doi:10.3969/j.issn.1672-5565.2015.02.02

基于高通量测序的长穗偃麦草功能分子标记发掘和分析

刘 赢,张 军,敖 游,宋丽莉,束永俊*
(黑龙江省分子细胞遗传与遗传育种重点实验室 哈尔滨师范大学生命科学与技术学院,哈尔滨 150025)

摘 要:长穗偃麦草是小麦重要的近源物种,含有丰富的抗逆基因,广泛地应用于小麦的遗传改良育种。本研究利用高通量测序,获得长穗偃麦草的转录组测序信息,利用比较基因组学方法研究其与小麦、水稻和玉米等作物的遗传关系,评估它们之间的亲缘关系。同时,将长穗偃麦草的高通量序列比对到小麦基因,利用软件 Freebayes 和 SAMtools/Bcftools 发掘功能基因的变异位点,并对这些含有变异位点的功能基因进行注释分析,揭示长穗偃麦草优异性状形成的分子机制,这将为长穗偃麦草优异基因资源的开发和应用奠定重要基础。

关键词:长穗偃麦草;转录组分析;功能基因;单核苷酸多态性
中图分类号:Q524+.3 **文献标志码**:A **文章编号**:1672-5565(2015)-02-082-06

Genome-wide identification and characterization of SNPs from *Thinopyrum elongatum* using high-thought sequencing

LIU Ying, ZHANG Jun, AO You, SONG Lili, SHU Yongjun*
(*Key Laboratory of Molecular Cytogenetics and Genetic Breeding of Heilongjiang Province, College of Life Science and Technology, Harbin Normal University, Harbin* 150025, *China*)

Abstract: *Thinopyrum elongatum* is an important relative of common wheat because it harbors numerous biotic and abiotic stress-resistance genes. In this study, we used the high-through sequencing technology and comparative genome strategies to analyze the *Th. elongatum* transcriptome, and evaluated the evolution relationship between *Th. elongatum* and other cereal crops. Meanwhile, all sequences were aligned to wheat genes, and SNP sites were identified by software Freebayes and SAMtools/Bcftools. The genes containing SNP sites were annotated using COG, which was helpful for exploring molecular mechanism of excellent traits formation in *Th. elongatum*, and it provides a valuable reference for future development and utilization of excellent genes in *Th. elongatum*.

Keywords: *Thinopyrum elongatum*; Transcriptomic analysis; Functional gene; Single nucleotide polymorphism

长穗偃麦草是一种多年生植物,属于小麦族偃麦草属,主要分布于温带和寒带,具有大量优异的性状,如长穗、多花、籽粒蛋白含量高、抗病[1-2]、抗寒、抗旱、抗盐碱等[3-5]。而且,长穗偃麦草基因组与小麦基因组亲缘关系较近,杂交后代易结实,成为小麦遗传改良育种的重要基因资源库。研究人员采用杂交的方法,将长穗偃麦草中抗病、抗逆以及优质性状导入到小麦基因组中,改良小麦性状和品质,取得了巨大的成功[6-10]。但是,由于基因信息未知,严重阻碍了长穗偃麦草基因资源的开发和应用。

分子标记是作物遗传育种常用的一种重要工具,筛选与优异性状连锁的分子标记,用于指导作物育种过程,能加速作物遗传育种进程。长穗偃麦草含有大量抗逆、抗病以及品质相关基因,因此,如何开发与优异基因相关的分子标记对开发和利用长穗偃麦草优异基因资源具有深远意义。但是,由于长穗偃麦草基因信息缺乏,使得长穗偃麦草分子标记开发工作进展缓慢,如何开发长穗偃麦草优异基因

收稿日期:2015-03-30;**修回日期**:2015-05-05.
基金项目:高等学校博士学科点专项科研基金(20122329120001)。
作者简介:刘赢,女,黑龙江人,在读硕士,研究方向:植物功能基因组学;E-mail:15114599330@163.com.
*通信作者:束永俊,男,安徽人,副教授,硕士生导师,研究方向:植物功能基因组学;E-mail:syjun2003@126.com.

相关的分子标记将成为其基因资源开发和应用亟需解决的问题。

高通量测序是利用边合成边测序,具有测序成本低、通量高、速度快等优点,成为基因组测序研究的热点工具。基因组测序的信息量较大,测序工作复杂,不易完成;而转录组测序主要研究细胞内基因表达信息,测序信息量较小,且既可以获得基因的序列信息,又可以检测基因表达的信息,成为获得植物基因组信息的重要手段,特别是非模式植物基因组研究的重要工具[11-12]。研究人员利用转录组测序研究基因的表达情况,发掘和鉴定了大量优异基因资源,并根据测序信息开发基因靶向的分子标记,在很多植物上,如水稻[13]、玉米[14]、小麦[15]等,取得了重要进展,成为植物基因功能鉴定和分子标记开发的一种重要工具。

本研究将长穗偃麦草的转录组序列比对到小麦的 mRNA,发掘 SNP 位点,并对 SNP 靶向的功能基因进行注释分析,揭示长穗偃麦草优异性状形成的分子机理,将为长穗偃麦草优异基因资源在小麦遗传改良中应用奠定基础。

1　材料和方法

1.1　材料

长穗偃麦草(PI 531718, $2x=14$)种子由美国农业部农业研究服务的国家种质中心(http://www.ars-grin.gov/)惠赠。将种子萌发,种植 24 周后,分别采集长穗偃麦草的茎叶和根部组织,放入-80 液氮中冷藏。将长穗偃麦草样品送交华大基因公司(中国, 深圳),委托对样品进行转录组测序,包括对总 RNA 提取、纯化、cDNA 文库构建以及转录测序等工作,返回高质量的测序数据,具体测序信息参见文献[16]描述。

1.2　长穗偃麦草转录组的比较基因组分析

去除转录组序列中低质量序列,利用软件 Trinity[17]和 iAssembler[18]将剩下的高质量序列进行组装,获得长穗偃麦草 UniGene 序列。分别从 IWGSC[19](http://www.wheatgenome.org/)和 Ensembl Plants(http://plants.ensembl.org/index.html)下载小麦(*Triticum aestivum*)、拟南芥(*Arabidopsis thaliana*)、水稻(*Oryza sativa*)、高粱(*Sorghum bicolor*)、谷子(*Setaria italica*)、玉米(*Zea mays*)、短柄草(*Brachypodium distachyon*)、大麦(*Hordeum vulgare*)、粗山羊草(*Aegilops tauschii*)和乌拉尔图小麦(*Triticum urartu*)的基因组序列和注释信息。利用程序 BLASTN 将长穗偃麦草 UniGene 序列比对到这些植物基因组,评估长穗偃麦草与以上物种基因组间相似性。

1.3　长穗偃麦草 SNP 发掘及其功能注释

利用软件 Bowtie2[20]将长穗偃麦草高通量序列匹配到小麦 mRNA 上,然后分别用软件 Freebayes[21]和 SAMtools/Bcftools[22]进行 SNP 位点发掘。将 Freebayes 设置为"-C 3 -i -q 20",SAMtools/ Bcftools 设置为"varFilter -d 3",筛选出现三次或以上,且两种软件都发掘的突变位点作为候选 SNP 位点。提取 SNP 位点靶向基因信息,统计 SNP 位点在小麦基因组的分析情况。同时,利用软件 BLASTX 比对拟南芥等植物蛋白质序列,提取比对蛋白质的 COG 信息对 SNP 靶向基因进行功能分类和注释[23]。

2　结果分析

2.1　长穗偃麦草转录组的比较基因组分析

将长穗偃麦草转录组序列组装,形成169 990条 UniGene 序列。将这些 UniGene 与拟南芥、小麦等植物基因组进行比对,发现 47.20%(80 236/169 990)的 UniGene 至少含有一条相似基因。其中,与小麦间相似基因最多,总计达到65 552条,然后依次是:粗山羊草(51 521)、乌拉尔图小麦(51 518)、大麦(48 819)、短柄草(35 888)、水稻(25 648)、谷子(25 132)、高粱(22 956)、玉米(21 917),最少的为拟南芥,只有 970 条,数据表明长穗偃麦草与双子叶植物间亲缘关系较远,与单子叶植物亲缘关系较近,特别是小麦族植物,乃至小麦基因组亲缘关系最近。同时,对序列的相似性分析发现,长穗偃麦草基因与小麦基因高度相似(见图 1、图 2),大多数基因(55 472, 84.6%)的相似性超过 90%,表明长穗偃麦草基因组与小麦基因组极其相似,可以用于小麦遗传育种过程。

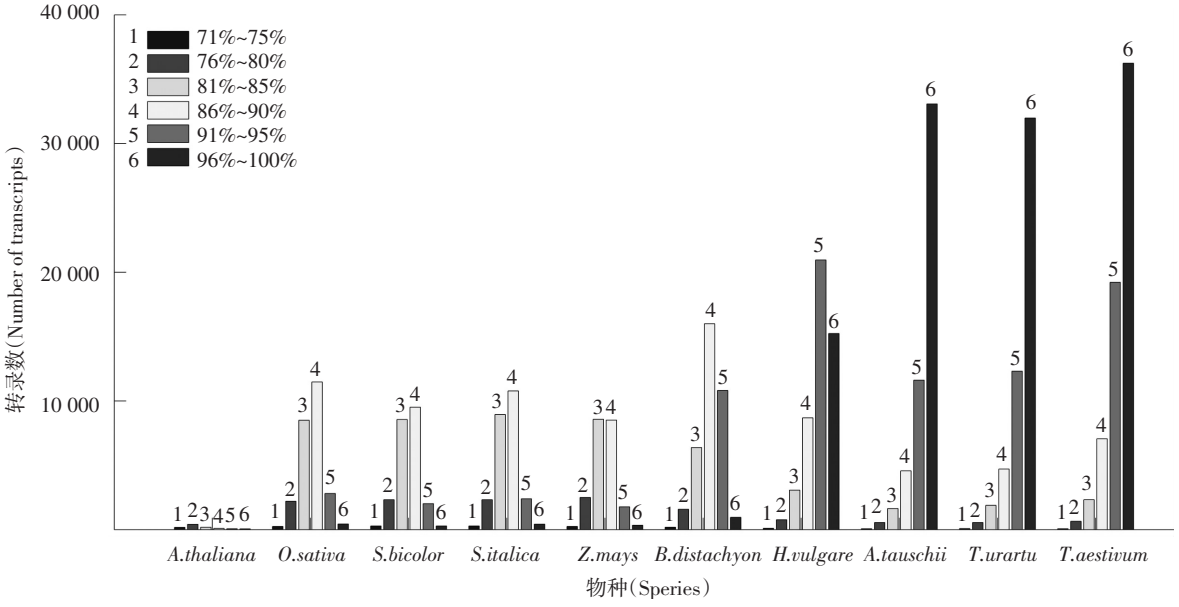

图 1 长穗偃麦草与其它植物基因组的亲缘关系分析

Fig.1 Relationship analysis of *Th. elongatum* genome with other plants

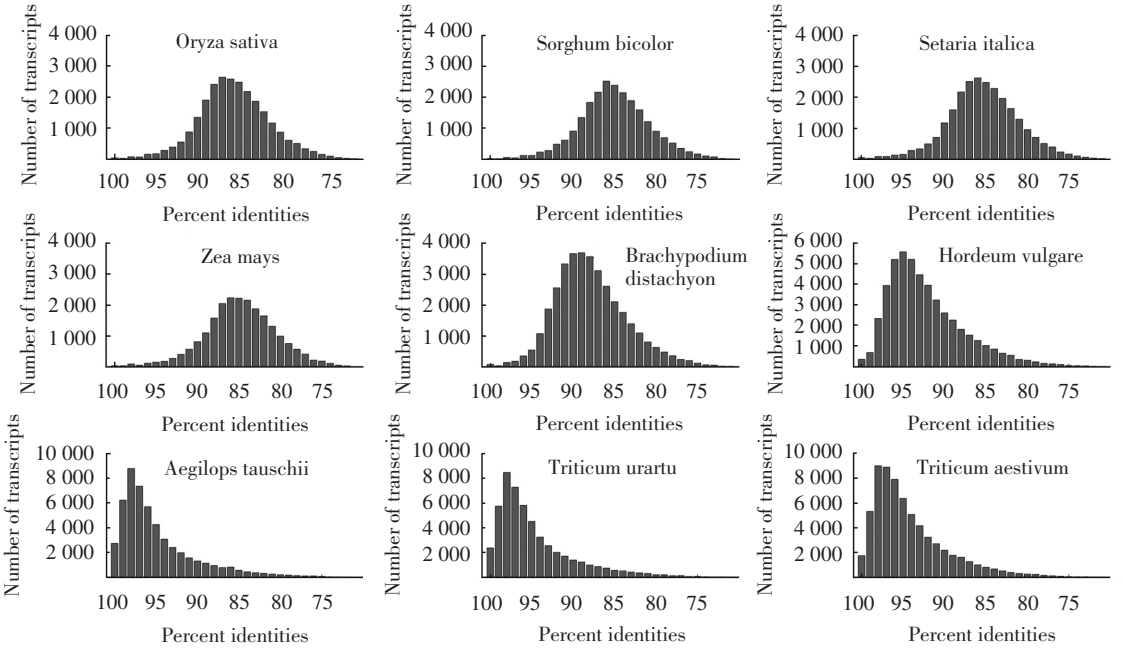

图 2 长穗偃麦草基因相似性分布

Fig.2 Similarity distribution of genes between *Th. elongatum* and other plants

2.2 长穗偃麦草 SNP 发掘和分析

将39 273 796条序列比对到小麦 mRNA 序列上,利用软件 Freebayes 发掘了606 660个 SNP 位点,软件 SAMtools/Bcftools 发掘了850 874个 SNP 位点,在两者中都出现的有561 147个 SNP 位点,将这些 SNP 位点确认为长穗偃麦草的候选 SNP 位点。在这些 SNP 位点中,主要是转换类型突变,高达68.82%(386 162/561 147),颠换突变较少,只占31.18%(174 985 /561 147),其中 C/T 突变最多,然后是 A/G、C/G、A/C、G/T 和 A/T,依次减少,如图 3 所示。通过 SNP 位点序列信息,将 SNP 定位到小麦基因组,发现每条染色体上分布有17 186~30 847个 SNP,其中染色体 5D 最多,染色体 2B 最少。进一步研究发现,同组染色体上 SNP 分布较均匀,差异不大,如图 4 所示,但是,3 号和 7 号染色体组除外,如染色体 3B 和染色体 7D 上 SNP 明显超过同组另外两条染色体。

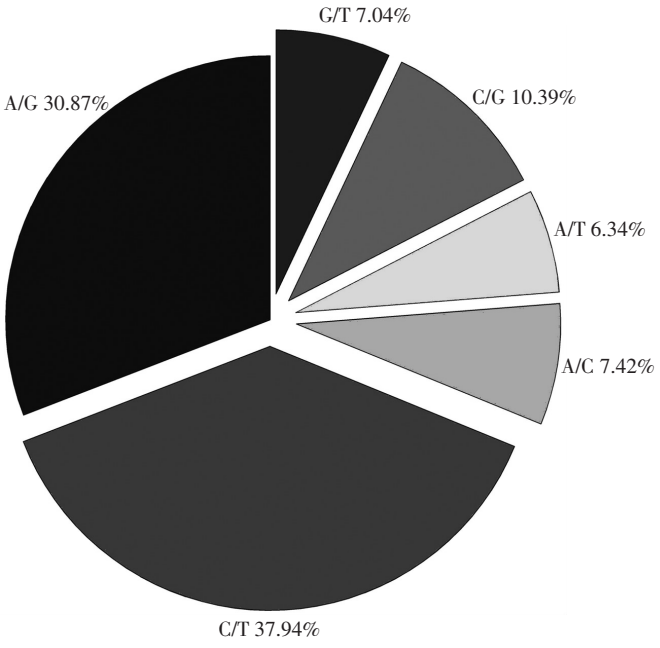

图 3　长穗偃麦草 SNP 分类信息

Fig.3　Classification information of SNP from *Th. elongatum*

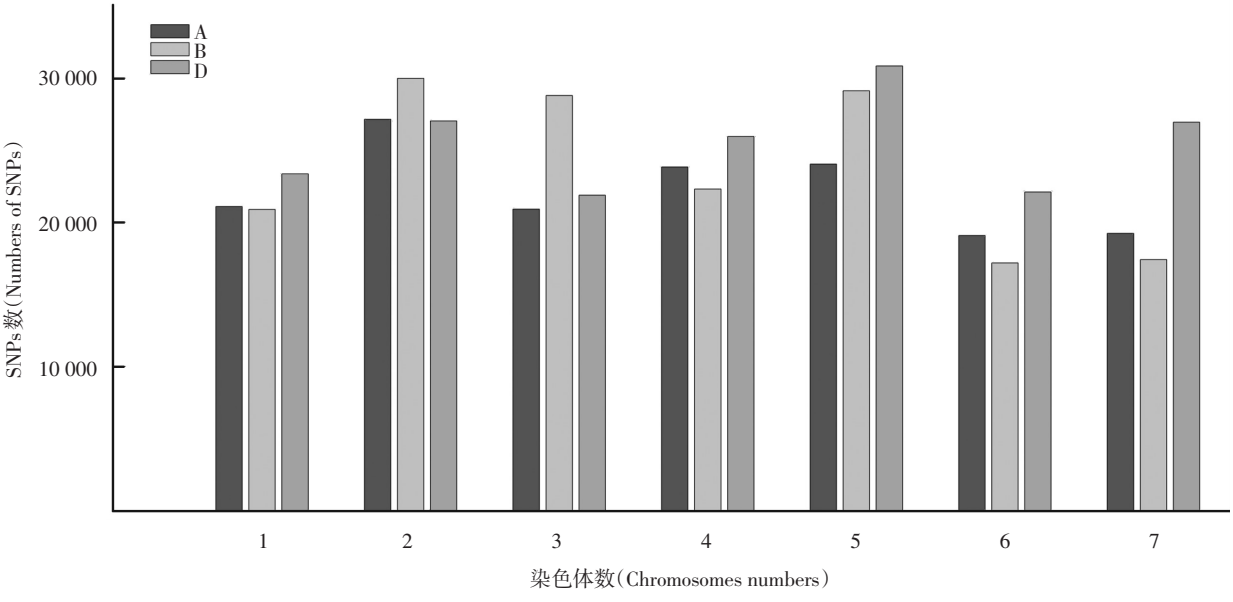

图 4　长穗偃麦草 SNP 的染色体分布

Fig.4　Chromosomes distribution of SNP from *Th. elongatum*

2.3　长穗偃麦草 SNP 靶向基因的功能分析

提取 SNP 位点的功能基因,发现这些 SNP 位于 47 712功能基因上。将这些基因序列,将其与拟南芥、水稻等基因组比对,对 SNP 靶向基因进行功能注释,KOG 注释结果如图 5 所示。这些基因主要参与 RNA 加工与修饰(A,1 191)、能量代谢与转化(C,1 095)、脂质转运与代谢(I,1 095)、转录调控(K,1 414)、蛋白质翻译后修饰与折叠(O,2 986)以及信号转导过程(T,3 563)。另外,这些 SNP 靶向的功能基因还参与一些特异的细胞过程,如染色质结构修饰、细胞周期调控、次生物质代谢和细胞间物质运输等过程,表明这些 SNP 靶向基因广泛地参与各种代谢反应过程, 影响着长穗偃麦草优异性状的形成。

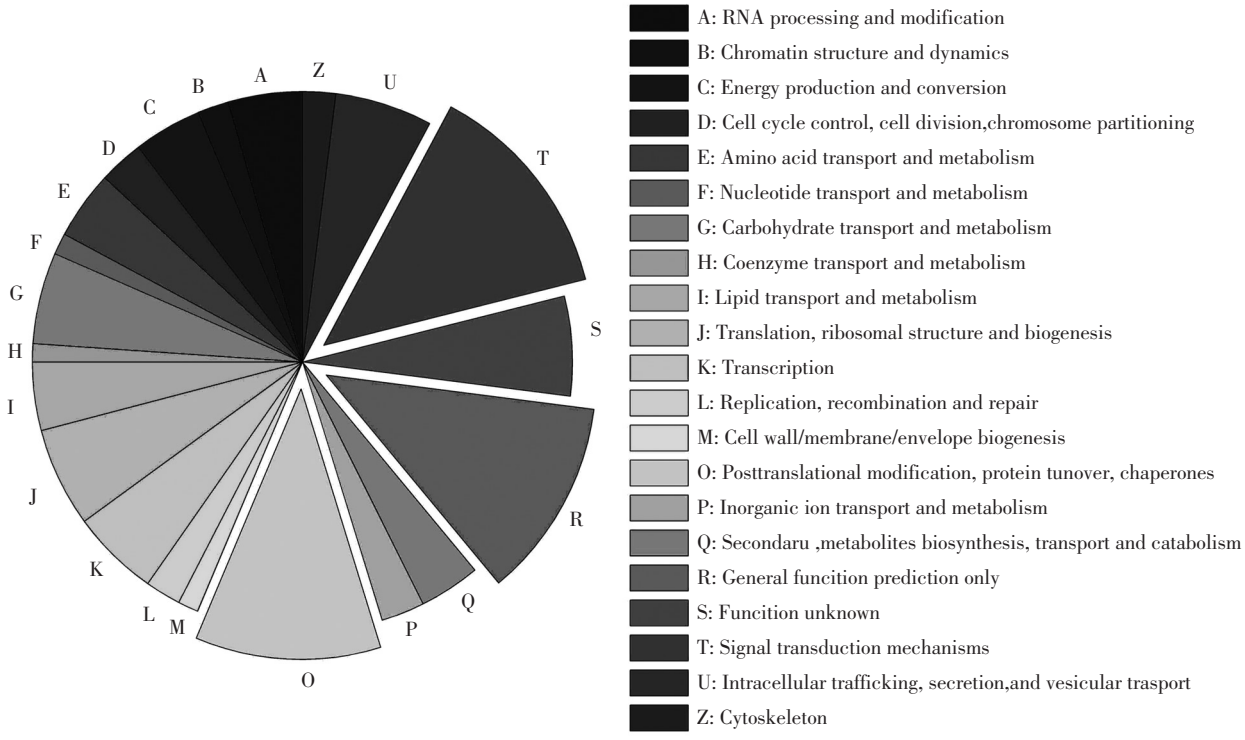

图 5 SNP 靶向基因的功能分析

Fig.5 Functional analysis of the genes from *Th. elongatum* containing SNP sites

3 讨 论

长穗偃麦草是栽培小麦的重要近缘物种,含有大量可用于小麦遗传改良的优异基因,广泛地运用于小麦的遗传改良,已经培育了一批含有偃麦草的优异小麦品种。但是,由于长穗偃麦草基因组信息极其匮乏,严重阻碍了长穗偃麦草基因资源在小麦遗传改良中的应用。本研究通过高通量测序获取长穗偃麦草的基因信息,通过比较基因组方法明确了长穗偃麦草基因组与小麦基因组间的亲缘进化关系,为其基因组在小麦改良中的应用提供理论基础。

分子标记是作物遗传改良的一种重要工具,可以用于优异性状的遗传连锁分析,指导作物遗传育种过程,提高育种效率,大大缩短育种进程,具有重要应用价值。但是,分子标记开发过程需要基因的序列信息,开发过程复杂,成本比较高。特别是一些非模式植物,由于基因组信息匮乏,导致分子标记开发滞后,严重阻碍了它们基因资源的开发和利用。长穗偃麦草与小麦亲缘关系较近,是小麦遗传改良重要基因资源来源,本研究利用高通量测序,获得长穗偃麦草基因信息,通过比较基因学方法发掘它与小麦间的 SNP 位点,获取了大量位于基因内部的功能分子标记,具有巨大潜在应用价值,这也为进一步利用长穗偃麦草的优异基因资源创造条件。

4 结 论

本研究通过高通量测序获得长穗偃麦草的基因序列信息,明确了它小麦以及其它小麦族植物间遗传进化关系。同时,利用高通量测序序列,发掘了长穗偃麦草与小麦间的突变位点,并对这些 SNP 靶向基因进行功能注释,这将为长穗偃麦草基因资源在小麦遗传改良中的应用提供理论支持。

参考文献(References)

[1] SCHACHERMAYR G M, MESSMER M M, FEUILLET C, et al. Identification of molecular markers linked to the Agropyron elongatum-derived leaf rust resistance gene Lr24 in wheat[J]. Theor Appl Genet, 1995, 90(7-8): 982-990.

[2] JIANG J, FRIEBE B, DHALIWAL H S, et al. Molecular cytogenetic analysis of Agropyron elongatum chromatin in wheat germplasm specifying resistance to wheat streak mosaic virus[J]. Theor Appl Genet, 1993, 86(1):41-48.

[3] JAUHAR P P. Synthesis and cytological characterization of trigeneric hybrids involving durum wheat, Thinopyrum bessarabicum, and Lophopyrum elongatum [J]. Theor

Appl Genet, 1992, 84(5-6):511-519.

[4] JAUHAR P P. Multidisciplinary approach to genome analysis in the diploid species, Thinopyrum bessarabicum and Th. elongatum (Lophopyrum elongatum), of the Triticeae[J]. Theor Appl Genet, 1990, 80(4):523-536.

[5] TAEB M, KOEBNER R M, FORSTER B P. Genetic variation for waterlogging tolerance in the Triticeae and the chromosomal location of genes conferring waterlogging tolerance in Thinopyrum elongatum[J]. Genome, 1993, 36(5):825-830.

[6] HUANG Q, LI X, CHEN W Q, et al. Genetic mapping of a putative Thinopyrum intermedium-derived stripe rust resistance gene on wheat chromosome 1B[J]. Theor Appl Genet, 2014, 127(4):843-853.

[7] PLACIDO D F, CAMPBELL M T, FOLSOM J J, et al. Introgression of novel traits from a wild wheat relative improves drought adaptation in wheat[J]. Plant Physiol, 2013, 161(4):1806-1819.

[8] JACOBY R P, MILLAR A H, TAYLOR N L. Investigating the role of respiration in plant salinity tolerance by analyzing mitochondrial proteomes from wheat and a salinity-tolerant Amphiploid (wheat x Lophopyrum elongatum)[J]. J Proteome Res, 2013, 12(11):4807-4829.

[9] HU L J, LI G R, ZENG Z X, et al. Molecular characterization of a wheat -Thinopyrum ponticum partial amphiploid and its derived substitution line for resistance to stripe rust[J]. J Appl Genet, 2011, 52(3):279-285.

[10] MONNEVEUX P, REYNOLDS M P, AGUILAR J G, et al. Effects of the 7DL.7Ag translocation from Lophopyrum elongatum on wheat yield and related morphophysiological traits under different environments[J]. Plant Breeding, 2003, 122(5):379-384.

[11] WANG Z, GERSTEIN M, SNYDER M. RNA-Seq: a revolutionary tool for transcriptomics[J]. Nat Rev Genet, 2009, 10(1):57-63.

[12] STRICKLER S R, BOMBARELY A, MUELLER L A. Designing a transcriptome next-generation sequencing project for a nonmodel plant species[J]. Am J Bot, 2012, 99(2):257-266.

[13] ZHANG G, GUO G, HU X, et al. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome[J]. Genome Research, 2010, 20(5):646-654.

[14] OPITZ N, PASCHOLD A, MARCON C, et al. Transcriptomic complexity in young maize primary roots in response to low water potentials[J]. BMC Genomics, 2014, 15:741.

[15] REDDY S K, LIU S, RUDD J C, et al. Physiology and transcriptomics of water-deficit stress responses in wheat cultivars TAM 111 and TAM 112[J]. J Plant Physiol, 2014, 171(14):1289-1298.

[16] SHU Yongjun, ZHANG Jun, AO You, et al. Analysis of the Thinopyrum elongatum Transcriptome under Water Deficit Stress[J]. International Journal of Genomics, 2015, 02:265791.

[17] HAAS B J, PAPANICOLAOU A, YASSOUR M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis[J]. Nat Protoc, 2013, 8(8):1494-1512.

[18] ZHENG Y, ZHAO L, GAO J, et al. iAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences[J]. BMC Bioinformatics, 2011, 12:453.

[19] BRENCHLEY R, SPANNAGL M, PFEIFER M, et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing[J]. Nature, 2012, 491(7426):705-710.

[20] LANGMEAD B, TRAPNELL C, POP M. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome[J]. Genome Biology, 2009, 10:R25.

[21] GARRISON E, MARTH G. Haplotype-based variant detection from short-read sequencing[J]. ArXiv, 2012, 1207: 3907.

[22] LI H, HANDSAKER B, WYSOKER A, et al. The Sequence Alignment/Map format and SAMtools[J]. Bioinformatics, 2009, 25(16):2078-2079.

[23] TATUSOV R L, GALPERIN M Y, NATALE D A, et al. The COG database: a tool for genome-scale analysis of protein functions and evolution[J]. Nucleic Acids Research, 2000, 28(1):33-36.