doi:10.3969/j.issn.1672-5565.2013.02.15

小麦长链非编码 RNA 的预测及功能分析

束永俊,张晶红,王明波,郭东林,王晓萍,郭长虹*
(黑龙江省分子细胞遗传与遗传育种重点实验室,生命科学与技术学院,哈尔滨师范大学,黑龙江 哈尔滨,150025)

摘 要:生物体有部分基因被转录成RNA,但是不编码相应蛋白质,称为长链非编码RNA(lncRNA)。它们参与基因的表观调控,这一过程对动物、植物的生长发育都有重要作用,但是,目前植物中发现和研究的lncRNA较少。为了研究lncRNA在植物中的功能,本研究建立了基于小麦全长cDNA的lncRNA识别程序。从6162条小麦全长cDNA中发现了231条lncRNAs,并从中鉴定出两个新miRNAs,这表明lncRNAs可以通过形成miRNAs前体基因形成其功能。此外,通过序列富集分析,我们从小麦lncRNAs中鉴定出三个保守的调控元件,结果显示小麦lncRNAs可能通过和其它蛋白质或DNA等分子作用,进而参与小麦生长、发育等过程的调控,这些结果对进一步研究植物体内的lncRNA的功能和作用机制具有重要意义。

关键词:小麦;长链非编码RNA;全长cDNA;微小RNA
中图分类号:Q518.2 **文献标识码**:A **文章编号**:1672-5565(2013)-02-153-05

Computational identification and functional analysis of long non-coding RNA in Triticum aestivum

SHU Yong-jun, ZHANG Jing-hong, WANG Ming-bo, GUO Dong-lin, WANG Xiao-ping, GUO Chang-hong*
(*Key Laboratory of Molecular Cytogenetics and Genetic Breeding of Heilongjiang Province, College of Life Science and Technology, Harbin Normal University, Harbin* 150025, *China*)

Abstract: There is a large portion of transcribed DNA, which does not code for a functional protein. These long non-coding RNAs (lncRNA) appear to have epigenetic regulatory function in animals. While epigenetic gene regulation is also an essential mechanism in plants, relatively little is known about the presence or function of lncRNAs in plants. To explore the function of lncRNA in plants, we have developed a computational pipeline for identified potential lncRNAs based wheat full length cDNA (fl-cDNA) sequences, and there are 231 lncRNAs identified from 6162 wheat fl-cDNAs. Meanwhile, two novel miRNAs are identified from these lncRNAs, which indicate that wheat lncRNAs would play their roles by acting as precursors for small RNA molecules. And by sequence analysis, there are three conservative motifs present enrichment in lncRNAs, which shows that they may interact with other protein or DNA with these motifs. These findings are useful for exploration of lncRNAs function mechanism in plant.

Key words: Triticum Aestivum; Long Non-coding RNA; Full Length cDNA; MicroRNA

在真核生物细胞内,遗传信息按照中心法则的规律进行传递,从DNA转录到RNA中,再翻译成相应的蛋白质,行使其功能,控制生物个体的表型性状。但是,近些年,研究表明:真核细胞内有大量的RNA具有一定生物功能,但是它们并不能翻译成相应的蛋白质分子,这类RNA称之为非编码RNA(non-coding RNA, ncRNA)。根据ncRNA的长度,将ncRNA分为小非编码RNA(small RNA)和长链非编码(long non-coding RNA, lncRNA)。其中,小非编码RNA包括miRNA, siRNA等,它们在真核细胞的各种生理生化反应,在基因表达调控中起重要作用。lncRNA在细胞内功能与小非编码RNA类似,

收稿日期:2013-03-17;**修回日期**:2013-04-22.
基金项目:黑龙江省教育厅科学技术研究项目(12521149)。
作者简介:束永俊,男,安徽人,博士/讲师,E-mail: syjun2003@126.com.
*通讯作者:郭长虹,女,黑龙江人,教授/博士生导师,研究方向:植物遗传学,E-mail:kaku_2008@163.com.

可以调控细胞内蛋白编码基因表达,但是其作用方式与 miRNA 等小非编码 RNA 不同,它主要通过调节邻近基因的染色质状态,调控目标基因的表达[1]。这种方式既可以下调抑制,也可以是诱导表达。同时,lncRNA 还可以通过选择性剪切、小非编码 RNA 的调控作用以及功能基因的 3'-UTR 等方式,参与真核生物功能基因的表达调控。

迄今为止,lncRNA 的研究主要集中在一些模式动物和人类中[2-3],如小鼠[4]、果蝇等[5-6],在植物中的研究报道较少,只发现了 Enod40 和 COLDAIR 等[7-8]。此外,植物 lncRNA 的鉴定也集中在一些模式植物,如拟南芥[9-10],或者模式作物,如水稻和玉米中[11-12],而其它重要作物中的 lncRNA 作用研究较少。

本研究拟对栽培小麦的全长 cDNA 文库序列进行分析,建立小麦 lncRNA 识别系统,筛选小麦特异的 lncRNA,并对预测的小麦 lncRNA 进行功能分析,阐明 lncRNA 在小麦生长和发育过程中的作用和意义。

1 材料与方法

1.1 序列来源

栽培的小麦品种"中国春"的全长 cDNA 序列(6 162 条)下载自 TriFLDB (Triticeae Full-Length CDS DataBase, Website: http://trifldb.psc.riken.jp)[13]。

1.2 小麦 lncRNA 的预测

对小麦的全长 cDNA 进行序列分析,去除长度在 200bp 以下的基因。剩余的 fl-cDNA 利用软件 UGENE[14](ugene.unipro.ru)进行开放读码框(open reading frame,ORF)的识别,参数如下:strand = direct,-min-length = 363,-require-init-codon = true,-require-stop-codon = true,选取最长 ORF 小于等于 120aa 的 fl-cDNA 作为候选 lncRNA。将候选 lncRNA 的 fl-cDNA 比对(BLASTX) SWISS-PROT 蛋白质数据库,参数如下:-strand plus-evalue 0.001-max_target_seqs 1-num_threads 8,去除与已知蛋白质同源的 fl-cDNA,剩余 fl-cDNA 即为小麦新发现的 lncRNA。

1.3 小麦 lncRNA 上相关 smRNA 的鉴定

小麦的 smRNA 数据下载于 NCBI 数据库(http://www.ncbi.nlm.nih.gov/,登录号为:GSE36867 和 GSE27327[17-18])。对下载的 smRNA 的高通量测序数据进行预处理,包括以下操作:去除接头序列、去除低质量序列、去除污染序列、去除载体序列和包含 polyA 的序列及去除小于 16 nt 的小片段等。将得到的干净序列(clean sequence)比对 Rfam、RepeatMasker 以及 Genbank 数据库,去除其中的 rRNA、tRNA、snRNA 和 snoRNA 等非编码 RNA 和重复序列。将剩下的高通量测序序列用 Bowtie 和 miRDeep2(http://www.mdc-berlin.de/)进行 smRNA 识别分析,将小麦已知的 smRNA 和预测的 smRNA 定位到预测小麦 lncRNA 上[19-20]。

1.4 小麦 lncRNA 的功能元件分析

用软件 DREME (http://meme.ebi.edu.au/meme/cgi-bin/dreme.cgi)分析 lncRNA 中富集的功能性 motif[21],对照序列选用 ChromDB 中的小麦蛋白质编码基因的序列(http://www.chromdb.org/,总计 183 条序列)[22],其它参数默认。

2 结果与分析

2.1 小麦 lncRNA 的预测

分析发现,6 162 条小麦 fl-cDNA 序列中有 10 条序列的长度小于 200bp,将其去除,如图 1 所示。剩余的 6 152 条 fl-cDNA 经 UGENE 鉴定,发现有 5 516条 fl-cDNA 含有长度超过 120aa 的 ORF,将它们去除。通过 BLASTX 分析发现,剩下的 636 条 fl-cDNA 中,有 405 条与已经蛋白质相似(evalue:0.001)。将这些同源的 fl-cDNA 序列去除,剩余的 231 条 fl-cDNA 即为小麦的候选 lncRNA。

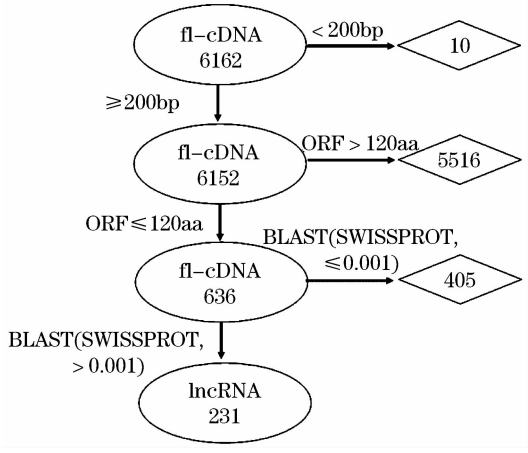

图 1 基于小麦全长 cDNA 的 lncRNA 预测流程

Fig. 1 Process pipeline results on wheat full-length cDNAs

2.2 小麦 lncRNA 的功能分析

为了鉴定筛选的 lncRNA 和 miRNA 间的关系,用小麦的已知 microRNA 和前体序列比对筛选得到的 lncRNA 序列,未发现有两者之间的相似性存在。下载小麦的 smRNA 高通量测序数据,经过去接头、摒弃低质量等过程处理,得到 99 077 707 条有效 smRNA 序列。将这些 smRNA 序列定位到候选的

lncRNA 序列上,发现总计有 543 333 条 smRNA 分别匹配到 225 条 lncRNA,其中,匹配 10 条以上 smRNA 的有 146 个,匹配 1 000 条以上有 19 个,匹配最多的是 Ta-lncR1061,总计匹配了 406 211 条 smRNA。对匹配到 10 条以上的 lncRNA 匹配位点进行折叠分析,发现 2 个存在潜在的茎环结构,分别位于 Ta-lncR960 和 Ta-lncR2770 上,可以形成 miRNA,如图 2 所示。

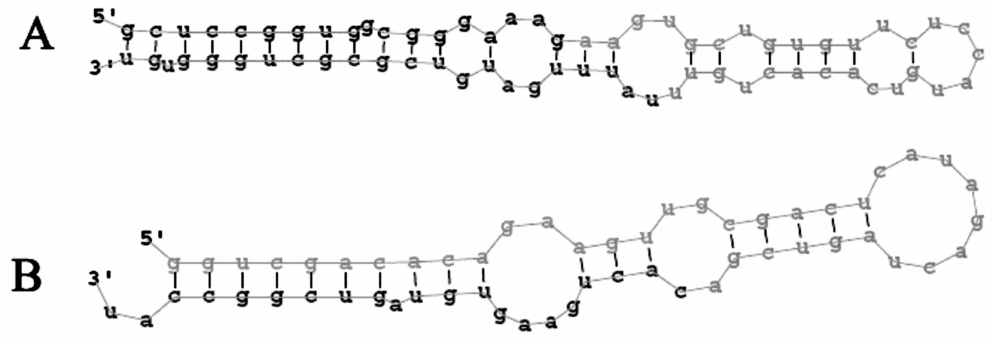

图 2　lncRNA 中新 microRNA 的识别

Fig. 2　Identification of novel microRNA from lncRNA in wheat

2.3　小麦 lncRNA 的功能元件分析

以 ChromDB 中的小麦基因序列作为对照,用 DREME 软件查找 lncRNA 中的保守功能元件(motif),发现 3 个重复元件,如图 3 所示。其中,序列 CCGCCWC 在 183 个 ChromDB 编码蛋白质基因中只出现 48 次,而在 231 个 lncRNA 中出现了 119 次,在小麦 lncRNA 中高度富集。类似的还有功能元件 CCRGCCGK 和 CGTCGYGC,它们分别在小麦蛋白质编码基因和 lncRNA 中出现的次数比为:6/55 和 1/28。结果显示,这些功能元件都是在小麦 lncRNA 基因中富集、甚至只在某些 lncRNA 中特异出现。

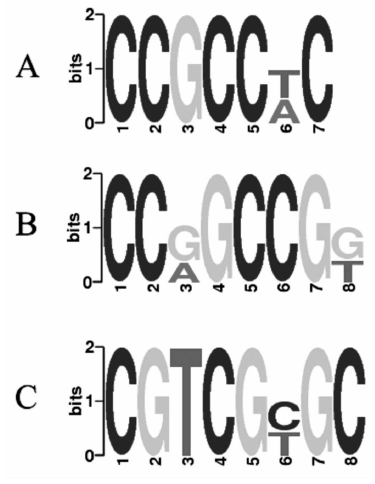

图 3　lncRNA 中富集的功能元件

Fig. 3　Enrichment analysis of functional motif in lncRNA

3　讨论

本研究在小麦中筛选得到 231 条候选 lncRNA 基因,占总数的 3.8%,远低于小鼠的 48.7% 和玉米的 13.3%。这是由于小麦的 fl-cDNA 只来源于一个测序文库,表达情况比较单一,检测到的 ncRNA 种类非常少,导致最终预测的 lncRNA 数量有限。而小鼠和玉米的 fl-cDNA 分别来源自 246 和 27 个表达文库,信息量丰富,可以检测到 lncRNA 数量较大[4, 11]。此外,本研究建立的 lncRNA 识别流程参数控制较严格,最终导致发现的 lncRNA 数量极少。

在植物中,miRNA 长度一般在 22 ~ 24bp 左右,参与调控植物生长和发育各个阶段的基因表达。它们都是先形成较长链的转录本,后经过 Dicer 的切割,形成成熟的 miRNA,行使其调控功能。一般来说,lncRNA 也是植物 miRNA 的一种重要来源,但是,我们研究发现:筛选的 lncRNA 与已知的小麦 miRNA 相似性极差,没有同源性序列。这可能是由于小麦 miRNA 研究较其它作物,或物种落后造成的。小麦中,已知的 miRNA 只有 48 条,远低于拟南芥、水稻和玉米等植物,这就为鉴定 lncRNA 中的 miRNA 元件增加了难度。另外,本研究中识别得到 lncRNA 只有 231 条,数量也是非常有限的,导致两者之间没有出现同源序列。但是,通过 smRNA 的高通量测序数据,结合 RNA 折叠分析,发现两条 lncRNA 可以折叠形成茎环结构,进而加工成相应的 miRNA。这说明,在小麦中 lncRNA 也是小麦的 miRNA 重要来源之一,同时,也可以利用 lncRNA 和 smRNA 的测序数据鉴定和发现新的 miRNA。

真核生物的基因一般含有大量的调控元件,它们可以和蛋白质结合进而调控基因的表达,常见的有转录因子的结合位点和染色质修饰蛋白结合位点,这两类作用元件一般都会影响基因的表达。为了明确 lncRNA 基因的表达模式,我们选取了

ChromDB 中的小麦基因作为对照(ChromDB 中的基因含有极少的调控元件),检测 lncRNA 中富集的调控元件[22]。通过比对 lncRNA 基因和对照的 ChromDB 中的基因发现,富集的调控元件主要有 CCGCCWC、CCRGCCGK 和 CGTCGYGC,这三条序列都富含 GC,可能涉及小麦 lncRNA 基因的甲基化调控,进而调控 lncRNA 的表达。同时,这些调控元件也可能是 lncRNA 基因调控其它基因的核心位点,具体其功能和作用方式有待进一步研究。

4 结论

本研究建立了小麦 lncRNA 识别程序,从小麦 6 162 条 fl-cDNA 中鉴定出 231 条 lncRNA。对 lncRNA 进行序列分析,发现小麦 lncRNA 可以折叠形成 miRNA。同时,从小麦 lncRNA 中识别出三个特异的调控元件,这对进一步解析小麦 lncRNA 的调控机制和作用机制具有重要的作用。

参考文献(References)

[1] Filomena De Lucia, Caroline Dean. Long non-coding RNAs and chromatin regulation [J]. Current Opinion in Plant Biology, 2011, 14(2):168-173.

[2] Paul Bertone, Viktor Stolc, Thomas E. Royce, Joel S. Rozowsky, Alexander E. Urban, Xiaowei Zhu, John L. Rinn, Waraporn Tongprasit, Manoj Samanta, Sherman Weissman, Mark Gerstein, Michael Snyder. Global identification of human transcribed sequences with genome tiling arrays [J]. Science, 2004, 306(5705):2242-2246.

[3] Yoshiyuki Sakuraba, Toru Kimura, Hiroshi Masuya, Hideki Noguchi, Hideki Sezutsu, K. Takahasi, Atsushi Toyoda, Ryutaro Fukumura, Takuya Murata, Yoshiyuki Sakaki, Masayuki Yamamura, Shigeharu Wakana, Tetsuo Noda, Toshihiko Shiroishi, Yoichi Gondo. Identification and characterization of new long conserved noncoding sequences in vertebrates [J]. Mammalian Genome, 2008, 19(10):703-712.

[4] The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs [J]. Nature, 2002, 420(6915):563-573.

[5] Alexey Soshnev, Hiroshi Ishimoto, Bryant McAllister, Xingguo Li, Misty Wehling, Toshihiro Kitamoto, Pamela Geyer. A conserved long noncoding RNA affects sleep behavior in Drosophila [J]. Genetics, 2011, 189(2):455-468.

[6] Robert Young, Ana Marques, Charlotte Tibbit, Wilfried Haerty, Andrew Bassett, Ji-Long Liu, Chris Ponting. Identification and properties of 1,119 candidate lincRNA Loci in the Drosophila melanogaster genome [J]. Genome Biology and Evolution, 2012, 4(4):427-442.

[7] Anna Campalans, Adam Kondorosi, Martin Crespi. Enod40, a short open reading frame - containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in Medicago truncatula [J]. Plant Cell, 2004, 16(4):1047-1059.

[8] Jae Bok Heo, Sibum Sung. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA [J]. Science, 2011, 331(6013):76-79.

[9] Besma Ben Amor, Sonia Wirth, Francisco Merchan, Philippe Laporte, Yves d'Aubenton-Carafa, Judith Hirsch, Alexis Maizel, Allison Mallory, Antoine Lucas, Jean Marc Deragon, Herve Vaucheret, Claude Thermes, Martin Crespi. Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses [J]. Genome research, 2009, 19(1):57-69.

[10] Yun Ju Kim, Binglian Zheng, Yu Yu, So Youn Won, Beixin Mo, Xuemei Chen. The role of Mediator in small and long noncoding RNA production in Arabidopsis thaliana [J]. EMBO J, 2011, 30(5):814-822

[11] Susan Boerner, Karen M. McGinnis. Computational identification and functional predictions of long noncoding RNA in Zea mays [J]. PLoS ONE, 2012, 7(8):e43047.

[12] Jihua Ding, Qing Lu, Yidan Ouyang, Hailiang Mao, Pingbo Zhang, Jialing Yao, Caiguo Xu, Xianghua Li, Jinghua Xiao, Qifa Zhang. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice [J]. Proceedings of the National Academy of Sciences, 2012, 109(7):2654-2659.

[13] Keiichi Mochida, Takuhiro Yoshida, Tetsuya Sakurai, Yasunari Ogihara, Kazuo Shinozaki. TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics [J]. Plant Physiology, 2009, 150(3):1135-1146.

[14] Konstantin Okonechnikov, Olga Golosova, Mikhail Fursov, the UGENE team. Unipro UGENE: a unified bioinformatics toolkit [J]. Bioinformatics, 2012, 28(8):1166-1167.

[15] Stephen Altschul, Thomas Madden, Alejandro Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, David Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. Nucleic Acids Research, 1997, 25(17):3389-3402.

[16] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J. Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, Sandrine Pilbout, Michel Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003 [J]. Nucleic Acids Research, 2003, 31(1):365-370.

[17] Zhonghui Tang, Liping Zhang, Chenguang Xu, Shaohua Yuan, Fengting Zhang, Yonglian Zheng, Changping Zhao. Uncovering small RNA - mediated responses to cold stress in a wheat thermosensitive genic male - sterile line by deep sequencing [J]. Plant Physiology, 2012, 159(2):721-738.

[18] Mingming Xin, Yu Wang, Yingyin Yao, Na Song, Zhaorong Hu, Dandan Qin, Chaojie Xie, Huiru Peng, Zhongfu Ni, Qixin Sun. Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing [J]. BMC Plant Biology, 2011, 11(1):61.

[19] Ben Langmead, Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2 [J]. Nature Method, 2012, 9(4):357-359.

[20] Marc R. Friedländer, Sebastian D. Mackowiak, Na Li, Wei Chen, Nikolaus Rajewsky. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades [J]. Nucleic Acids Research, 2012, 40(1):37-52.

[21] Timothy L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data [J]. Bioinformatics, 2011, 27(12):1653-1659.

[22] Karla Gendler, Tara Paulsen, Carolyn Napoli. ChromDB: the chromatin database [J]. Nucleic Acids Research, 2008, 36 (suppl 1):D298-D302.