Non-coding RNAs에 대한 생물정보학 접근

국립암센터 ▌남 승 윤

1. 서 론

1.1 Non-coding RNA의 역사 및 연구 동향

2,000년도에 들어서면서, 인간의 유전체(genome) 지도를 완성했지만, protein coding gene 이외의 영역에는 어떤 생명 현상이 일어나는 지는 여전히 미지에 가깝다[1]. 특히, protein(단백질)을 만들지 못하는 non-coding RNAs (ncRNAs)의 등장은 생물학계에서 “dark matters (암흑물질)”라는 이름으로 불리기도 한다[2]. 유전체 상에서 그 기능 및 실체가 잘 알려지지 않은 dark matters의 존재는 현재 next-generation sequencing(NGS)의 도래로 인하여, 생명과학 및 의과학 뿐만 아니라 생물정보학계에도 큰 관심을 일으키고 있다. 학계의 대다수 기존 의견은 인간의 유전체 상에서 주로 protein coding gene 영역에서만 전사가 일어나고, 나머지 영역은 일종의 junk로 생각되었다. 하지만, NGS로 생성된 대용량의 데이터는 생물학자들에게 불편한 진실을 요구하고 있다. 유전체의 거의 전 영역에서 발현양은 적지만, 전사가 일어나고 있으며, 실험상의 오류가 아니라는 증거들이 나오고 있다[1,3,4]. 이러한 관찰을 pervasive transcription 현상으로 학계에서는 인식되고 있다. 이러한 현상의 중요한 특징으로, dark matter로 불리우는 non-coding RNAs가 광범위하게 있다는 사실이 드러나고 있다. 현재 이를 규명하기 위한 생물학 및 생물정보학의 협업 연구가 NGS데이터를 이용하여 대두되고 있다. 더불어, 이러한 ncRNA에도 다양한 종류의 class가 존재한다는 사실도 밝혀지고 있다.

그림 1에서 보듯이, 1961년 protein coding gene이 전사되어 생성된 mRNA의 발견 이후에, 1990년대 초반까지는 non-coding RNA의 발견은 거의 이루어지지 않았다. 하지만, 1993년 이후에 miRNA(ncRNAs의 한 종류) 발견 및 high-throughput 기술의 발달과 더불어 ncRNA분야의 연구는 크게 발전하고 있다[5,6]. 본 논문에서는, 다양한 ncRNA의 class중에서도 H19와 Xist와 같은 long non-coding RNA(lncRNA)를 주로 다루게 될 것이다.

1.2 lncRNA의 정의 및 간략한 기능 소개

앞서 언급한 바와 같이, pervasive transcription 현상 뒤에 다양한 종류의 ncRNA가 protein-coding gene 영역 밖에서 전사되고 있다. 현재까지 대부분의 ncRNA의 연구는 miRNA[5,6]에 집중되어 있다. 그에 반해 lncRNA는 정의조차 완전히 확립되지 않았다. 사실 최초의 lncRNA는 miRNA보다 앞서 발견되었지만, 그 실체는 최근에서야 NGS의 발달로 드러나고 있다[1,7-9].

lncRNA는 일반적으로 200 bp 이상의 protein coding potential(단백질 번역 능력)이 없는 ncRNA를 통칭하고 있다[1,7-9]. 이러한 lncRNA는 다양한 RNA binding protein(e.g., Argonaute, EED, EZH2)과 결합하여 생물학적 현상을 조절하고 있다. lncRNA는 현재 많이 연구되고 있는 ncRNA인 miRNA의 중요한 기능(e.g., mRNA stability, translational inhibition, RNA editing, splicing)

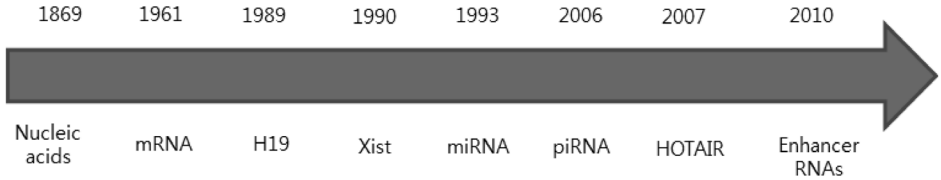

그림 1 주요 RNA관련 발견 역사[1]. 화살표 위의 숫자는 연도를 의미하고, 아래에는 해당 연도에 발견된 여러 RNA를 의미한다.

† 본 연구는 국립암센터 기관고유연구사업 지원(NCC 1210460)으로 이루어진 것입니다.

도 할 수 있을 뿐만 아니라, chromatin 구조를 변화시킴으로써 다양한 유전자의 발현을 제어할 수 있다고 보고된다[1]. 따라서 우리는 본론에서 현재까지 알려진 lncRNA의 서열 및 유전적 정보를 저장하거나 또는 curation한 생물정보학 기반의 도구 및 데이터베이스를 소개하고자 한다. 결론에서는 향후 lncRNA에서 생물정보학이 나아갈 방향에 대해 간단히 토의한다.

2. 본 론

2.1 lncRNAs에 대한 생물정보학 데이터베이스

lncRNA의 genomic characterization을 위해서 사용할 수 있는 데이터베이스를 표 1에 정리하였다. 현재까지 lncRNA 데이터베이스는 11여개에 불과한데[10-16,18-21], 대부분 lncRNA의 서열, 2차원 구조, 관련 논문 citation과 같은 단편적인 자료들을 제공하고 있다. 물론 이와 같은 이유는 데이터베이스의 curator에 있기 보다는 lncRNA의 생물학적 지식이 아직 많이 축적되지 않았기 때문이다. 그럼에도 몇몇 데이터베이스는 이러한 한계를 벗어나기 위한 시도를 하고 있는데, ChIPBase[12]는 lncRNA의 전사가 어떠한 transcription factor에 의해 조절되는지를 ChIP-Seq과 같은 대용량 sequencing 자료를 활용하여 정량적으로 접근하고 있다. LNCipedia는 "competing endogenous RNA(ceRNA)" [22]라고 알려진 현상에 관여하는 lncRNA에 대한 계산학적인 자료를 공개하고 있다. 현재 이처럼, 생물학자들에게 직관적인 생물학적 메커니즘을 제시하도록 기획된 데이터베이스[12,14]가 향후 lncRNA의 데이터베이스 연구가 지향할 방향으로 보인다. 또한, 표 1에서 보듯이 computational prediction의 결과를 저장한 lncRNA 데이터베이스뿐만 아니라, 생물학자들에게는 manually-curated 데이터베이스도 중요한 자원으로 인식됨을 볼 수 있다.

2.2 lncRNAs에 대한 도구

현재, lncRNA의 발현양을 조사하는 것은 여전히 어

표 1 lncRNA와 관련한 생물정보학 데이터베이스

Methods	웹 주소	설명	참고 문헌
ncRNome	http://genome.igib.res.in/lncRNome	18,000여개의 lncRNA에 대한 서열, 2차원 구조, 문헌 정보를 저장한 knowledge 데이터베이스.	[10]
lncRNAdb	http://www.lncrnadb.org	Pervasive transcription을 주창한 Mattick 그룹에서 만든 lncRNA 데이터베이스. 각종 서열 및 발현 정보를 기록함.	[11]
ChIPBase	http://deepbase.sysu.edu.cn/chipbase	Transcription factor에 의한 lncRNA의 조절을 밝히기 위한 다양한 ChIP-Seq 자료들을 통합한 데이터베이스.	[12]
The Functional lncRNA Database	http://www.valadkhanlab.org/database	lncRNA와 mRNA의 3'UTR 사이의 유사성을 밝힘과 동시에 lncRNA에서 miRNA가 생성됨을 제시함.	[13]
LNCipedia	http://www.lncipedia.org	lncRNA의 2차 구조, 단백질 번역 능력(protein coding potential), miRNA binding sites를 정리한 데이터베이스.	[14]
LncRNADisease	http://cmbi.bjmu.edu.cn/lncrnadisease	lncRNA가 최근에 질병 유병과 관련이 있음에 밝혀짐에 따라, 질병과 연관성이 높은 478개의 lncRNA를 manually curation하였으며, 해당 lncRNA의 (실험적으로 증명된) interacting partner를 제시함.	[15]
DIANA-LncBase	http://www.microrna.gr/LncBase	lncRNA의 target에 대한 manually curated information 뿐만 아니라 computationally predicted information을 저장해놓은 데이터베이스.	[16]
NONCODE v3.0	http://www.noncode.org	lncRNA에 대한 초창기 데이터베이스들 중의 한 종류. protein coding potential 및 FANTOM 프로젝트[17]의 cDNA microarray에서 발견된 lncRNA의 발현량을 제시함.	[18]
Noncoding RNA Expression Database(NRED)	http://jsm-research.imb.uq.edu.au/NRED	lncRNA의 초기 데이터베이스들 중의 하나임. human과 mouse에서 lncRNA의 발현량을 보여줌.	[19]
GeneCards v3	http://www.genecards.org	이스라엘의 Weizmann Institute of Science에서 제공하고 있는 lncRNA를 포함한 ~80,000 human ncRNA에 대한 서열 및 간략한 정보를 제공함.	[20]
PLncDB	http://chualab.rockefeller.edu/gbrowse2/homepage.html	식물에 대한 lncRNA 정보를 제공함.	[21]

표 2 lncRNA와 관련한 생물정보학 도구들

Methods	웹 주소	설명	참고 문헌
ncFANs	http://ebiomed.org/ncfans/	Affymetrix에서 제작한 chip에서 lncRNA의 발현양을 조사해주는 도구. Stand-alone program 및 web server 모두 제공.	[24]
Linc2GO	http://www.bioinfo.tsinghua.edu.cn/~liuke/Linc2GO/index.html	lncRNA의 sub-class인 lincRNA의 기능을 예측하기 위하여, lincRNA를 타겟팅하는 miRNA를 조사함. 이 도구는 ceRNA 가설[22]을 이용한 첫 번째 도구임.	[25]
Software.ncrna.org	http://software.ncrna.org/	ncRNA연구에 사용되는 다양한 alignment 도구들과 web server 도구들을 정리함.	[26]
iSeeRNA	http://www.myogenesisdb.org/iSeeRNA	Support Vector Machine을 이용하여 transcriptome data로부터 lncRNA의 일종인 lincRNA를 발견하는 web server.	[27]
Noncoder	http://noncoder.mpi-bn.mpg.de/#	Affymetrix GeneChip Exon 1.0 ST arrays(exon arrays)을 이용하여서, lncRNA의 발현량을 조사하는 web server.	[28]
ncPRO-seq	http://ncpro.curie.fr	NGS에서 small RNA-Seq 분석을 위한 stand-alone pipe 라인임. fastq, fasta, color space 포맷과 더불어 bam format의 파일을 input으로 받음. web server도 지원함.	[29]
lncRScan (long non-coding RNA Scan)	http://code.google.com/p/lncrscan/	NGS에서 RNA-Seq분석을 통하여 lncRNA를 찾아내고 발현양을 보고하는 도구.	[30]
Pipe-R	http://tcoffee.crg.cat/apps/piper/do#	lncRNA의 genomic location을 찾아주는 web server.	-
PhyloCSF	http://compbio.mit.edu/PhyloCSF	다양한 종(種)간의 Phylogenetics를 이용하여 단백질 번역 능력을 측정해주는 stand-alone 도구.	[31]
miRcode	http://www.mircode.org	lncRNA를 targeting하는 miRNA를 찾아주는 도구들. ceRNA가설[22]에 기반하여 만든 web server.	[32]
CoRAL	http://wanglab.pcbi.upenn.edu/coral	ncRNA의 다양한 종류(lncRNA를 포함하여)를 기계학습에 의하여 classification 해주는 도구.	[33]

려운 문제로 남아있다. 최근에, NGS의 sequencing 데이터를 이용해서 lncRNA의 발현양을 조사할 수 있는 도구들이 속속 등장하고 있다. 또한, 기존에 많이 사용해왔던 microarray의 probes들을 re-annotation해서, lncRNA 발현양을 측정할 수 있는 방법도 등장했다. 이에 따라서 GEO[23]와 같은 public microarray repository에 저장된 데이터를 이용하여, 기존의 mRNA 발현양 뿐만 아니라 lncRNA의 발현양도 함께 조사할 수 있게 되었다. 기존의 발현 데이터를 표 2와 같은 도구들[24-33]을 사용함으로써, lncRNA와 mRNA의 통합 분석이 가능할 수 있게 된 것이다.

3. 결 론

최근 lncRNA가 유전자 발현을 조절하고, 또한 chromatin 구조를 변화시킬 수 있다고 밝혀지고 있다. 다양한 질병에서도 lncRNA가 연관되어 있다고 알려지고 있다. lncRNA가 다른 therapeutic intervention에 비해 우수한 점은 유전체상에 위치한 특정 영역의 유전자들을 정확하게 조절할 수 있다[9]. 이는 약(drug)으로서의 중요한 요구 사항인 target specificity 측면에서 이점이 될 수 있다. 따라서 이러한 장점을 극대화하기 위해서는 lncRNA의 분자 수준의 기작 이해와 phenotype의 변화에 대한 연구가 진행되어야 한다. 이를 위해서는 lncRNA가 어떤 유전자를 타겟하여 조절하고, 이 과정에서 어떤 RNA-binding protein 및 chromatin remodeling enzyme과 상호 작용하는지가 먼저 규명되어야 한다. 이를 위하여, 기존에 알려진 lncRNA와 타겟 유전자사이의 서열 및 구조 정보를 이용하여 feature를 정의하고, 이를 적합한 모델로 설계하는 것이 중요하다. 특히, (miRNA와 달리) lncRNA는 구조에 의존적이기 때문에 이러한 문제 해결에 큰 도전을 받고 있다.

참고문헌

[1] Rinn J. L., Chang H. Y., “Genome regulation by long noncoding RNAs”, Annu Rev Biochem, Vol. 81, pp. 145-166, 2012

[2] Michalak P., “RNA world-the dark matter of evolutionary genomics”, J Evol Biol, Vol. 19, No. 6, pp. 1768-1774, 2006

[3] Dinger M. E., Amaral P. P., Mercer T. R., Mattick J. S., “Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications”, Brief Funct Genomic Proteomic, Vol. 8, pp. 407-423, 2009

[4] Kapranov P., St Laurent G., Raz T., Ozsolak F., and et al., “The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA”, BMC Biol, Vol. 8, pp.149, 2010

[5] Zamore P. D., Haley B., “Ribo-gnome: the big world of small RNAs”, Science, Vol. 309, pp. 1519-1524, 2005

[6] Bartel D. P., “MicroRNAs: genomics, biogenesis, mechanism, and function”, Cell, Vol. 116, No. 2, pp. 281-297, 2004

[7] Baker M., “Long noncoding RNAs: the search for function”, Nature Methods, Vol. 8, pp. 379-383, 2011

[8] Chowdhury D., Choi Y. E., Brault M. E., “Charity begins at home: non-coding RNA functions in DNA repair”, Nat Rev Mol Cell Biol, Vol. 10, No. 3, pp. 155-159, 2009

[9] Wahlestedt C., “Targeting long non-coding RNA to therapeutically upregulate gene expression”, Nat Rev Drug Discov, Vol. 12, No. 6, pp. 433-446, 2013

[10] Bhartiya D. Pal K., ghosh S., Kapoor S. and et al., “lncRNome: a comprehensive knowledgebase of human long noncoding RNAs”, Database(Oxford), article ID: bat034, 2013

[11] Amaral P. P., Clark M. B., Gascoigne D. K., Dinger M. E., and et al., “lncRNAdb: a reference database for long noncoding RNAs”, Nucleic Acids Res, Vol. 39, pp. D146-D151, 2011

[12] Yang J. H., Li J. H., Jiang S., Zhou H. and et al., “ChIPBase: A database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data”, Nucleic Acids Res, Vol. 41, pp. D177-D187, 2013

[13] Niazi F., Valadkhan S., “Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3'UTRs”, RNA, Vol. 18, pp. 825-843, 2012

[14] Volders P. J., Helsens K., Wang X., Menten B., and et al., “LNCipedia: a database for annotated human lncRNA transcript sequences and structures”, Nucleic Acids Res, Vol. 41, pp. D246-D251, 2013

[15] Chen G., Wang Z., Wang D., Qiu C., and et al., “LncRNADisease: a database for long-non-coding RNA-associated diseases”, Nucleic Acids Res, Vol. 41, pp. D983-D986, 2013

[16] Paraskevopoulou M. D., Georgakilas G., Kostoulas N., Reczko M., and et al., “DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs”, Nucleic Acids Res, Vol. 41, pp. D239-D245, 2013

[17] Bono H., Yagi K., Kasukawa T., Nikaido I. and et al., “Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarrays”, Genome Res, Vol. 13, pp. 1318-1323, 2003

[18] Bu D., Yu K., Sun S., Xie C., and et al., “NONCODE v3.0: integrative annotation of long noncoding RNAs”, Nucleic Acids Res, Vol. 40, pp. D210-D215, 2012

[19] Dinger M. E., Pang K. C., Mercer T. R., Crowe M. L., and et al., “NRED: a database of long noncoding RNA expression”, Nucleic Acids Res, Vol. 37, pp. D122-D126, 2009

[20] Belinky F., Bahir I., Stelzer G., Zimmerman S., and et al., “Non-redundant compendium of human ncRNA genes in GeneCards”, Vol. 29, No. 2, pp. 255-261, 2012

[21] Jin J., Liu J., Wang H., Wong L., and et al., “PLncDB: Plant Long noncoding RNA Database”, Bioinformatics, Vol. 29, No. 8, pp. 1068-1071, 2013

[22] Salmena L., Poliseno L., Tay Y., Kats L., and et al., “A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?”, Cell, Vol. 146, No. 3, pp. 353-358, 2011

[23] Barrett T., Suzek T.O., Troup D. B., Wilhite S. E., and et al., “NCBI GEO: mining millions of expression profiles-database and tools”, Nucleic Acids Res, Vol. 33, pp. D562-D563, 2005

[24] Liao Q., Xiao H., Bu D., Xie C., and et al., “ncFANs: a web server for functional annotation of long non-coding RNAs”, Nucleic Acids Res, Vol. 39, pp. W118-W124, 2011

[25] Liu K., Yan Z., Li Y., Sun Z., “Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis”, Bioinformatics, Epub ahead of print, 2013

[26] Asai K., Kiryu H., Hamada M., Tabei Y., “Software. ncrna.org: web servers for analyses of RNA sequences”, Nucleic Acids Res, Vol. 36, pp. W75-W78, 2008

[27] Sun K., Chen X., Jiang P., Song X., and et al., “iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data”, BMC

Genomics, Vol. 14, pp. Suppl 2:S7, 2013
[28] Gellert P., Ponomareva Y., Braun T., Uchida S., “Non-coder: a web interface for exon array-based detection of long non-coding RNAs”, Nucleic Acids Res, Vol. 41, No. 1, pp. e20, 2013
[29] Chen C. J., Servant N., Toedling J., Sarazin A., and et al., “ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data”, Bioinformatics, Vol. 28, pp. 3147-3149, 2012
[30] Sun L., Zhang Z., Bailey T. L., Perkins A. C., and et al., “Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study”, BMC Bioinformatics, Vol. 13, pp. 331, 2012
[31] Lin M. F., Jungreis I., Kellis M., “PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions”, Bioinformatics, Vol. 27, No. 13, pp. i275-i282, 2011
[32] Jeggari A., Marks D. S., Larsson E., “miRcode: a map of putative microRNA target sites in the long non-coding transcriptome”, Bioinformatics, Vol. 28, No. 15, pp. 2062-2063, 2012
[33] Leung Y. Y., Ryvkin P., Ungar L. H., Gregory B. D., and et al., “CoRAL: predicting non-coding RNAs from small RNA-sequencing data”, Nucleic Acids Res, Epub ahead of print, 2013

약 력

남 승 윤

2002 서울대학교 화학 (이학사)
2004 서울대학교 협동과정 생물정보 (이학석사)
2008 서울대학교 협동과정 생물정보 (이학박사)
2008~2010 Indiana Univ. School of Medicine (postdoc)
2009~2011 한국과학기술정보연구원 (선임연구원)
2011 현재 국립암센터 융합기술연구부 (선임연구원)
관심분야 : non-coding RNA, 암화 후성유전체, 계산생물학 모델링
E-mail : seungyoon.nam@ncc.re.kr