Mitigating the lack of knowledge about long noncoding RNA: extracting biological functions

Yagoub A. I. Adam[1], Evandro Eduardo Seron Ruiz[2], and Alessandra Alaniz Macedo[2]

[1] Inter-institutional Graduate Program on Bioinformatics, Department of Computer Science and Mathematics, FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil, yagoubali@gmail.com

[2] Department of Computer Science and Mathematics, FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil.

Abstract. Long non-coding RNAs (lncRNAs) play crucial roles in diverse biological processes. For example, they help to regulate gene expression and to shape the 3D organization of the cellular nucleus. However, their functions in these processes are not well known. Most of the ongoing research efforts regarding lncRNAs have focused on predicting their properties and their functions. Here, we aimed to extract the biological functions of lncRNAs referred in the biomedical literature. To this end, we performed gene enrichment tests for 417 verified gene names extracted from the scientific literature. We also succeeded in extracting biologically significant functions from 2455 single sentences of 2513 scientific abstracts retrieved from the PubMed digital library. These sentences had to be filtered to eliminate those that did not mention a biological function. They were also split into 2-3-grams and the meaningful function annotations based on Gene Ontology terms were extracted. The results provided by the gene enrichment test and the knowledge gained by automatic extraction of information from the scientific literature showed that lncRNAs play a critical role in all biological and cellular processes. Although simple natural language processing techniques fulfilled our purposes of extracting concepts concerning the biological functions of lncRNAs, in the future, we aim to use more sophisticated methods, such as grammatical rules, to extract more functions and to generate a full database for the biological functions related to lncRNAs.

Keywords: Long noncoding RNA; Biological Function; Information Extraction; Natural Language Processing

1 Introduction

According to the central dogma of gene expression, RNA molecules mediate the information transfer between DNA and proteins. Nevertheless, RNA molecules have been recently discovered to perform other biological functions: ribozymes can act as enzymes, and RNA can participate in the translation machinery [6].

Furthermore, RNA molecules with regulatory functions have been discovered [39], which is further evidence of the diverse functions of RNA molecules and the importance of RNA in complex biological functions. Several studies have revealed that a large amount of RNA is transcribed, but not translated into protein [39, 30, 35, 34, 29]. These types of RNA are defined as noncoding RNA; hence, they are transcribed from noncoding regions. The five major groups of noncoding RNA are:

1. Small nuclear RNAs (snRNAs), which catalyze RNA splicing;
2. Micro RNAs (miRNAs). They regulate the level of targeted gene expression and have characteristic length of between 22 and 26 nucleotides;
3. Small interfering RNAs (siRNAs), which inhibit the transcription of targeted genes with complementary sequences and have epigenetic regulation functions such as defense against infections and chromatin remodeling [3]. These RNAs typically display between 20 and 25 nucleotides and form double-stranded RNA molecules;
4. Piwi-interacting RNAs (piRNAs). They play a role in epigenetic regulations. The length of these RNAs ranges from 24 to 31 nucleotides [16];
5. Long non-coding RNA (lncRNAs). Numerous studies have revealed that many lncNAs containing over 200 nucleotides are transcribed. These diverse noncoding RNAs are classified as lncRNAs [40, 5]. LncRNAs take part in countless processes, including evolutionary functions, regulation of biological functions, stem cell differentiation, regulation of the cell cycle, and regulation of a variety of cellular pathways. Recent studies have shown that lncRNAs are emerging as important regulators of disease processes such as cancer [30, 29, 14, 13, 28, 4, 25, 38, 2, 27, 21–24].

As lncRNAs have a critical role in all the biological and cellular processes, many efforts have been made to characterize and to annotate lncRNAs. lncRNAs have been considered as part of the "dark matter of the genome" due to the limited understanding of their functions and activities. Understanding the functions of lncRNAs has become a holy grail in the field of bioinformatics and molecular biology, and new approaches to annotate lncRNAs are necessary.
In this study, we aim to investigate the role lncRNAs play in all cellular functions by using natural language processing (NLP) methods to further understand their role in genome as its is presented in the biomedical literature. We also intend to mitigate the lack of knowledge about lncRNAs by extracting information from specialized biomedical literature.
The remaining sections of this paper are organized as follows: Section 2 presents some related work in textual processing of biomedical information. Section 3 describes the proposed method to mitigate the lack of knowledge about lncRNAs. This method extracts biological functions from the biomedical literature. Section 4 discusses the results obtained. Section 5 presents our final remarks and suggestions for future work.

2 Textual Processing and Extraction

There are versatile approaches to extract interesting biomedical information from the literature [7]. These approaches mostly include:

- Text mining to named entity recognition (NER) and terminology extraction;
- Text classification; and
- Relationship extraction and hypothesis generation.

As our aim is to correlate lncRNA and their biological functions in the literature, co-occurrence of targeted terms in sentences is a simple method to extract targeted sentences. This method provides wide coverage of the sentences; however, it also extracts high numbers of unrelated sentences. As a result, this approach provides high sensitivity and low specificity for the targeted sentences [11].
On the other hand, the extraction of information by targeted patterns (also called pattern-based techniques) increases the specificity but lowers the sensitivity [36]. Pattern-based techniques are either supervised techniques, in which domain experts manually define patterns through analysis of reliable literature, or unsupervised techniques, whereas patterns are automatically defined and learned through well-annotated literature. To increase the accuracy of the information extracted by the pattern-based approach, the grammatical structure of the sentences should be considered [1].
A well-known NLP method to extract information from biomedical literature is the shallow parsing of sentences into word groups, also called n-grams [15, 20]. The outputs from parsers may be processed by either a supervised rule-based learning method [26, 8, 17, 33] or by an unsupervised learning approach, such as a Support Vector Machine [12, 19, 18].

3 Material and Method

To extract biological functions of lncRNAs from the biomedical literature we designed a pipeline that encompasses the following steps: (i) Abstract text retrieval from PubMed; (ii) Extraction of official gene symbols for gene enrichment test; (iii) Processing of the retrieved primary raw data set and the extraction of relevant sentences; and (iv) Extraction of biological functions from relevant sentences. Figure (1) summarizes this proposal.

3.1 Text Retrieval from PubMed

The first pipeline task was the retrieval of relevant abstracts from PudMed by querying with the term 'noncoding RNA [All Fields]' as a single query term in order to achieve higher sensitivity and a high recall. See Figure 1, **AB**.

3.2 Extraction of Gene Symbols and Gene Enrichment Test

To extract the gene names from the texts, we noticed that sentences enclosed in parentheses and words between three and six characters were extracted as they show a potential of being gene symbols. Afterward the Gene Ontology (GO) terms related to the gene symbols were verified using the `g:rProfile` web-tool version r1536_e83_eg30 [32, 31]. The Gene Ontology enRIchment anaLysis and visuaLizAtion tool (GOrilla) was also used to identify and visualize the enriched GO terms for the verified genes [10, 9]. See Figure 1, **B1** and **B2**. Figures 2 depict the results that are commented in Section 4.

3.3 Processing of the Primary Retrieved Text Raw Data

At this stage, the main target was to simplify the collected sentences by eliminating the potential errors that could occur during the following steps. The text processing task for error prevention consists of:

- Conversion of all characters to lowercase;
- Removal of parenthetical remarks, which release the terms inside a pair of parentheses;
- Normalization of compound words and extended sentences containing the signs '+','-', or '" into single words by removing non-letter characters;
- Replacement of multiple spaces with single space; and finally
- Replacement of the synonymous terms that refer to lncRNAs with a new defined abbreviation. See Table (1).

See Figure 1, **A1**.

3.4 Extraction of relevant sentences

The abstracts were split into sentences and the relevant sentences were selected. The relevant sentences were defined as sentences that contained a predefined abbreviation of long noncoding RNA, according to Table 1 and which were extracted for further analysis. See Figure 1.

3.5 Extraction of the biological functions of lncRNAs from relevant sentences

From the relevant sentences, eventual lncRNAs biological functions were extracted from up to 10 words that came after the keyword 'lncRNA'. This was accomplished by shallow parsing the collection of words into bi-grams and tri-grams. See Figure 1, **A4**.

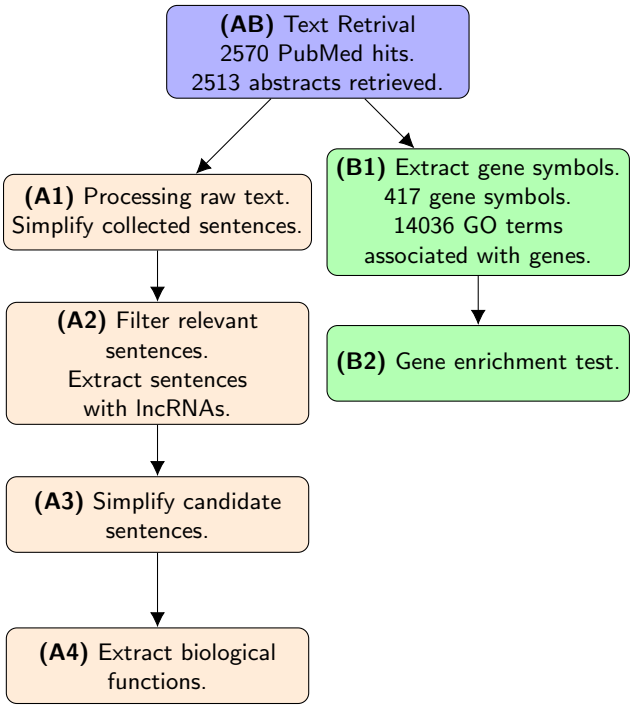

Fig. 1. Flowchart showing the pipeline of our research method.

Table 1. Predefined lncRNAs synonyms which replaced with single term 'lncRNAs'.

long noncoding rnas — long non coding rnas — long noncoding rna — long non coding rna — long notcoding rnas — long not coding rnas — long notcoding rna — long not coding rna — lncrnas — lncrna — lnc rnas — lnc rna — long ncrnas — long ncrna — long nc rnas — long nc rna — long nontranslated rnas — long nontranslated rna — long non translated rnas — long non translated rna — long nottranslated rnas — long nottranslated rna — long not translated rnas — long not translated rna — long untranslated rnas — long untranslated rna — long un translated rnas — long un translated rna — long intergenic nonprotein coding rnas — long intergenic nonprotein coding rna — long intergenic non protein coding rnas — long intergenic non protein coding rna — linc rnas — linc rna — long nonprotein coding rnas — long nonprotein coding rna — long non protein coding rnas — long non protein coding rna

4 Results and Discussion

Querying the PubMed search engine using "noncoding RNA [All Fields]" as the query term, our method retrieved 2513 abstracts out of 2570 related PubMed Id (see Attachment S1 Query result). The remaining 57 abstracts failed to be retrieved mainly due to problems such as format issues: abstracts were in PDF or as image (or other non-XML format), or the documents had no abstract section,

such as laboratory notes, clinical records, patents, etc.
The gene enrichment task was conducted for 417 verified genes by using the Gorilla tool. The results of this test showed that lncRNAs are associated with genes present in larger amount in various important biological processes such as:

- Oncogenesis or tumorigenesis;
- development and differentiation of stem cells as well as regulation of stem cell proliferation;
- Regulation of biological processes (such as transcription) that also regulate the immune system through positive regulation of the production of Type I Interferon;
- Communication within cellular environment through signaling pathways and transport of celular materials; and
- Response to external and internal stress.

To extract the biological functions of lncRNAs from the collected abstracts, the first step was to extract 2455 sentences as candidate sentences containing biological functions. Second, the proposed method collected up 10 words that followed the predefined lncRNA keywords as a final text. This final text was filtered by removing non-relevant high frequency terms. Bigrams and trigrams were then filtered, revealing that lncRNAs are associated with various biological processes such as:

1. Cancer initiation/inhibition, cancer development, cancer metastasis, and cancer progression for various types of cancers including: bladder cancer, gastric cancer, lung cancer, prostate cancer, breast cancer, colorectal cancer, pancreatic cancer, cervical cancer, thyroid cancer, esophageal cancer, liver cancer, colon cancer, urothelial cancer, and brain cancer, cancer growth, cancer invasion, adenocarcinoma, urothelial carcinoma, esophageal carcinomas, neuroblastoma, squamous carcinomas, pancreatic gliomas, lymphoma, lung carcinomas, hepatoblastoma, hepatocellular carcinomas, melanoma, adenocarcinoma, colorectal carcinomas, meningiomas, and glioma vascular cancer;
2. Regulation of biological processes such as: gene expression and transcription, chromatin remodeling and interactions, and regulation of genome;
3. Cellular processes like cell growth, cell migration, cell proliferation, and cell differentiation; and
4. Response to external and internal stress through cell signaling and response to diseases.

Figures 2, 3, and 4 and the table in the S2 attachments depict our results. Figure 2 shows the the p-values of 16 statistically significance GO terms. These terms are related to lncRNA functions. These significant GO terms were considered associated with biological function as they presented p-values ≤ 0.05, according to a hypergeometric distribution applied using the GOrilla software, as explained in subsection 3.2. See Table 4 for the list of the 16 GO terms. Refer to Table S2 in the attachments for more details about the gene enrichment task.

Figure 3 shows the network of biologically significance bi-grams around one of the following terms: Chromatin, Cell, Genome, Expressions, RNA, Regulations,

DNA, Genes, Cancer and Biomarkers. As one of these terms appears in a relevant sentence, obtained as explained in Section 3.4, all the co-occurring bi-grams were collected and linked sequentially as they were found. The bi-gram network formed by this procedure was visualized using Cytoscape [37]. This figure reinforces the number of possible significant functions associated with lncRNAs.

Now considering only single terms, words, Figure 4 shows the 100 top most frequent words in the relevant sentences, i.e. sentences which contain the lncRNA predefined term. Although many words in this list may be considered very ordinal words, by the looking closer for biologically meaningful words the general roles of lncRNAs within the cell could be predicted.

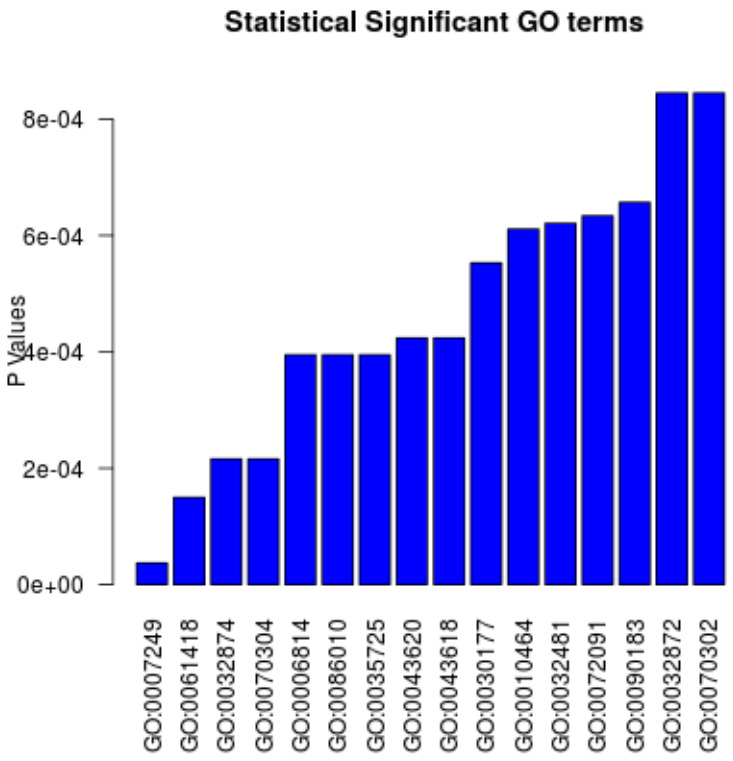

Fig. 2. Significant GO terms from the gene enrichment test.

GO Term	Description
GO:0007249	I-kappaB kinase/NF-kappaB signaling
GO:0061418	regulation of transcription from RNA polymerase II promoter in response to hypoxia
GO:0032874	positive regulation of stress-activated MAPK cascade
GO:0070304	positive regulation of stress-activated protein kinase signaling cascade
GO:0006814	sodium ion transport
GO:0086010	membrane depolarization during action potential
GO:0035725	sodium ion transmembrane transport
GO:0043620	regulation of DNA-templated transcription in response to stress
GO:0043618	regulation of transcription from RNA polymerase II promoter in response to stress
GO:0030177	positive regulation of Wnt signaling pathway
GO:0010464	regulation of mesenchymal cell proliferation
GO:0032481	positive regulation of type I interferon production
GO:0072091	regulation of stem cell proliferation
GO:0090183	regulation of kidney development
GO:0032872	regulation of stress-activated MAPK cascade
GO:0070302	regulation of stress-activated protein kinase signaling cascade

Table 2. Description of the selected GO terms.

5 Final Remarks

The amount of scientific literature available on health subjects is always increasing. For instance, PubMed currently comprises over twenty million MEDLINE citations. Despite the positive effects of having a large amount of information available, some negative aspects arise. The huge volume of scientific information burdens professionals because searches for accurate information are complex and time consuming.

This paper is an initial effort to demonstrate how we have started to discover knowledge by automatically extracting information from a scientific digital repository and, more specifically, extracting terms related to lncRNAs biological function. We have used simple NLP methods to extract concepts related to the biological functions of lncRNAs and to extract genes associated with lncRNAs. We have found that RNA molecules not only function as mediators but also have more complex biological functions. lncRNAs are emerging as important regulators of cellular process and functions and disease processes. Further investigations with more elaborate natural language processing methods, such as the use grammatical rules, might help to extract even more functions and knowledge related to lncRNAs.

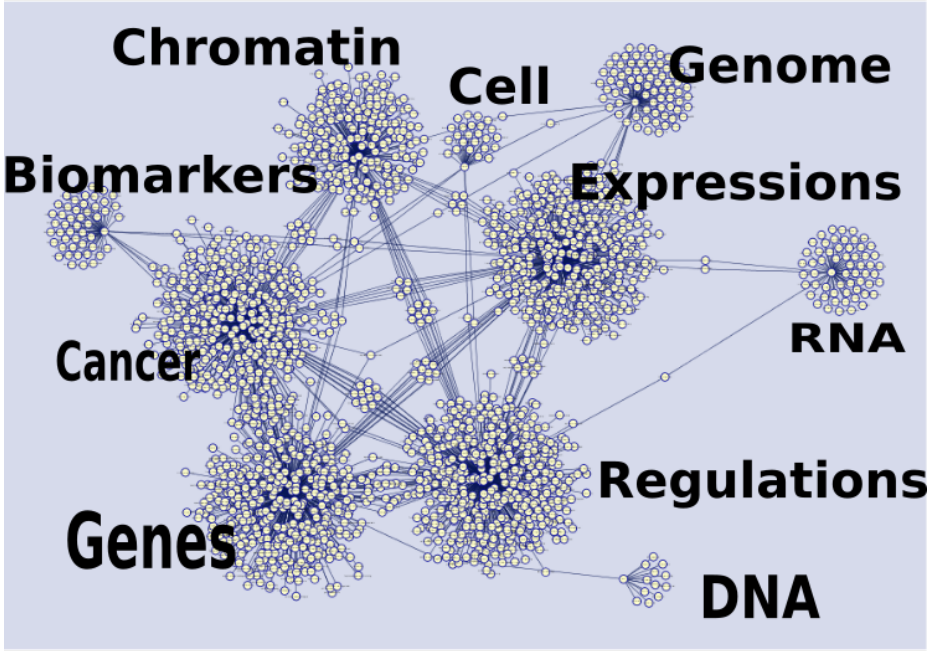

Fig. 3. Networks of bigrams terms in the final text using DNA, Genes, Cancer, Biomarkers, RNA, Expressions, Regulations, Cell, and Chromatin as nodes.

6 Acknowledgments

Yagoub A. I. Adam gratefully acknowledges the financial support of CAPES.

References

1. S. Ananiadou, D. B. Kell, and J.-I. Tsujii. Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24(12):571–579, 2006.
2. D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.
3. E. Bernstein and C. D. Allis. RNA meets chromatin. *Genes & Development*, 19(14):1635–1655, 2005.
4. E. Bernstein, E. M. Duncan, O. Masui, J. Gil, E. Heard, and C. D. Allis. Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Molecular and Cellular Biology*, 26(7):2560–2569, 2006.
5. M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, 25(18):1915–1927, 2011.

6. T. R. Cech. The ribosome is a ribozyme. *Science*, 289(5481):878–879, 2000.
7. A. M. Cohen and W. R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005.
8. J. Czarnecki, I. Nobeli, A. M. Smith, and A. J. Shepherd. A text-mining system for extracting metabolic reactions from full-text articles. *BMC Bioinformatics*, 13(1):172, 2012.
9. E. Eden, D. Lipson, S. Yogev, and Z. Yakhini. Discovering motifs in ranked lists of dna sequences. *PLoS Computational Biology*, 3(3):e39, 2007.
10. E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48, 2009.
11. R. A. Erhardt, R. Schneider, and C. Blaschke. Status of text-mining techniques applied to biomedical text. *Drug Discovery Today*, 11(7):315–325, 2006.
12. G. Erkan, A. Özgür, and D. R. Radev. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *EMNLP-CoNLL*, volume 7, pages 228–237. Prague, Czech Republic, 2007.
13. M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5):503–510, 2010.
14. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, 2012.
15. M. Huang, X. Zhu, and M. Li. A hybrid method for relation extraction from biomedical literature. *International Journal of Medical Informatics*, 75(6):443–455, 2006.
16. Y. W. Iwasaki, M. C. Siomi, and H. Siomi. PIWI-interacting RNA: its biogenesis and functions. *Annual Review of Biochemistry*, 84:405–433, 2015.
17. H. Jang, J. Lim, J.-H. Lim, S.-J. Park, K.-C. Lee, and S.-H. Park. Finding the evidence for protein-protein interactions from PubMed abstracts. *Bioinformatics*, 22(14):e220–e226, 2006.
18. S. Katrenko and P. Adriaans. Learning relations from biomedical corpora using dependency trees. In *Knowledge Discovery and Emergent Complexity in Bioinformatics*, pages 61–80. Springer, 2007.
19. S. Katrenko, P. Adriaans, et al. Learning relations from biomedical corpora using dependency tree levels. In *Proceedings of the Fifteenth Dutch-Belgian Conference on Machine Learning*, 2006.
20. A. Koike, Y. Niwa, and T. Takagi. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, 21(7):1227–1236, 2004.
21. M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–858, 2001.
22. N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–862, 2001.
23. R. C. Lee and V. Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–864, 2001.
24. R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

25. C. Maison, D. Bailly, A. H. Peters, J.-P. Quivy, D. Roche, A. Taddei, M. Lachner, T. Jenuwein, and G. Almouzni. Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nature Genetics*, 30(3):329, 2002.
26. E. K. Mallory, C. Zhang, C. Ré, and R. B. Altman. Large-scale extraction of gene interactions from full-text literature using DeepDive. *Bioinformatics*, 32(1):106–113, 2015.
27. T. Nagano and P. Fraser. No-nonsense functions for long noncoding RNAs. *Cell*, 145(2):178–181, 2011.
28. A. Necsulea, M. Soumillon, M. Warnefors, A. Liechti, T. Daish, U. Zeller, J. C. Baker, F. Grützner, and H. Kaessmann. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505(7485):635, 2014.
29. J. A. Nickerson, G. Krochmalnic, K. M. Wan, and S. Penman. Chromatin architecture and nuclear RNA. *Proceedings of the National Academy of Sciences*, 86(1):177–181, 1989.
30. I. J. Paul and J. D. Duerksen. Chromatin-associated RNA content of heterochromatin and euchromatin. *Molecular and Cellular Biochemistry*, 9(1):9–16, 1975.
31. J. Reimand, T. Arak, P. Adler, L. Kolberg, S. Reisberg, H. Peterson, and J. Vilo. g: Profilera web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, 44(W1):W83–W89, 2016.
32. J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g: Prolife-a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, 35(suppl_2):W193–W200, 2007.
33. F. Rinaldi, G. Schneider, K. Kaljurand, M. Hess, C. Andronis, O. Konstandi, and A. Persidis. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial Intelligence in Medicine*, 39(2):127–136, 2007.
34. M. Salditt-Georgieff and J. Darnell. Further evidence that the majority of primary nuclear RNA transcripts in mammalian cells do not contribute to mRNA. *Molecular and Cellular Biology*, 2(6):701–707, 1982.
35. M. Salditt-Georgieff, M. Harpold, M. Wilson, and J. Darnell. Large heterogeneous nuclear ribonucleic acid has three times as many 5'caps as polyadenylic acid segments, and most caps do not enter polyribosomes. *Molecular and Cellular Biology*, 1(2):179–187, 1981.
36. J. Šarić, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. Extraction of regulatory gene/protein networks from Medline, 2005.
37. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.
38. J. R. Warner, R. Soeiro, H. Birnboim, M. Girard, and J. E. Darnell. Rapidly labeled HeLa cell nuclear RNA: I. Identification by zone sedimentation of a heterogeneous fraction separate from ribosomal precursor RNA. *Journal of Molecular Biology*, 19(2):349–361, 1966.
39. R. A. Weinberg and S. Penman. Small molecular weight monodisperse nuclear RNA. *Journal of Molecular Biology*, 38(3):289–304, 1968.
40. L. Yang, J. E. Froberg, and J. T. Lee. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends in Biochemical Sciences*, 39(1):35–43, 2014.

A Attachments

All the following attachments can be found in goo.gl/oxK1HX

1. S1 Query result in XML format
2. S2 Table of significant GO terms
3. S3 Annotated Genes
4. S4 Bigram networks

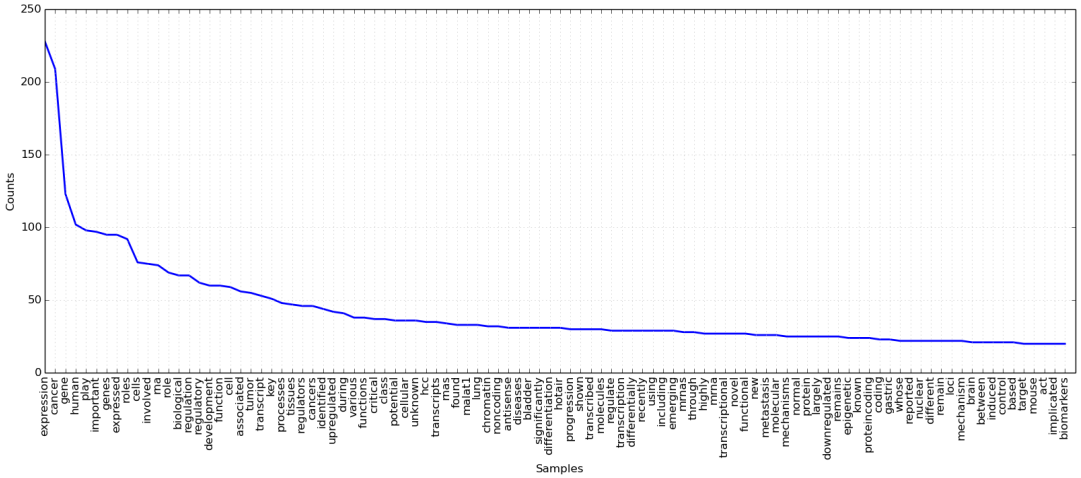

Fig. 4. The 100 most frequency words in the final text.