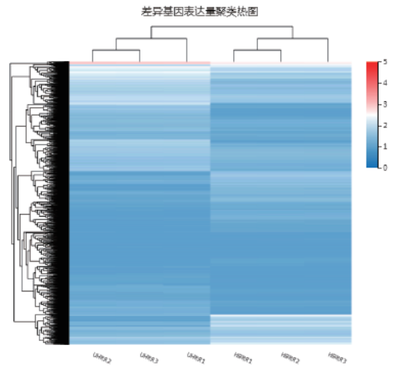


图5 差异表达基因层次聚类热图

2.4.3差异表达基因GO分析

根据差异基因检测结果，我们对其进行GO功能分类以及富集分析。GO分为分子功能、细胞组分和生物过程三大功能类，我们将对三大功能类单独进行进一步的分类以及富集分析。差异基因GO功能分类与富集结果见图6。

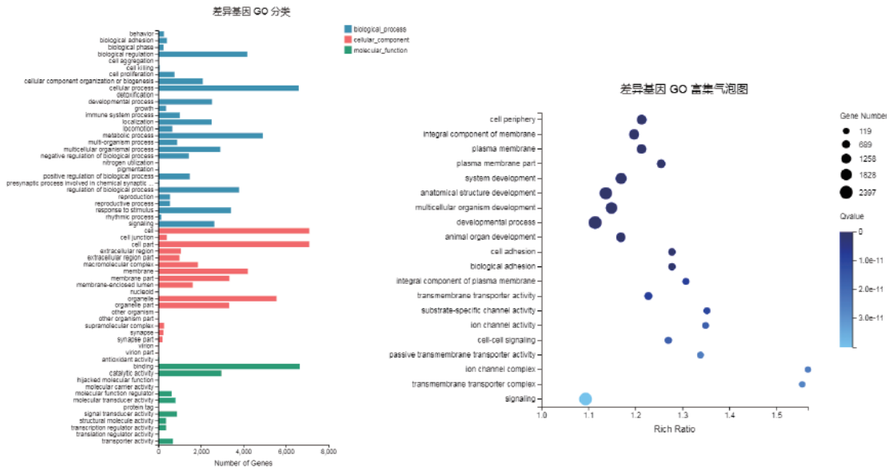


图6 差异基因GO功能分类与富集结果

2.4.4差异表达基因Pathway分析

KEGG（Kyoto Encyclopedia of Genes and Genomes）是有关Pathway的主要公共数据库，该数据库整合了基因组、化学以及系统功能信息，特别是测序得到的基因集与细胞、生物体以及生态环境的系统性功能相关联。根据差异基因检测结果，我们对其进行KEGG生物通路分类以及富集分析。差异基因Pathway分类与富集结果见图7。

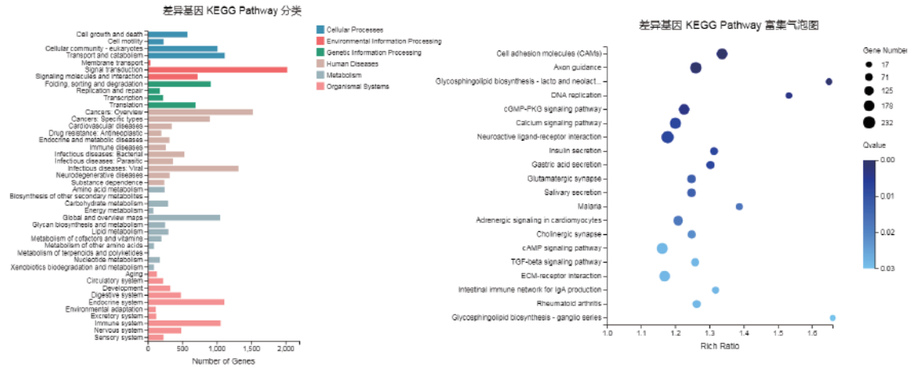


图7 差异基因Pathway分类与富集结果

2.5 项目周期

样品检测合格后，建库+测序+标准信息分析：转录组约24个工作日，RNA-Seq约15个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.6 预期结果

利用RNA研究手段，借助高通量测序平台，通过对不同发育时期的器官进行比较，从基因表达层面对器官发育的生理机制进行深入挖掘。

2.7 后期验证手段

挑选目标或者候选基因，通过qPCR进行验证，看测序结果是否和qPCR结果一致。

应用案例

3.1 案例一：转录组测序研究艾草和萜类化合物生物合成（华大参与）[1]

文章中艾草采集自安徽中医药大学中药园，然后对艾草的不同组织（叶片、根、茎）分别进行转录组测序。三个样本混合一起组装得到99,807个unigenes，N50长度 1456 bp，其中67,446个unigenes得到公共数据库注释。因为叶片中合成萜类化合物生物有重要药用价值，因此文章着重比较叶片和其他两种组织的差异表达基因，发现叶片中的许多特异表达或者上调表达的基因，另外发现了很多与萜类化合物生物合成有关的酶或转录因子的编码基因特异表达。

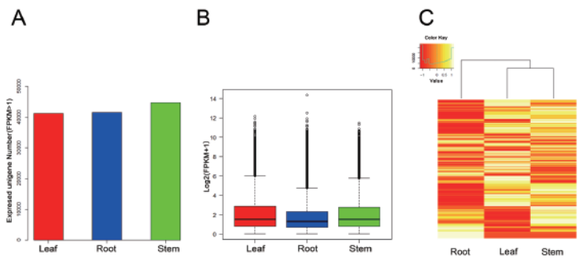


图8 不同组织unigene表达分析

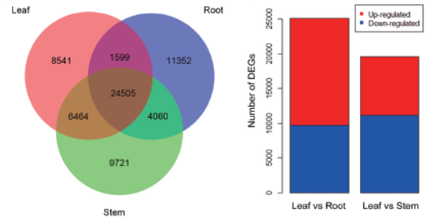


图9 不同组织unigene差异表达分析

表1 叶子中特异表达的萜类、聚酮化合物相关的通路的基因

Terpenoid and polyketide metabolic pathway	Pathway ID	Number of up-regulated genes	
		Leaf vs Root	Leaf vs Stem
Terpenoid backbone biosynthesis	Ko00900	51	17
Limonene and pinene degradation	Ko00903	54	47
Carotenoid biosynthesis	Ko00906	70	35
Diterpenoid biosynthesis	Ko00904	25	9
Zeatin biosynthesis	Ko00908	17	10
Sesquiterpenoid and triterpenoid biosynthesis	Ko00909	16	16
Monoterpenoid biosynthesis	Ko00902	1	11
Brassinosteroid biosynthesis	Ko00905	17	3

3.2 案例二：转录组测序研究大鼠转录组图谱[2]

文章研究了11种器官、4个发育阶段（幼年、青春期、成年及老年）、2种不同性别的320个大鼠样本。每个条件设置雄雌各4个个体，消除生物学变异（实际共用了32只大鼠）。样本利用SE50测序，约40M reads/样本。大量的转录本分析获得了器官特异性、年龄依赖性或性别特异性的差异表达模式，并构建一个与其他广泛应用的数据库交联、可通过网络开放获取的大鼠BodyMap的数据库。

图10 器官特异性表达基因

图中显示两两器官差异表达基因的数量，橘色为过表达，绿色为抑制表达。A：是被比较的器官；B：11个器官分别和A器官相比；C-F：在四个时期差异表达基因的数量，比其他器官多（橘色）或少（绿色）。发现一些器官富集、差异性表达的基因反映出已知的器官特异性生物功能。

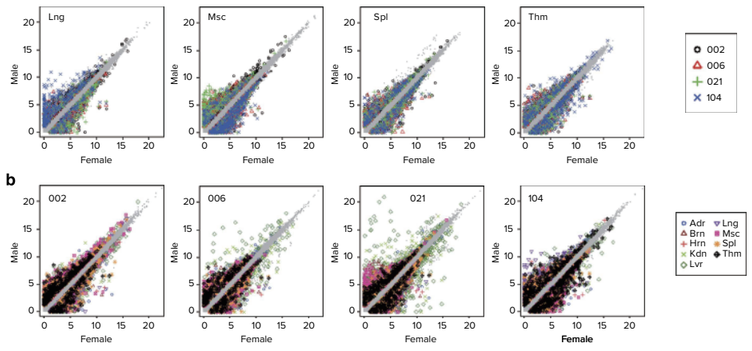


图11 性别特异性表达基因

雌雄之间在不同器官的、在不同年龄的差异表达基因。

3.3 案例三：单细胞RNA测序揭示人和小鼠早期胚胎发育进程[3]

利用单细胞RNA测序技术，研究了人、鼠的33个早期胚胎细胞，包括人和小鼠成熟的卵细胞、原核融合细胞、受精卵、二细胞期、四细胞期和八细胞期的卵裂球细胞。全面分析了人和小鼠从卵细胞到桑椹胚发育期，基因表达模式的动态变化。分析25个共表达模式中，有9个能特异性地代表不同的发育阶段。

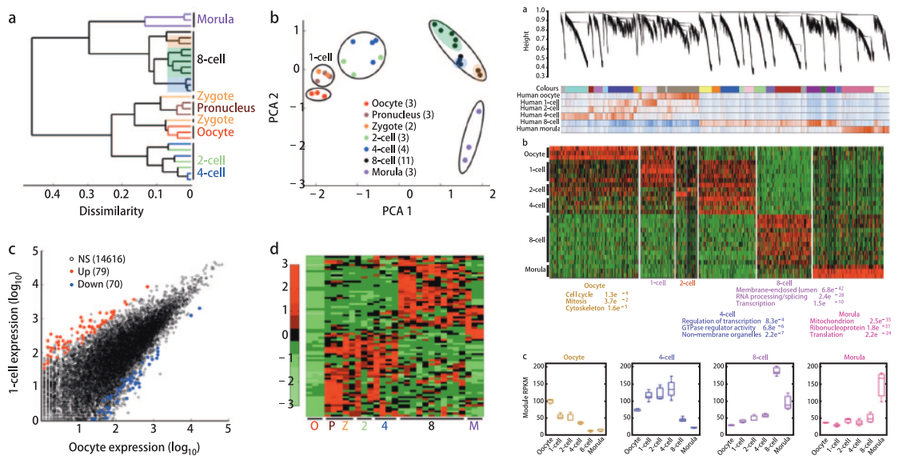


图12 人胚胎植入前各个发育阶段的表达基因分析和网络分析

3.4 案例四：赤霉素信号传导新机制调控水稻分蘖的氮素响应[4]

以半矮秆育种为代表的“绿色革命”极大地提高了作物产量，但也带来氮营养利用效率降低的问题，需要大量地氮肥投入，不仅破坏环境，而且违背了农业的可持续发展原则。

《Science》杂志以封面文章形式报道水稻赤霉素与氮素响应方面取得的重要进展。研究发现，NGR5是植物响应氮素的正调控因子，它与PRC2蛋白复合物互作，通过介导组蛋白H3K27me3甲基化修饰水平来调节靶基因的表达，进而调控植物生长发育（例如分蘖）对土壤氮素水平的响应。他们还发现NGR5是赤霉素信号传导途径的一个新的关键元件。赤霉素通过促进NGR5蛋白降解，导致表观遗传修饰降低，进而促进靶基因表达，实现赤霉素抑制植物分枝生长发育。因此，赤霉素信号传导新机制的发现从分子水平揭示了“绿色革命”品种在高肥条件下增产的原因。其中华大DNBSEQ平台再次提供了RNA-Seq和ChIP-Seq。

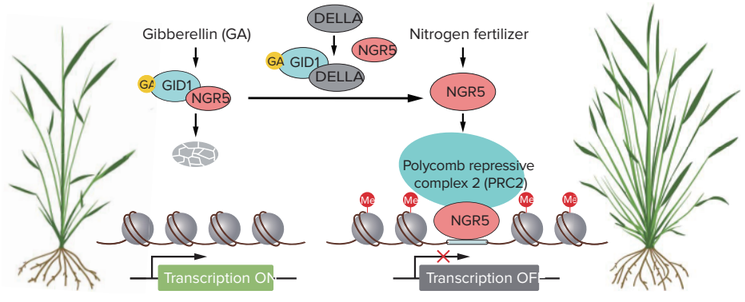


图13 赤霉素信号传导新机制调控水稻分蘖的氮素响应

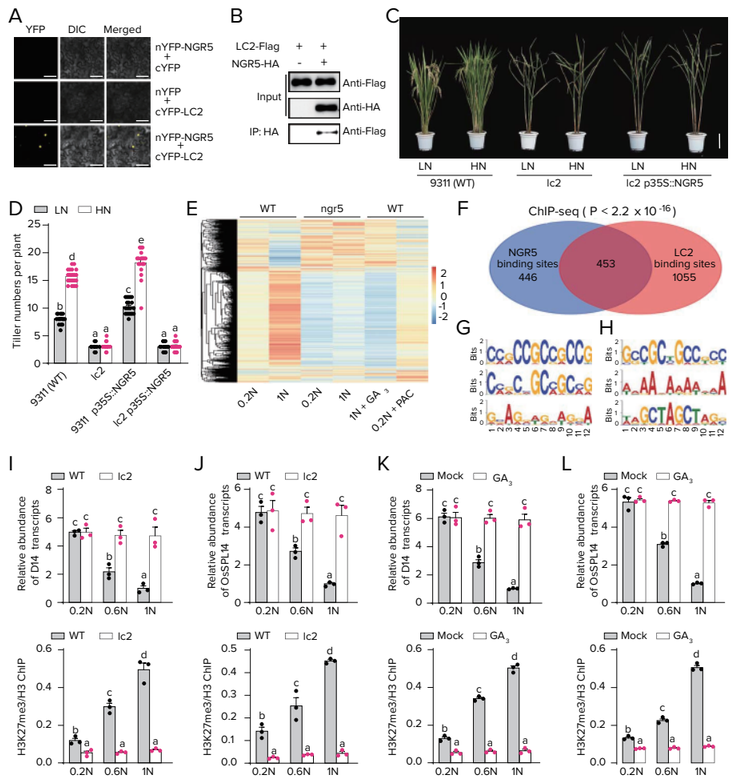


图14 氮素通过H3K27me3修饰调控分蘖

3.5 案例五：反刍动物的角的起源与鹿角再生[5]

Science以封面文章形式报道了华大等团队合作的反刍动物基因组计划的3篇系列研究论文。其中一篇为转录组测序研究角，反刍动物多种多样的角具有共同的基因、细胞和组织起源，其中一个控制角发育的关键基因在麝科和獐亚科趋同丢失功能。研究还发现鹿角的快速再生生长过程中招募了大量癌基因参与，其中多个抑癌基因特别是p53途径基因受到了强烈正选择，可能与鹿科动物癌症发生率很低有关。该研究为未来通过基因编辑手段培育无角牛、羊优良品种提供参考靶点，并能够为人类再生医学及癌症相关研究提供重要借鉴。

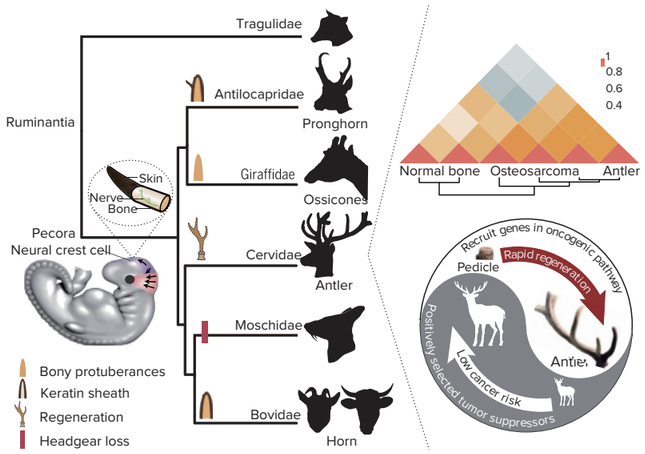


图15 反刍动物头饰神经嵴细胞起源及鹿角快速再生和低肿瘤风险的严格控制

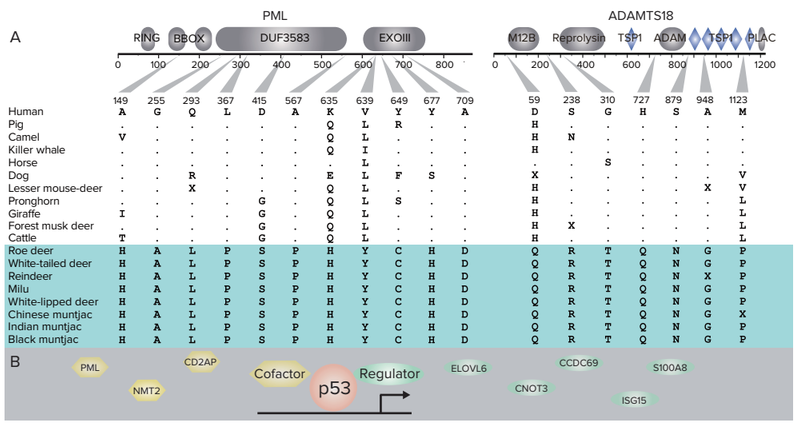


图16 鹿角肿瘤抑制基因正选择的例子

可能存在的风险

做完转录组/RNA-Seq以后，可能有部分基因测序的结果和qPCR的结果不一致，这是正常的，因为两种方法本身存在差异，特别是针对低表达量的基因。所以转录组测序建议关注两种基因表达的检测结果变化趋势总体是不是一致，具体表达量的值及差异倍数数据作为参考。如果变化趋势不一致，可能的需要考虑以下几点：

1）测序时所用样品同qPCR实验中所用是否为同一批材料且处理条件一致；

2）尽量选取表达量高的基因进行验证，同时差异倍数在5~10倍的基因更合适；

3）考虑qPCR的实验方案、引物序列及原始结果。比如设计探针是否考虑多转录本情况，转录组测序是对转录本。如该基因对应多个转录本，则可能有偏差；

4）该基因是否存在新的可变剪切。

常见问题

1、是否需要生物学重复?重复几次?

答：是的，至少需要2次生物学重复，3次以上的生物学重复更好。2011年7月Hansen[4]发表的文章表明生物学差异是基因自身表达的特性，与检测技术的选择以及数据处理的方式无关。如果不设生物学重复，高影响因子的杂志可能会因此而拒稿。

2、转录组测序一般推荐多大的数据量?

答：一般推荐6G的数据量，如果该物种的基因组较大，推荐8G-10G的数据量。

华大优势

高品质的自主测序平台：DNBSEQ测序仪是华大基因自主研发的高通量测序系统，提供多个高性价比测序服务，优质、稳定、高效，低dup无需人为干预，无index hopping风险。从2016年11月以来该平台连续发表多篇高水平文章，包括Nature、Science、Cell等顶级期刊；

样品起始量更低：常规建库，人鼠样品200ng起，其他物种样品1ug起，可微量定制化建库，低至200pg；可单细胞定制化建库，低至单个细胞；

全面的样品类型：提供单细胞、FFPE、微量等特殊样品服务，完全解决样品准备的后顾之忧。

更精准的分析结果：独创的插入片段文库，完美匹配PE150的长读长，转录本组装长度更长，可变剪接、基因融合鉴定更敏感更准确；

Dr. Tom系统解决个性化难题：无需生信分析基础，只需鼠标一键点击操作，随时随地即兴交互，玩转个性化分析，只需任意一种RNA测序数据，就可给您多组学关联信息，多数据库联合分析，多维度数据展示，循环挖掘数据，在成千上万的候选基因中轻松锁定解释生物学问题的核心基因；

极速交付：无需等待到样品检测合格，从提取和检测开始，到Dr.Tom个性化分析，每个环节都是极速交付。

参考文献

[1] Liu M, Zhu J, Wu S, et al. De novo assembly and analysis of the Artemisia argyi transcriptome and identification of genes involved in terpenoid biosynthesis. Scientific Reports. 2018 Apr 11;8(1):5824.

[2] Yu Y, Fuscoe J C, Zhao C, et al. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages [J]. Nature Communications, 2014, 5.

[3] Xue Z, Huang K, Cai C, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA [thinsp] sequencing[J]. Nature, 2013, 500(7464): 593-597.

[4] Wu K, Wang S, Song W, et al. Enhanced sustainable green revolution yield via nitrogen-responsive chromatin modulation in rice[J]. Science, 2020, 367(6478).

[5] Wang Y, Zhang C, Wang N, et al. Genetic basis of ruminant headgear and rapid antler regeneration[J]. Science, 2019, 364(6446): eaav6335.

[6] Hansen K D, Wu Z, Irizarry R A, et al. Sequencing technology does not eliminate biological variability[J]. Nature Biotechnology, 2011, 29(7): 572-573.

植物抗逆基因表达、调控研究方案

研究背景

1.1 抗逆研究背景

在自然界条件下，由于不同的地理位置和气候条件以及人类活动等多方面原因造成了各种不良环境，植物生长过程中，受到环境胁迫，超出了植物正常生长、发育所能忍受的范围，致使植物受到伤害甚至死亡，造成农业损失巨大。在此过程中，植物也会逐渐对不良环境产生适应性和抵抗力，俗称抗性。在植物生长发育阶段，会呈现不同程度的抗逆表型，适应环境变化。目前，培育抗性作物种质是解决农产品安全和环境恶化问题的重要资源的重要手段之一，因此，研究并提高植物抗逆能力，对于农业增长增收具有重要的理论意义和现实意义。

植物所受的胁迫环境因素包含物理、化学、生物三大类；物理胁迫包含干旱、水涝、高温、低温、强风、空气污染等；化学胁迫包含盐碱、生长元素匮乏或者过剩，杀虫剂或者除草剂等；生物胁迫包含真菌、病毒、虫害等；随着高通量测序（NGS）的发展，植物抗性研究更加深入，遂利用转录组、表达谱、非编码RNA高通量测序等手段，实现RNA水平全转录组表达调控研究，针对抗逆性状进行基因表达、抗逆形成机制、抗逆调控机制进行深入挖掘，以辅助抗逆农作物育种培育，助力农学研究和实际生产应用。

1.2 抗逆分子生物学研究进展

截止2019年12月，对NCBI中植物抗逆胁迫RNA分子生物学水平研究文献进行统计，总计1692篇，且逐年呈上升趋势。《2019研究前沿报告》农业、植物学和动物学中，植物抗逆研究依然被列为热点新兴前沿。且RNA表达调控水平，已拓展到全转录组研究方向。

图1 植物抗逆研究RNA相关发表文章统计

方案设计

2.1 样本设计方案

样本选择前，选择表型差异明显的，且有生理生化实验数据支持的样本，切勿盲目选择，以免后期文章撰写和实验验证中出现问题。

通过“R”-抗性品种和“S”-易感品系的转录组、表达谱、Small RNA及LncRNA测序等，进一步筛选、鉴定可能与“R”抗性相关的基因以及可能的代谢通路和关键酶以及调控靶基因关联分析等。

建议方案A

样品建议：易“S”群体个体（建议每个样品制备重复样品）。

实验方法：采集个体以及对照样品，可采用不同剂量逆毒侵染等，对发育不同时期进行跟踪记录，调查研究，并收集各时期样品，进行RNA提取，建库。

建议方案B

样品建议：“R”和“S”群体个体；

实验方法：对“R”和“S”品种，对同一时期不同个体进行跟踪记录，调查研究，并收集典型样品，进行RNA提取，建库；

方案A和B高通量测序：全转录组测序研究。注意样本选择时收集足够同时做LncRNA和Small RNA建库测序的RNA总量，如总量过低，一是有可能造成实验失败，二是有可能影响数据分析结果的准确性。

推荐两种测序策略方案

a）经济型模式-两种文库测序：LncRNA + Small RNA

大RNA研究：采用LncRNA测序模式，即采用去核糖体链特异性建库，进行PE100测序，10G clean data；实现lncRNA、mRNA、circRNA的鉴定、定量和功能分析等。

小RNA研究：采用富集small RNA片段建库方法，进行有方向性的SE50测序，20M clean reads，实现miRNA、siRNA、piRNA的鉴定、定量和功能分析等。

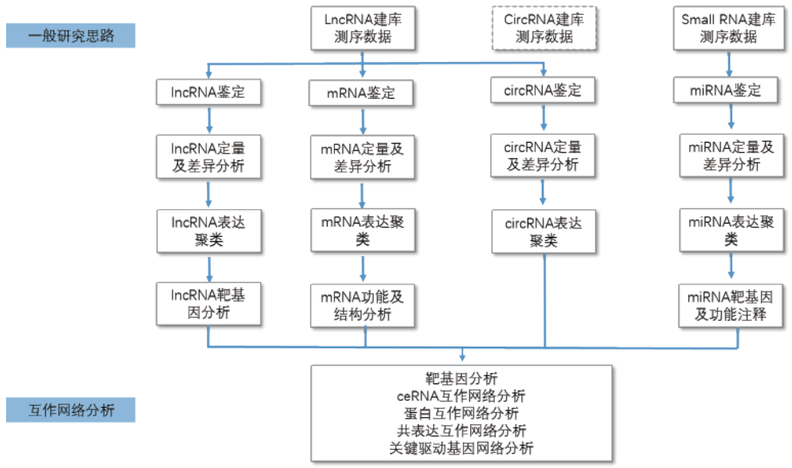
b）大容量模式-三种文库测序：LncRNA + Small RNA+ CircRNA

大RNA研究：采用LncRNA测序模式，即采用去核糖体链特异性建库，进行PE100测序，10G clean data；实现lncRNA、mRNA的鉴定、定量和功能分析等。

环状RNA研究：采用去除线性RNA，反向富集环状RNA建库方法，进行PE100测序，10G clean data，对环状RNA高深度测序，实现对低丰度环状RNA的鉴定、定量及功能分析。

小RNA研究：采用富集小RNA片段建库方法，进行有方向性的SE50测序，20M clean reads，实现miRNA、siRNA、piRNA的鉴定、定量和功能分析等。

2.2 详细分析方案



2.3 实验验证方案

功能验证：

a)定量验证：qPCR/茎环qPCR定量验证

b)体外验证：miRNA敲除/miRNA拮抗物/target protecter technology、荧光素酶标记等

c)体内验证：构建小鼠模型

总结常用验证方法，根据验证关系归结以下三类：

2.3.1 miRNA与mRNA作用关系验证：

a)敲除目的miRNA：Knockdown目的miRNA，qPCR验证靶向目的mRNA表达情况。

b)加入Antogomirs：通过导入化学合成的小分子miRNA mimics或拮抗miRNA 的antagomir，采用过表达或干涉(抑制、敲除)等方法观察靶mRNA及编码蛋白表达进行信号通路研究，是目前miRNA功能研究的常用方法。

c)Target protector technology：通过加入目标基因的保护物，Northern Blot检验mRNA表达是否受miRNA降解，进行反向验证。

d)荧光素酶报告系统检测方法：通过生物信息学分析获得候选的miRNA结合区域，将野生型和结合位点突变的3' -UTR序列克隆入商品化的萤光素酶报告载体，通过观察miRNA对发光强度的影响对结合位点加以验证。

e)其他miRNA定量验证方法：qPCR技术可以用来检测miRNA以及其靶mRNA，主要方法有茎环法和加尾法等。

2.3.2 lncRNA与miRNA相互关系验证：

a)荧光素酶报告系统检测方法：通过生物信息学分析获得候选的lncRNA结合区域，将野生型和结合位点突变的3' -UTR序列克隆入商品化的萤光素酶报告载体，通过观察lncRNA对发光强度的影响对结合位点加以验证。

b)生物素－亲合素系统（Biotin-Avidin pull down System）：通过pull down富集到的lncRNA进行qPCR定量验证miRNA是否特异结合lncRNA。

2.3.3 lncRNA与基因表达及表型关系验证：

a)采用siRNA干扰敲除方法：通过敲除目的lncRNA，观察表型行为进行验证。

b)其他lncRNA验证方法：可通过人工导入lncRNA，利用qPCR验证目的靶基因定量表达情况以及表型行为情况，进行验证实验[1]。

2.4 方案设计注意事项

样本选择方面：

前期表型及相应生化指标等测量要精准。

设置生物学重复。

数据分析及验证方面：

推荐UMI Small RNA测序，提高检测灵敏度。

验证尽量全面具体。

2.5 分析结果

2.5.1 差异表达分析

对不同发育时期的样品，进行样品间及重复样品间比较，通过差异分析和显著性差异筛选等，查找“R”和“S”相关基因以及lncRNA、miRNA表达差异。

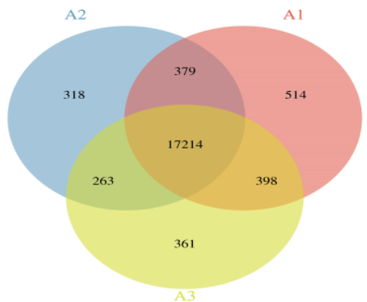


图2 条件特异表达基因维恩图

2.5.2 差异基因表达模式聚类分析

通过此项分析，可对各时期共有差异基因进行变化趋势统计分析，寻找“R”基因表达关键点等有效信息，为抗逆机制研究提供新的思路。

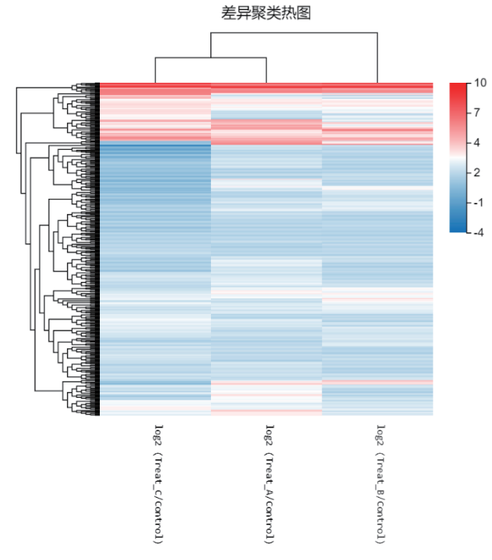


图3 差异表达基因聚类分析图

2.5.3 GO功能显著性富集分析

对已找到的各组间显著差异基因，根据GO分类信息进行统计分类，有助于开拓潜在的抗逆基因。

分析中，对差异表达miRNA或lncRNA的靶基因进行GO显著性富集分析，此项分析可结合表达谱分析中的差异基因数据联合分析，结合生物学现象相互验证。

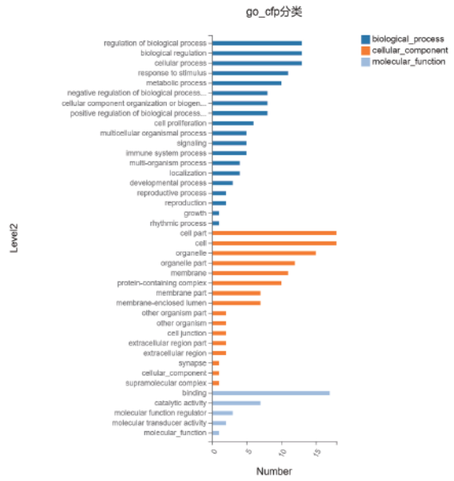


图4 差异基因的GO功能注释分类统计图

2.5.4 Pathway显著性富集分析

对转录组或表达谱已找到的各组间显著差异基因进行代谢通路图研究，不同基因相互协调行使其生物学，基于Pathway的分析有助于更进一步了解基因的生物学功能。通过Pathway显著性富集能查找差异表达基因参与的最主要生化代谢途径和信号转导途径，进一步明确抗逆机理等。

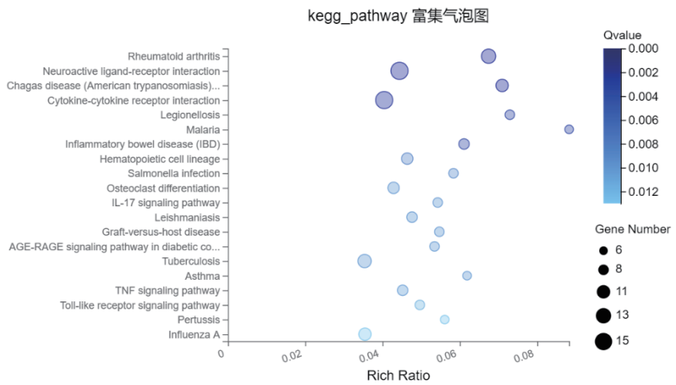


图5 差异基因pathway富集统计气泡图

2.5.5 miRNA/lncRNA靶基因分析

通过miRNA、lncRNA靶基因分析，进一步确定miRNA、lncRNA作用方式。全面描述抗逆条件下，非编码RNA参与的调控机制。

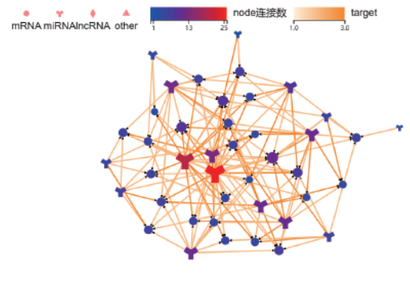


图6 miRNA靶基因分析

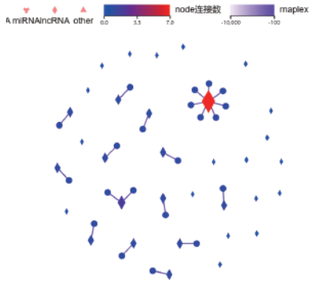


图7 lncRNA靶基因分析

2.5.6 mRNA与miRNA、lncRNA互作分析

通过研究miRNA和lncRNA与其靶基因的互作模型构建，研究抗逆机理。通过查找相关文献，进一步获得调控信息。

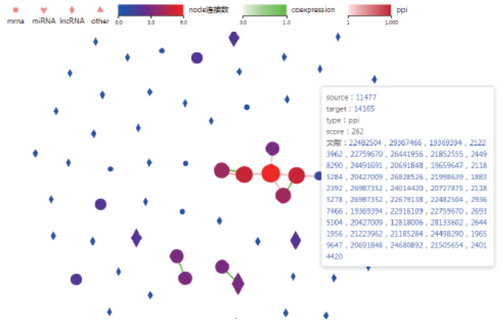


图8 差异表达lncRNA与靶基因的蛋白互作和共表达互作网络图

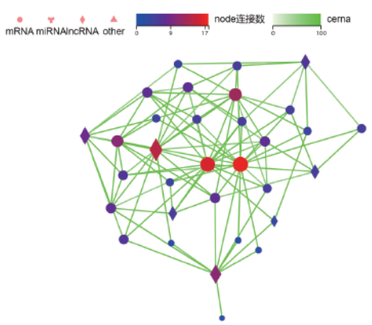


图9 差异ceRNA—mRNA互作网络图

2.6 项目周期

表1 不同研究策略交付周期

类型	周期
经济型模式：LncRNA + Small RNA	50个工作日
大容量模式：LncRNA + Small RNA+ CircRNA	55个工作日

应用案例

案例一：盐胁迫下棉花的长链非编码RNA鉴定 [2]

研究概要：

本研究对盐胁迫（S_NaCl1,2）和对照（S_CK1,2）处理的三叶期棉花幼苗，进行全转录组链特异性RNA测序。鉴定了1117种lncRNA，其中差异表达的lncRNA有44种。对差异表达的基因间lncRNA（lincRNA）进行功能分析，做顺式调控靶基因的富集，主要富集于应激相关类别。RT-qPCR验证了所有选择的lincRNA对盐胁迫有反应。发现lnc_388可能参与调节Gh_A09G1182，并且lnc_883可通过调节Gh_D03G0339 MS_channel的表达参与调节对盐胁迫的耐受性。同时预测棉花中miRNA的靶标模拟物（miRNA mimics），鉴定了6种miRNA。RT-qPCR与lncRNA和miRNA的结果均表明在盐胁迫下，lnc_973和lnc_253可以调节ghr-miR399和ghr-156e作为内源靶标模拟物的表达。

研究设计：

选择耐盐棉花品种SN91-11，培育至三叶期幼苗，随机分两组，一组250mM NaCl处理，一组淡水处理对照，24h → 每个棉花幼苗样品分离总RNA，去rRNA链特异性建库 → lncRNA预测、鉴定，表达分析 → lncRNA功能分析，靶标模拟物miRNA预测 → qRT-PCR验证

研究结果展示：

1、盐胁迫下棉花lncRNA的表达

分析转录组测序结果，预测、鉴定了1117种特有的lncRNA，分析保守性，发现lincRNA特殊的表达模式并聚类。随机选择11种lincRNA进行RT-qPCR验证，结果与测序大多一致，4种上调，2种下调。此外，发现Gh_A05G3489，Gh_A01G0321，Gh_A01G0639和Gh_A11G0366与lincRNA共表达。

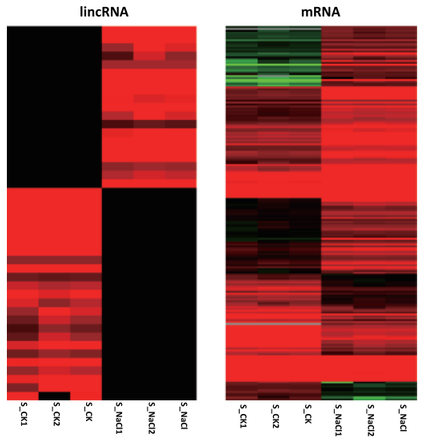


图10 盐胁迫处理组和对照组样本中特异性表达的lincRNA和mRNA聚类图

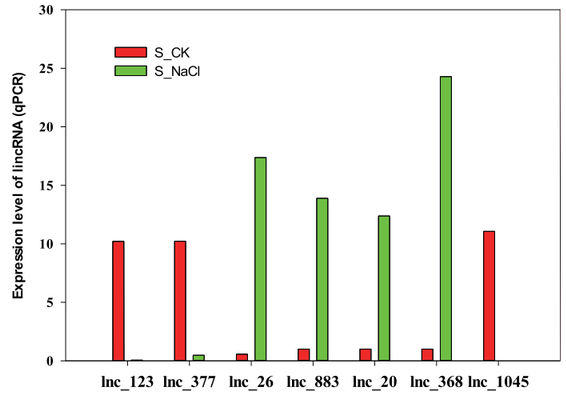


图11 盐胁迫下7种lincRNA的差异表达RT-qPCR验证

2、盐胁迫下差异表达的lncRNA的功能分析

选择与差异表达lincRNA间隔小于20kb的共表达mRNA进行GO富集，发现差异表达lincRNA参与调解重要的过程，如碳水化合物代谢，解毒，能量合成，转录，染色质修饰和响应盐胁迫的转录后调节。RT-qPCR定量lincRNA及推定的顺式调控靶基因，发现在盐胁迫下lnc_883和LRR8（Gh_A09G1182）共表达并显著上调。

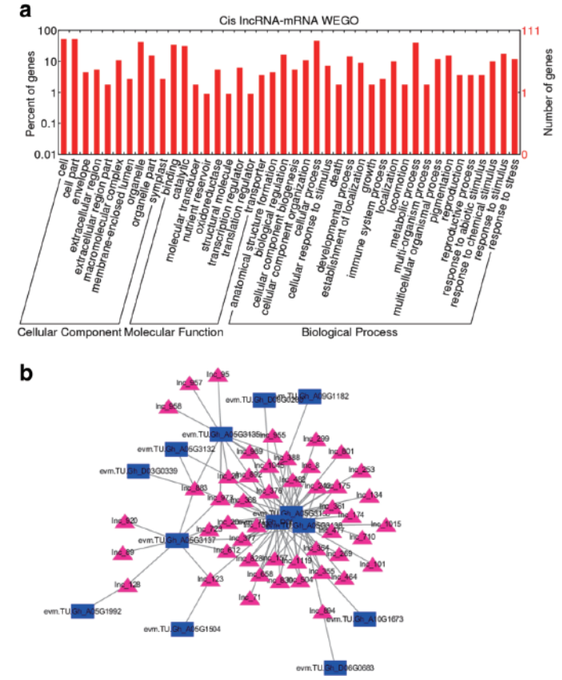


图12 盐胁迫下差异表达lincRNA功能分析；a. 差异表达lincRNA的功能富集；b. lincRNA-mRNA网络互作图

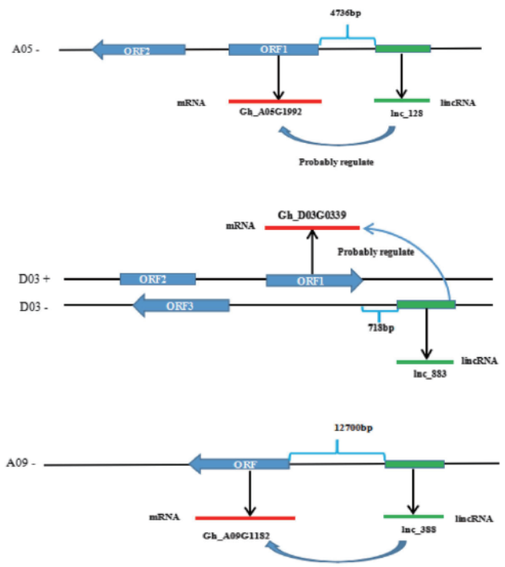


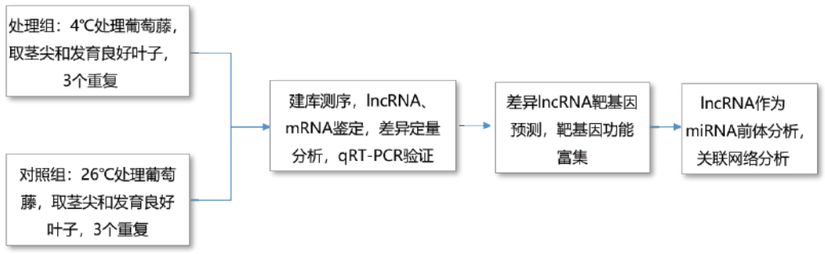
图13 lincRNAs的结构及推定的靶基因

案例二：葡萄中冷胁迫相关lncRNA的鉴定和功能分析[3]

研究概要：

本研究发掘了葡萄中冷胁迫相关的lncRNA，解析这些lncRNA在冷胁迫下的表达模式，并研究了其与靶基因的表达相关性。研究结果为深入了解葡萄冷胁迫耐受机制奠定了基础。研究发现，在葡萄中203个已知lncRNA在冷胁迫下上调，144个已知lncRNA在冷胁迫下下调。这些冷胁迫响应的lncRNA的靶基因涉及植物非生物胁迫耐受，例如CBF转录因子、LEA蛋白、WRKY转录因子等。结果显示，这些lncRNA正调控靶基因表达。推测这些lncRNA在葡萄冷胁迫响应过程中具有调控基因表达的作用，并可能在葡萄冷胁迫耐受中发挥作用。此外，2088个新lncRNA在葡萄中被鉴定。其中，部分lncRNA作为miRNA前体，而部分为miRNA的靶基因。

研究设计：



主要结果展示：

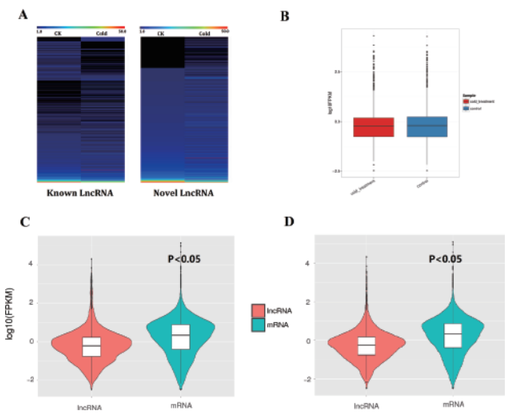


图14 葡萄中lncRNA和mRNA的表达情况

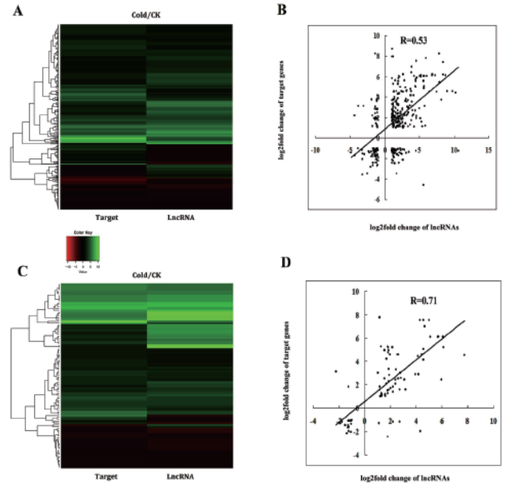


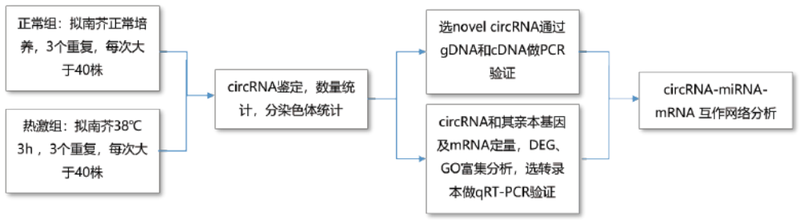
图15 显示葡萄冷胁迫响应lncRNA与其顺式靶基因、反式靶基因在冷胁迫下的表达均为正相关

案例三：拟南芥中环状RNA参与热应激反应研究[4]

研究概要：

热应激会阻碍植物生长减少作物产量，circRNA参与应激研究较少。本研究通过RNA测序和生物信息学分析在拟南芥中鉴定1599个先前未知的circRNA和1583个热特异性circRNA。结果表明，热应激显着增强circRNA的积累，增加circRNA的长度和循环外显子的数量，增加circRNA和转录本的可变剪接。并观察到一些circRNA和其亲本基因的表达模式呈正相关。此外，ceRNA（竞争性内源RNA）网络的预测表明，热诱导的circRNA可能通过circRNA介导的ceRNA网络参与植物对热应激的反应。

研究设计：



主要结果展示：

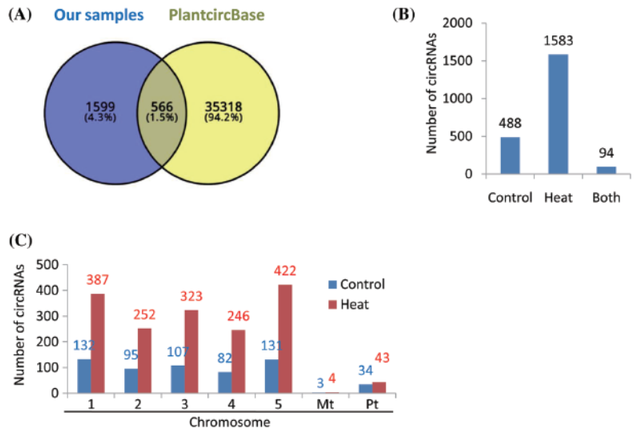


图16 circRNA鉴定，热激增加circRNA积累

A、维恩图选出1599个新的circRNA；B、热激后发现1583个特异性circRNA；C、分染色体表达情况

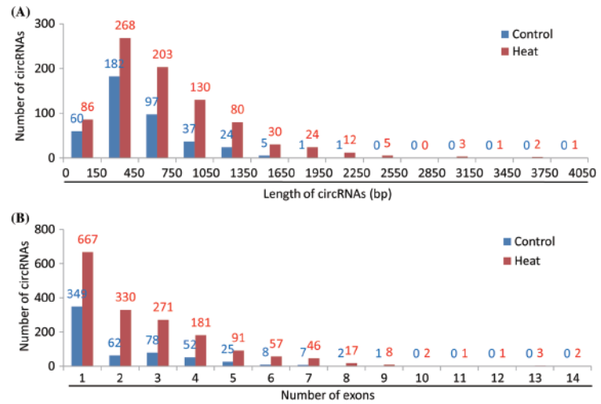


图17 热激增加circRNA的长度和循环外显子的数量，主要是通过更多的外显子成环实现的

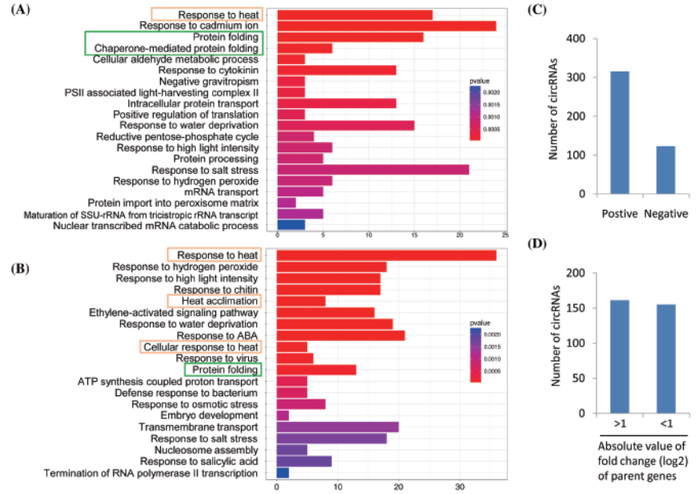


图18 circRNA和亲本基因表达模式比较

circRNA与亲本基因呈正调控作用（一同上调或一同下调）。439个circRNA中70%与亲本基因表达趋势一致，其中50% 的亲本基因是显著差异。

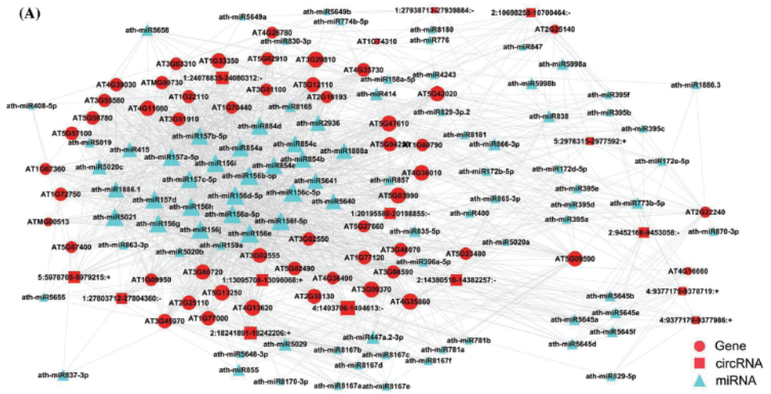


图19 ceRNA网络

热激反应下circRNA和基因上调（红色标注）

可能存在的风险

在样品选择的过程中，需要注意：首先需要对选取样品进行生理生化试验，根据结果利用数据统计学方法筛选具有生物学重复意义的样品，以免在高通量分析结果中数据偏离较大。

另外，在分析过程中可能出现重点关注的基因由于表达量过低，而未在高通量测序结果中分析得到的情况。一般情况下，需要提高测序数据量等方式进行解决。

常见问题

1、两种测序策略方案如何选择？

答：两种模式，一种为经济型，即采用两种建库测序模式；一种为大容量型，即采用三种建库测序模式；主要区别在于环状RNA研究方式不同，一个是对环状RNA广谱研究，即对环状RNA表达情况初步研究；二是对环状RNA单独富集建库测序，即可实现对环状RNA高深度研究，以及低表达环状RNA的鉴定和定量分析等。所以如何选择两种模式，是需要根据研究目的和需求决定的，以及对RNA研究深度需求决定的。

2、两种测序策略方案送样量要求？

答：满足两种建库测序模式（LncRNA+Small RNA）总量需求：Total RNA最低起始量2ug；满足三种建库测序模式（LncRNA+Small RNA+CircRNA）总量需求：Total RNA最低起始量7ug。

3、长链非编码RNA建库测序结果可以用于分析环状RNA吗？

答：可以，长链非编码RNA建库模式采用去核糖体方式建库，对所获得的RNA进行打断测序，相当于获得全部RNA，即可以进行分析环状RNA表达情况，可应用于分析各物种环状RNA表达谱。

4、LncRNA测序数据中mRNA的定量效果如何？

答：人标品：已知mRNA定量与qPCR定量斯皮尔曼系数达到0.88。可作为参考。

华大优势

实验严谨：建库工艺稳定，技术重复性高，可接受多种类型样本，total RNA最低起始量低至2μg；

技术领先：UMI Small RNA建库，定量精准，1ng低起始量，更多有效数据，成功率高。

分析全面：独特的Dr.Tom多组学数据挖掘系统交付。数据图表循环挖掘，多维度结果图展示，10大注释数据库，12种分析小工具，多组学关联分析，互作网络可视化，随时更新文献信息，查基因得文献，便于文章撰写；

质控严格：实验操作及信息分析操作采用全方位及现金的质量管理体系标准。

参考文献

[1] Merry C R, Forrest M E, Sabers J N, et al. DNMT1-associated long non-coding RNAs regulate global gene expression and DNA methylation in colon cancer[J]. Human molecular genetics, 2015, 24(21): 6240-6253.

[2] Deng Fenni,Zhang Xiaopei,Wang Wei et al. Identification of Gossypium hirsutum long non-coding RNAs (lncRNAs) under salt stress.[J] .BMC Plant Biol., 2018, 18: 23.

[3] Wang Pengfei,Dai Lingmin,Ai Jun et al. Identification and functional prediction of cold-related long non-coding RNA (lncRNA) in grapevine.[J] .Sci Rep, 2019, 9: 6638.

[4] Pan T, Sun X, Liu Y, et al. Heat stress alters genome-wide profiles of circular RNAs in Arabidopsis.[J]. Plant Molecular Biology, 2018, 96(3):217-229.

全长转录组在动植物研究中的应用

研究背景

近年来，随着高通量测序技术的发展，转录组测序已经成为研究基因表达调控的主要手段。通过转录组测序可挖掘基因功能、探究基因表达及调控模式、最终揭示表型信息背后的分子学机制，已经广泛应用于基础研究、临床诊断及药物研发等领域。

从原理上来说，转录组测序本应该是一个简单的过程，只需要分离得到RNA样本，然后对RNA进行高通量测序，最后拼接出来的RNA应该具有很高的准确性和可重复性。但是，国际RNA测序基因组注释评价项目协会（international RGASP consortium）在2013年发表的两篇论文[1-2]报道了一场竞争程度相当激烈的、大规模的RNA测序热潮，各国的科学家们都在寻求最佳的RNA测序分析算法，而结果却是出人意料的丰富多样。即便是对于人类基因组，甚至没有哪个转录本重构（transcript reconstruction）方法的准确率能够达到60%。只有线虫和果蝇的结果相对好一些，但是要注意的是，这两种生物的基因组要比人类的小得多，也简单得多。况且这3个物种都是迄今为止被研究得最充分、最深入的3个物种。还有很多其它物种只是近几年才完成基因组测序工作，并没有太多时间完善这些物种的基因组序列，因此对这些物种而言，它们的转录本重构工作会更加困难，准确性会更低。

总体来说，基于短读长测序平台的转录组测序产品由于读长的限制（PE100/PE150），在转录本组装的过程中存在较多的嵌合体，并且不能准确地得到完整转录本的信息，因而对后面的表达量分析、可变剪接、基因融合等分析造成了较大的影响。基于PacBio的单分子实时测序技术，目前平均读长已经达到20Kb以上，最长可达80Kb，其长度已经超过一般转录组中典型的基因的长度，所以利用PacBio测序平台进行转录组的研究，可以直接得到全长转录本信息，而无需组装，从而最大限度的保证了转录组测序结果的准确性。

1.1 PacBio测序原理介绍

在一个单分子实时反应管（SMRTCell）中有许多圆形纳米小孔，即ZMW，外径100多纳米，激光从底部打上去后不能穿透小孔进入上方溶液区，能量被限制在一个小范围里，使得信号仅来自这个小反应区域，孔外过多游离核苷酸单体依然留在黑暗中，将背景降到最低。

单个ZMW底部固定有一个结合了模板DNA的聚合酶，这个DNA聚合酶是实现超长读长的关键之一，读长主要和酶的活性保持有关。当加入测序反应试剂后，4色荧光标记4种碱基，会发出不同光，根据光的波长与峰值可判断进入的碱基类型。PacBio RSII一个SMRTCell中有15万个ZMWs，每个孔中有一个单分子DNA链在高速合成，如众星闪烁，每合成一个碱基即显示为一个脉冲峰，配上高分辨率的光学检测系统，就能实时进行检测。

短读长测序的碱基荧光标记都标记在5 ‘端甲基上，在合成过程中，荧光标记物保留在DNA链上，随着DNA链的延伸会产生三维空间阻力导致DNA链延长到一定程度后会出现错读。这是短读长测序读长仅能达到100多bp到200bp的一个原因。PacBio平台的碱基荧光标记在3’ 端磷酸键，在DNA合成过程中正确的碱基进入时，在3’ 端磷酸键的标记是会随磷酸键断裂自动被打断，标记物被弃去，即合成的DNA链不带荧光标记，和天然的DNA链合成产物一致，所以PacBio测序可以达到很长的读长。

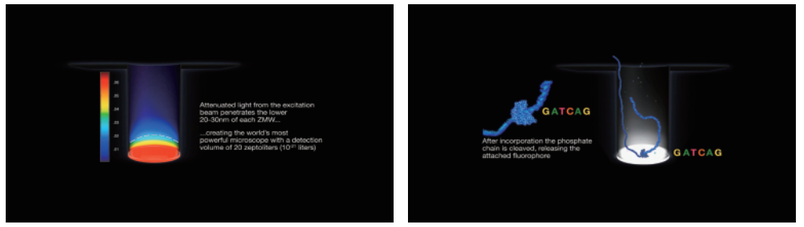


图1 PacBio单分子实时测序

1.2 全长转录组建库流程

质量合格的RNA 样品通过Clontech SMARTer PCR cDNA Synthesis Kit和UMI引物合成全长的UMI+cDNA。采用1+0.4X AMPure PB bead进行片段筛选，然后经DNA损伤修复、末端修复、接头连接、酶反应消化、纯化等步骤，最终得到哑铃形的标准全长转录组文库。经Qubit仪器和Agilent2100生物分析仪检测合格后，进行Sequel 上机测序和后续数据分析解读。

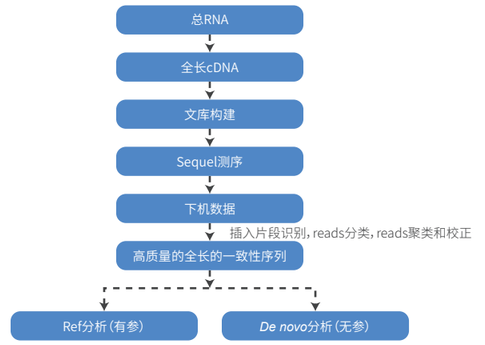


图2 全长转录组技术流程

方案设计

适用范围	推荐测序方案	信息分析思路
• **无参考序列物种：**菊花、青蒿、人参、玫瑰、银杏、牡丹、松鼠、对虾等	• 全长转录组测序：不同组织/不同发育时期样本单独测序或混合测序； • 每个样本单独进行RNA-Seq测序，设置生物学重复	• 全长转录组测序构建此物种参考基因集； • CDS预测，SSR分析，LncRNA及功能分析等； • isofrom定量分析；以Sequel测序得到的全长转录本的序列为参考序列，对RNA-Seq测序进行定量及差异分析及功能分析。 • 辅助基因组注释

图3 无参考序列物种研究方案

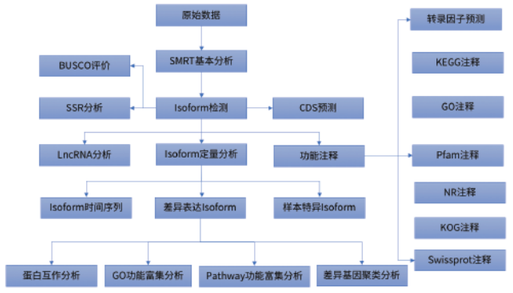


图4 基于UMI的Iso-Seq *de novo*的信息分析条款

适用范围	推荐测序方案	信息分析思路
•**有参考序列物种：**人、鼠、猪、马、鸡、牛、羊、线虫、拟南芥、玉米、水稻等	•全长转录组测序：不同组织/不同发育时期样本单独测序或混合测序； •每个样本单独进行RNA-Seq测序，设置生物学重复	•Sequel测序结果与参考基因组序列进行比对，从而发现新的基因和转录本，并且可以补充基因组的注释信息； •基因结构和功能分析等； •isofrom定量分析 •RNA-Seq数据进行不同样品间的差异分析。

图5 有参考序列物种研究方案

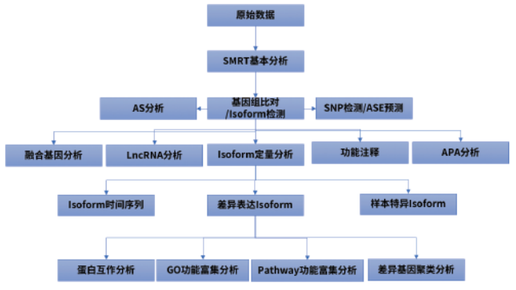


图6 基于UMI的ISO-Seq ref的信息分析条款

2.1 样本选择

1）构建参考基因集，结合RNA-SEQ对基因定量

全长转录组：不同组织，不同发育时期样本进行Sequel测序，无需生物学重复；为了节约成本也可将不同组织和不同发育时期的样本混合进行测序，在不增加测序成本的基础上可得到不同组织特异高表达的基因，这样得到的总体基因比例相对较低。

测序策略：推荐每个样本构建一个文库，测序20G以上。

RNA-Seq：不同组织或不同发育时期样本单独测序，推荐3个以上生物学重复。

2）研究不同组织间/不同发育时期转录本结构/定量差异：

不同组织/不同发育时期分别取样，进行Sequel建库测序；推荐每个组织测序数据量20G以上（Wang B et al. Genome research, 2018.）。

3）辅助基因组注释

选择不同组织不同发育时期样本混合测序

2.2 测序策略

全长转录组：Sequel 平台，推荐每个样本构建一个文库（可根据需求增加一个4.5-10Kb文库），测序20G以上。该数据量是基于目前Sequel通量性价比较高的推荐，该数据量还远远未达到饱和，通过增加数据量（提高有效reads数）可得到更多的全长转录本的数目；

RNA-seq测序：BGISEQ平台，SE50测序，20Mb Reads/sample。

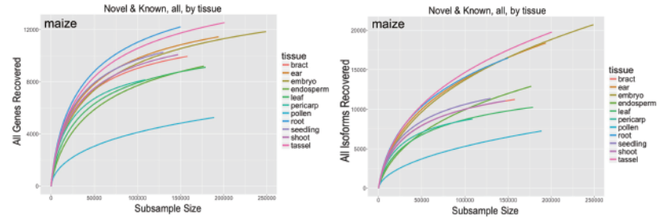


图7 三代测序基因/isoform饱和度曲线

2.3 分析结果

2.3.1 全长转录本的获得

对Sequel测序的下机数据进行处理，其中Reads of insert处理后的序列可以分成四类，分类结果见图8。在原始序列中的5'端接头、3'接头、poly A的检测中，全长转录本必须同时检测到这三者。

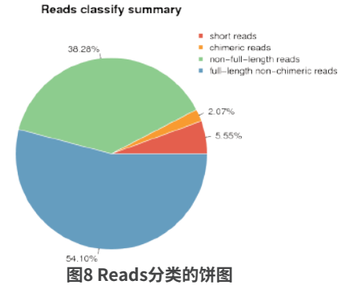


图8 Reads分类的饼图

全长非嵌合序列（full-length non-chimeric）嵌合的全长序列（full-length chimeric），非全长序列（non-full-lenth）和短序列（short reads）。从图中可以看到只有少量的reads被鉴定为嵌合，这说明构建的SMRTbellTM文库质量较好。

2.3.2 全长转录本的注释

对于无参考序列的物种，对得到的全长转录本进行七大功能数据库注释(NR、NT、GO、COG、KEGG、Swissprot和InterPro)，并且用维恩图来展示NR、COG、KEGG、Swissprot以及InterPro的注释结果，结果见图9。

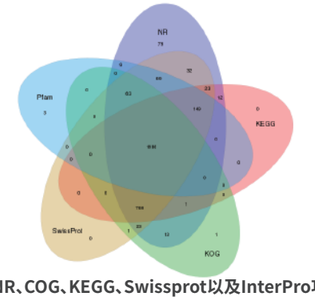


图9 NR、COG、KEGG、Swissprot以及InterPro功能注释维恩图

2.3.3差异表达基因分析

对于无参考序列的物种，将转录组或者RNA-Seq测序得到的Reads比对到全长转录本上，从而检测样品之间的差异表达基因。图10为不同处理之间的差异表达基因数量统计图，图11为差异表达基因层次聚类热图。后续也可以对筛选到的差异表达的基因进行GO和Pathway的富集分析，从而知道这些基因的功能及参与的信号通路，解释相关的生物学现象。

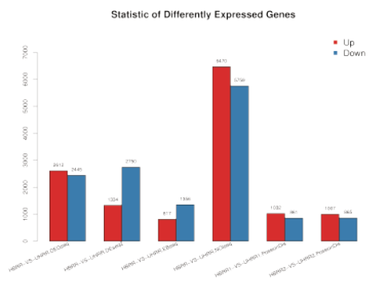


图10 差异表达基因数量统计图

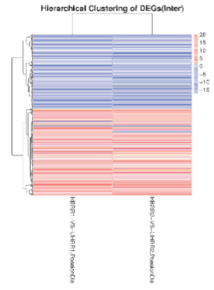


图11 差异表达基因层次聚类热图

1.4.4全长转录组比对到参考基因组

将全长转录本的序列比对到参考基因组，可以找到新的转录本，鉴定可变剪接和基因融合的现象，图12为转录本分类的结果，图13为融合基因的结果。

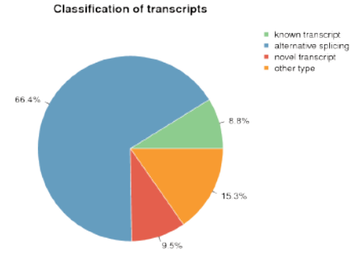


图12 转录本分类图

已知转录本（known transcript）、可变剪切转录本（alternative splicing）、新转录本（novel transcript）和其它类型（other type）。从图中可以看到全长转录组测序能鉴定到更多的可变剪接转录本。

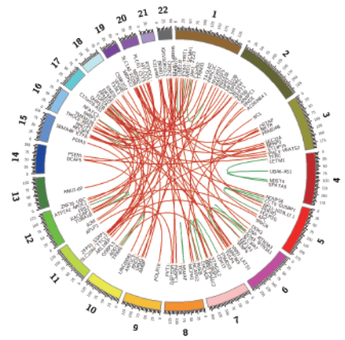


图13 融合基因的Circos图

对检测到的所有的融合基因提供可视化的展示结果，从图中可以得出哪些基因发生了融合，以及这些基因的位置关系。绿色代表同一个染色体内的基因发生了融合，红色代表染色体之间的基因发生了融合。

2.4 项目周期

样品检测合格后，建库+测序+标准信息分析：全长转录组的交付周期约40个工作日。实际项目完成时间根据所选具体样本数以及信息分析条款决定。

应用案例

3.1 案例一：玉米和高粱全长转录组测序 [3]

材料选择：

1）玉米（B73）：根、茎、叶、幼苗（14日龄），雌穗（ears，v8期），雄穗（tassels，v7期），花粉（r1期）、胚，胚乳和果皮（授粉后20天从种子获得），须（silk,r1期），苞片（最内层）。

2）高粱（BTx623）：根、茎、叶、幼苗（组织萌芽后14天）、胚、胚乳和果皮（授粉后20天）、花粉（9-10周阶段）、花序-1（1-5mm）花序-2（5-10mm）、花序-3（1-2cm）

测序策略：

1）PacBio平台

建库：Barcode建库；文库大小：<1 kb、1–2 kb、2–3 kb、3–5 kb、>5 kb 文库

测序：PacBio RS II：130 SMRT Cells；PacBio Sequel：5 SMRT Cells；

2）Illumina平台

Illumina HiSeq 2500 PE125

研究内容：

1）Isoform 统计和分类

2）Isoform 和可变剪接的组织特异性

3）伴随无义介导的mRNA衰变的AS事件

4）可变剪接的保守性

5）转录因子Isoform分析

6）可变多聚腺苷酸化

7）LncRNA分析

8）基因表达聚类分析

9）基因进化分析

主要结果：

1. 测序得到总数据量（玉米+高粱）：6,893,280 reads；长度分布：256–6643 bp；玉米有1,570,093(96.7%)条高质量转录本能比对到玉米参考基因组，非冗余序列136,745；高粱有979,305 (89.5%)条高质量转录本能比对到参考基因组，非冗余序列95,380。

2. 玉米的11个组织中有1659 个共有isoforms，高粱的11个组织中有1069个共有 isoforms；在玉米中，花粉的特异性isoforms占比最高（27.2%），根的特异性isoforms占比最低（14.5%）；在高粱中，花序-1的特异性isoforms占比最高（35.8%），花粉次之（34.1%），根的特异性isoforms占比最低（20.8%）。

3. 在玉米和高粱中，分别有18,741 (45%) 和 13,327 (38.5%)个基因存在可变剪接事件。可变剪接类型包括：(ES) Exon skipping; (A5) alternative 5′ splice-site; (A3) alternative 3′ splice-site; (IR) intron retention; (AF) alternative first exon; (AL) alternative last exon.；其中IR在所有组织中占比最高，AL占比最少。

4. 玉米isoform中有55,080(40.3%)个无义介导的的mRNA衰变（NMD）；高粱中有34,322(36%)个NMD；玉米和高粱中，Non-NMD isoforms均比NMD isoforms表达量高。

5. 玉米和高粱同源基因的可变剪接形式类似；玉米同源基因的可变剪接数目比高粱多；玉米和高粱保守isoform中IR可变剪接形式产生NMD的可能性最高；ES产生 NMD的可能性最低。

6. 不同组织富集的基因GO功能差异较大；玉米和高粱同一组织富集的基因GO功能也有差异。

7. 玉米和高粱中检测到的转录因子家族数目和参考基因集一致；玉米新发现179个转录因子isoforms,高粱中新发现129个转录因子isoforms。

8. 玉米和高粱的APA motif 占比最高的是AATAAA motifs，排名前三的motifs在两个物种中数量相同。

9. 通过PLEK构建了RNA分类模型，获得lncRNA序列；得到1706个高粱新lncRNA，平均长度1241bp，长于前期三代研究结果（平均长度880bp）；高粱lncRNA的平均长度比玉米（535bp）长。

10. 高粱不同组织间基因表达相关性跟组织间的进化关系相关，进化上越相近的组织，相关性越高；玉米不同组织间基因表达相关性跟组织进化无明显的相关性；玉米和高粱同一组织基因表达相关性比同一物种不同组织相关性高。

基因进化分析发现，玉米比高粱有更多的“年轻”基因；玉米的蛋白编码基因中有59.2%是“古老”基因，3%是“年轻”基因；高粱的蛋白编码基因中有66%是“古老”基因，6.7%是“年轻”基因；和进化“年轻”的基因相比，“古老”基因长度和ORF长度更长，Isoform种类更多。生殖组织比营养组织的进化年龄更高；一般情况，进化年龄指数（TAI）越高的组织，转录组多样性指数（TDI）越高；玉米和高粱同一组织的TAI和TDI指数有细微差别。

3.2 案例二：PacBio测序研究玉米转录组的复杂性[4]

材料选择：玉米自交系B73不同发育阶段的6个组织（根、花粉、胚芽、胚乳、幼雌穗、幼雄穗），提取RNA；

测序策略：

Illumina平台：6个组织进行RNA-Seq测序，每个样品三个重复；

PacBio平台：每个样本反转录之前加入特异性barcode，后续进行等量混合，上机测序47cell；

分析方案：检测玉米可变剪接现象；转录因子分析；lncRNA分析；融合基因分析；甲基化分析。

主要结论：

1. 构建5种不同片段大小的文库，上机测序47cell，总共产生3,716,604条reads，过滤掉低质量的reads，总共获得1,553,692 条全长的转录本序列(FL)。

2. 和RefGen-V3的isoforms进行长度比较，发现全长转录本预测出来的转录本整体上比V3基因集的要长。在目前的V3基因集中一个基因平均有2.84个isoforms，而全长转录组数据显示，一个基因平均有6.56个isoforms，是前者的两倍多。Isoforms的组织特异性分析显示：花粉有更高的组织特异性，而根的特异性最低。

3. 在玉米的V3参考基因集中，转录因子数目为2,624，分为57个家族。全长转录组解决方案将转录因子的数据增加到5,423个，几乎是两倍。

4. 鉴定出878个lncRNA，其中11个是以前报道过的，867个是新的lncRNA。这些lncRNA的平均长度为1.1kb（范围为0.2kb-6.6kb），比之前的认知的lncRNA要长很多(平均400+kb)。花粉拥有最多的特异的lnc（238个），穗是最少的lnc（68）个。

5. 从Pacbio数据中鉴定出1,430个融合转录本。其中143个被Illumina数据支持。结果表明，融合事件多发生在染色体间。

3.3 案例三：全长转录组研究揭示丹参药用成分合成机理[5]

对丹参酮根部的周皮、韧皮、木质3种类型的根部组织进行了mRNA测序，测序平台为HiSeq 2500和PacBio平台。利用HiSeq 2500平台检测丹参酮合成途径的相关基因的表达水平，利用PacBio平台的数据进行可变剪接的分析。

结果展示：

1) 采用HiSeq 2500数据对PacBio RSII平台所产生的subreads进行了校正，最后得到了16,241个高质量非冗余isoforms。

2) 基于HiSeq 2500产生的mRNA数据的差异表达分析，发现了在根部周皮部特异表达与高表达的丹参酮合成相关基因，包括SmCPS1、SmKSL1、GGPS、IPI、CYP等；

3) 另外基于PacBio的数据，发现了大约有40%检测基因位点发生了可变剪接现象，其中有些基因参与了萜类化合物代谢及类异戊二烯代谢。

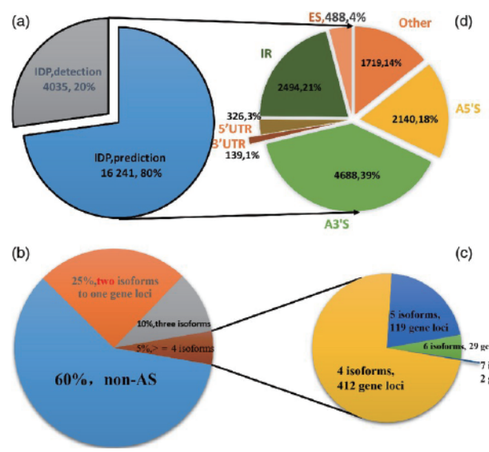


图14 采用IDP方法检测和预测丹参基因Isoform

a图：检测和预测的isoform的个数；b和c图：每个基因位点的检测到的Isoform的分布；d图：每种可变剪切类型的分布。

可能存在的风险

由于全长转录组测序较贵，推荐的方法数据量不够饱和，表达量较低的基因可能检测不到。

常见问题

1、PacBio Sequel转录组推荐的测序方案？

答：建库：0-5kb文库（1+0.4X磁珠纯化文库）；测序：1-2个Cell。转录组0-5K文库也包含5K以上片段，测序结果反映转录本的真实情况。如果特别关注5K以上的长转录本，也可以增加一个4.5-10K的文库。

华大优势

通量大：华大基因拥有12台Sequel测序仪，通量大，测序成本低，周期短。

样品起始量低：华大基因全长转录组样本需求仅1μg，远低于同行样本量需求。

信息分析内容全面：实时跟进科学研究前沿，不断升级信息分析内容。

个性化分析：具有丰富个性化分析经验，可根据项目需要选择最适宜的分析软件，只为保障最精准结果。

无需组装：长读长无需要组装，即可得到准确的全长转录本的序列信息。

精准基因集：借助Sequel平台读长的优势获得更精准基因集，可以改善基因表达定量的结果。

更多新发现：可以发现新的基因和转录异构体。

结构变异分析：可准确的鉴定可变剪接及基因融合现象。

辅助基因注释：可辅助基因组*de novo*基因注释，获得更好的基因注释结果。

经验丰富：华大基因自2015年推出全长转录组产品以来，已完成800+个全长转录组测序。目前华大基因的Sequel平台运行良好，实验及信息分析人员上机及问题处理经验丰富。

参考文献

[1] Steijger T, Abril J F, Engström P G, et al. Assessment of transcript reconstruction methods for RNA-seq[J]. Nature methods, 2013, 10(12): 1177-1184.

[2] Engström P G, Steijger T, Sipos B, et al. Systematic evaluation of spliced alignment programs for RNA-seq data[J]. Nature methods, 2013, 10(12): 1185-1191.

[3] Wang B, Regulski M, Tseng E, et al. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing[J]. Genome research, 2018, 28(6): 921-932.

[4] Wang B, Tseng E, Regulski M, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing[J]. Nature communications, 2016, 7: 11708.

[5] Xu Z, Peters R J, Weirather J, et al. Full‐length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of Salvia miltiorrhiza and tanshinone biosynthesis[J]. The Plant Journal, 2015, 82(6): 951-961.

动植物生理机制的蛋白组与转录组关联分析研究方案

研究背景

生命体是一个多层次，多功能的复杂结构体系，从DNA、RNA、蛋白质到代谢物的过程中涉及到一整套精细的表达调控机制，如转录调控、转录后调控、翻译调控、翻译后调控等。高通量技术的发展积累了大量的组学数据，这使得由精细的分解研究转向系统的整体研究成为可能[1]。整合多组学数据能够实现对生物系统的全面了解，建立有效指示表型的模型，揭示重要的生物标志物。

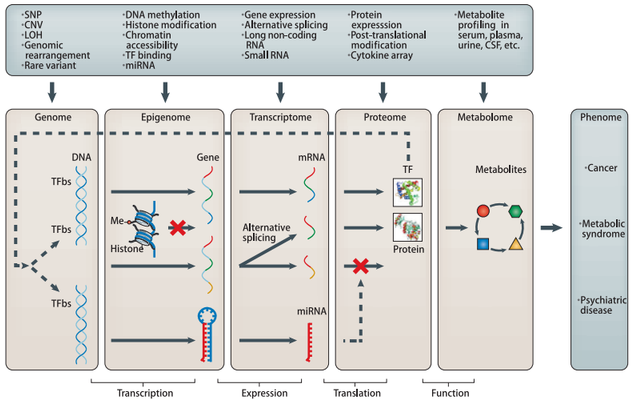


图1 系统生物学之基因组、表观组、转录组、蛋白组、代谢组与表型的关系[1]

该图从系统生物学角度描述了基因组、表观组、转录组、蛋白组、代谢组学的研究对象和能够获得的主要结论，以及组学和表型之间的关系。

蛋白质是生命功能的执行者，其含量的变化在生物体的生长发育[2]、环境应激[3,4]、疾病发生发展[5]等过程中发挥着重要的作用，对于蛋白质的表达量进行深入研究是十分重要而又关键的。蛋白质组学包含基因组和转录组所不曾有的功能性相关信息：

•基因的表达呈现时空和丰度高低的特征；

•许多蛋白质的修饰形式具备特定的生物学功能；

•大多数蛋白质可形成具有功能性的复合物，诸如蛋白质/蛋白质、蛋白质/核酸、蛋白质/脂类等。

mRNA的表达量是影响蛋白表达最为重要和直接的因素，通过分析蛋白组与转录组的关联性，可以系统全面地了解生物体内基因表达调控途径[6]。

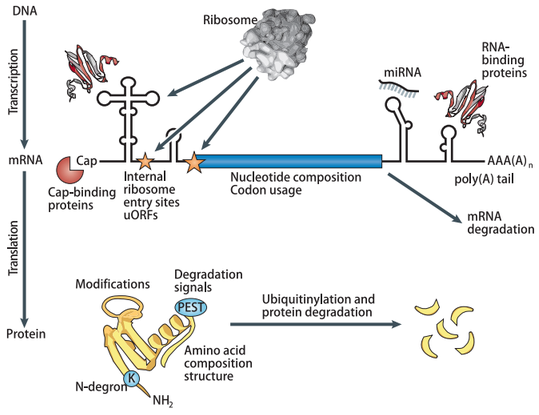


图2 转录调控和蛋白翻译后修饰模式[6]

该图展示了蛋白质的丰度取决于RNA和蛋白产生和降解之间的平衡，图中上半部分主要解释转录和翻译过程的合成及稳定状态，下半部分主要阐述蛋白降解。

因此，要全面探究生物体生长发育、环境应激机制，疾病发生发展规律，精准描绘关键基因的表达模式，同步检测mRNA和蛋白质的表达量并进行联合分析已成为当前研究的必然趋势。本方案重点介绍蛋白组和转录组关联分析的方法，以及涉及到的分析内容，旨在帮助广大科研工作者进行动植物生理机制的研究，为全面了解动植物生长发育、环境应激性、致病机理等生理机制提供基础全面的数据结果。

方案设计

2.1 拟解决的关键科学问题

1. 动植物生长发育过程中生理机制的研究，为动植物遗传品质改良，培育新品种提供理论数据；
2. 动植物环境应激性研究，为动植物生理机制提供理论数据，提高农牧业动植物环境适应性；
3. 动植物致病机理研究，提高农牧业动植物的抗病能力，为人类疾病研究提供模式基础。

2.2 拟采取的研究方案

2.2.1 整体思路图

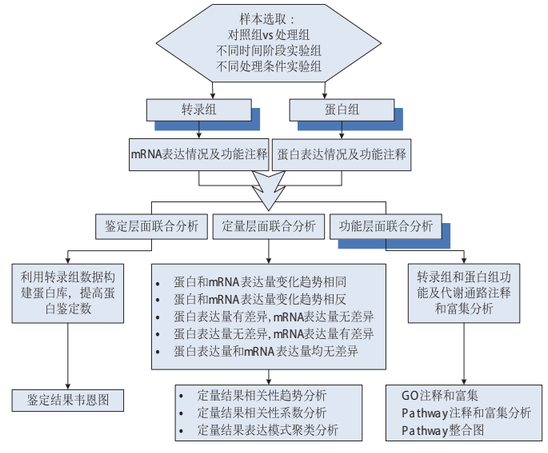


图3 蛋白组与转录组关联分析方案整体思路

2.2.2 样本选取

动植物生理机制研究，主要关注的是动植物的生长发育、环境应激性、致病机理等方面的问题。按照主要的三个应用方向，样本选取一般可参考如下原则：

1）生长发育研究：根据发育周期进行取样，设置不同时间阶段的实验组[2]；

2）环境应激性研究：根据环境刺激因子的剂量、浓度、处理时间进行取样，设置不同处理梯度的实验组[3、4]；

3）致病机理研究：选取抗病株vs易感株进行取样，同时设置不同的处理，进行组内和组间比较[5]；

转录组和蛋白组学的样本选择尽量保持一致，即遵循同时、同类型、同部位取样的原则。推荐转录组和蛋白组的取样一 一对应，如果不能保证样本和生物重复完全一致，则有可能出现不同时期不同样本的转录组和蛋白组相关性系数整体偏低的情况。

2.2.3 转录组分析

本阶段旨在通过转录组技术获得mRNA表达情况和功能分析结果，初步筛选转录组层面与表型相关的生物标志物。

采用的技术：转录组测序、RNA芯片、RT-PCR技术[7]，通过不同处理组与对照组样本间的比较及筛选，寻找差异表达的基因，并对差异表达的基因进行GO和Pathway的富集分析。

样本要求：1）针对动植物的生理机制研究内容，设计不同类型样本，详见2.2.2样本选取；

2）每组至少2个生物重复，推荐3个以上的生物重复；3）对于无参考序列的物种，需要对所有样本的转录组测序结果进行拼接，从而得到参考序列，然后作为基因表达定量的参考序列；对于有参考序列的物种，以基因组序列为基因组表达定量的参考序列。

2.2.4 蛋白组分析

本阶段旨在利用蛋白质组技术获得蛋白表达情况和功能分析结果，初步筛选蛋白组层面与表型相关的生物标志物。

采用的技术：采用iTRAQ/IBT蛋白定量、DIA蛋白定量[7]，并利用转录组测序构建蛋白质序列数据库，进而通过样品间的比较及筛选，获得蛋白表达数据，并对差异蛋白进行功能富集分析。

样本要求：1）设置与转录组分析相同的动植物样本；2）建议至少2个生物重复，推荐3个以上的生物重复。

2.2.5 关联分析

本阶段旨在对转录组和蛋白组分别获得的表达量情况和功能分析结果进行关联分析，获得两个组学上的定性、定量和功能层面上的关联分析结果。

1）鉴定关联分析：利用转录组数据库构建蛋白质数据库，提高蛋白鉴定数。

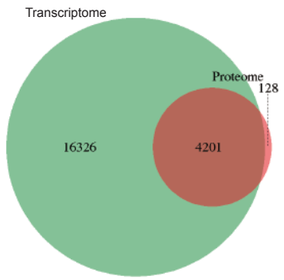


图4 定性关联韦恩图

该图绿色代表转录组鉴定到的基因个数，红色代表蛋白质组鉴定到的蛋白个数，两个圆圈重叠的部分即为转录组与蛋白组共同鉴定到的基因个数。

对所有鉴定到的蛋白和mRNA进行关联，一般情况下，转录组鉴定到的基因表达情况的覆盖度高于蛋白组，故利用转录组数据库构建蛋白质数据库，可提高蛋白鉴定数。

通过鉴定关联分析韦恩图，可从整体上分析鉴定到的mRNA和蛋白质的情况。

2）定量关联分析：对于转录组和蛋白组数据在表达量的层面上进行关联，获得两个组学层面表达量趋势和相关性进行分析，对表达情况进行聚类，并对于表达趋势一致或者不同的基因进行深入分析。

首先，按照蛋白和mRNA表达量的变化将所有关联到的基因分成5类，以便细致描绘基因表达调控模式：

A. 蛋白和mRNA表达趋势相同－DEPs_DEGs_Same Trend

B. 蛋白和mRNA表达趋势相反－DEPs_DEGs_Opposite

C. 蛋白表达有差异，mRNA表达无差异－DEPs_NDEGs

D. 蛋白表达无差异，mRNA表达有差异－NDEPs_DEGs

E. 蛋白表达和mRNA表达均无差异－NDEPs_NDEGs

对细分的5类表达关联类型，有助于验证表达一致性（A-正相关），补充（C/D/E-仅蛋白或RNA差异或无差异）、揭示特殊（B-负相关）的生物调控和代谢机制。

然后，为了深入了解各种表达情况的相关性趋势和相关性系数，需要对5类情况分别进行分析：

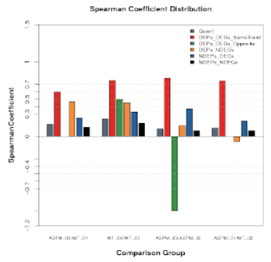


图5 相关性系数统计分析

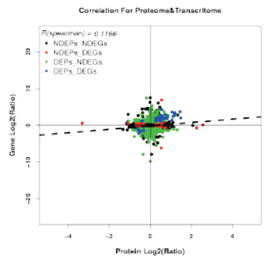


图6 相关性趋势分析

图5、6中，各种颜色分别代表上述mRNA和蛋白质的表达趋势情况。

最后，在定量关联层面，分析转录组和蛋白组的表达情况聚类模式：

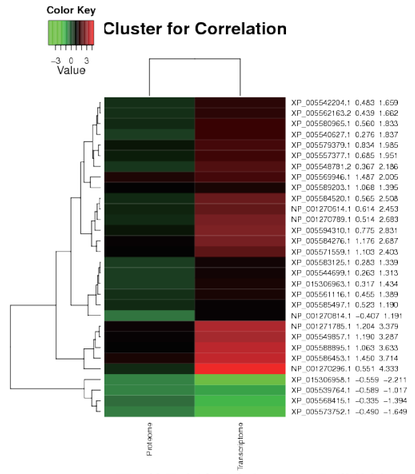


图7 差异表达蛋白和mRNA聚类分析

该图横坐标展示比较组间的转录组和蛋白组分类，纵坐标展示关联到的差异表达基因的名称。其中绿色代表基因下调表达，红色代表基因上调表达。

在动植物生理机制的研究中，最为关注的是转录组和蛋白组表达趋势一致的情况，表达情况正相关，有利于说明关键基因的表达情况在两个组学层面都得到了验证；对于转录组和蛋白组表达趋势相反的情况，一般是为了说明一些特殊的抑制调控方式，可根据具体的蛋白功能进行分析。另外，单一组学表达发生变化的情况，可通过后续的功能关联分析，寻找基因上下调的关系进行调控关键基因的深入挖掘。

3）功能关联分析：通过两个组学的鉴定、定量层面的关联，可以分析基因表达产物mRNA和蛋白一对一的关联方式，但对于某一类基因或者具有上下游调控关系的基因，仅通过一对一的关联方式无法进一步分析，需要通过功能和代谢通路的分析进行解释和分析。

通过转录组和蛋白组的功能注释和富集信息，获得两个组学层面的功能关联结果；通过GO注释和富集，找到转录组和蛋白组GO分类条目上共同关联到的基因，以及共同显著富集的GO条目（如图8）；通过Pathway注释和富集，找到转录组和蛋白组Pathway代谢通路上共同关联到的基因，以及共同显著富集的Pathway；最终通过两个组学层面的Pathway整合图（如图8），将检测到的mRNA和蛋白同时注释在同一条Pathway中，从一条代谢通路上研究基因和蛋白的上下游调控关系。

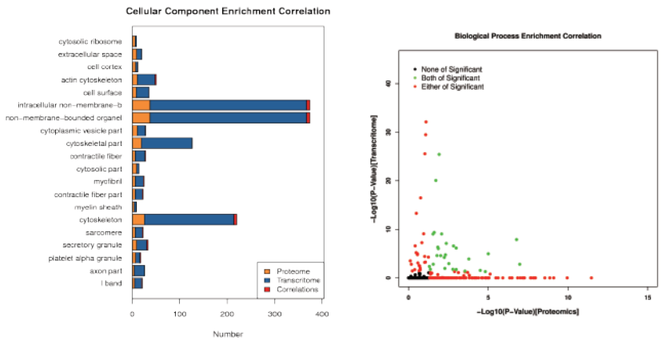


图8 左图差异表达mRNA、蛋白及共同关联到的基因在GO term上注释到的个数；右图差异转录组和蛋白组GO term-cellular component 上的富集关联情况

左图差异表达mRNA、蛋白及共同关联到的基因在GO term上注释到的个数；右图每个点代表一个go term，横坐标显示蛋白组GO term的p-value值，纵坐标显示转录组GO term的p-value值；红色的点代表转录组和蛋白组同时在这个GO term上显著富集，绿色的点代表只有一个组学层面在这个GO term上显著富集，黑色的点代表两个组学层面在这个GO term上都不显著富集。

根据蛋白组和转录组差异表达基因的分析结果，对其进行KEGG生物通路分类以及富集分析，最终将蛋白组和转录组差异表达基因的信息汇总在一张通路图中，分别显示mRNA上下调，蛋白上下调，mRNA和蛋白同时改变或者一方改变，mRNA和蛋白同时不变等情况，直观展示一条通路中的所有转录组和蛋白组鉴定和定量到的数据，更加方便地展示关键调控基因。KEGG (Kyoto Encyclopedia of Genes and Genomes)是有关Pathway的主要公共数据库，该数据库整合了基因组、化学以及系统功能信息，特别是测序得到的基因集与细胞、生物体以及生态环境的系统性功能相关联。

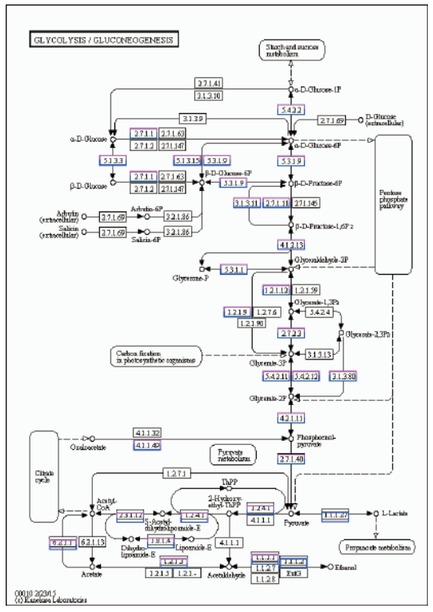


图9 差异蛋白和差异基因Pathway整合图（红框代表差异蛋白，蓝框代表差异基因）

2.3 项目执行周期

转录组分析周期：样品检测合格后，建库+测序+标准信息分析：约25个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

蛋白组分析周期：样品检测合格后，蛋白定量iTRAQ技术标准周期约25个工作日完成。实际项目完成时间根据所选具体样本数以及信息分析条款决定。

关联分析周期：数据分析，约12个工作日完成。实际项目完成时间根据所选具体样本数以及信息分析条款决定。

项目总时间：约40个工作日完成。实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.4 预期的结果

利用蛋白组和转录组研究技术，基于完整的蛋白质和mRNA表达量数据，对转录组和蛋白组进行定量、定性、功能层面的关联分析，期望能够深刻阐述动植物生理机制，为推动模式动物研究[8]、动植物育种[2, 7]、农作物防虫抗旱[4]等工作提供基础广泛的理论数据。

2.5 后期验证手段

利用MRM、western blotting技术对与表型相关的关键蛋白进行验证，进而确认该蛋白在相应的动植物生理机制中的作用；寻找表达模式相似的基因簇进行转录调控研究，可以通过比对找到调控同一类基因簇表达的启动子及其对应的转录因子，验证该转录因子的调控表达模式[10]；联合磷酸化蛋白定量分析，研究关键调控基因的蛋白修饰形式改变对动植物生理状态的影响[11]。

应用案例

3.1 案例一 黑腹果蝇发育过程中转录后调控定量分析[8]

复杂的转录后调控的存在，使得mRNA水平和蛋白质丰度之间的相关性降低。为了研究转录组和蛋白组的相关性，本研究以果蝇胚胎发生为研究对象，生成了14个时间点的成对转录组-蛋白组时间进程数据集。结果显示，mRNA-蛋白质相关性有限（ρ=0.54），但在没有复杂的转录后调控的前提下，蛋白质翻译和降解的数学模型能描述达84%的基于mRNA动力学的蛋白质时间序列。本研究假设RNA结合蛋白Hrb98DE参与糖的转录后控制早期胚胎发育中的代谢，并使用Hrb98DE敲出炎症了这一假设。本研究提出了一个可用于从大规模、时间序列的转录组和蛋白组数据中鉴定转录后基因调控的系统生物学框架。

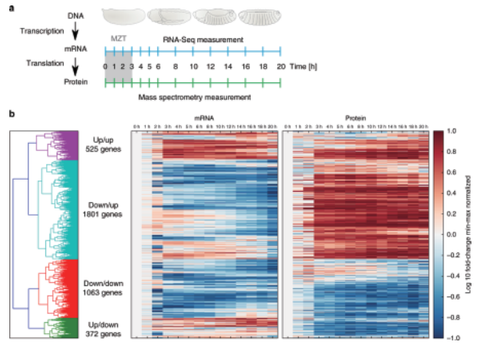


图1 果蝇胚胎发生的转录组&蛋白组分析结果

a.运用RNA-Seq和质谱检测的方法在果蝇胚胎发育期间测定配对的mRNA和蛋白质的时间点；b. mRNA和蛋白质时间序列的热图。

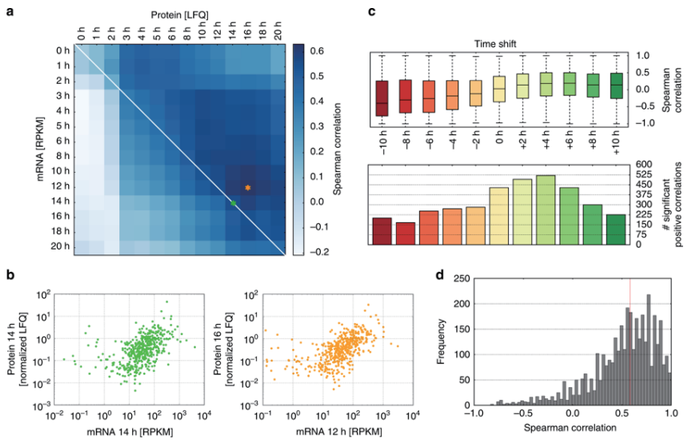


图2 mRNA和蛋白质相关性分析结果

a. 所有样品的RNA-蛋白质全谱相关性分析；b. 最大的RNA-蛋白质全谱相关性分析；c. 个体基因mRNA和蛋白质时间序列局部相关性；d. 时间序列下，最大的斯皮尔曼相关系数分布。

3.2 案例二 通过多组学数据展示不同发育阶段小鼠胃的分子数据[9]

哺乳动物的胃在结构上呈现高度多样化，其功能严重依赖于正常的胚胎发育。虽然前人已经报道对胃发育期间形态变化的研究，但缺乏对潜在分子变化进行系统化分析。本课题展示了小鼠胃在多个发育阶段的全面转录组和蛋白组图谱。课题基于蛋白质和RNA的变化水平，对三个不同阶段的12108个基因产物进行了定量分析，获得纵向时间尺度上胃功能分子标记。转录组分析发现了与发育相关的重要亚型，并在肽段水平上功能验证了未注释的新可变剪切转录本，蛋白组分析发现了胃发育中差异表达的蛋白质在弥漫型胃癌中也显著表达。最后，得出：胃的发育和胃癌肿瘤发生的信号通路密切相关的结论。

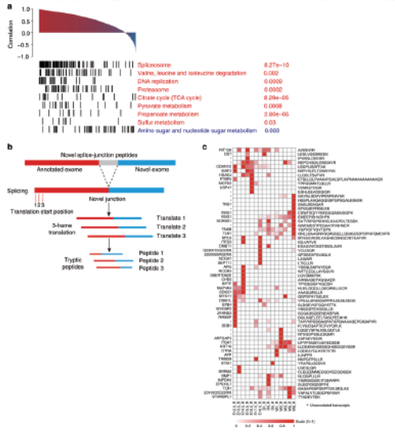


图3 小鼠胃蛋白质组和转录组整合分析

a.根据所有时间点下每个基因产物的标准化分数（z-score），通过Pearson相关系数，计算蛋白质-RNA相关性。红色表示正相关基因及其富集途径，蓝色表示负相关基因及其富集途径；b.发现新型剪接点多肽的途径；c.所有15个时间点上，60种新型剪接体多肽的相对丰度分布。

可能存在的风险

蛋白组和转录组关联分析后，发现蛋白组和转录组相关性不高，约为27%~40%[12]。这种情况是比较常见的，主要原因是转录和翻译的速率不同，mRNA和蛋白质的半衰期也不同，mRNA和蛋白质的丰度并非总是线性相关。研究者可通过代谢通路分析，研究mRNA和蛋白质的上下调关系，分析关键的调控基因，深入研究代谢通路的精细调控关系，结合基因组层面的突变信息，寻找与表型相关的完整代谢调控通路。

常见问题

1、蛋白组和转录组生物重复是否需要一致？

答：是的，推荐蛋白组和转录组选取的生物重复一致。如果项目设计无法保证二者的生物重复完全一致，也可以开展关联分析，需要分别提供两个组学层面的相关信息，如鉴定定量到的差异表达基因，表达量，p-value等。

2、非华大的转录组数据或者蛋白组数据是否能用华大的分析流程进行关联分析？

答：可以，其他公司的转录组数据有基因序列文件和转录组定量文件可进行关联分析；蛋白组数据需要蛋白组鉴定序列文件、蛋白组定量文件可进行关联分析。

3、转录组和蛋白组相关性不高，应该怎样解释和开展后续分析？

答：需要区分具体是哪种类型的相关性系数不高，建议按照常见的几大类开展解释和后续分析：

1）蛋白和mRNA表达趋势相同的相关性不高：说明表达一致性不高，这种情况很常见，主要原因是转录和翻译的速率不同，mRNA和蛋白质的半衰期也不同，mRNA和蛋白质的丰度并非总是线性相关；后续可通过其他实验方式进行关键基因表达产物的验证；

2）蛋白和mRNA表达趋势相反的相关性不高：此类表达情况主要说明一些特殊的调控关系，如反馈抑制调节方式的蛋白，相关性系数并无高低好坏的判定；后续分析可根据差异富集的关键代谢通路，逐一查看相关的蛋白和mRNA的表达情况，帮助分析上下游基因的代谢调控关系；

3）仅单一组学表达有差异，或者两组学层面均无差异的情况相关性不高：此类表达情况主要是前两种情况的补充，相关性系数也并无高低好坏的判定；后续分析也是需要结合代谢通路的上下游调控关系进行代谢调控关键因子的分析，避免基因表达产物mRNA和蛋白一对一的相关性关系，拓展到更大范围的对通路或者功能条目的影响中分析将会获得更多可能的结论。

华大优势

强大的数据兼容性：无障碍兼容多种蛋白组和转录组数据类型，测序、芯片等具有定量信息的数据均可分析，推荐采用蛋白定量iTRAQ、IBT、DIA和RNA-Resequencing数据进行关联分析；

更细致的关联分类：按照蛋白组和转录组基因表达相关性趋势将数据细分为5大类，对于每类相关联的结果提供更完整的相关性分析和功能注释信息；

升级的通路关联方式：借助最新版KEGG数据库丰富的通路信息，将原本的蛋白组和转录组鉴定、定量层面的关联拓展到功能和代谢通路的关联，对分析关键基因对上下游相关蛋白的调控作用，以及迅速找到两个组学层面差异富集的关联代谢通路起到决定性作用；

华大一贯的高品质体验：完善的多组学研究技术，可提供一站式解决方案及相关技术服务；提供高质量的分析结果，报告结构科学、图片清晰可直接用于文章发表，有效提升报告阅读体验。

参考文献

[1] Marylyn D. Ritchie, Emily R. Holzinger, Ruowang Li, et al. Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics. 13 January 2015.

[2] Wang XC, Li Q, Jin X, et al. Quantitative proteomics and transcriptomics reveal key metabolic processes associated with cotton fiber initiation. J Proteomics. 2015 Jan 30; 114:16-27.

[3] Trevisan S, Manoli A, Ravazzolo L, et al. Nitrate sensing by the maize root apex transition zone: a merged transcriptomic and proteomic survey. J Exp Bot. 2015 Apr 23.

[4] Yang N, Xie W, Yang X, et al. Transcriptomic and proteomic responses of sweetpotato whitefly, Bemisia tabaci, to thiamethoxam. PLoS One. 2013 May 9;8(5): e61820.

[5] Chen Q, Guo W, Feng L, et al. Transcriptome and proteome analysis of Eucalyptus infected with Calonectria pseudoreteaudii. J Proteomics. 2015 Feb 6; 115:117-31.

[6] Christine Vogel and Edward M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Reviews Genetics. 13 March 2012.

[7] Wang SH, You ZY, Ye LP, et al. Quantitative proteomic and transcriptomic analyses of molecular mechanisms associated with low silk production in silkworm Bombyx mori. J Proteome Res. 2014 Feb 7; 13(2): 735-51.

[8] Becker K, Bluhm A, Casas-Vila N, et al. Quantifying post-transcriptional regulation in the development of Drosophila melanogaster[J]. Nature communications, 2018, 9(1): 4970.

[9] Li X, Zhang C, Gong T, et al. A time-resolved multi-omic atlas of the developing mouse stomach[J]. Nature communications, 2018, 9(1): 4910.

[10] Chen Z, Wen B, Wang Q, et al. Quantitative proteomics reveals the temperature-dependent proteins encoded by a series of cluster genes in thermoanaerobacter tengcongensis. Mol Cell Proteomics. 2013 Aug; 12(8): 2266-77.

[11] Long M, Zhao J, Li T, et al. Transcriptomic and proteomic analyses of splenic immune mechanisms of rainbow trout (Oncorhynchus mykiss) infected by Aeromonas salmonicida subsp. salmonicida. J Proteomics. 2015 Jun 3; 122:41-54.

[12] Muers M. Gene expression: Transcriptome to proteome and back to genome. Nat Rev Genet. 2011 Jun 28; 12(8):518.

环境微生物群落多样性研究方案

研究背景

微生物是地球上已知种类最多、数量最大、分布最广的生物类群，仅原核微生物的总量大约就达4×10^{30}-6×10^{30}个。由于大多数微生物尚不能纯培养,传统的微生物研究方法,如显微镜微形态观察、选择性培养基计数、纯菌种分离和生理生化鉴定等,在微生物多样性研究中都存在很大的局限性。基于非培养基础上的分子生物学方法可以使人们快速、系统地分析环境样品中微生物组成、结构和多样性,极大地促进了微生物生态学的发展。Zuckerkandl 等首次提出使用基因序列作为分子钟来分析生物间的亲缘关系[1]。Woese 和Fox 基于16S rRNA 基因序列对原核生物的系统进化关系进行了分析,提出了著名的 “三域学说”[2] 。从此,16S rRNA 基因成为了最常用的生物标志物,广泛应用于微生物的系统进化、分类及多样性研究中。基于16S rRNA信息的系统分类结果与基于全基因组信息的分类结果很相似[3] 。随着测序技术的发展,人们可以更加快捷地获得环境样品中的16S rRNA 基因序列,这些序列信息可以和数据库中的已知信息进行比对,以研究环境样品中微生物群落的特点。

基于16S，18S，ITS或功能基因的物种多样性和丰度分析主要应用于宿主肠道、土壤、水体等环境中，而绝大部分的研究均是比较不同环境或不同条件下物种组成和丰度上的差异性[4-8]，所以针对不同环境来源的样品，物种组成差异始终是分析的重点。

在物种组成和丰度差异上，请关注能够明显显示样品间差异的分类水平，如“科”水平，进而进行相应科的背景的阐述和总结，从而关联物种组成差异和环境适应性。

在样品复杂度的分析上，主要关注单个样品内部的多样性，样品两两之间多样性的比较，以及样品间聚类的分析结果。

如果样品进行了分组分析，OTU丰度PCA分析，物种丰度热图以及Beta多样性热图，都可以直观反映出来组间有无差异。如果属于同一组别的样品在这些分析中均聚类在一起，说明了同一组别的样品在物种组成上具有相似性。可从统计学的角度找出两个组别差异的物种。

解决的科学问题

1) 医学领域：人体微生物与人体健康/疾病的关系，人体微生物对疾病干预过程的影响；

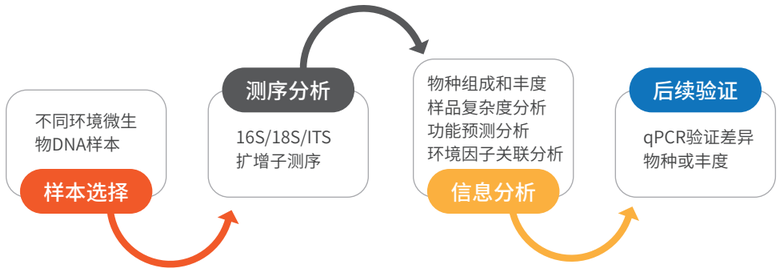
2) 动物领域：肠道、瘤胃（如产甲烷菌类群）与动物健康/营养消化研究等；

3) 农业领域：根际微生物与植物互作、农业耕作/施肥处理与土壤微生物群落等；

4) 环境领域：微生物与雾霾处理、污水治理、石油降解、酸性矿水处理、海洋环境的关系；

5) 特殊极端环境：极端环境条件下的微生物类群研究，如冰川、火山等。

方案设计



2.1 样本选择建议

2.1.1针对不同环境样本推荐不同的取样方法

如不能立即进行基因组DNA提取，将样品迅速置于-80度冰箱保存，每次取一小份进行提取，避免反复冻融 。

1）肠道组织样本采集

a) 组织样本用无菌磷酸盐缓冲液轻轻清洗，直到没有内容物流出；

b) 用无菌的显微镜玻片刮取附着在表面的组织细菌，转移到无菌的2.0mL离心管中；

c) 立即转入-80℃低温保存，送样时选择干冰运输寄送。

2）土壤meta样品采集

a) 根据研究目的确定采样范围，取样器具要事先消毒灭菌处理，开始采样；

b) 去除表面浮土，使用乙醇火烧的铲子挖取地下5~20cm的土层；

c) 去除可见杂质后，土壤过2mm筛网，建议每个样品从3个及以上采样点采集并混合而成，把土样装入无菌2.0mL离心管中，每管取约50~100mg（约花生米大小，装入2.0mL离心管不超过1/3体积）到无菌的2.0mL离心管中，每个样本取3-5管备份；

d) 分装好后，立即转入-80°C低温保存，送样时选择干冰运输寄送。

3）水体样本采集

a) 根据研究目的确定采样深度和范围；

b) 采集好的水样需要通过滤膜进行过滤，可以根据水样的浑浊程度选择相应孔径的滤膜；

c) 将滤膜转移到2.0mL离心管中，立即转移至-80°C低温保存，送样时选择干冰运输寄送。

清亮水样：可选择小孔径的滤膜，一般选0.22μm或0.45μm的滤膜，过滤水样体积大于10L；

浑浊水样：过滤前静置分离悬浮颗粒，也可以用大孔径的滤膜预过滤一遍，再用小孔径的滤膜进行过滤。

2.1.2样品组及样本个数需满足特定分析要求

样本间多样性分析（样品数≥4），样品组间显著性差异分析（组别≥2，每个组样品数≥3）。

2.2 采用的技术

采用16S/18S/ITS扩增子测序或全长16S测序技术，分析物种组成和丰度及样品复杂度并根据需求进行功能预测、环境因子关联分析等。

2.3 测序参数

16S/18S/ITS扩增子测序建议简单环境（如肠道、发酵液等）项目每个样本≥50,000tags，复杂环境（土壤、海水等）项目每个样本≥100,000tags；

全长16S测序建议每个样本≥6000 reads。

2.4 分析结果

2.4.1 物种及其丰度分析

通过与数据库进行比对，对OTU进行物种分类并分别在门、纲、目、科、属、种几个分类等级对各个样品作物种profiling面积图和柱状图,可以直观看出不同物种在每个样品中所占的比例。

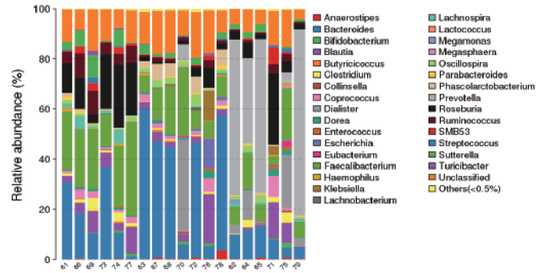


图1 样品Genus分类水平中物种profiling柱状图

根据物种或样品丰度相似性进行聚类得到heatmap图，并在聚类结果加上样品的处理或取样环境分组信息，可以直观的观察到相同处理或相似环境样品的聚类情况，直接反映样品群落组成的相似性和差异性。

基于样品中的物种以及丰度的分析结果构建物种的进化树，可以更深一步了解样品中物种的进化关系。枝长的长短表示进化距离的差异，系统关系越近的物种，在进化树种距离越近。

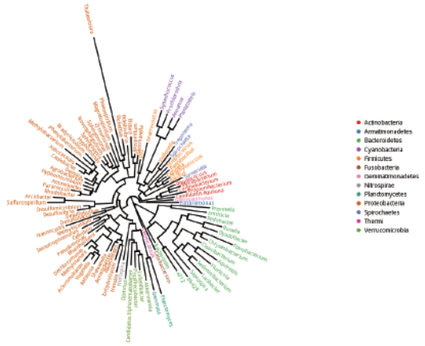


图2 属水平物种系统进化树(相同颜色属名代表相同的门)

2.4.2单个样品多样性分析

通过Alpha多样性分析研究单个样品中物种多样性，包括observed species指数（sobs）、chao1指数、ace指数、shannon指数和simpson指数等。前四个指数越大，最后一个指数越小，说明样品中的物种越丰富，多样性也越高。

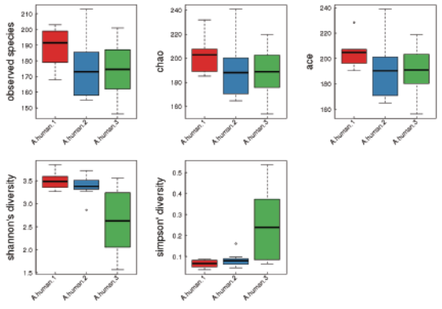


图3 组间Alpha多样性盒形图

2.4.3样品间多样性比较分析（n>=4）

Beta多样性（Beta diversity）分析是用来比较一对样品在物种多样性方面存在的差异大小。分析各类群在样品中的含量，进而计算出不同样品间的Beta多样性值。多种指数可以衡量Beta多样性，常用的为Bray-Curtis，weighted UniFrac，unweighted UniFrac，这些指数值越大表示样品间的差异越大。Bray-Curtis距离是反映两个群落之间差异性的常用指标，该计算不考虑序列间的进化距离，只考虑样品中物种存在情况；UniFrac是通过利用系统进化的信息来比较样品间的物种群落差异，其计算结果可以作为一种衡量beta diversity的指数，考虑了序列间的进化距离，其中weighted UniFrac考虑了序列的丰度，unweighted UniFrac不考虑序列丰度。

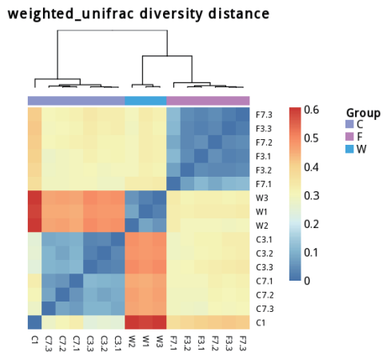


图3 组间Alpha多样性盒形图

主坐标分析（Principal coordinates analysis，PCoA）可进一步展示样品间物种多样性差异大小。PcoA结果中两个样品距离较近，则表示这两个样品的物种组成较相似。

2.4.4LEfSe组间群落差异分析

检测分组样品中最显著丰度差异的物种作为Biomarker。识别不同丰度的特征以及相关联的类别。通过生物学统计差异使其具有强大的识别功能，并执行额外的测试，以评估这些差异是否符合预期的生物学行为。

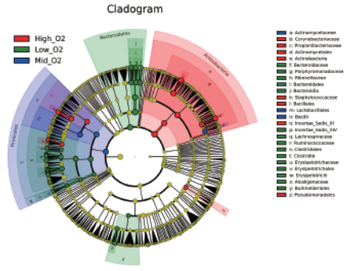


图5 LEfSe分析结果

2.4.5 16S测序样品功能预测

很多细菌自身具有不同数目的16SrRNA 拷贝数，我们利用PICRUSt 软件对由16S测序分析得到的OTU丰度进行拷贝数均一化，由此得到样品中可能出现的细菌及数目，从细菌的基因组信息得到对应的基因信息及注释信息，再结合均一化的OTU丰度来预测样品中可能存在的各级KEGG通路及丰度值以及COG功能信息及丰度值。也可以通过样品间比较分析功能差异。

由于该功能信息是通过多步骤预测得到的，准确性较低，特别是差异分析往往并不能得到有效结果；如果要深入挖掘微生物群落基因功能，建议结合生物学验证或通过宏基因组学进行研究。

2.4.6物种间相关系数网络图分析

在组数不多但样品数较多的情况下，可通过物种间相互关系网络图分析展示组间/样品间核心物种和物种间相关性，参考推测核心物种在样品特征维持上所起的作用。该分析基于物种在单个样品中的相对丰度信息，利用CCREPE算法计算各物种之间的相关系数，展示菌群之间的关系与核心物种。

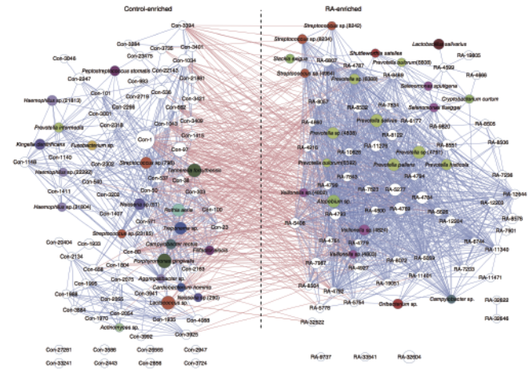


图6 物种间相互关系网络图[9]

2.5 项目执行周期

样品检测合格后，建库+测序+标准信息分析：约30个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.6 预期的结果

利用扩增子研究手段，借助高通量测序平台，通过对不同组别、不同样本进行比较分析，研究物种丰度及样品复杂度，基于环境微生物与各种分类因素之间的关系进行深入挖掘。

2.7 后期验证手段

挑选目标或者差异物种，通过qPCR进行验证，看看测序结果是否和qPCR结果一致[10]。

应用案例

案例一：肠道菌群衍生物细菌素可增强早期断奶仔猪腹泻抵抗力[11]

在猪养殖过程中，早期断奶可以缩短猪的屠宰周期并改善母猪的繁殖性能。然而早期断奶容易导致应激性腹泻，仔猪死亡率上升，生长性能降低。使用抗生素可以预防仔猪断奶腹泻，降低饲养成本，但是由于病原菌抗生素抗性和抗生素残留问题，欧盟已完全禁止在动物饲养中使用抗生素。因此，寻找抗生素替代品以预防早期断奶仔猪的腹泻对于畜牧业和粮食安全至关重要。哺乳动物肠道菌群与宿主健康密切相关，通过粪菌移植或益生菌/益生元调控肠道菌群已成为有前景的胃肠道疾病治疗策略。

与商业杂交LY仔猪相比，CM仔猪（中国本土品种）对早期断奶应激诱导的腹泻抵抗力更强。本研究在早期断奶之前给LY仔猪口服CM仔猪粪便微生物群，LY仔猪的腹泻抗性增强。通过比较粪菌移植组和对照组LY仔猪肠道微生物群的相对丰度，鉴定到两个可能跟腹泻抗性相关的菌种加氏乳杆菌LA39（Lactobacillus gasseri LA39）和乳酸杆菌（Lacto-bacillus frumenti），并通过qPCR进行验证。腹泻抵抗依赖于细菌素gassericin A，gassericin A与角蛋白19（KRT19）在肠上皮细胞质膜上的结合对于增强液体吸收和减少分泌至关重要。本研究结果表明L. gasseri LA39和L. frumenti可能是预防哺乳动物腹泻的有效抗生素替代品。

方案设计：

对LY仔猪和CM仔猪按不同处理进行分组如下：

①LY：LY仔猪，不经任何处理，n=3；

②LY（saline）：LY仔猪，day10-day18隔日口服生理盐水，n=3

③LY（high dose）：LY仔猪，day10-day18隔日口服高浓度CM仔猪粪菌悬液，n=3；

④LY（low dose）：LY仔猪，day10-day18隔日口服低浓度CM仔猪粪菌悬液，n=3；

⑤LY（oxytetracycline）：LY仔猪，断奶日（day21）肌肉注射长效土霉素，n=3

⑥CM：CM仔猪，不经任何处理，n=3

以上各组仔猪在断奶后第3，5，6，8，11天收集粪便样本，进行16S V4和ITS2测序。

主要结果：

1. 粪菌移植仔猪腹泻症状缓解，肠道菌群结构和功能发生改变

1）接受CM粪菌移植的LY仔猪早期断奶症状缓解。

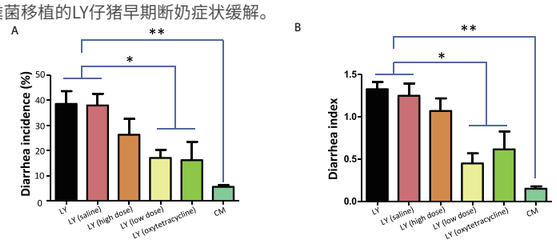


图7 各组仔猪腹泻发病率（A）和腹泻指数（B）。

2）LY(saline)组，LY(low dose)组和CM组粪便细菌有明显差异；真菌无明显分群。

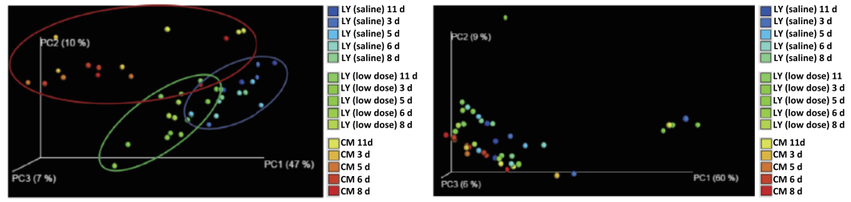


图8 细菌和真菌群落PCoA 结果

3）接受CM粪菌移植的LY仔猪细菌/真菌群落多样性发生了改变。

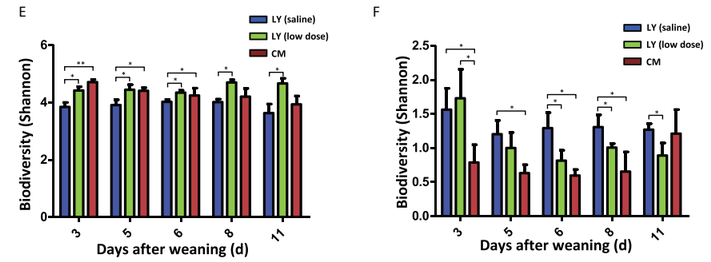


图9 细菌和真菌alpha diversity 结果

4）LY(saline)组，LY(low dose)组和CM组粪便细菌功能结构有明显差异；

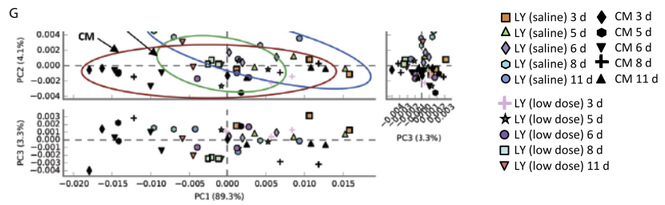


图10 仔猪肠道菌群功能pCoA图（基于PICRUSt软件 KEGG pathway分析）

2. 口服腹泻抗性相关的肠道微生物可预防仔猪早期断奶应激引起的腹泻。鉴定得到5个跟腹泻抗性相关的菌种，其中加氏乳杆菌LA39（*Lactobacillus gasseri LA39*）和乳酸杆菌（*Lactobacillus frumenti*）单独服用即可增强仔猪腹泻抗性。

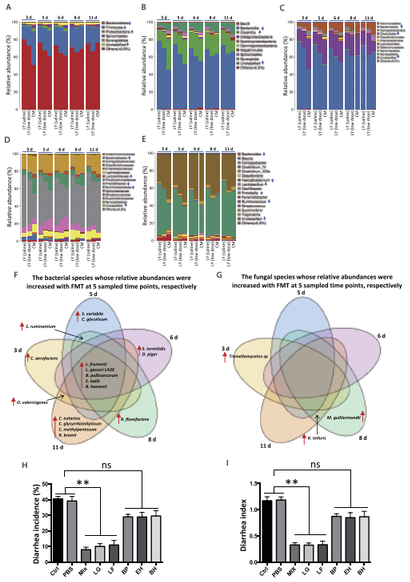


图11 A-E, 三组仔猪肠道菌群物种注释结果（门、纲、目、科、属）；F, G, 粪菌移植仔猪肠道菌群变化物种韦恩图（F, 细菌；G, 真菌）；H, I, 特定菌株对仔猪腹泻发生率（H）和腹泻抗性（I）的影响。

3. 肠道微生物介导的抗腹泻功能依赖于分泌型Gassericin A，混合菌群移植和单独服用LG/LF菌株的LY仔猪肠道菌群Gassericin A编码基因GaaA丰度升高；口服GaaA缺陷LG菌株不能改善LY仔猪腹泻症状。

4. Gassericin A与肠上皮细胞中的细胞质膜结合，增强肠液吸收，减少肠液分泌。

5. KRT19介导的Gassericin A与细胞质膜结合，对于Gassericin A介导的肠液吸收增强和液体分泌减少至关重要。

6. Gassericin A通过激活由雷帕霉素机制靶标介导的磷酸二酯酶活性降低细胞周期核苷酸水平，增加肠液吸收并减少肠液分泌。

案例二：黄海海岸线地下水的细菌多样性[12]

通过16S MiSeq测序的方法研究黄海沿海井水和孔隙水中细菌的多样性，发掘潜在能够降解污染物的细菌群体。

海岸线地下水是沿海与大海交换营养物、金属、碳等物质的重要载体，这一交换过程对沿海的生态系统会产生有害影响，而微生物在这一生物地球化学过程中发挥着重要的作用。本研究选取4个井水样本和3个孔隙水样本进行扩增子测序（16S, V4-V5），并分析每个样品的含盐量、温度、pH值等等。共得到1078个OTUs，其中23个OTU普遍存在于井水样本中，169 OTU存在于所有孔隙水样品中，Alpha多样性结果表明井水中细菌多样性低于孔隙水。放线菌（Actinobacteria）和beta变形菌（Betaproteobacteria）主要存在于井水样本中；而Gamma-变形菌(Gammaproteobacteria)、蓝细菌（Cyanobacteria）和浮霉菌（Planctomycetes）主要存在于孔隙水样本中；鞘脂杆菌（Sphingobacteriia），纤维粘网菌（Cytophagia）和黄杆菌（Flavobacterria）在两组样本中均存在。微生物群落和环境因子关联分析结果表明井水和孔隙水为沿海水域提供不同的营养成分，而井水中的潜在关键菌群睾丸酮丛毛单胞菌（Comamonas testosteroni）是一种很好的海岸线地下水污染物生物降解候选菌。

案例三：双重氧化酶基因BdDuox调节果蝇肠道菌群平衡[13]

双重氧化酶基因BdDuox在调节果蝇肠道菌群平衡中发挥着重要的作用。

果蝇肠道中存在种类丰富的微生物、包括有益共生菌、非共生菌、食源性微生物和病原菌等种类，正常情况下，这些微生物保持相互制约的平衡关系；本研究将探索宿主免疫系统对肠道菌群平衡的影响。昆虫肠道免疫系统主要依靠两种主要的效应分子协同作用来抑制外来入侵微生物的扩增和繁殖：一种是由双重氧化酶催化产生的微生物杀伤活性氧类（Reactive oxygen species, ROS），另一种是免疫缺陷信号转导途径（Imd）产生的抑菌肽（antimicrobial peptides, AMPs）。本研究中作者主要研究双重氧化酶基因（BdDuox）在调节果蝇肠道菌群免疫平衡中的作用。研究结果显示BdDuox基因的敲低会导致一定程度的果蝇肠道菌群失调：总体微生物含量增加，肠杆菌属和明串珠菌科菌群丰度相对降低；肠道菌群失调会反馈调节免疫系统：活化BdDuox，促进ROS产生，抑制肠道有害细菌的过度繁殖，调节肠道菌群组成和结构趋于正常。

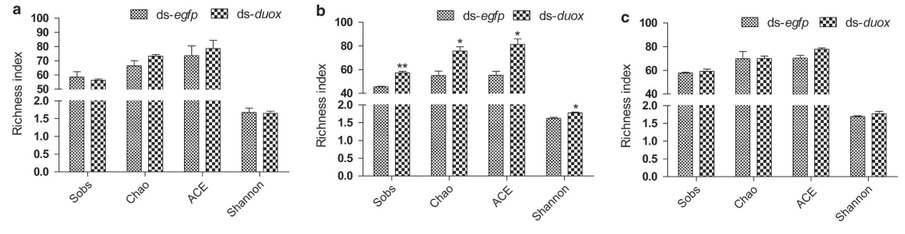


图12 BdDuox基因缺陷影响果蝇肠道菌群多样性

可能存在的风险

分类物种可能比较少，该分类跟样品复杂度、数据库的完整程度有关；物种定量结果有偏差，各物种16S拷贝数不同会直接引起这一偏差；损失低丰度物种信息，低丰度序列可能会由于PCR偏向性或在分析过程中被当作错误人为丢掉导致损失。

虽然16S rRNA 基因被广泛应用于微生物多样性分析中,然而对于某些属的菌群分辨效果较差,如Vibrio、Pseudomonas等[12]。人们一般设定16S rRNA 基因序列相似性≥97%的原核生物为同一个种,但很多不同种的微生物其16S rRNA 基因序列相似性却高于97%。另外,许多细菌的16S rRNA 基因是多拷贝的,而且各拷贝的序列组成存在一些差异;16S rRNA 基因也存在着水平转移问题,这些都直接影响着对原核微生物群落结构和多样性的分析。

常见问题

1. 16S 产品有样品数量的要求，需要所有样品准备好了才能进行测序分析吗？

答：从科学的角度来讲，最好能够整批样品同时测序分析，既可以减少不同批次间的系统误差，还能节省项目周期。若样品准备有困难，也可以分批次启动，数据分析中需要注意由此可能带来的系统误差。

2. 16S测序一般推荐多大数据量？

答：对特定区域扩增子测序推荐数据量：简单环境（如肠道、发酵液等）一般推荐≥50,000tags，复杂环境（如土壤、海洋等）需相应增加数据量，一般推荐≥100,000tags。

全长16S测序推荐数据量：≥6000 reads。

3. 16S测序一般推荐多少样本量?

答:16S样本多样性及组间差异分析是基于统计结果进行的分析,一般样本数越多,统计结果越准确。最低样本数要求如下:样本间多样性分析(n≥4),组间多样性分析样本(组别≥2,每组样本数n≥3)。

4. 16S测序能不能进行功能分析?

答:16S测序主要是基于16S rDNA 序列相似性进行OTU聚类进而进行物种注释及相关多样性研究。因为16S测序并没有测到对应物种的基因组信息,不能直接基于测序结果进行功能注释。

利用软件PICRUSt可以进行16S功能预测,该软件的原理是:对由16S测序分析得到的OTU丰度进行拷贝数均一化,得到样品中可能出现的细菌及数目,从细菌的基因组信息得到对应的基因信息及注释信息,再结合均一化的OTU丰度来预测样品中可能存在的各级KEGG通路及丰度值以及COG功能信息及丰度值。

基于16S的功能预测可以作为后续功能研究提供参考,但由于该分析不能反映群落中因基因表达差异导致的功能差异。如果主要关注功能差异,最好选择宏基因组测序来进行功能研究。

5. 常用的扩增区域、扩增引物及测序策略(供参考)

	扩增区域	引物名称	引物对	测序策略*
细菌	V4	515F	GTGCCAGCMGCCGCGGTAA	PE250
		806R	GGACTACHVGGGTWTCTAAT	
	V3-V4	341F	ACTCCTACGGGAGGCAGCAG	PE300
		806R	GGACTACHVGGGTWTCTAAT	
真菌	ITS1	its1	CTTGGTCATTTAGAGGAAGTAA	PE250
		its2	GCTGCGTTCTTCATCGATGC	
	ITS2	its3	GCATCGATGAAGAACGCAGC	PE300
		its4	TCCTCCGCTTATTGATATGC	

注:只有两端完全测通的Reads (Tags)才能用于进一步的分析,因此不同的扩增区域请严格遵循对应的测序类型。

6. 16S相关文章中选择的测序区域各不相同(如V4, V3-V4等),选择的依据是什么,选择哪个区域比较好?

答:不同物种不同区域多样性不同,选择不同区域测序结果会有不同,可能会造成物种多样性的低估或高估。在非全长16S测序的情况下,测序区域也并非越长越好,跟全长16S结果最相近的测序区域即是最优选择。

根据我们大量的项目经验,目前测序项目较多的区域为V4和V3-V4,具体项目测序区域建议参考相关文献进行选择。

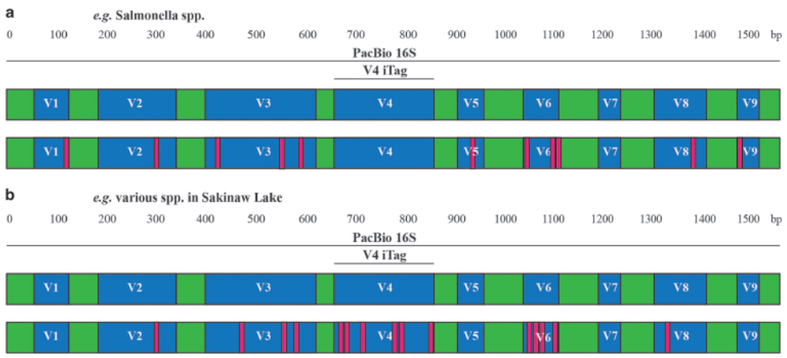


图13 不同物种16S各可变区变异程度不同

7. 16S样本要求有哪些,样本准备有哪些注意事项?

答:华大基因对16S样本要求如下:

Meta 扩增子测序					
样本类型		总量	浓度	完整性(胶图)	纯度
Meta rDNA Amplicon	Genomic DNA	≥0ng(推荐 50ng以上)	≥0ng/μl	必须为基因组样本	无蛋白,RNA/盐离子等污染,样本无色透明不粘稠
Meta rDNA Amplicon PCR-free Library	PCR products	≥3μg	≥30ng/μL	条带清晰无弥散	

一般16S建库选择基因组样本,推荐样本量在50ng以上,如果样本准备困难,大于0ng也可以尝试建库。

16S产品建库受模板DNA和体系纯度(含有杂质,盐离子,色素,腐殖酸等)等因素影响,在取样过程中,尽量减少宿主细胞含量及其他杂质的影响。

样本采集后尽快放入-80°C冻存,干冰运输,减少样本降解导致的微生物群落结构变化。

8. 我的样本检测合格了,样本量也达到了推荐样本量要求,却建库失败了,可能是什么原因造成的,有没有什么解决方案?

答:这种情况常见于宿主微生物样本,该类样本通常还有大量的宿主DNA,由于样本检测中不能区分宿主和微生物DNA,实际检测到的DNA其实绝大多数都是宿主DNA,导致检测到的样本量很高,却建库失败的情况。

对于这类样本,推荐在样本制备过程中进行特殊处理,尽量减少宿主DNA含量。目前市面上有可以去宿主DNA的试剂盒(如QIAamp DNA Microbiome Kit),对于宿主含量较高的样本如唾液、粘膜样本等,可以选择对应的试剂盒处理。

华大优势

策略多样: 不同来源样本采用不同提取方法和建库测序策略,满足多种环境研究需求

平台多样: 有不同的短读长和长读长测序平台,可满足16S/18S/ITS不同高变区域或全长测序的需求。

经验丰富: 已测序样品类型涉及粪便、土壤、水体、唾液、牙菌斑、体腔、胃液、白带、空气、血液、皮屑等。

样本需求量低: 华大基因16S产品推荐DNA样本量50ng以上,样本量需求低于同行其他公司要求;对于样本获取困难的样本,只要样本量高于0 ng也有可能建库成功。

便捷的信息分析系统: 华大微生物扩增子分析流程化繁为简,支持单个项目、多个项目联合分析,获取原始数据更加便利;方案提交自由掌控,支持多次修改方案和分组任务操作,自动化运行产出整套结果;分析结果美化精进,支持新建重分析任务,结果图片个性化修改,展示更加灵活。不需具备生信基础也能轻松投递任务,并且兼容16S、ITS、18S及多种定制化扩增子项目。

参考文献

[1] Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. Journal of Theoretical Biology, 1965, 8(2): 357-366.

[2] Woese C R, Fox G E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America, 1977, 74(11): 5088-5090.

[3] Rinke C, Schwientek P, Sczyrba A, Ivanova N N, Anderson I J, Cheng J F, Darling A, Malfatti S, Swan B K, Gies E A, Dodsworth J A,Hedlund B P, Tsiamis G, Sievert S M, Liu W T, Eisen J A, Hallam S J, Kyrpides N C, Stepanauskas R, Rubin E M, Hugenholtz P, Woyke T. Insights into the phylogeny and coding potential of microbial dark matter. Nature, 2013, 499(7459): 431-437.

[4] McCafferty J, Mühlbauer M, Gharaibeh RZ, et al. (2013) Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. ISME J. 7(11):2116-25.

[5] Zhao L, Wang G, Siegel P, et al. (2013) Quantitative genetic background of the host influences gut microbiomes in chickens. Sci Rep. 3:1163.

[6] Rubin BE, Gibbons SM, Kennedy S, et al. (2013) Investigating the impact of storage conditions on microbial community composition in soil samples. PLoS One. 8(7):1:6.

[7] Mao Y, Xia Y, Zhang T. (2013) Characterization of Thauera-dominated hydrogen-oxidizing autotrophic denitrifying microbial communities by using high-throughput sequencing. Bioresour Technol. 128:703-10.

[8] Peng X, Yu KQ, Deng GH, et al. (2013) Comparison of direct boiling method with commercial kits for extracting fecal microbiome DNA by Illumina sequencing of 16S rRNA tags. J Microbiol Methods. 95(3):455-62.

[9] Zhang, X., et al., The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med, 2015. 21(8): p. 895-905.

[10] Xu J, Lian F, Zhao L, et al. Structural modulation of gut microbiota during alleviation of type 2 diabetes with a Chinese herbal formula[J]. The ISME journal, 2015, 9(3): 552.

[11] Hu J, Ma L, Nie Y, et al. A Microbiota-Derived Bacteriocin Targets the Host to Confer Diarrhea Resistance in Early-Weaned Piglets[J]. Cell host & microbe, 2018, 24(6): 817-832. e8.

[12] Ye, Q., et al., Bacterial Diversity in Submarine Groundwater along the Coasts of the Yellow Sea. Front Microbiol, 2015. 6: p. 1519.

[13] Yao, Z., et al., The dual oxidase gene BdDuox regulates the intestinal bacterial community homeostasis of Bactrocera dorsalis. ISME J, 2016. 10(5): p. 1037-50.

[14] Pascual J, Macian M C, Arahal D R, Garay E, Pujalte M J. Multilocus sequence analysis of the central clade of the genus Vibrio by using the 16S rRNA, recA, pyrH, rpoD, gyrB, rctB and toxR genes. International Journal of Systematic and Evolutionary Microbiology, 2010, 60(1): 154-165.