조절 Long Noncoding RNA의 생물정보학적 주석과 도전과제

한양대학교 ▌남 진 우

1. 서 론

최근 next-generation sequencing(NGS_ 기술의 발전에 힘입어 인간 전장 유전체의 다양한 신호를 바탕으로 그 기능을 자세히 조사한 ENCODE 프로젝트가 그 막대한 양의 데이터와 결과를 발표하였다[1,2]. 약 150여개의 세포주에서 DNaseI 신호, Histone modification 신호, RNA-seq, ChIP-seq 등의 NGS 데이터를 생산, 분석하여 인간 전장 유전체의 약 80%가 기능적 요소를 가지고 있을 것으로 예상하였다[1]. 이것은 인간 유전체의 대부분이 세포 활동을 위해 중요한 부분임을 새로이 밝힌 것으로, 만약 유전체의 변이가 단백질 유전자 부분에 생기지 않고, 유전체의 다른 영역에 변이가 생겨도 심각한 세포 활동에 영향을 미칠 수 있음을 설명하는 것이다. 그러므로 단백질 생성과 관련되는 mRNA 전사체 외에도 기능적으로 중요한 모든 기능 인자의 동정, 분류, 특성 분석 연구가 시급히 진행되어야 한다.

인간 유전체 지도가 완성된 2000년 초, 유전체의 약 2%만이 단백질을 만드는 exon으로 이루어져 있으며, 나머지는 intron, intergenic region으로 구성된 noncoding region 또는 Junk region일거라 추측하였다[3]. 하지만 최근 ENCODE 프로젝트 결과로 인간 유전체의 62%의 영역에서 두 개 이상의 세포 주에서 전사를 통해 RNA를 생성하고 있다고 보고하였다[1](그림 1). 그 나머지의 상당수가 이미 밝혀졌거나, 밝혀지지 않은 noncoding RNA(ncRNA)임이 밝혀졌으며, 그 중에서도 유전자의 발현을 조절하는데 기여하는 조절 ncRNA임이 대다수를 차지함이 밝혀졌다.

~~lncRNA는~~ miRNA precursor나 다른 structural noncoding RNA와의 구별을 위해 일반적으로 200nt 이상

† 본 연구는 미래창조과학부에서 제공하고 한국연구재단이 지원하는 연구비로 (NRF-2013R1A1A1010185) 수행되었음.

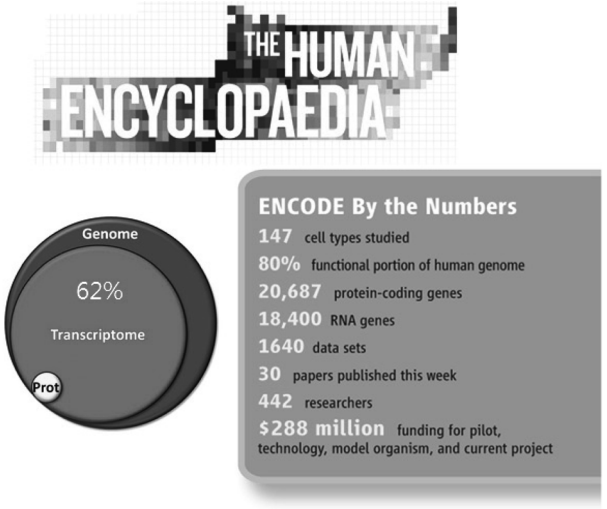

그림 1 인간 ENCODE project 결과 요약

인 noncoding RNA로 분류한다[4]. mRNA-like lncRNA는 mRNA처럼 RNA polymerase II에 의해서 전사되어, 5'capping, splicing, polyadenylation의 성숙 과정을 거치게 된다[4,5]. mRNA와 같이 세포내에서 안정되어 있으며, 약 40%의 lncRNA가 multi-exonic structure를 갖는다. 하지만, open reading frame(ORF)가 없거나 상당히 짧은 ORF를 가지며, coding potential을 갖고 있지 않다[5]. mRNA와 마찬가지로 lncRNA도 promoter 구조를 가지고 있으며, 다양한 세포 특이적 전사인자에 의해 발현이 조절되고 있으며, lncRNA 전사 시 H3K4me3 신호를 promoter 영역에서 H3K36me3 신호를 유전자 영역에서 확인할 수 있다[4,6].

ENCODE 프로젝트의 gene annotation 프로젝트(GENCODE)에서 약 2만개의 인간 lncRNA을 동정하였으며[7], 별도로 진행된 lncRNA human bodymap 프로젝트에서 14000여개의 인간 lncRNA를 주석하였으며[8], 현재까지 약 25000여개의 lncRNA가 데이터베이스에

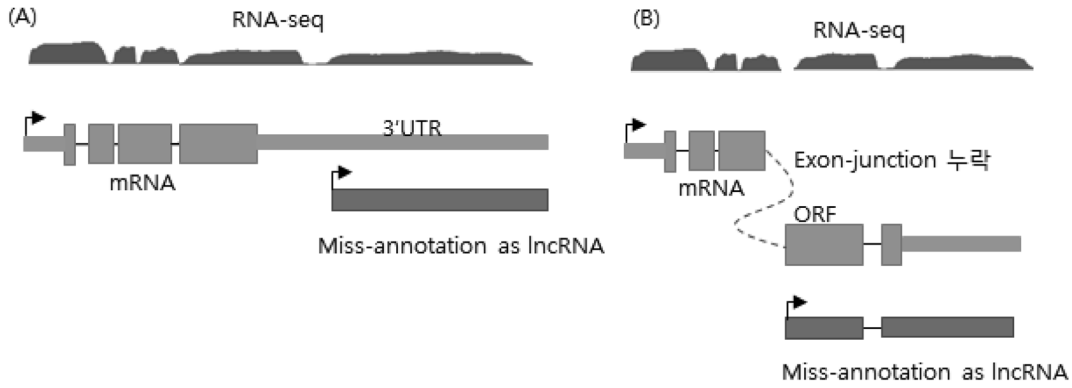

그림 2 대표적으로 잘못 주석된 lncRNA 예들

보고되어 있다. 생쥐의 경우 현재까지 약 6400여개의 lncRNA가 동정되었다[4]. 하지만 최근 조사에서 현 lncRNA의 주석의 30-40% 정도가 잘못 주석된 것으로 알려지고 있으며, 특히 알려진 단백질 유전자의 조각들이 lncRNA로 오인되는 경우가 상당히 발생하고 있다. 그리하여 좀더 정확한 lncRNA 전사체 구조를 예측하고 단백질 유전자와의 분류를 더 정확하게 하기 위한 많은 연구들이 진행되고 있다.

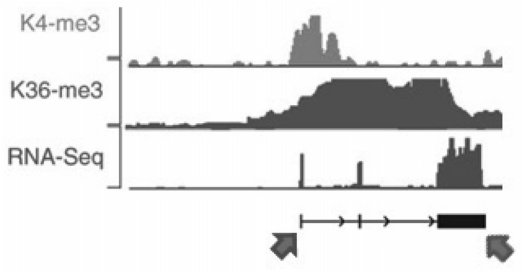

그림 3 LncRNA 주변의 histone modification 신호

2. lncRNA 주석 연구의 어려움들

1) Nonsense of noncoding: 주석이 잘못된 lncRNA 예

현재 주석된 lncRNA중 약 10%가 알려진 또는 새롭게 발견된 단백질 유전자의 아래쪽 10kb 내에 존재하여, 단백질 유전자의 3'UTR 조각으로 의심되고 있다.

2) Transcript의 전사 시작 위치와 3' 말단 부분이 명확하지 않음

RNA-seq 데이터를 기반한 전사체 재구성과 유전자 주석의 또 다른 어려움은 전사 시작 지점과 전사체 끝 지점을 정확히 할 수 없다는 것이다. H3K4me3, H3K-36me3 histone modification 데이터는 대략의 유전자 경계 정보를 제공해 줄 수 있지만, 정확한 위치 정보를 제공해 주지는 못한다(그림 3). 이 문제는 lncRNA 주석에도 같은 문제를 일으키며, 정확하지 않은 lncRNA의 5' 3' 말단 정보를 갖게된다. 이러한 문제로 한 개의 lncRNA 유전자가 두 개의 별도 lncRNA로 잘못 annotation 되는 경우가 생길 수 있다.

3) Intronic lncRNA과 antisense lncRNA 동정의 어려움

lncRNA의 주석의 또 다른 어려움은, 많은 lncRNA 유전자 구조가 단백질 유전자의 구조와 복잡하게 얽혀 있어, 전사되는 RNA가 독립적인 RNA인지 아니면 mRNA isoform의 RNA인지 구별하기 쉽지 않다는 점이다[6]. 특히, sense overlap되거나 sense intronic, tandem array로 위치하는 경우 RNA-seq 데이터만으로 mRNA와 lncRNA를 구별하기 쉽지 않다. 또한 lncRNA가 반대쪽 strand에 존재하는 경우에는 반드시 strand 정보가 제공되는 RNA-seq 데이터를 사용해야만 RNA 구조를 정확히 구별할 수 있게 된다.

4) 짧은 펩타이드를 만드는 lncRNA

최근 Weissman 그룹에서 개발된 Ribosome profiling의 NGS 기술[10]과 글로벌 mass-spectrometry를 이용하여 해당 RNA의 번역(translation) 효율을 측정하는 기술을[11] 이용하여 알려진 lncRNA의 단백질 합성 가능성을 확인하는 연구가 발표 되었다[12]. Ribosome profiling은 RNA상에 결합되어 있는 ribosome의 양을 정량함으로써 번역(translation)의 효율을 관측할 수 있는 실험으로 널리 사용되고 있다[10]. 최근 생쥐의 배아 줄기 세포에서 진행된 ribosome profiling 실험은 알려진 lncRNA의 일부가 짧은 펩타이드를 생성하고 있음을 밝혔으며, 특별히 잘 알려진 open read frame 구조와 다른 ORF 구조가 발견되었다[12]. 이러한 사실은 현재 알려진 lncRNA의 일부가 실제로는 'coding' RNA임을 보여주며, 이렇게 잘못 annotation 된 lncRNA를 배제하기 위해 글로벌 단백질 번역 효율을 측정한 데이터의 사용이 필수적이다. 그러나 ribosome profiling의 경우 잘 알려진 noncoding RNA(예를 들어 miRNA

나 snoRNA)에도 불특정 신호가 나타나, ribosome profiling 데이터의 효과적인 활용을 위해서는 노이즈 백그라운드를 구분하는 새로운 분류자를 개발할 필요성이 대두되고 있다.

3. 전사체 재구성과 lncRNA 분류를 위한 새로운 생물정보학적 접근

1) multimodal NGS 데이터를 이용한 ab initio 전사체 재구성 방법

Cufflinks과 같은 기존 RNA-seq 기반 전사체 재구성 프로그램은[13] 그림 4의 I 단계에서 보듯 재구성된 전사체가 독립적인지 아니면 근접한 다른 전사체의 조각인지 판단하기 어려울 때가 많다. 이러한 경우 전사 시작점을 시퀀싱 하기 위해 개발된 Deep Cap Analysis Gene Expression Sequencing(CAGE-seq)을 [14] 분석하여, 유전체 수준에서 RNA의 전사 시작점을 결정할 수 있다(II 단계). 다음 전사체의 3' 말단을 찾기 위해 PolyA-seq[15] 데이터를 분석하여 유전체 수준에서 RNA의 3' 말단을 결정할 수 있다(III 단계). II, III 단계에서 결정된 전사체 말단 정보를 이용하여 각 전사 RNA가 독립적인지 아니면 근접한 또 다른 전사 RNA의 조각인지를 최종 판정하게 된다(IV 단계).

그림 5는 재구성된 독립 전사체들에 대해 lncRNA인지를 분류하기 위한 전산학적인 coding potential을 측정 방법을 보여준다. 기존에 알려진 모든 단백질의 아미노산 서열을 이용하여, lncRNA 후보군에서 나타날 수 있는 아미노산 서열이 있는지를 확인한다. 이때 blastx를 이용하여 유사 서열이 존재하는 지를 조사하여 그 통계적 유의성을 계산한다. coding potential을 측정하며(Coding potential calculator) 높은 coding potential 점수를 갖는 lncRNA 후보는 단백질 유전자로 분류한다. 두 번째 단계로, 가능성이 있는 ORF 지역의 Synonymouse(아미노산 변이가 없는) / Nonsynonymous (아미노산 변이가 있는) 변이율을 측정하여 Nonsynonymous 변이가 적은 RNA는 Coding potential이 높은 것으로 추정하고, 이를 단백질 유전자로 분류한다.

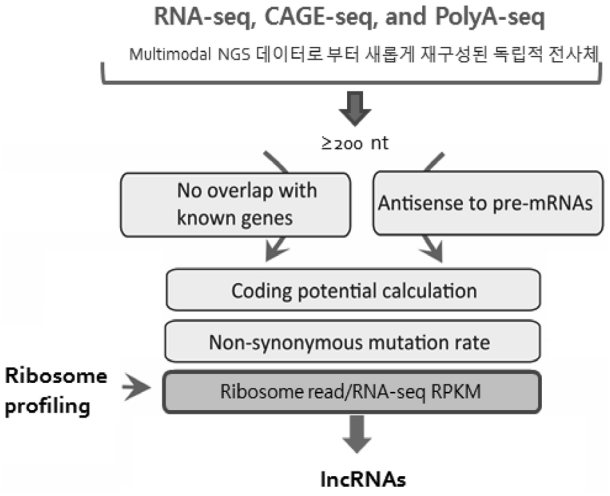

그림 5 lncRNA 분류를 위한 알고리즘들

3) Ribosome profiling 데이터를 이용한 'coding-potential' 전사체 제거 방법

전산학적인 coding potential 측정 방법 외에 Ribosome profiling 데이터를 이용하여 lncRNA 후보 중 자신의 발현양에 비해 ribosome 신호가 강한 것들은 번역 효율이 높다고 판단하여 제거한다[9]. 하지만, 전사체의 발현양에 따라 Ribosome 신호의 background가 동시에 높아지는 경향이 있기 때문에, 발현양에 따라 ribosome 신호를 보정해주지만, 여전히 백그라운드 노이즈가 있는 lncRNA 후보군들이 존재하는 것으로 알려졌다[9].

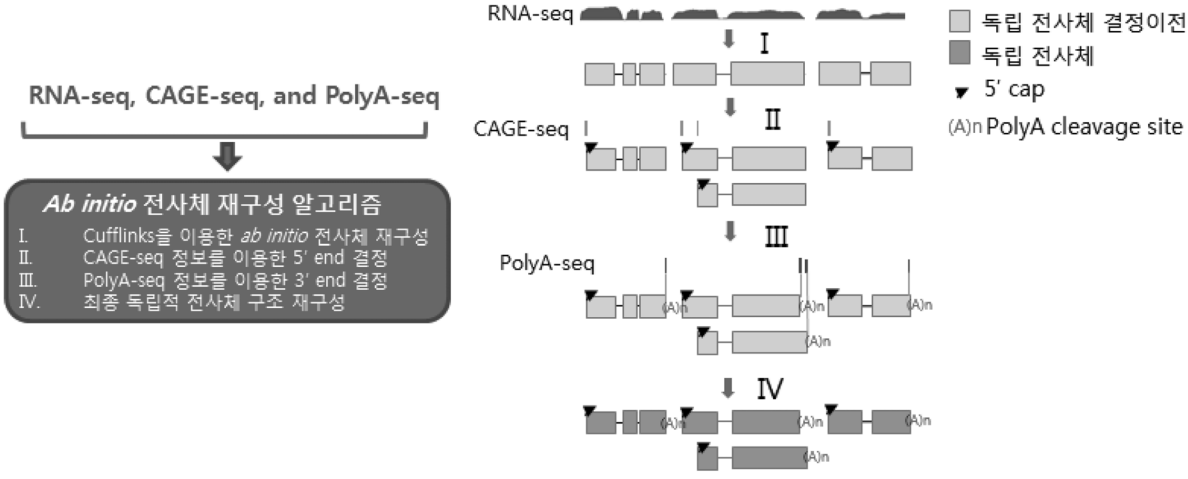

그림 4 새롭게 제안된 전사체 재구성 알고리즘

표 1 기계학습을 이용한 lncRNA 분류 연구들

연구 논문	사용한 특질	기계학습 알고리즘
[16]	PhastCons score, ORF length, ORF proportion, frequencies of seven di- or tri- nucleotide sequences(GC, CT, TAG, TGT, ACG and TCG)	Support Vector Machine (SVM)
[17]	sequence length, CPC, minimum free energy(MFE), frequency of three neighboring bases, content G and C, frequency of a single base, MFE/length, MFE/GC-content	Genetic Algorithm and SVM combined (GA-SVM)
[18]	translation efficiency(TE), inside versus outside, fraction length, disengagement score	Random Forest
[19]	maximun length of ORF, ORF coverage (proportion), Fickett TESTCODE score, hexamer score	Logistic Regression

4) 기계학습을 이용한 lncRNA 분류 방법

최근 lncRNA를 보다 정확하고 효율적인 분류하기 위한 다양한 기계학습 알고리즘들이 개발되고 있다. lncRNA 서열, 보존성, 구조등 다양한 특질들이 학습에 이용되고 있으며, 특히 CPC 값 PhyloCSF 같이 alignment에 의존한 방법들뿐 아니라, ORF 길이 nucleotide 빈도, minimum free energy(MFE), GC ratio, codon bias 같은 alignment-free 방법들도 소개되고 있다(표 2). 일부 방법에서는 translation efficiency를 측정하기도 하였다. 각 프로그램 마다 다양한 특질을 이용하여, Support vector machine(SVM), GA-SVM, logistic regression, random forest와 같은 기계학습 알고리즘이 적용하고 성공적인 lncRNA 분류를 진행하였다. 하지만 여전히 많은 잘못된 lncRNA 분류가 나타나고 있는 실정이다.

4. 결 론

유전체의 noncoding 영역을 이해하고 새로운 중요인자를 발굴하는 것은 인간 유전체에 변이에 의해 생기는 다양한 질병의 발생 기전을 이해하고, 이를 진단하고, 더 나아가 치료의 길을 여는데 상당히 중요한 시점이다. 특히 인간 유전체를 모두 해독하고도 여전히 질병을 정복을 못하고 있는 가장 큰 이유가, 현재까지도 유전체의 98%에 해당하는 noncoding 영역을 정확히 이해하지 못하고 있기 때문이다. 본 논문에서 소개한 다양한 전사체 재구성의 도전과제와 이를 해결할 수 있는 몇가지 제시된 방법은 앞으로 식물, 동물 뿐 아니라 인간 유전체의 전사체를 해독하고 이해하는데 큰 도움을 줄 것으로 기대한다. 특히 noncoding 영역에서 많이 생성되는 lncRNA의 정확한 분류와 주석은 앞으로 더욱더 중요해질 것이며, 이를 위해 다양한 전산학적, 생물정보학적 방법이 적용될 것으로 기대된다.

참고문헌

[1] Gerstein, M.B., et al., Architecture of the human regulatory network derived from ENCODE data. Nature, 2012. 489(7414): p. 91-100.

[2] Gerstein, M., Genomics: ENCODE leads the way on big data. Nature, 2012. 489(7415): p. 208.

[3] Lander, E.S., et al., Initial sequencing and analysis of the human genome. Nature, 2001. 409(6822): p. 860-921.

[4] Khalil, A.M., et al., Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A, 2009. 106(28): p. 11667-72.

[5] Garber, M., et al., Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods, 2011. 8(6): p. 469-77.

[6] Ulitsky, I., et al., Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell, 2011. 147(7): p. 1537-50.

[7] Harrow, J., et al., GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res, 2012. 22(9): p. 1760-74.

[8] Cabili, M.N., et al., Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev, 2011. 25(18): p. 1915-27.

[9] Nam, J.W. and D.P. Bartel, Long noncoding RNAs in C. elegans. Genome Res, 2012. 22(12): p. 2529-40.

[10] Ingolia, N.T., et al., Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 2009. 324(5924): p. 218-23.

[11] Tanaka, T., Y. Yamada, and M. Ikehara, Polymer support synthesis of oligodeoxyribonucleotide with an aminoethyl or aminohexyl group at the 5' end by the phosphite-triester approach. Chem Pharm Bull(Tokyo), 1988. 36(4): p. 1386-92.

[12] Ingolia, N.T., L.F. Lareau, and J.S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.

Cell, 2011. 147(4): p. 789-802.
[13] Trapnell, C., et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010. 28(5): p. 511-5.
[14] Kodzius, R., et al., CAGE: cap analysis of gene expression. Nat Methods, 2006. 3(3): p. 211-22.
[15] Jan, C.H., et al., Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature, 2011. 469(7328): p. 97-101.
[16] Sun, K., et al., iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics, 2013. 14 Suppl 2: p. S7.
[17] Wang, Y., et al., Computational identification of human long intergenic non-coding RNAs using a GA-SVM algorithm. Gene, 2014. 533(1): p. 94-9.
[18] Chew, G.L., et al., Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. Development, 2013. 140(13): p. 2828-34.
[19] Wang, L., et al., CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res, 2013. 41(6): p. e74.

약 력

남 진 우

2001 연세대학교 생물학과 졸업(학사)
2004 서울대학교 협동과정 생물정보학 전공 졸업(석사)
2007 서울대학교 협동과정 생물정보학 전공 졸업(박사)
2007~2008 서울대학교 microRNA ~~창의 연구소~~ Postdoctoral Associate
2008~2012 Whitehead Institute for Biomedical Research/MIT
2012~2014 한양대학교 의생명공학전문대학원
2014~현재 한양대학교 자연과학대학 생명과학과
관심분야 : ncRNA연구, 전사체/유전체 연구, 질병/암 유전체 연구, 생물정보학, 기계학습
E-mail : jwnam@hanyang.ac.kr