APPLICATIONS OF BIOINFORMATICS TO NON-CODING RNAS IN THE ERA OF NEXT-GENERATION SEQUENCING

CHAO CHENG
Department of Genetics, Institute for Quantitative Biomedical Sciences,
Norris Cotton Cancer Center, Geisel School of Medicine, Dartmouth College
Hanover, NH 03755, USA
Email: chao.cheng@dartmouth.edu

JASON MOORE
Department of Genetics, Institute for Quantitative Biomedical Sciences,
Norris Cotton Cancer Center, Geisel School of Medicine, Dartmouth College
Hanover, NH 03755, USA
Email: jason.moore@dartmouth.edu

CASEY GREENE
Department of Genetics, Institute for Quantitative Biomedical Sciences,
Norris Cotton Cancer Center, Geisel School of Medicine, Dartmouth College
Hanover, NH 03755, USA
Email: casey.greene@dartmouth.edu

The human genome encodes a large number of non-coding RNAs, which employ a new and crucial layer of biological regulation in addition to proteins. Technical advancement in recent years, particularly, the wide application of next generation sequencing analysis, provide an unprecedented opportunity to identify new non-coding RNAs and investigate their functions and regulatory mechanisms. The aim of this workshop is to bring together experimental and computational biologist to exchange ideas on non-coding RNA studies.

1. Background

Non-coding RNAs (ncRNAs) are RNA molecules encoded by genes in the genome that are transcribed and functional but not translated into proteins. Recent studies have shown that more than 90% human genome is transcribed but coding sequences occupy only a small fraction of the genome (<2%) [1]. This suggests the existence of a large number of non-coding RNAs [2]. In fact, the FANTOM3 (Functional Annotation of Mammalian cDNA) project has identified ~35,000 non-coding transcripts with similar processing as mRNAs, including 5' capping, splicing, and poly-adenylation, but with little or no open reading frame (ORF) [3]. Given the large number of non-coding RNAs, it is reasonable to assume that these molecules are critical players in biological processes. At present, we are just starting to understand the functions of non-coding RNA.

1.1. *Classifications of non-coding RNAs*

Non-coding RNA genes include highly abundant and functionally important RNAs such as transfer RNA (tRNA) and ribosomal RNA (rRNA), as well as RNAs such as snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, and piRNAs among other types.

Based on the size of the mature version of non-coding RNAs, we can divide them into long non-coding RNAs (lncRNAs) and small non-coding RNAs. The cutoff value for size is arbitrarily determined with non-coding RNAs longer than 200 nucleotides categorized as lncRNAs and the rest as small. Compared to the small non-coding RNAs, existing knowledge about lncRNAs is even more limited.

According to their genomic locations, lncRNAs can be grouped into stand-alone lncRNAs, natural antisense transcripts, long intronic RNAs, transcribed pseudogenes and other lncRNAs (e.g. promoter associated RNAs, enhancer RNAs). Importantly, stand-alone lncRNAs are transcription units that do not overlap protein-coding genes. Some of these are referred to as lincRNAs (large intergenic noncoding RNAs). A recent study indicates that the human genome produce tens of thousands of lincRNAs [4].

1.2. *Functions of non-coding RNAs*

The functions of certain non-coding RNA types such as microRNAs have been intensively studied under a variety of biological contexts. However the functions of most of the lncRNAs including lincRNAs remain elusive or unclear. Despite of this, the functionality of lncRNAs is suggested by (1) the conservation of their promoters, splice junctions, exons, predicted structures, genomic; (2) their association with particular chromatin signatures that are indicative of active transcription; (3) their regulation by key molecular signals and transcription factors; (4) their dynamic expression and alternative splicing during differentiation; (5) their tissue- and cell-specific expression patterns and subcellular localization; (6) their altered expression or splicing patterns in cancer and other diseases [5].

In fact, lncRNAs are known to be able to exert regulatory functions at the transcriptional, post-transcriptional and epigenetic levels by different mechanisms. At the transcriptional level lncRNAs target transcriptional activators or repressors, different components of the transcription reaction including RNA polymerase II and the DNA duplex to regulate gene transcription and expression [6]. At the post-transcriptional level they participate in pre-mRNA processing, splicing, transport, translation, and degradation. At the epigenetic level they are involved in gene imprinting, X-chromosome inactivation and many other biological processes.

Several regulatory mechanisms of lncRNAs have been elucidated [7]. First, some lncRNAs can serve as decoys to prevent regulatory proteins from binding to DNA. For example, the lncRNA Gas5 contains a hairpin sequence motif in its secondary structure that resembles the DNA-binding site of the glucocorticoid receptor (GR) and decoy GR to inhibit the transcription of its target genes [8]. Second, some lncRNAs can serve as adaptors to bring two or more proteins into complexes. Third, some lncRNAs are required for guiding the proper localization of specific protein complexes; Finally, some linRNAs can compete with miRNAs for miRNA-binding sites or serve as “sponges” to sequester miRNAs away from their mRNA targets [9].

2. Major directions and challenges

The goal of this workshop is to encourage the development of advanced methods for identification and functional characterization of ncRNAs through a combination of experimental and bioinformatics approaches.

2.1. *Application of bioinformatics to studies of non-coding RNAs*

Computational and bioinformatics techniques have been applied to study non-coding RNA mainly in the following directions: (1) prediction and identification of new non-coding RNAs from genome sequence analysis or by combining computational analysis with experimental data (e.g. tiling array, RNA-seq data); (2) prediction of miRNA target genes; (3) prediction the secondary and tertiary structures of RNAs; (4) investigation on the conservation and evolution of non-coding RNA genes or miRNA target genes; (5) non-coding RNA function prediction by computational analysis such as "guilt by association"; (6) construction of integrated regulatory networks that include non-coding RNA regulatory layers; (7) construction of databases and webservers to facilitate non-coding RNA studies.

2.2. *Main challenges in computational analysis*

Compared to protein studies, application of computational methods to non-coding RNA field is still in its infancy. There are several challenges that limit its application. First, non-coding RNAs represent heterogeneous classes of molecules; each has their specific characteristics and regulatory mechanisms. Second, non-coding RNA genes are non-conserved or less than conserved than protein coding genes; many of them have low expression levels and no obvious knockout phenotypes. Third, the knowledge about non-coding RNAs is still limited, can consequently there is no training data large enough for implementing machine learning techniques or statistical models. Fourth, the quality of non-coding RNAs gene annotation is relatively low. With the technical advancement and accumulation of data, we expect these challenges would be overcome in a short future.

2.3. *Main topics of this workshop*

2.3.1 Identification, annotation, classification and the evolution of lncRNAs.

Computational and experimental methods have been proposed to annotate lincRNAs with special consideration to their lower expression profile. Phylogenetic analysis of lincRNAs in mammalian has demonstrated an interesting evolutionary history of them.

2.3.2 Prediction RNA Secondary Structure

Secondary structure is highly important to the correct processing and function of many non-coding RNAs. Many computational methods have been proposed for modeling and understanding RNA structure.

2.3.3 Expression analysis of lncRNAs

To gain insight into the potential cellular functions of lncRNAs, systematic gene expression profiling has been performed by RNA-seq or tiling array. In particular, disease associated lncRNAs have been predicted by integrative analysis.

2.3.4 Complexity of RNA regulatory mechanism

The regulatory mechanism of non-coding RNAs is very diverse and complicated. With the advancement of non-coding RNA studies, we would expect the discovery of more regulatory mechanisms.

3. Workshop contributions

The workshop includes six invited speakers.

Dr. Runsheng Chen is a Professor in the Institute of Biophysics of Chinese Academy of Sciences. He is a member of Chinese Academy of Sciences. His research focuses on the identification of non-coding RNA genes in multiple organisms, function prediction and annotation of long non-coding RNAs, and the construction of non-coding RNA annotation databases. His lab has developed computational methods and tools for predicting, annotating and classifying non-coding RNAs.

Dr. Yiwen Chen received his PhD in physics from the University of North Carolina at Chapel Hill and is currently a Postdoctoral Fellow in Dr. Shirley Liu's Lab at the Dana-Farber Cancer Institute at Harvard School of Public Health. Dr. Liu's research focuses on developing bioinformatics methods and tools for analyzing high throughput data, using the dynamics of histone mark ChIP-seq and DNase-seq to infer in vivo transcription factor binding and regulation, employing genome wide approaches to understand the specificity and mechanism of epigenetic enzymes and lncRNAs, as well as integrating publicly available high throughput data to better understand cancer mechanisms.

Dr. David Corey is a Professor in the Department of Pharmacology at University of Texas Southwestern Medical Center. He received his PhD in Chemistry from the University of California, Berkeley. His research focuses on the mechanism of promoter-targeted antigene RNAs, the function of Argonaute and small RNA-dependent pathways in mammalian cell nuclei, the allele-selective inhibition of Huntington protein expression as well as the recognition of RNA and DNA by chemically modified nucleic acids and locked nucleic acids.

Dr. Manuel Garber is an Associate Professor in the Program in Bioinformatics and Integrative Biology, and the Director of the Bioinformatics core at University of Massachusetts Medical School. He received his PhD in Mathematics from Brandeis University. His research focuses on the evolutionary history of non-coding genes as well as the systematic dissection of the transcriptional regulation of the immune response. His lab has also been developing the tools to analyze, integrate and fully leverage the advancements in genome wide experimental technologies.

Dr. John Hogenesch is an Associate Professor of Pharmacology and the Associate Director of the Penn Genome Frontiers Institute at the University of Pennsylvania. He

received his PhD in Neuroscience from Northwestern University. His research focuses on the study of the mammalian circadian clock using genomic and computational tools. His lab has a longstanding interest in understanding noncoding RNA function through global gene expression analysis, and functional screening to gain insight into the potential cellular functions of lincRNAs and microRNAs.

Dr. David Mathews is an Associate Professor in the Department of Biochemistry and Biophysics at University of Rochester Medical Center. He received his PhD in Chemistry and MD in Medicine from University of Rochester. His research focuses on predicting RNA structure and developing computational tools for targeting RNA with pharmaceuticals and for using RNA as a pharmaceutical. His lab has developed software for predicting secondary structure of RNAs, software for predicting base pairing probabilities using a partition function and methods for predicting a secondary structure common to multiple sequences.

4. Acknowledgements

We would like to thank all of the speakers for kindly accepting our invitation and generously supporting our workshop, as well as the PSB organizers for their assistance arranging this workshop and providing a venue for these discussions. J.M was supported by the NIH grants LM009012 and LM010098. C.C. and C.G. were supported by the NIH COBRE (Center of Biomedical Research Excellence) grant GM103534.

References

1. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316:**1484-1488.
3. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1563.
4. Hangauer MJ, Vaughn IW, McManus MT: **Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs.** *PLoS Genet* 2013, **9:**e1003569.
5. Mattick JS: **The genetic signatures of noncoding RNAs.** *PLoS Genet* 2009, **5:**e1000459.
6. Goodrich JA, Kugel JF: **Non-coding-RNA regulators of RNA polymerase II transcription.** *Nat Rev Mol Cell Biol* 2006, **7:**612-616.
7. Rinn JL, Chang HY: **Genome regulation by long noncoding RNAs.** *Annu Rev Biochem* 2012, **81:**145-166.
8. Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP: **Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor.** *Sci Signal* 2010, **3:**ra8.
9. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J: **Natural RNA circles function as efficient microRNA sponges.** *Nature* 2013, **495:**384-388.