植物学报 *Chinese Bulletin of Botany* 2017, **52** (4): 530–538, www.chinbullbotany.com
doi: 10.11983/CBB16113

·专题论坛·

新疆天山雪莲转录组注解知识库

李锦[1], 严潘瑶[2], 钱飞箭[2], 邱宝义[2], 夏雯雯[1], 牟晓威[2], 邱丽娟[3]
林忠平[4], 陈铭[5], 祝建波[1*], 陈贤丰[1, 2*]

[1]石河子大学生命科学学院, 石河子 832003; [2]圣庭生物信息研究所, 台州 318020; [3]中国农业科学院作物科学研究所, 北京 100081; [4]北京大学生命科学学院, 北京 100871; [5]浙江大学生命科学学院, 杭州 310058

摘要 新疆天山雪莲(*Sasussured involucrata*)具有较高的极端低温耐受特性, 为低温耐受机制研究提供了一种非常好的模式植物。新疆天山雪莲转录组注解知识库(http://www.shengtingbiology.com/SaussureaKBase/index.jsp)是基于网络数据资源的综合性数据库, 由html、Perl、Perl CGI/DBD/DBI、Java和JavaScript编程所设计的前端界面和用于数据存取、注释及管理的后端数据库管理系统PostgrelSQL构成。知识库包含基因组数据、转录组原始数据、质量控制数据、GC含量、功能基因序列及注释、功能基因代谢通路、功能基因的注释统计、雪莲与其它物种的转录组或基因组比较分析数据和生物分析软件包等资源。该数据库不仅有利于低温功能基因组学及低温耐受机制研究, 而且为冷耐受性状物种的分子育种提供基因资源平台和理论依据。

关键词 天山雪莲, 知识库, 转录组, 低温, 抗逆, 功能基因

李锦, 严潘瑶, 钱飞箭, 邱宝义, 夏雯雯, 牟晓威, 邱丽娟, 林忠平, 陈铭, 祝建波, 陈贤丰 (2017). 新疆天山雪莲转录组注解知识库. 植物学报 **52**, 530–538.

冷害和冻害是限制农业生产的重要自然灾害, 全世界每年由于低温造成的农业经济损失达数千亿元。低温不仅限制植物的种植范围, 而且还会造成其减产和品质下降, 严重时甚至导致绝收。迄今为止, 由于大多数农作物来源于热带和亚热带, 不具备抗寒特性, 极易面临极端天气下的冷害或冻害的威胁。因此, 植物的耐低温机制研究具有重要的经济和科技效益。目前研究最为透彻的是模式植物拟南芥(*Arabidopsis thaliana*)的冷驯化机制(Jaglo-Ottosen et al., 1998; Song et al., 2012)。然而, 拟南芥作为一种低温非敏感型物种, 在研究上具有一定的局限性。不同生态型物种的直系同源基因在功能和调控机制上往往具有一定的差异, 这些差异的大小直接影响植物的抗寒能力。番茄(*Lycopersicon esculentum*)作为一种冷敏感物种, 与拟南芥相比, 虽然也具有完整功能性的CBF(一种冷驯化调控通路的上游转录因子), 但是却有着完全不同的下游基因(Zhang et al., 2004)。雪莲(*Sasussured involucrata*)生长在极端环境, 其基因往往具有更强的抗逆能力。如*SikPIP3*和*SikSCAPD*在转化烟草(*Nicotiana tabacum*)实验中, 前者表现出高抗干旱和低温的能力, 后者的抗低温能力也较强(焦天奇等, 2012; Liu et al., 2015)。目前, 极地物种的低温耐受机制研究鲜有报道, 新疆天山雪莲作为一种在极端低温环境下生存的物种, 可以很好地补充拟南芥作为研究材料在低温耐受基因功能和信号调控机制研究中的不足。

新疆雪莲隶属菊科凤毛菊属, 为多年生一次性开花结实的高山冰缘草本植物, 生长在海拔2 400–4 100 m高山悬崖峭壁的雪线附近(庄丽等, 2006)。雪莲生境的月平均气温很低, 日温较季节性温度变化幅度大, 通常在一天之内要经历类似夏天到冬天的急剧温度变化。生理生态研究表明, 雪莲具有极强的耐寒、抗辐射和抗缺氧的能力, 能在月平均温度3–5°C的环境下快速生长(陈发菊, 1999; 何涛等, 2007)。

新一代测序(next generation sequencing, NGS) (Mardis, 2011)技术的出现给生命科学的发展带来巨大变化, 测序的低成本、高精度和高准确度使得一些小型和中型实验室也可以独立完成。这种新技术的变

收稿日期: 2016-05-25; 接受日期: 2016-09-18
基金项目: 国家自然科学基金(No.C020406)
* 通讯作者。E-mail: zjbshz@126.com; jeff_chen@shengtinggroup.com

革为一些非模式生物的测序提供了可能, 如棉花(*Gossypium arboreum*)的基因组测序(Li et al., 2014)、人参(*Panax ginseng*)的转录组测序(Jayakodi et al., 2014)和辣椒(*Capsicum annuum*)的基因组测序(Kim et al., 2014)等。测序技术的成熟使得越来越多的物种序列信息得以揭示, 出现了许多专一性物种基因信息数据库, 其中包括牦牛(*Bos mutus*)基因组数据库(The Yak Genome Database, YGD) (Hu et al., 2012)、油菜(*Brassica campestris*)基因组数据库(The Genetics and Genomics Database for Brassica Plants, BRAD) (Cheng et al., 2011)、豇豆(*Vigna unguiculata*)基因组数据库(An Annotation Knowledge Base for Cowpea Genomic, CGKB) (Chen et al., 2007)、烟草转录因子数据库(The Database of Tobacco Transcription Factors, TOBFAC) (Rushton et al., 2008)和杨树(*Populus*)基因组数据库(The Salinity Tolerant Poplar Database, STPD) (Ma et al., 2015)等。这些数据库为科研工作者获取各种有用信息提供了很大的帮助。当今气候的极端变化使得植物低温响应研究成为一大热点, 同时也取得了重要成果, 但是有关低温响应机制研究的生物数据库却很少, 目前只有美国的ColdArrayDB (https://cold.dpb.carnegiescience.edu/cgi-bin/data.cgi)。而新疆天山雪莲独有的生理特征可以填补国内相关研究信息的不足。

根据我们前期的预实验, 测得新疆天山雪莲的基因组大小约为2 GB, 转录组约为200 MB, 共计199 758条转录本, 其中臆测蛋白数量为60 501 (Li et al., 2017)。数据的注释是基于blastn和blastx两种水平下与目前常用的6个公共生物数据库(NCBI NR Peptide Dataset (ftp://ftp.ncbi.nlm.nih.gov/blast/db/)、EBISwissProt (http://www.uniprot.org/downloads)、Pfam (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release)、KEGG (http://www.kegg.jp/kegg/download/)、GO (http://geneontology.org/page/downloads)和COG (http://www.ncbi.nlm.nih.gov/COG/))中的相关序列进行比对分析。目前已经有多个物种的基因组数据公开发表, 并且已经建立了相关的网络资源数据库。通过在蛋白或核苷酸水平上对新疆天山雪莲的转录组与上述物种的基因组数据进行一对一的比对及注释, 可以为不同物种在转录组或蛋白组水平的比较提供更精确的思路。新疆天山雪莲转录组注解知识库(*Sasussured involucrata* Transcriptome Knowledge Base, SITKB: http://www.shengtingbiology.com/SaussureaKBase/index.jsp)提供了基于序列同源性和HMM (Hidden Markov Model, http://hmmer.janelia.org/)所获得的基因注释数据与序列等资源, 用户可以按需提取天山雪莲相关的资源信息。

1 SITKB的构建与数据库的内容

1.1 数据的处理与分析

新疆天山雪莲(*Sasussured involucrata* Kar. et Kir.)材料来源于实验室组织培养的无菌苗, 所有处理以冷驯化、非冷驯化与极端高温3种方式进行, 具体低温处理方式以4°C (24小时)、4°C (7天)、从4°C (7天)移至0°C (24小时)、从0°C (24小时)移至–2°C (24小时)为冷驯化处理; 从常温(20°C)直接移至–2°C (12小时)和–10°C (12小时)作为非冷驯化处理; 以及从常温(20°C)移至40°C (24小时)为极端高温处理。利用Illumina公司的HiSeq 2500平台对天山雪莲总RNA进行测序。初始序列经过FastQC (参数为默认设置) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)去除低质量和污染的序列得到clean read序列后, 通过Trinity (Version: 2014-07-17; 参数为默认设置) (Grabherr et al., 2011)从头组装共获得199 758条转录本, 其中60 501条属于臆测蛋白。

转录组的注释通过blast (参数为1×10^{-10})和基于HMM (Hidden Markov Model; 参数为默认设置) (https://en.wikipedia.org/wiki/Hidden_Markov_model)算法来达到蛋白和核苷酸2种水平的注释效果。在核苷酸水平上, 采用Trinity软件中的TransDecoder (参数为默认设置)来预测基因的CDS区域, 并通过InterProScan来检测功能域的注释。在上述6个公共生物数据库中利用Blastn和Blastp两种方式进行比对分析和注释, 其中NCBI NR注释结果用于GO注释的获取, Pfam和InterPro共同注释的数据用于KEGG注释。KOG或COG的直系同源基因分析用于检测整个转录组数据的完整性和基因的功能性质。其中, 多种数据处理与分析通过Perl (http://www.perl.org/)编程语言来实现。所有数据处理和分析的运算均在圣庭集

团与石河子大学共建的中国冷应激耐受系统生物学研究中心的超级计算机系统中完成。

1.2 转录组比较

目前，许多植物的转录组或基因组数据已经公开发表，如拟南芥、水稻(*Oryza sativa*)的基因组和转录组，以及豌豆(*Pisum sativum*)基因组数据等。新疆天山雪莲的基因组测序尚未完成，只是前期完成了一个10倍数据量的测定，从而获得天山雪莲基因组大小和GC含量(未发表数据)。对于雪莲的转录组数据，通过blastx (期望值为$1e^{-8}$)对拟南芥蛋白质组、水稻蛋白质组和大豆(*Glycine max*)蛋白质组等进行比对分析和数据整合。除了利用六大公共数据库(NCBI NR、SwissProt、InterPro、RefSeq、Gene Ontology和UniProt)进行功能注释外，新疆天山雪莲数据库(SITKB)还整合了10种植物的基因组数据(表1)，初步实现了雪莲基因与其它物种基因直系同源关系的鉴定。

1.3 基于同源比对的注释

测序获得的新疆天山雪莲转录组序列通过blastx (期望值为$1e^{-8}$)与NCBI GenBank Proteins、UniProtKB-TrEMBL、UniProtKB-Swiss-Prot和UniProtKB-PIR (Protein Information Resource, http://pir.georgetown.edu/)中的fasta格式化的无冗余蛋白质序列进行比对分析。共有59 629条功能基因的转录本得到注释，98.6%的功能基因转录本得到精确的比对结果。

1.4 基于HMM的基因模型和功能域分析

为了验证组装后的转录组的完整程度和注释的精确性，利用HMMER软件包(参数为默认设置)，通过Pfam和InterPro数据库对新疆天山雪莲转录组的功能基因进行功能域分析。Pfam和InterPro数据库是一种用于蛋白质功能域分析的资源数据库，通过结合在2种数据库中的分析结果，可以获得更为准确的功能域分析结果。

1.5 SITKB数据库的结构和设计

新疆天山雪莲转录组注释知识库通过组织PostgreSQL (http://www.postgresql.org/)的相关数据库管理系统和梳理雪莲转录组序列信息而构建。数据库利用层次结构的方法进行设计，以图表和序列文件进行呈现(图1)。整个数据库采用MVC模式，依靠Java编程语言和Tomcat服务器呈现一种便捷易懂的可视化界面。通过整合各种注释结果进行基因信息展示，从而给用户展现一个有条理的可视化界面。

2 新疆天山雪莲转录组注解知识库(SITKB)功能介绍

为了使整个数据库的数据源更为整洁有序，源数据包

表1 新疆天山雪莲转录组注解知识库的物种特异性比对注释统计

Table 1 The statistic of the annotation by blast against special species of *Sasussured involucrata* transcriptome knowledge base

物种	比对上的基因数量	注释率(%)
大豆(*Glycine max*)	164940	82.57
水稻(*Oryza sativa*)	18101 (pep)	29.92
甜菜(*Beta vulgaris*)	58361/50002 (pep)	29.216/82.648
拟南芥(*Arabidopsis thaliana*)	11344 (pep)	18.75
莱茵衣藻(*Chlamydomonas reinhardtii*)	31638 (pep)	52.293
卷柏(*Selaginella moellendorffii*)	40281 (pep)	66.58
小立碗藓(*Physcomitrella patens*)	43298/52976 (pep)	26.52/71.557
无油樟(*Amborellaceae trichopoda*)	58449/46912 (pep)	29.26/77.54
芜菁(*Brassica rapa*)	64103/49989 (pep)	32.104/82.625
蒺藜苜蓿(*Medicago truncatula*)	64103/49989 (pep)	32.104/82.625

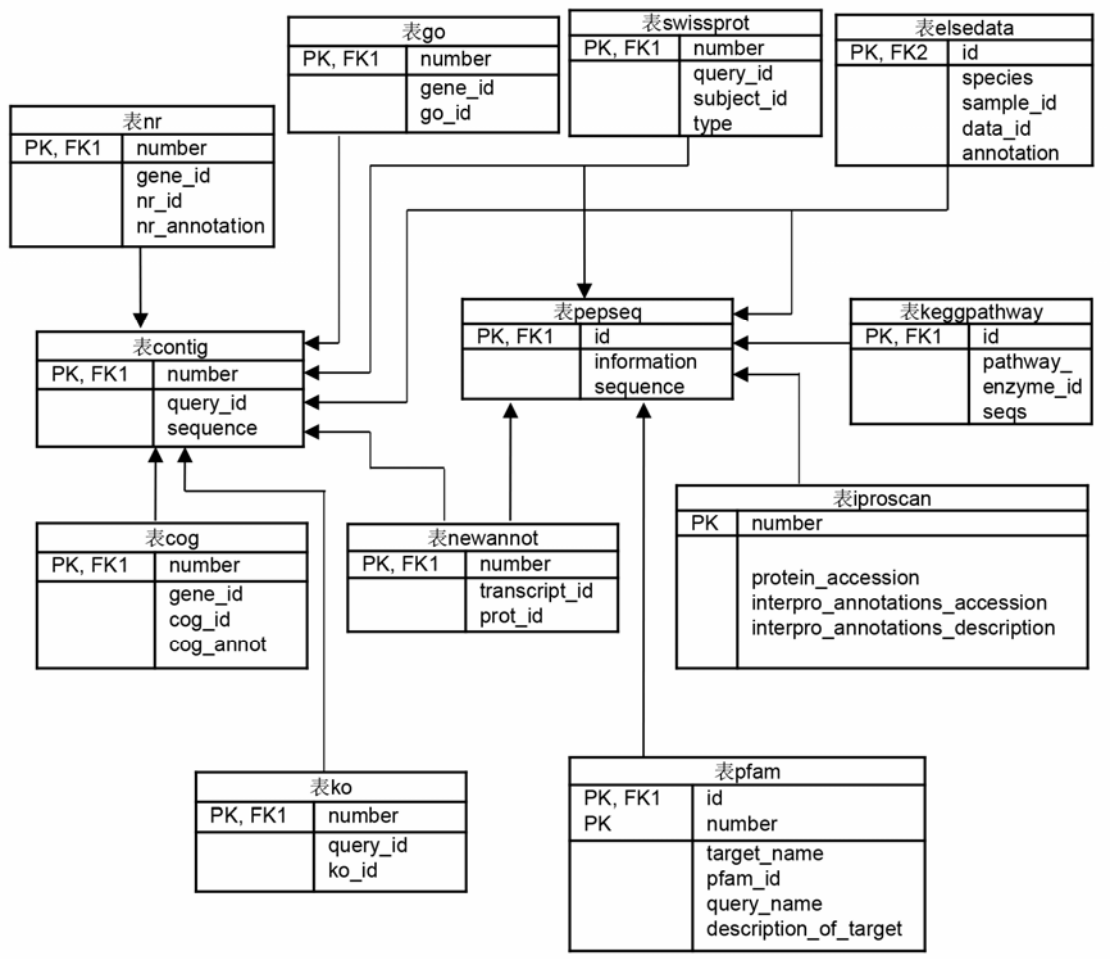

图1 新疆天山雪莲转录组注解知识库逻辑关系图

Figure 1 The logical relationship model of the management system within the *Sasussured involucrata* transcriptome knowledge base

括以下几种功能: 雪莲目的基因功能检索、雪莲目的基因代谢途径检索、雪莲基因注释统计及雪莲原始数据质量控制; 此外, 还包括雪莲的比较基因组学数据和相关单基因或基因组数据分析软件包等。通过Java编程语言将目标数据与Web界面相连接, 即MVC架构。

整个数据采用逻辑层次分布的方式, 将整个数据分为4个板块: 雪莲数据库(雪莲基因功能信息、代谢途径和源数据质量控制等)、相关数据(比较转录组或基因组学数据)、软件使用(生物信息分析相关软件)及其它(植物系统生物学知识库、豆科植物基因组学数据库和水稻数据库等, 见相关链接) (图2)。

2.1 雪莲数据库

将获得的新疆雪莲转录组序列利用blast和HMM的方法在上述6个公共生物数据库进行比对和分析, 从而获得雪莲基因转录本序列注释, 将注释结果进行整合梳理, 分成以家族和代谢途径(图3) 2种方式。

雪莲数据库具有优良的可视化界面, 用户可按照特定的需求检索相关信息。数据库中主要包含雪莲转录组注释的综合性结果, 具体信息根据不同功能家族或功能途径进行二级分类。此外, 数据库还包含新疆天山雪莲的转录组测序序列的碱基质量控制结果。雪莲转录组注释结果统计及整个转录组基因的聚类图表也在这里呈现, 方便用户了解雪莲转录组测序组装结果的完整性及注释的成功率。用户也可根据需求先检索基因的相关信息, 然后根据基因的相关信息获得自己所需的功能基因序列。具体的子板块有: (1) 雪莲目的基因的功能检索; (2) 雪莲目的基因代谢途径

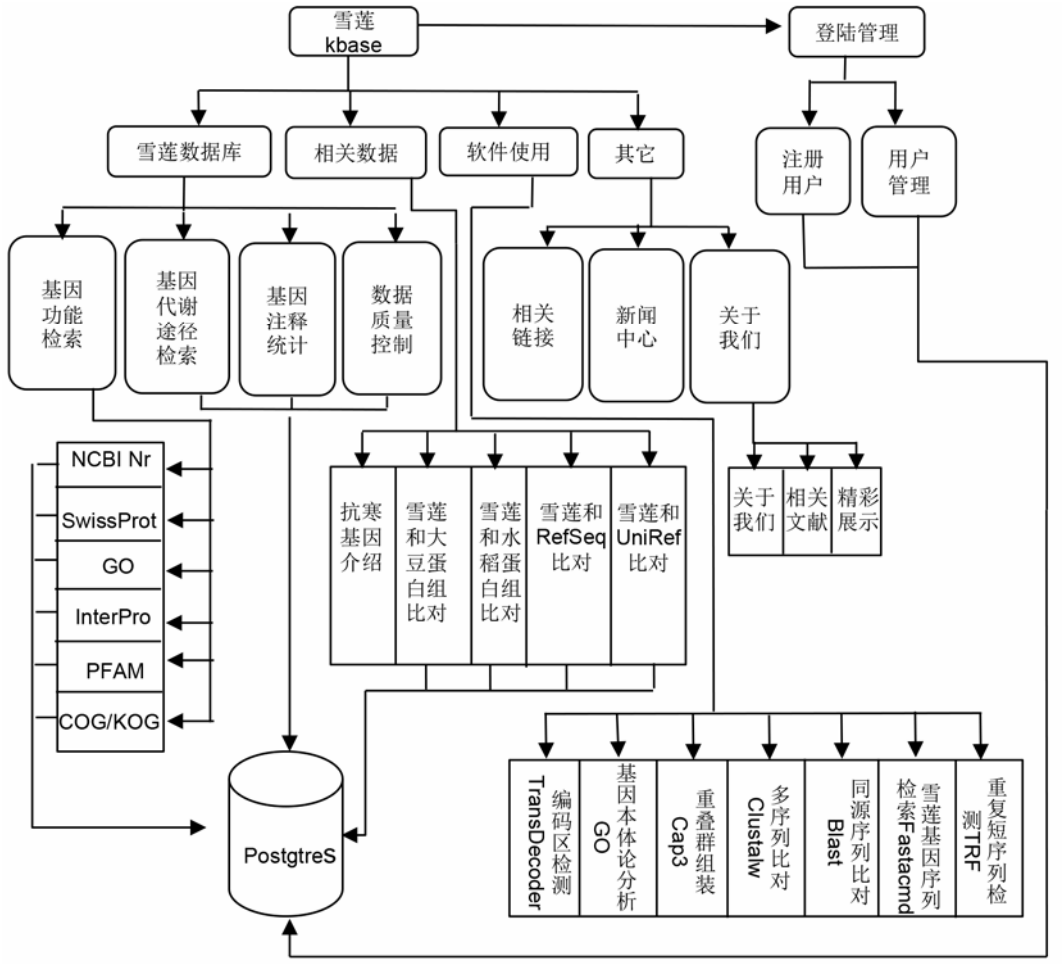

图2 新疆天山雪莲转录组注解知识库功能模块关系

Figure 2 The architecture of data functional model in the *Sasussured involucrata* transcriptome knowledge base

检索; (3) 雪莲基因空间聚类/注释统计; (4) 雪莲原始数据的质量控制。

2.2 相关数据

相关数据主要包括雪莲转录组与其它物种基因组的比对注释的结果。

2.2.1 植物抗寒基因的介绍

大多数植物需要经过驯化才能具有耐寒的性状，这些耐寒物种大多来源于温带或亚热带，热带植物则不具有耐寒特性。根据温度的不同可将驯化过程分为冷驯化和冻驯化。其中，温带和亚热带的物种主要受冷害的影响，而温带的作物则主要受冻害的影响。植物的耐寒机制研究中，对冷驯化机制的研究最为透彻，即*CBFs*调控机制。除此之外，植物的耐低温性状还与细胞的信号转导、细胞的转录调控(各种转录因子的低温表达模式)及细胞的转运蛋白(PIPs)等有关。

2.2.2 基于blast的物种基因组比对注释

基于blast的算法，将新疆天山雪莲的转录组序列与各大物种数据库的基因组序列进行比对分析，获得一对一的注释信息。具体的物种基因组数据库、物种名和数据的注释情况见表1。

2.3 基于web序列分析的软件包

SITKB提供了一些基于web的序列分析软件，用户可

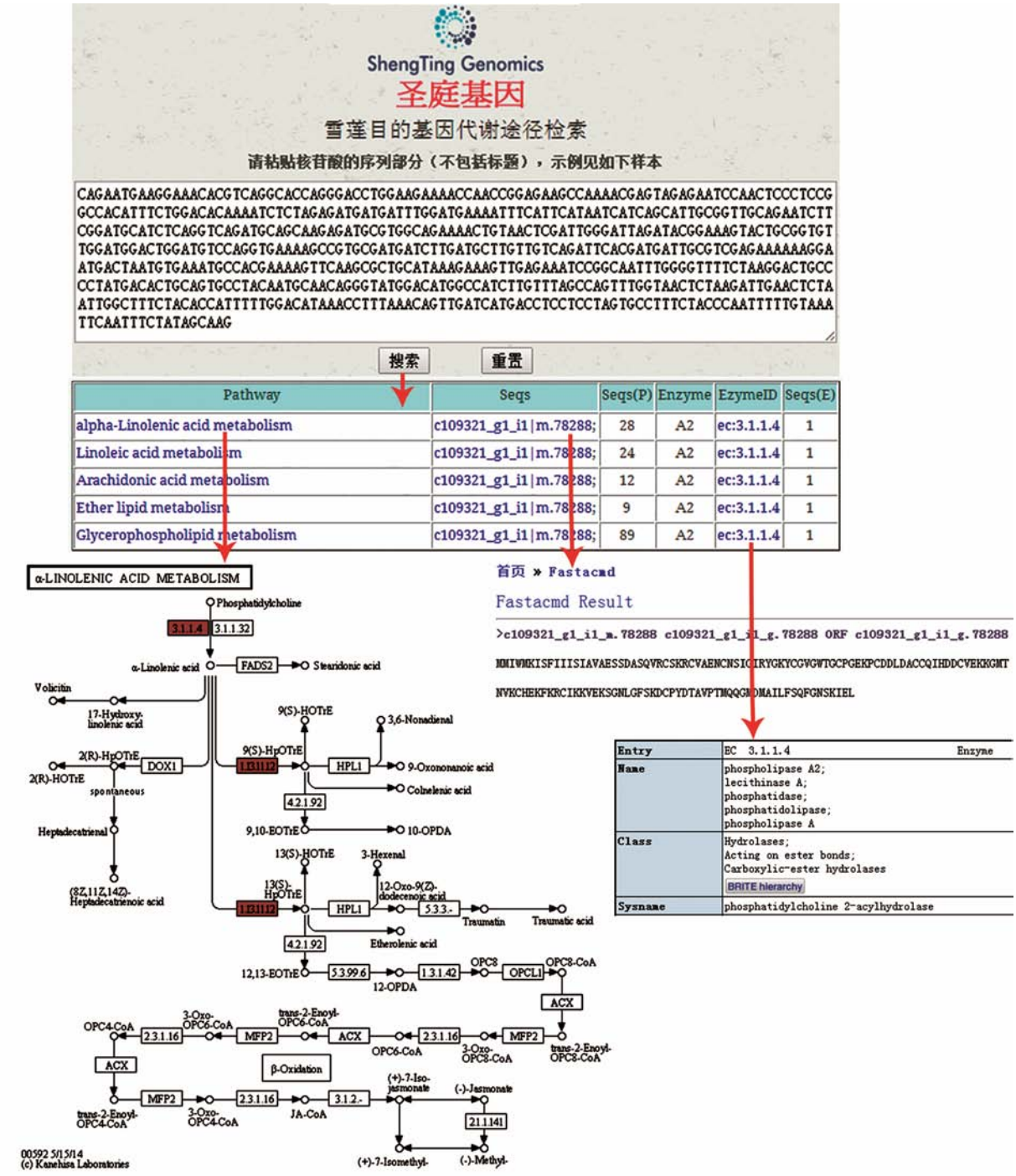

图3 新疆天山雪莲转录组注解知识库基因代谢途径检索示例

Figure 3 An illustration of gene metabolic pathway search in the *Sasussured involucrata* transcriptome knowledge base

以根据各自需求进行相关序列分析，并且可以选择特定的数据库进行分析(包括雪莲数据库)。这些软件包可以帮助用户快速分析相关基因的序列。

2.3.1 基因GO Slim检索

GO Slim软件由圣庭集团生物信息中心设计，主要目的是方便用户根据不同目的检索自己所需要的雪莲

基因GO注释信息(包括生物功能、生物进程和细胞组成的注释信息)。

2.3.2 编码区检测(TransDecoder)

TransDecoder由澳大利亚的邦科学与工业研究组织和美国的博大研究院共同开发，目的是寻找基因的开放阅读框，主要针对由Trinity软件拼接的基因序列，可以根据用户的需要对特定的雪莲基因进行开放阅读框的分析。通过对雪莲基因序列的编码区进行分析，使用户了解所获得基因序列的完整性。

2.3.3 同源序列搜索(BlastALL)

BlastALL是基于序列同源的比对软件，由美国的NCBI开发，目的是寻找基因序列间的一致性。利用该软件可以比对分析蛋白或核苷酸等雪莲数据。

2.3.4 雪莲基因检索(Fastacmd)

Fastacmd软件由圣庭集团生物信息中心设计，主要目的是方便用户检索所需的雪莲基因序列信息。用户可以在雪莲数据库模块中首先检索雪莲的基因ID编号和注释信息，进而对雪莲相关基因的序列信息进行检索。

2.3.5 多序列比对(ClustalW)

ClustalW是一种多序列比对软件，可以进行多条基因序列的比对，进而预测出各基因间的进化关系(即进化树的构建)，同时也可以根据序列的一致性区域寻找基因共有的保守区域。用户可以根据多序列比对结果对雪莲整个家族基因进行分析，获得家族基因的亲缘关系。

2.3.6 重叠群组装(Cap3)

Cap3是一套用于序列拼接的软件，此软件适用于小的数据集或EST拼接，可以更正拼接错误和连接contig。用户可以依据NCBI数据库中检索到的大量EST，通过Cap3软件将高度同源的EST进行拼接，从而获得完整和无冗余的EST基因集。

2.3.7 重复短序列检测(TRF)

TRF用于检测基因组序列的短重复序列，用户可以根据特定的参数来设定，如定位参数、匹配概率和插入概率等。在基因序列中，一般会存在不同程度的短序列重复，这种重复在基因拼接时必须先摒除，防止由于短序列重复造成拼接错误。此外，由于有些重复序列是基因功能所必需的，所以首先对重复序列进行检测，然后进一步分析其功能特性。

2.3.8 组学数据分析资源

基因组学分析工具里包含了全部二代测序所需要的基本基因组学分析工具，包括基因组组装软件包、 从头测序所需的组装软件包、基因组序列分析工具及各种文件格式转化软件。本数据库中提供了contig拼接、scaffold的组装以及各种基因注释等软件，用户可以根据数据库中所提供的URL超链接下载相关分析工具。

3 展望

基因组或转录组数据库对于科研工作非常重要，利用数据库数据，科研工作者可根据实验目的进行实验设计，并达到事半功倍的效果。目前，已经建立了许多重要经济作物的数据库。虽然目前低温生物学研究已经成为热点，但是研究的目标物种往往不具有在极端生境下生长的特性，或者只具有短时间的极端生境生长适应能力，这是低温生物学研究的不足。

本科研团队立足于新疆天山雪莲的低温生物学研究，弥补了前人研究的不足。新疆天山雪莲是一种极端低温生境下生存的物种，长期生活在高海拔的雪线附近赋予了天山雪莲一种特殊的耐寒机制，这种耐受性机制可能与已被广泛报道的拟南芥的冷驯化机制有很大差异。新疆天山雪莲转录组注释知识库的建立，为广大的低温生物学研究者提供了极端低温生境下的植物基因资源，方便科研工作者综合分析植物的低温耐受机制。

参考文献

陈发菊 (1999). 我国雪莲植物的种类、生境分布及化学成分的研究进展. 植物学通报 **16**, 561–566.

何涛，吴学明，贾敬芬 (2007). 青藏高原高山植物的形态和解剖结构及其对环境的适应性研究进展. 生态学报 **27**, 2574–2583.

焦天奇，孙辉，刘瑞娜，王爱英，祝建波 (2012). 转天山雪莲*sikPIP*$_3$基因烟草的获得及抗逆性鉴定. 西北植物学报 **32**, 431–438.

庄丽, 李卫红, 孟丽红 (2006). 新疆雪莲的资源利用、研发与保护. 干旱资源与环境 **20**, 195–202.

Chen XF, Laudeman TW, Rushton PJ, Spraggins TA, Timko MP (2007). CGKB: an annotation knowledge base for Cowpea (*Vigna unguiculata* L.) methylation filtered genomic genespace sequences. *BMC Bioinformatics* **8**, 129–137.

Cheng F, Liu SY, Wu J, Fang L, Sun S, Liu B, Li P, Hua W, Wang X (2011). BRAD, the genetics and genomics database for *Brassica* plants. *BMC Plant Biol* **11**, 136–141.

Grabherr MG, Brian J, Haas BJ, Yassour M, Joshua Z, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol* **29**, 644–652.

Hu QJ, Ma T, Wang K, Xu T, Liu JQ, Qiu Q (2012). The Yak genome database: an integrative database for studying yak biology and high-altitude adaption. *BMC Genomics* **13**, 600–604.

Jaglo-Ottosen KR, Gilmour SJ, Zarka DG, Schabenberger O, Thomashow MF (1998). Arabidopsis CBF1 overexpression induces COR genes and enhances freezing tolerance. *Science* **280**, 104–106.

Jayakodi M, Lee SC, Park HS, Jang W, Lee YS, Choi BS, Nah GJ, Kim DS, Natesan S, Sun C, Yang TJ (2014). Transcriptome profiling and comparative analysis of *Panax ginseng* adventitious roots. *J Ginseng Res* **38**, 278–288.

Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, Seo E, Choi J, Cheong K, Kim KT, Jung K, Lee GW, Oh SK, Bae C, Kim SB, Lee HY, Kim SY, Kim MS, Kang BC, Jo YD, Yang HB, Jeong HJ, Kang WH, Kwon JK, Shin C, Lim JY, Park JH, Huh JH, Kim JS, Kim BD, Cohen O, Paran I, Suh MC, Lee SB, Kim YK, Shin Y, Noh SJ, Park J, Seo YS, Kwon SY, Kim HA, Park JM, Kim HJ, Choi SB, Bosland PW, Reeves G, Jo SH, Lee BW, Cho HT, Choi HS, Lee MS, Yu Y, Do Choi Y, Park BS, van Deynze A, Ashrafi H, Hill T, Kim WT, Pai HS, Ahn HK, Yeam I, Giovannoni JJ, Rose JK, Sørensen I, Lee SJ, Kim RW, Choi IY, Choi BS, Lim JS, Lee YH, Choi D (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat Genet* **46**, 270–278.

Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, Chen W, Liang X, Shang H, Liu W, Shi C, Xiao G, Gou C, Ye W, Xu X, Zhang X, Wei H, Li Z, Zhang G, Wang J, Liu K, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu S (2014). Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat Genet* **46**, 567–572.

Li J, Liu H, Xia W, Mu J, Feng Y, Liu R, Yan P, Wang A, Lin Z, Guo Y, Zhu J, Chen X (2017). *De novo* transcriptome sequencing and the hypothetical cold response mode of *Saussurea involucrata* in extreme cold environments. *Int J Mol Sci* **18**, 1155.

Liu HL, Shen HT, Chen C, Zhou XR, Liu HF, Zhu JB (2015). Identification of a putative stearoyl acyl-carrier-protein desaturase gene from *Saussurea involucrata* Kar. et Kir. *Biol Plant* **59**, 316–324.

Ma YZ, Xu T, Wan DS, Ma T, Shi S, Liu JQ, Hu QJ (2015). The salinity tolerant poplar database (STPD): a comprehensive database for studying tree salt-tolerant adaption and poplar genomics. *BMC Genomics* **16**, 205–211.

Mardis ER (2011). A decade's perspective on DNA sequencing technology. *Nature* **470**, 198–203.

Rushton PJ, Bokowiec MT, Laudeman TW, Brannock JF, Chen XF, Timko MP (2008). TOBFAC: the database of tobacco transcription factors. *BMC Bioinformatics* **9**, 53–59.

Song SY, Chen Y, Zhao M, Zhang WH (2012). A novel *Medicago truncatula* HD-Zip gene, *MtHB2*, is involved in abiotic stress responses. *Environ Exp Bot* **80**, 1–9.

Zhang X, Fowler SG, Cheng H, Lou Y, Rhee SY, Stockinger EJ, Thomashow MF (2004). Freezing-sensitive tomato has a functional CBF cold response pathway, but a CBF regulon that differs from that of freezing-tolerant Arabidopsis. *Plant J* **39**, 905–919.

The *Sasussured involucrata* Transcriptome Knowledge Base

Jin Li[1], Panyao Yan[2], Feijian Qian[2] , Baoyi Qiu[2], Wenwen Xia[1], Xiaowei Mou[2], Lijuan Qiu[3]
Zhongping Lin[4], Ming Chen[5], Jianbo Zhu[1*], Xianfeng Chen[1, 2*]

[1]College of Life Sciences, Shihezi University, Shihezi 832003, China; [2]ShengTing Bioinformatics Institute, Taizhou 318020, China; [3]Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China; [4]College of Life Sciences, Perking University, Beijing 100871, China; [5]College of Life Sciences, Zhejiang University, Hangzhou 310058, China

Abstract The *Sasussured involucrata* possesses high cold-tolerance capacity, which could be a good experimental model for investigating the cold tolerance mechanism. The *S. involucrata* Transcriptome Knowledge Base (SITKB) (http://www.shengtingsystemsbiology.com/SaussureaKBase/) is a web-based comprehensive database system implemented via html, Perl and Perl CGI/DBI/DBD, Java and JavaScript programming languages on the web front end and PostgrelSQL Relational Database Management System as the backend for data, annotation and knowledge management. SITKB systems was designated to facilitate cold-tolerance gene mining and knowledge discovery and hosts useful knowledge sets such as datasets of raw sequence data quality control, GC contents of raw and assembled DNA sequences, functional annotation of the assembled contigs, assignment of metabolic pathways to the contigs with coding potential, statistics on the rate of meaningful annotation, and blast comparison against other plant transcriptomes or genomes. The knowledge base system also contains a set of software packages for bioinformatics analysis against its stored datasets. This new knowledge base would provide a resource for data mining and knowledge discovery of the functional genes and pathways responsive to cold perturbation and also enhance the capability for elucidating the cold tolerance mechanism. It also offers genomic data resources and a theoretical foundation to facilitate the molecular breeding of cold tolerant crops.

Key words *Sasussured involucrata*, knowledge base, transcriptome, cold, stress resistance, functional gene

Li J, Yan PY, Qian FJ, Qiu BY, Xia WW, Mou XW, Qiu LJ, Lin ZP, Chen M, Zhu JB, Chen XF (2017). The *Sasussured involucrata* transcriptome knowledge base. *Chin Bull Bot* **52**, 530–538.

---

* Authors for correspondence. E-mail: zjbshz@126.com; jeff_chen@shengtinggroup.com

(责任编辑: 朱亚娜)