doi:10.3969/j.issn.1672-5565.2015.02.01

nc2Cancer:一个研究与癌症相关人类非编码 RNA 的数据库

程　卓,刘　珂,严章明,向书念,孙之荣*
(清华大学生命科学学院教育部生物信息学重点实验室,北京 100084)

摘　要:伴随着高通量测序技术的飞速发展,许多新型的非编码 RNA 陆续被发现,比如长链非编码 RNA(lncRNA)和环状 RNA(Circular RNA)。先前的研究已经表明这些非编码 RNA 在基因表达调控过程中起着很重要的作用,并且与癌症的发生有着很密切的联系。但是,由于研究者们仍然对它们行使何种功能知之甚少,鉴定这些非编码 RNA 是否与人类癌症存在密切的相互关系仍然是一个巨大的挑战。为了促进这一领域的研究,这篇文章的作者分析了大规模的 RNA 相互作用数据,然后建立了数据库 nc2Cancer (http://www.bioinfo.tsinghua.edu.cn/nc2Cancer/index.php)。这个数据库的目标便是提供非编码 RNA 与癌症之间的全面关系。现在,该 nc2Cancer 数据库包括了三种类型的非编码 RNA 分子:长链非编码 RNA,环状 RNA 以及由假基因转录而成的 RNA。这项研究将有助于研究者更好地去理解非编码 RNA 的功能以及它们在人类癌症发生过程中所起到的作用。

关键词:非编码 RNA;癌症;MicroRNA;RNA 相互作用网络
中图分类号:TP392;Q522　　**文献标志码**:A　　**文章编号**:1672-5565(2015)-02-077-05

nc2Cancer: a database for cancer-associated human ncRNAs

CHENG Zhuo, LIU Ke, YAN Zhangming, XIANG Shunian, SUN Zhirong*
(*MOE Key Laboratory of Bioinformatics*, *School of Life Sciences*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: With the rapid development of high-throughput sequencing technology, many new types of non-coding RNAs (such as lncRNAs and circular RNAs) have been identified. Previous studies have shown that these non-coding RNAs play critical roles in gene regulation and have strong associations with cancers. However, since we know little on how the ncRNAs excute their functions, it remains a challenge for us to fully identify the associations between ncRNAs and human cancer. To facilitate the study of this field, we analyzed large-scale RNA-RNA interaction data and constructed a database named nc2Cancer (http://www.bioinfo.tsinghua.edu.cn/nc2Cancer/index.php), a web resource which aims to provide comprehensive associations between ncRNAs and cancer. Currently, nc2Cancer covers three types of non-coding RNAs (lncRNAs, circular RNAs and RNAs transcribed from pseudogenes). This study will greatly expand the understanding of ncRNA functions and their possible roles in cancer development.

Keywords: ncRNA; Cancer; microRNA; RNA-RNA network

近些年来,伴随着高通量测序技术的发展,许多新的非编码 RNA 类型被陆续鉴定出来。比如长链非编码 RNA(Long non-coding RNAs)和环状 RNA(Circular RNAs)[1-3]。越来越多的研究已经暗示了这些非编码 RNA 分子在许多生物学过程中起着重要的调控作用,比如细胞分化,表观遗传调控,细胞凋亡以及细胞周期控制等,并且这些非编码 RNA 还被认为与肿瘤的发生有着密切的联系[4-5]。但是,我们至今仍然缺少一个专注于全面提供非编码 RNA 与癌症相互关系的数据库。

现在已经证明,非编码 RNA 所具有的一个基本调控规则是,它们可以通过与 microRNA 相互作用,以“RNA 海绵”的方式去调控 microRNA 的活性[6-8],从而影响相应的基因表达。举例说明:PTEN 已被证明在许多种类癌症的发生过程中是一个重要的肿瘤抑制基因。Laura Poliseno 等人研究

收稿日期:2015-03-31;**修回日期**:2015-04-14.
基金项目:国家 973 计划 (No. 2012CB925203),国家自然科学基金(No.31171274)。
作者简介:程卓,男,硕士研究生,研究方向:系统生物学、生物信息学;E-mail:chinachengzhuo@gmail.com.
*通信作者:孙之荣,男,教授,博士生导师,研究方向:系统生物学,生物信息学;E-mail:sunzhr@mail.tsinghua.edu.cn.

发现,一个假基因 PTENP1,可以通过“RNA 海绵”的方式调控两种 microRNA:miR-17 和 miR-19 在细胞中的量,而这两种 microRNA 都可以通过与 PTEN 的结合从而调控该抑癌基因的表达[9]。

相较于其他非编码 RNA,microRNA 的功能已经得到了相对充分的研究。许多收集 microRNA 与疾病相互关系的数据库也已被开发出来[10-12]。由于 microRNA 对于许多非编码 RNA 来说是重要的功能调节因子,所以我们也可以很自然地认为如果一个非编码 RNA 倾向于与跟某些 microRNA 相互作用,而这些 microRNA 又与某种特异的疾病有较密切的联系。则这个非编码 RNA 也倾向于与该疾病相关联。基于这个推想,我们综合了大量的 microRNA 与非编码 RNA 相互作用数据,以及经过实验验证的 microRNA 与癌症相互关联数据,然后构建了一个数据库 nc2Cancer。这个数据库的网址是 http://www.bioinfo.tsinghua.edu.cn/nc2Cancer/index.php。据我们所知,nc2Cancer 是第一个全面的通过 RNA-RNA 相互作用侧重研究非编码 RNA 与癌症相互关系的数据库。

1 数据库构建与数据库内容

经过实验验证的 microRNA-癌症相互作用数据是从以下三个数据库下载的:miR2Disease,miRCancer 以及 HMDD v2.0[10-12]。我们将这三项数据整合成了一个数据集。为了保证研究的准确性,我们只保留了数据集中与不少于 10 种 microRNA 有相互关系的癌症种类。最后,我们得到了 3 058 条 microRNA 与癌症的相互关系数据,其中包括有 42 种癌症类型。

MicroRNA 与非编码 RNA 的相互作用数据则是从数据库 starBase[13] 下载的,这个数据库提供了大量高质量的 RNA 相互作用数据。最后,我们一共得到了 9 112 条 microRNA-环状 RNA 相互作用数据,16 123 条 microRNA-假基因相互作用数据以及 10 203条 microRNA-lncRNA 相互作用数据。

我们进一步筛选了非编码 RNA 与癌症的数据对,保证其中任一条数据对中非编码 RNA 与癌症都至少由三个及以上的 microRNA 相互连结。我们使用了下面的超几何分布公式来计算非编码 RNA 与癌症之间是否显著地共享 microRNA。其中的 P 值是由以下方法得出:

$$P = \sum_{i=x}^{\min(K,n)} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}。$$

在这个公式里,N 代表数据集中所有 microRNA 的数量,n 代表与该非编码 RNA 相互作用的 microRNA 数量,K 则是与某种特定类型的癌症有关的 microRNA 数量,而 x 则是与该非编码 RNA 以及该类型癌症都有联系的 microRNA 数量。我们仅仅保留了足够小 P 值(FDR<0.05)的非编码 RNA-癌症数据对。构建数据库的流程图(见图 1)。

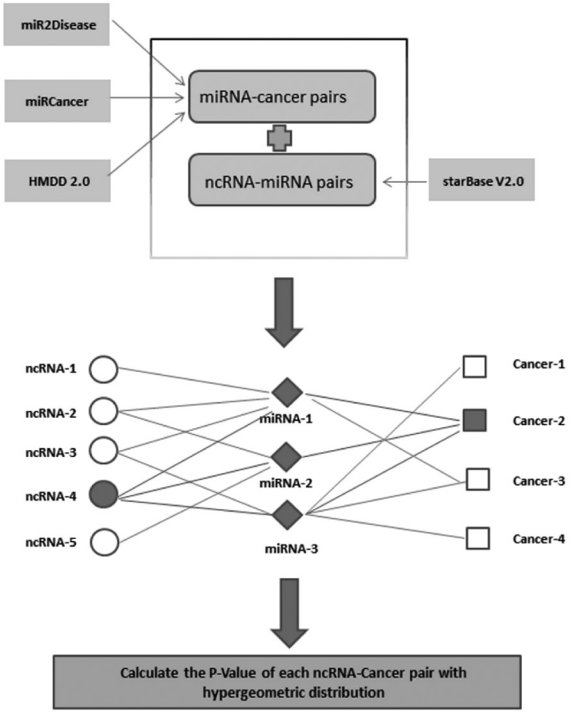

图 1 生成非编码 RNA 与癌症相互关系的流程图

Fig.1 Workflow of generating ncRNA-cancer associations

PHP 语言和 SQLite 被用于开发该 nc2Cancer 数据库。目前服务器支持用户通过任一非编码 RNA 的 ID 或癌症名称进行搜索。用户界面被表示如下(见图 2)。

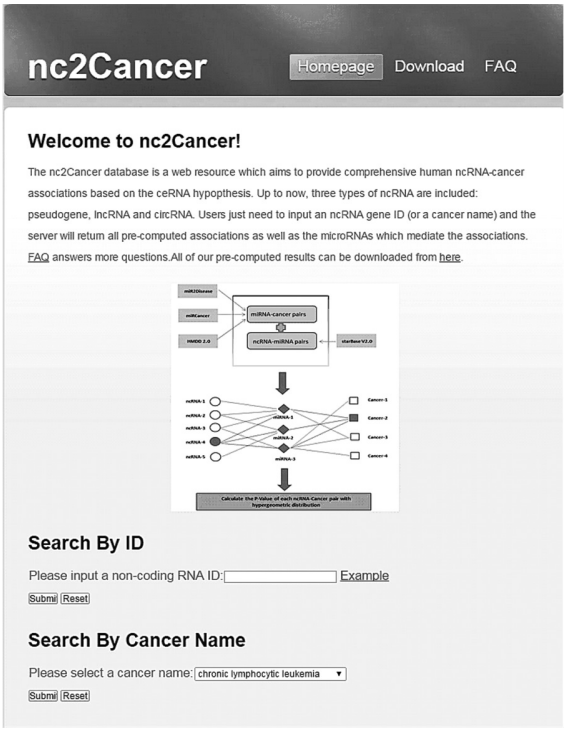

图 2　nc2Cancer 的用户界面

Fig.2　The user interface of nc2Cancer

2　结果与结论

我们最终呈递了 1 176 条非编码 RNA 与癌症的相互关系数据,其涵盖 31 种癌症以及 725 条非编码 RNA (172 条环状 RNA,192 条 lncRNA 和 361 条假基因)。通过与我们最初收集的数据相比较,初始数据中非编码 RNA 的总量有 4 880 条,而收录在我们 nc2Cancer 数据库中的与癌症相关的非编码 RNA 为 725 条,比例为 14.9%。在与较多条非编码 RNA 关联的癌症类型中,排名前三位的分别是慢性淋巴细胞白血病(202 条关联),肺癌(170 条关联)以及肾癌(103 条关联)。平均来说,一条非编码 RNA 所参与的癌症关联数为 1.6,统计图见图 3。

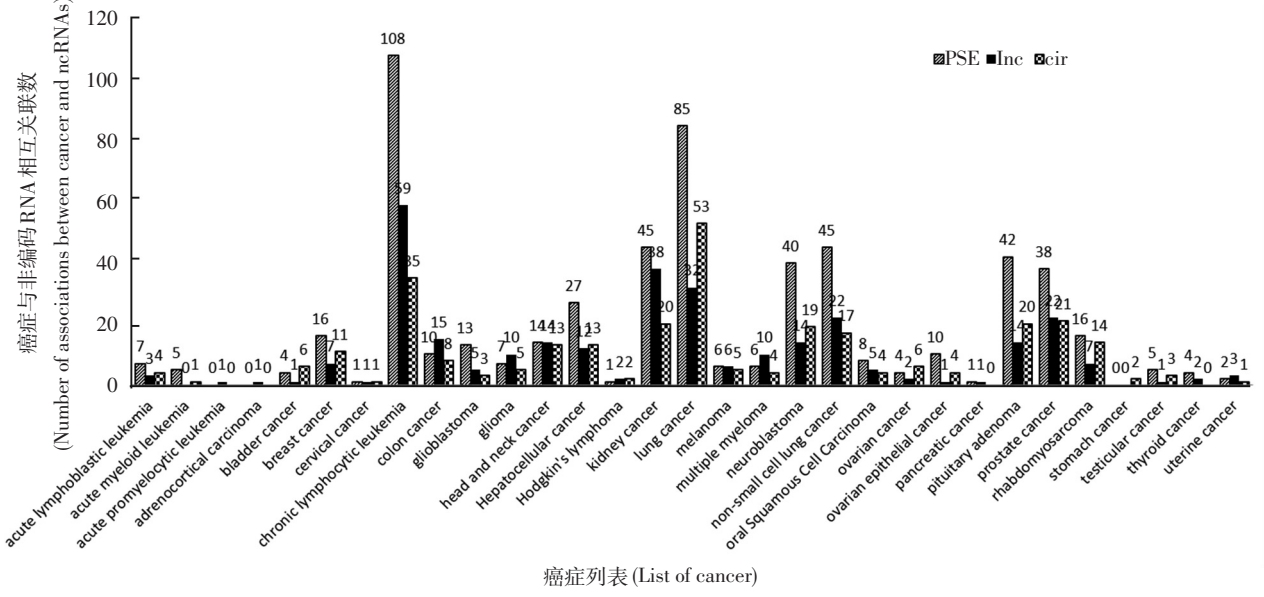

图 3　与特异癌症类型相关的非编码 RNA 数量统计

Fig.3　The number of ncRNAs been associated with individual cancers

为了评估我们的方法,我们将非编码 RNA 与 microRNA 之间的关系在总量保持不变的情况下随机打乱,然后分别计算每次非编码 RNA 与癌症的相互关系数量。重复上述过程 500 次后,我们发现随机非编码 RNA 与癌症的相互关系预期数约为 500,这比我们在真实数据下推断的相互关系数量要小得多(见图 4)。在该图中黑色的曲线表示随机情况下我们的方法推断的非编码 RNA 与肿瘤相互关系的数目;箭头指示的是从真实数据中推断的非编码 RNA 与癌症相互关系的数目。结果显示真实数据的结果具有明显的显著性。

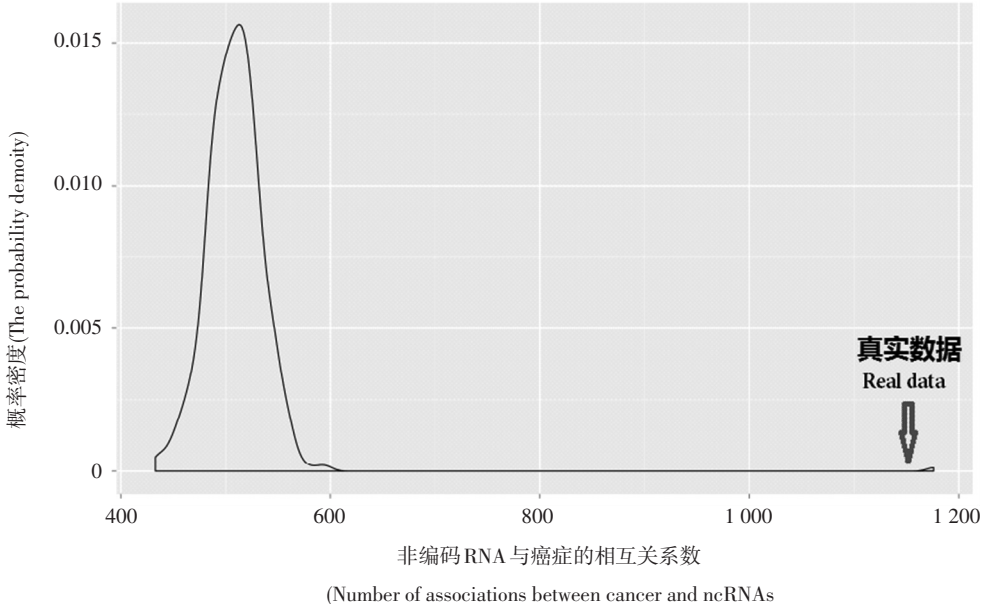

图 4 方法的计算性评价

Fig.4 Computational evaluation of the method

除了计算评估的方法,我们也发现了很多的实验证据可以用来支持我们的结果。例如,lncRNA 中 CDKN2B-AS1 的异常表达可引起各种人类疾病,如乳腺癌和白血病[13]. 而在我们的结果中,我们预计该 lncRNA 与乳腺癌,肺癌和多发性骨髓瘤有相互关联。H19 和 MALAT1 均被报道在结肠癌细胞的转移表型中起着关键性作用[14-15]。而在我们的数据库中,计算结果认为这两条 lncRNA 与结肠癌存在相互关联。XIST 被报道与染色质修饰相关复合物有着紧密的联系,并且可以影响乳腺癌细胞中的基因表达[16],而 XIST-乳腺癌相互关系也被我们的数据库所计算并收录。假基因被大量的证据证实为存在生物学活性,特别是它们可以通过各种机制来调节其亲本基因的表达[17]。通过浏览我们数据库中假基因与癌症的相互关系,我们发现了一些与癌症相关的假基因,而且它们的亲本基因已被实验证实为原癌基因,如 ACTB, ANXA2, CDC42, DDX6 和 RPL7A 等(见表 1)。该例证也可以用来评估我们在 nc2Cancer 中所做的预测。

表 1 与癌症相关的假基因以及它们原癌亲本基因

Table 1 Examples of cancer-associated pseudogenes and their parental oncogenes

假基因	癌症名称	错误发现率	原癌母基因
ACTBP11	垂体腺瘤	0.006 154 052	
ACTBP11	急性淋巴细胞白血病	0.045 471 645	
ACTBP7	垂体腺瘤	0.013 292 355	ACTB
ACTBP7	肺癌	0.035 815 725	
ACTBP8	垂体腺瘤	0.025 696 649	
ANXA2P1	睾丸癌	0.028 220 982	ANXA2
CDC42P4	慢性淋巴细胞性白血病	0.012 840 171	CDC42
DDX6P2	前列腺癌	0.024 482 898	DDX6
RPL7AP11	肾癌	0.046 716 514	RPL7A
RPL7AP66	肾癌	0.019 399 655	

3　展　望

我们的研究之前,科学家们已建成有两个数据库(LncRNADisease 和 Circ2traits)用于探讨非编码 RNA 和疾病之间的关联。LncRNADisease 为疾病相关的长非编码 RNA 数据库。该数据库运用了一种生物信息学的方法,通过使用 lncRNA 位置信息来预测新的 lncRNA-疾病相互关系[1]。Circ2traits 则是计算环状 RNA 是否与疾病显著相关的数据库,然后他们还寻找了环状 RNA 上与疾病表型相关的 SNP 位点[18]。与上述两个数据库相比,nc2Cancer 中最大的区别在于我们覆盖了较常见的三种类型的非编码 RNA(假基因,lncRNA,以及环状 RNA),而前两个数据库只集中在其中的某一类非编码 RNA 上面。在 LncRNADisease 中,研究人员使用的是 lncRNA 的位置信息,而在我们的研究中则采用了一个完全不同的策略来确定 lncRNA 与疾病的关联。我们认为,这两种的结果都是有意义并且彼此互补的。此外,与他们的数据相比,我们所收集的 microRNA 与靶基因的相互作用数据是通过高通量的 Clip-Seq 方法得到,具有更高的精度和准确性。

综上所述,我们开发了一个基于网络的数据库名为"nc2Cancer",它可以使研究人员能够轻松地找到与癌症相关的非编码 RNA,包括假基因,lncRNA,以及环状 RNA。nc2Cancer 提供了一个有用的非编码 RNA 的功能参考来帮助生物学家探索在癌症中非编码 RNA 的角色。我们所有的结果均可以被下载来为他人的计算项目提供方便。

参考文献(References)

[1] CHEN G, WANG Z, WANG D, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases[J]. Nucleic Acids Res, 2013, 41: D983-986.

[2] LAU E. Non-coding RNA: Zooming in on lncRNA functions [J]. Nature Reviews. Genetics, 2014, 15: 574-575.

[3] PERKEL J. Assume nothing: the tale of circular RNA[J]. BioTechniques, 2013, 55: 55-57.

[4] SUMAZIN P, YANG X, CHIU H, et al. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma [J]. Cell, 2011, 147: 370-381.

[5] SWAMI M. Small RNAs: Pseudogenes act as microRNA decoys[J]. Nature Reviews Genetics, 2010, 11: 530-531.

[6] GODLEWSKI J, NOWICKI M, BRONISZ A, et al. Targeting of the Bmi-1 oncogene/stem cell renewal factor by microRNA-128 inhibits glioma proliferation and self-renewal[J]. Cancer Research, 2008, 68: 9125-9130.

[7] KHALIL A, GUTTMAN M, HUARTE M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression [J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106: 11667-11672.

[8] POLISENO L, SALMENA L, ZHANG J, et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology[J]. Nature, 2010, 465: 1033-U1090.

[9] JIANG Q, WANG Y, HAO Y, et al. miR2Disease: a manually curated database for microRNA deregulation in human disease[J]. Nucleic Acids Res, 2009, 37: D98-D104.

[10] LI Y, QIU C, TU J, et al. HMDD v2.0: a database for experimentally supported human microRNA and disease associations[J]. Nucleic Acids Res, 2014, 42: D1070-D1074.

[11] XIE B, DING Q, HAN H, et al. miRCancer: a microRNA-cancer association database constructed by text mining on literature[J]. Bioinformatics, 2013, 29: 638-644.

[12] LI J, LIU S, ZHOU H, et al. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data[J]. Nucleic Acids Res, 2014, 42: D92-D97.

[13] SHI X, SUN M, LIU H, et al. Long non-coding RNAs: a new frontier in the study of human diseases[J]. Cancer Letters, 2013, 339: 159-166.

[14] CUI H, ONYANGO P, BRANDENBURG S, et al. Loss of imprinting in colorectal cancer linked to hypomethylation of H19 and IGF2[J]. Cancer Research, 2002, 62: 6442-6446.

[15] GUTSCHNER T, HAMMERLE M, DIEDERICHS S. MALAT1—a paradigm for long noncoding RNA function in cancer[J]. Journal of Molecular Medicine, 2013, 91: 791-801.

[16] SIRCHIA S, TABANO S, MONTI L, et al. Misbehaviour of XIST RNA in breast cancer cells[J]. PloS One, 2009, 4: e5559.

[17] HAN L, YUAN Y, ZHENG S, et al. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes [J]. Nature Communications, 2014, 5: 3963.

[18] GHOSAL S, DAS S, SEN R, et al. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits [J]. Frontiers in Genetics, 2013, 4: 283.