Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ESSAY

Ribogenomics: the Science and Knowledge of RNA

Jiayan Wu, Jingfa Xiao, Zhang Zhang, Xumin Wang, Songnian Hu, Jun Yu *

CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

Received 4 April 2014; accepted 14 April 2014
Available online 24 April 2014

KEYWORDS

RNA;
Ribogenomics;
Operational;
Informational;
Transcriptome

Abstract Ribonucleic acid (RNA) deserves not only a dedicated field of biological research — a discipline or branch of knowledge — but also explicit definitions of its roles in cellular processes and molecular mechanisms. Ribogenomics is to study the biology of cellular RNAs, including their origin, biogenesis, structure and function. On the informational track, messenger RNAs (mRNAs) are the major component of ribogenomes, which encode proteins and serve as one of the four major components of the translation machinery and whose expression is regulated at multiple levels by other operational RNAs. On the operational track, there are several diverse types of RNAs — their length distribution is perhaps the most simplistic stratification — involving in major cellular activities, such as chromosomal structure and organization, DNA replication and repair, transcriptional/post-transcriptional regulation, RNA processing and routing, translation and cellular energy/metabolism regulation. An all-out effort exceeding the magnitude of the Human Genome Project is of essence to construct just mammalian transcriptomes in multiple contexts including embryonic development, circadian and seasonal rhythms, defined life-span stages, pathological conditions and anatomy-driven tissue/organ/cell types.

Introduction

Ribogenomics is the science and knowledge about ribonucleic acid (RNA). As one of the four major macromolecules (percentage weight in mammalian cell: DNA, ~7 pg, 0.3%; RNA, ~20 pg, 1%; protein, ~500 pg, 20%; and polysaccharide, ~2 μg, 78.7% [1,2]) of cellular life forms, RNA deserves not only a dedicated research field but also definitions of its roles in cellular processes and molecular mechanisms. Therefore, ribogenomics, at least in a sense of cellular mass, in terms of research focus and priority, may not be more imperative than proteomics but certainly has no reason to draw less attention than genomics.

RNA molecules can be divided into two essential functional categories: operational (including what have been defined as catalytic) and informational. At the center of the informational RNAs (other types of informational RNAs including those guiding processes that change mRNA sequences) is messenger RNA (mRNA). In a typical mammalian cell, mRNA takes ~4% of the total RNA mass and aside from 80% ribosomal RNA (rRNA), other operational RNAs make up the rest. If we take the constitutive RNAs — transfer RNAs (tRNAs) and rRNAs — out of the total, the ratio of the operational RNA *vs.* the informational RNA [3,4] is about four.

* Corresponding author.
E-mail: junyu@big.ac.cn (Yu J).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.gpb.2014.04.002

What are the operational RNA types in the dynamic portion of the total RNA? First, all RNA macromolecules are operational. Only the protein-coding portion of all mRNAs is relatively informational, together with certain sequence-specific guiding RNAs (including small sequence-matching RNAs) that may actually aid RNA editing and splicing, while the chemical entity of them remains operational. Second, all non-coding RNAs (ncRNAs) are exclusively operational, including mRNA-like transient transcripts that are often generated from gene duplications — genome-wide, segmental or individual — but may not be translated into functional proteins [5–7]. Transcripts of such kind have been confusing as some of them neither are conserved across closely-related species nor contain normal reading frames albeit often polyadenylated [8,9]. Third, all small RNAs (sRNAs), such as microRNAs (miRNAs), small nuclear RNAs (snRNAs), tRNA-derived sRNAs (tsRNAs), small nucleolar RNAs (snoRNAs) and small interfering RNAs (siRNAs), are operational although they may be processed to become functional in different ways [10–12]. Fourth, long ncRNAs (lncRNAs), intron-encoded or intergenic sequence-encoded, are also all operational, which may act on different aspects of cellular activities and mechanisms [12]. In this article, we first divide ribogenomes into informational and operational tracks, pointing out the obvious differences and intricate relationships between the two tracks, and then provide insights on the research scopes and fundamental scientific questions of ribogenomics under such a scheme.

Ribogenomics on the operational track

Life had started with RNAs [13–15]. Molecular mechanisms and cellular processes of the operational ribogenomic track have to be created earlier than those of the informational track until the genetic code was created [16–20]. We have argued before that early RNA-built life forms may have begun as eukaryote-like organisms since simple life forms might not be able to utilize DNA at all initially and bacteria might be too greedy to keep all complicated molecular mechanisms going, such as RNA splicing and polyadenylation [16]. Fundamentally, RNA macromolecules and their intermediates, as well as building blocks, must have performed all essential cellular functions but some may have lost to proteins over evolutionary time scales. Therefore, active searches for the function of various RNA macromolecules should focus on systematic discovery at all levels and for all facets of molecule mechanisms and cellular processes rather than taking the attitude of "guarding the stump for dumb hare to hit on".

Operational RNAs are diverse in function (**Figure 1**) as well as in sequence length and genomic origin [12,21]. In function-seeking studies, any effort should include both size classes,

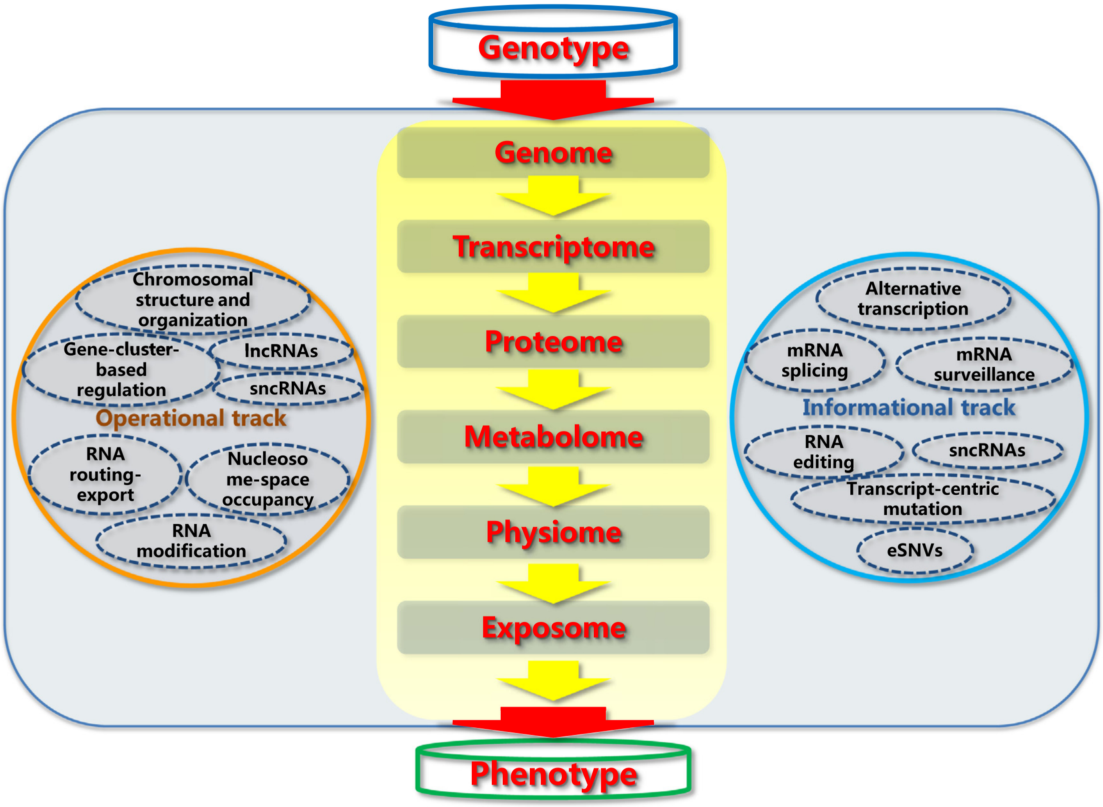

Figure 1 Schematic view from genotype to phenotype with informational and operational tracks

Ribogenomics in the context of a genotype-to-genotype view. Genotype becomes one of the deterministic factors that include ribogenomic, epigenomic, homoeostatic, compartmental and plastic tracks. For the sake of discussion, here we simply classify non-coding RNAs (ncRNAs) into small ncRNAs (sncRNAs) and long ncRNAs (lncRNAs). To emphasize the influence of transcript-centric mutations, we identify expression-related simple nucleotide variations (eSNVs) as an important class of sequence variations that are not yet considered in the context of traditional population genetics and evolution.

large and small, of ncRNAs and choose complementary methods [22,23]. The first level of functional studies is chromosomal structure, conformation and organization as exemplified by the X chromosome inactivation in recent years [24]. A striking recent discovery is the sweeping partition of house-keeping and tissue-specific genes between early and late replicating genes, respectively [25,26]. Aside from the well-known RNA involvement in replication, some sRNAs are also engaged in DNA repair mechanisms, such as DNA double-strand break repair-induced sRNA (diRNA) [27] and tRNA-derived miRNA [28]. The second level to look for RNA-centric regulation is gene-cluster organization that is regulated at multiple levels, such as organization of chromosomal elements or sequence repeats, nucleosome positioning, histone marks, antisense RNAs and transcriptional regulation [29–33]. There have also been many new elements in transcriptional regulation discovered by in-depth transcription level studies [34] and we will be discussing more details in the context of the informational track of ribogenomics. The third level to classify the RNA-centric mechanisms and processes is RNA processing — splicing-routing-exporting. Routing is a novel concept that proposes a possible function for a particular class of introns — minimal introns [34,35]. While the minimal intron is universal and stringent in size, the fraction of minimal intron-containing genes appears stable within and across taxonomic scales [36]. Among mammals, about 10% of the total introns are minimal introns, which are shared by one third of the genes [37]. The fourth level concerns the metabolism or stability of RNAs, including integrity surveillance and clearance/degradation, which is related to RNA damage, synthetic error and chemical modification [38]. Finally, RNA modification deserves a new discipline, as it has been proposed as epitranscriptomics [39]. Although it is classified in our scheme as part of the operational ribogenomic track, we have no intention to de-emphasize the importance of RNA modification and its roles in RNA function. However, it does have its own dilemma — only a minor fraction of the RNA macromolecules and a limited number of nucleotide residues of them are actually modified.

Ribogenomics and its research are still at their infancy — the discovery phase. Cheering for discovering a new class of sRNAs is only the beginning of a long search for their functions and mechanisms (**Table 1**). The measurable parameters for components in an operational ribogenomic track are multi-fold. The first parameter is specificity since nucleotide pairing by hybridization provides a powerful apparatus. In addition, the imperfect base pairing leaves opportunities for other protein apparatuses to be engaged. The second parameter is sensitivity that is also chemical in nature. For instance, a miRNA has to overwhelm its mRNA targets in apparent (effective) concentration. It is hard to image how a miRNA with a low copy number, such as dozens, inhibits an mRNA that has hundreds of copies and multiple target sites. Although the relationship between an mRNA target and an inhibitory miRNA may be more complex than what we have anticipated, quantitative studies on both parties are of essence. The third parameter is stability that is usually measured by half-life. RNA degrades very fast in general since the process does not need energy. RNA stability and structural dynamics are known to relate to its post-transcriptional modifications [45,46]. The fourth is the *ad hoc* creation of ncRNAs from a large candidate pool and one of the problems is the fact that a significant fraction of ncRNAs lacks sequence conservation across taxonomic scales [9]. In summary, since operational RNA macromolecules are rather massive in numbers and diverse in functions, our efforts to understand them have to be significant enough.

Ribogenomics on the informational track

It may not be easy to define a transcriptome if informational and operational RNAs are not differentiated; the former is largely mRNA that is large and more characteristic, and the latter is largely sRNA together with size-variable ncRNAs. Even protein-coding transcripts are not easily identified when it comes to plants and vertebrates – whose genomes of different lineages are not only structured differently but also fast-evolving after genome-wide duplication (GWD) [47–53]. Although mRNAs and their precursors are the only informational RNA class, their evolutionary transients are difficult to be thoroughly defined [5,53].

Classic transcriptomics has been focusing on identification of mRNAs in a given cell type or tissue, often for a comparative analysis. Such a study is based on the assumption that mRNAs dominate the functional content and are relevant to the function and functional changes of a given cell type. For a typical mammalian cell, the estimated mRNA is 1–2 pg in mass, 2 kb on average in length and about 0.5–1 million in number. Based on our theoretical model [54] and estimation on the total number of mRNAs in different cell types [55,56], a transcript-rich cell (such as stem cells and cells from cerebrum and testis) may have as many as 1 million mRNAs

Table 1 Some informational and operational RNAs summarized in the literature for human, mouse, rice and *Arabidopsis*

RNA types		Human	Mouse	Rice	*Arabidopsis*
Informational RNA					
mRNA		130,029	80,383	44,118	30,633
guideRNA		210 [40]	NA	NA	NA
Operational RNA					
Large	lncRNA	53,000 [12]	NA	NA	13,000 [41]
	lincRNA	27,500 [12]	NA	NA	6480 [41]
Small	miRNA [42]	4450	3094	1305	635
	snoRNA	403 [40]	NA	46 [43]	587 [43]
	piRNA	114 [44]	2710 [44]	NA	NA

Note: We do not estimate the number of genes here but merely count the number of mRNAs recorded in the UniGene database (http://www.ncbi.nlm.nih.gov/unigene/statistics/). The different numbers of RNAs identified in the databases or publications reflect the incomplete nature of the studies and collections. Only some representative classes of RNAs are listed here. NA, not yet available.

Table 2 Distribution of mRNA abundance in different cell types based on a theoretical model

mRNA expression level (copies/cell)	500 K		1000 K		5000 K	
	No. of mRNA	Percentage (%)	No. of mRNA	Percentage (%)	No. of mRNA	Percentage (%)
< 1	42,665	55	26,977	35	1114	1.40
1–5	22,866	30	31,104	40	25,863	34
5–10	4822	6	7450	10	15,688	20
10–50	5089	6	8415	11	22,866	30
50–100	855	1	1496	2	4822	6.30
100–500	763	1	1421	2	5089	6.60
> 500	92		289		1710	2.20
Mean copies per mRNA	6.48		12.96		64.81	
Median copies per mRNA	0.83		1.66		8.28	

Note: The total number of mRNAs per cell in different cell types is estimated based on our theoretical model or previous studies [54–56]. 500 K indicates the total number of mRNA copies in a transcript-rich cell, such as stem cells and cells from cerebrum and testis; 1000 K indicates the total number of mRNA copies in a transcript-poor cell, such as various cell lines and epithelial cells; 5000 K indicates the total number of mRNA copies in a hypothetical cell used for data analysis in this theoretical model [54].

whereas a transcript-poor cell (such as various cell lines and epithelial cells) may have only half of the number or 0.5 million (**Table 2**). Therefore, a deep-sampling strategy allows a reasonable description of a transcriptome and its transcript distribution. Empirically, we acquire over 20 million mapped sequence tags (raw data usually are not good ground for such estimation) for a given transcriptome [57–59], regardless what mRNA preparation methods are used [22].

The ultimate goal of an essential transcriptomic effort is to map all transcripts from all cell types of a given organism (species), such as human and mouse for biomedical research. By using the current technical platforms, this goal is not easily fulfilled for the following reasons. First, for a thorough transcript discovery effort, we do not have a series of protocols to purify RNAs of different sizes and copy numbers for consistent library construction. For copy number validation, our current techniques either are too expensive (sequencing) or have poor dynamic range (microarray) in detection so that low-copy transcripts are often lost in the process. Second, a basic sampling of different tissues/organs/cells already significantly exceeds hundreds of libraries [60]. In addition, longitudinal studies can easily add up the libraries to thousands. If diseases and other pathological conditions such as tumors are to be concerned, the libraries for such an endeavour can exceed hundreds of thousands. The cost therefore can easily reach billions. Third, a pilot project to pave a way for a large-scale discovery of mammalian transcriptomes is of essence, where cell and tissue types are well defined. Nevertheless, an effort to map all transcripts in representative mammals and full mapping for humans under various physiological and pathological conditions has to be launched in the near future. Some of the parameters to be defined and theoretical concerns are detailed in **Box 1**.

Along the informational ribogenomic track, an integrated view of genetics and evolution are also critical. There have been several critical points to be made clear and integrated into the current paradigm. First, we have, in the recent years, discovered a novel phenomenon — transcript-centric mutations, originally in the rice genome [3,61] and later in all organisms from bacteria to human [25,26,62,63]. The universality of this mutation spectrum is attributed to its underlying mechanism, namely transcription-coupled DNA repair [3,4]. The new mechanism will undoubtedly reset the traditional way of mutation assessment and interpretation. Second, we have also learnt how to differentiate two replication-centric mutations based on a simple partition between house-keeping and tissue-specific genes, where the latter class of genes tends to have ~30% more mutations than the former class [25]. The mechanism is related to chromosomal structure in which house-keeping genes are organized in such a way that they are always replicated earlier. Further stratification of genes, partition of chromosomes and detailed analysis on population data are all necessary for revealing mechanisms at chromosomal organization and transcriptional levels. Third, distributed along all transcripts, the spectrum of the transcript-centric mutations not only shows a gradient effect when aligned from the transcription starting site but also displays a periodicity reflecting nucleosome-space occupancy [29,64]. We also anticipate more concerns and discoveries on well-defined and confounding factors of inheritable genetic variations.

Acquisition of an organism's transcriptomes and its technical challenges

How could we experimentally define a transcriptome and transcriptomes of an organism? Conceptually, a transcriptome can be defined as all transcripts in a cell type under a defined condition. It includes transcripts of all sizes and essentially two categories of RNA macromolecules: informational and operational RNAs. Alternatively, the two can be categorized into three groups: mRNAs, lncRNAs and small ncRNAs (sncRNAs). Obviously, some of the transcripts are part of protein-coding genes [47] and others are just transcripts or transcribed RNA elements. Since not all transcripts are easily validated for precise functions, novel RNA elements remain a viable group of candidate operational RNAs.

The empirical definition of transcriptomic components in full remains a tough challenge due to several technical hurdles. First, its thorough discovery relies on sequencing technology that is essentially a sampling strategy, involving many parameters such as sample quality and instrument efficiency (*e.g.*, read length, error rate, throughput and per sample cost) [65,66]. Second, experimental protocols are also highly relevant. Although there are endless choices of commercial kits, their reliability and reproducibility remain to be systematically

Box 1 The complexity of defining transcriptomes

Occurrence-definition: universality and specificity

- Universal: shared by all tissues/organs/cells
- Tissue-specific: shared by a single or limited number of tissues (such as nerves, muscles and epithelia)
- Cell-specific: unique to a single cell type
- Near universal: shared by most tissues but not all
- Rationally shared: genes that are shared between unrelated tissues or cell types based on function

Expression-definition: variability and magnitude

- Expression-variable (majority; genes vary in expression among tissues)
- Expression-constant (minority; genes are expressed constantly in all cell types)
- Highly-expressed (> 1000s of copies)
- Moderately-expressed (10s–100s of copies)
- Lowly-expressed (< 10 copies)

Function-informational: gene composition, structure, organization and variation

- Size: large (> 500 kb) *vs*. small, median size
- GC/purine content: GC-rich *vs*. GC-poor
- CpG islands: high, moderate and low density
- Minimal-intron-containing
- Biologically-defined repetitive sequence element associated
- Gene cluster-associated
- Transcript-centric variation
- Germline-specific
- Purifying (Ka/Ks < 1) and positively-selected (Ka/Ks > 1) genes

Function-operational: cellular structure, process and mechanism

- Mitochondrion-associated
- Chloroplast-associated
- Nucleolus-associated
- Circadian-regulated
- Cell cycle-regulated
- Stem cell-differentiation
- Translation machinery
- Splicing machinery
- Nuclear exporting machinery

Condition-definition

- Embryonic development
- Epidermal differentiation
- Phenotypic plasticity: *e.g*., hibernation
- Pathological conditions

Note: Ka/Ks indicates the ratio of nonsynonymous to synonymous substitution rates.

evaluated before large-scale applications. Most delicate of all is sample handling and preparation: RNA quantitation, internal control for cross-library normalization, parallel sampling strategy, appropriate controls, *etc*. Third, a software package needs to be put together for data normalization across libraries, comparative expression analysis, gene annotations and functional classifications. Fourth, implementation and coordination of an international and large-scale effort is also a great challenge. A consortium for International Mammalian Ribogenome Project should be formed first. Different transcriptomic projects as its components should be carefully planned and physicians and clinical researchers are to be involved also to justify the usefulness of each transcriptome to be produced.

One important question remains: can we avoid transcriptomic projects being open-ended? Although the answer is definite, cautions have to be taken seriously. There are good news and bad news for transcriptomic projects. The good news includes unlimited number of projects to be proposed, multiple choices for technical platforms and approaches, and similar data types for centralization and integration. The bad news includes the presence of cellular heterogeneity, unavoidable cross-contamination of cell types/tissues/organs and instability of cellular gene expression. We have to be careful in defining and organizing transcriptome projects in that both real-time and transgenerational measurement are of importance.

Conclusion

For decades into the business of genomics and transcriptomics, we have never before felt the need of systematic data acquisition, rational parameter analysis, and comprehensive understanding of the intricate relationship among pathways and networks of ribogenomics and other macromolecule-centric "omics", including genomics, epigenomics, proteomics and metablomics on a multiple-track system [4]. We do have a long to-be-done list here for the ribogenomic basics: mammalian transcriptomes (development, circadian, seasonal, life span, *etc*.) and human disease transcriptomes (cancers, cardiovascular diseases, metabolic disease, autoimmune diseases, *etc*.). The complexity also comes from both anatomic and longitudinal partitions, let alone the growing catalogue of RNA types. However, "a march of a thousand miles always has to begin with a single step", a large-scale project, in the context of international collaboration and magnitude of the Human Genome Project, is of essence; or maybe multiple projects of the kind are to be organized. The Human Genome Project was proposed about 30 years ago, its profound and unprecedented influence is still alive. If we believe that its legacy should live on and is carried by the current generation of genomicists and ribogenomicists, it is about time to act.

Authors' contributions

JW and JX carried out data analysis. JY drafted the manuscript. ZZ, SH and XW provide insights. All authors had read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

Acknowledgements

This work was supported by the National Basic Research Program of China (Grant Nos. 2011CB944100 and

2011CB944101), the National Natural Science Foundation of China (Grant No. 90919024), the Special Foundation Work Program (Grant No. 2009FY120100) and the National High Technology Research and Development Program of Ministry of Science and Technology of China (Grant No. 2012AA020409).

References

[1] Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular biology of the cell. 3rd ed. New York: Garland Publishing; 1994.

[2] Lodish H, Berk A, Zipursky AL, Matsudaira P, Baltimore D, Darnell J. Molecular cell biology. 4th ed. New York: W. H. Freeman; 2000.

[3] Yu J. Challenges to the common dogma. Genomics Proteomics Bioinformatics 2012;10:55–7.

[4] Yu J. Life on two tracks. Genomics Proteomics Bioinformatics 2012;10:123–6.

[5] Wang J, Zhang J, Li R, Zheng H, Li J, Zhang Y, et al. Evolutionary transients in the rice transcriptome. Genomics Proteomics Bioinformatics 2010;8:8223–8.

[6] Yu J, Wang J, Lin W, Li S, Li H, Zhou J, et al. The genomes of *Oryza sativa*: a history of duplications. PLoS Biol 2005;3:e38.

[7] Zhou Y, Tang J, Walker MG, Zhang X, Wang J, Hu S, et al. Gene identification and expression analysis of 86,136 Eexpressed Sequence Tags (EST) from the rice genome. Genomics Proteomics Bioinformatics 2003;1:26–42.

[8] Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 2002;420:563–73.

[9] Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, et al. Mouse transcriptome: neutral evolution of "non-coding" complementary DNAs. Nature 2004;431:1038–9.

[10] Axtell MJ. Classification and comparison of small RNAs from plants. Ann Rev Plant Biol 2013;64:137–59.

[11] Kim VN, Han J, Siomi MC. Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol 2009;10:126–39.

[12] Ma L, Bajic VB, Zhang Z. On the classification of long non-coding RNAs. RNA Biol 2013;10:925–33.

[13] Gilbert W. Origin of life: the RNA world. Nature 1986;319:618.

[14] Woese CR, Olsen GJ, Ibba M, Söll D. Aminoacyl-tRNA synethetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 2000;64:202–36.

[15] Cech TR. Crawling out of the RNA world. Cell 2009;136:599–602.

[16] Zhang Z, Yu J. Does the genetic code have a eukaryotic origin? Genomics Proteomics Bioinformatics 2013;11:41–55.

[17] Zhang Z, Yu J. On the organizational dynamics of the genetic code. Genomics Proteomics Bioinformatics 2011;9:1–9.

[18] Xiao J, Yu J. A scenario on the stepwise evolution of the genetic code. Genomics Proteomics Bioinformatics 2007;5:143–51.

[19] Yu J. A content-centric organization of the genetic code. Genomics Proteomics Bioinformatics 2007;5:1–6.

[20] Zhang Z, Yu J. The pendulum model for genome compositional dynamics: from the four nucleotides to the twenty amino acids. Genomics Proteomics Bioinformatics 2012;10:175–80.

[21] Cai Y, Yu X, Hu S, Yu J. A brief review on the mechanisms of miRNA regulation. Genomics Proteomics Bioinformatics 2009;7:147–54.

[22] Cui P, Lin Q, Ding F, Xin C, Gong W, Zhang L, et al. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics 2010;96:259–65.

[23] Liu W, Zhao Y, Cui P, Lin Q, Ding F, Xin C, et al. Thousands of novel transcripts identified in mouse cerebrum, testis, and ES cells based on ribo-minus RNA sequencing. Front Genet 2011;2:93.

[24] Pollex T, Heard E. Recent advances in X-chromosome inactivation research. Curr Opin Cell Biol 2012;24:825–32.

[25] Cui P, Ding F, Zhang L, Hu S, Yu J. Replication and transcription contribute differently in mutation rates of human genome. Genomics Proteomics Bioinformatics 2011;10:4–10.

[26] Cui P, Lin Q, Ding F, Hu S, Yu J. The transcript-centric mutations in human genomes. Genomics Proteomics Bioinformatics 2012;10:11–22.

[27] Wei W, Ba Z, Gao M, Wu Y, Ma Y, Amiard S, et al. A role for small RNAs in DNA double-strand break repair. Cell 2012;149:101–12.

[28] Maute R, Schneidera C, Sumazin P, Holmes A, Califano A, Basso K, et al. TRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in B cell lymphoma. Proc Natl Acad Sci U S A 2013;110:1404–9.

[29] Cui P, Lin Q, Zhang L, Ding F, Xin C, Zhang D, et al. The disequilibrium of nucleosomes distribution along chromosomes plays a functional and evolutionarily role in regulating gene expression. PLoS One 2011;6:e23219.

[30] Cui P, Liu W, Zhao Y, Lin Q, Zhang D, Ding F, et al. Comparative analyses of H3K4 and H3K27 trimethylations between the mouse cerebrum and testis. Genomics Proteomics Bioinformatics 2012;10:82–93.

[31] Cui P, Liu W, Zhao Y, Lin Q, Ding F, Xin C, et al. The association between H3K4me3 and antisense transcription. Genomics Proteomics Bioinformatics 2012;10:74–81.

[32] Yang L, Yu J. A comparative analysis of divergently-paired genes (DPGs) among *Drosophila* and vertebrate genomes. BMC Evol Biol 2009;9:55.

[33] Xie B, Wang D, Duan Y, Yu J, Lei H. Functional networking of human divergently paired genes (DPGs). PLoS One 2013;8:e78896.

[34] Haberle V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, et al. Two independent transcription initiation codes overlap on vertebrate core promoters. Nature 2014;507:381–5.

[35] Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GK. Minimal introns are not "junk". Genome Res 2002;12:1185–9.

[36] Wang D, Yu J. Both size and GC-content of minimal introns are selected in human population. PLoS One 2011;6:e17945.

[37] Zhu J, He F, Wang D, Liu K, Huang D, Xiao J, et al. A novel role for minimal introns: routing mRNAs to the cytosol. PLoS One 2010;5:e10144.

[38] Wolin SL, Sim S, Chen X. Nuclear noncoding RNA surveillance: is the end in sight? Trends Genet 2012;28:306–13.

[39] Saletore Y, Meyer K, Korlach J, Vilfan ID, Jaffrey S, Mason CE. The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome Biol 2012;13:175.

[40] Mitch L. Databases: let it sno. Science 2005;310:27.

[41] Jin J, Liu J, Wong L, Chua NH. PLncDB: plant long noncoding RNA database. Bioinformatics 2013;29:1068–71.

[42] Kozomara A, Griffiths-Jones S. MiRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 2014;42:D68–73.

[43] Brown JW, Echeverria M, Qu LH, Lowe TM, Bachellerie JP, Hüttenhofer A, et al. Plant snoRNA database. Nucleic Acids Res 2003;31:432–5.

[44] Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res 2008;36:D173–7.

[45] Helm M. Post-transcriptional nucleotide modification and alternative folding of RNA. Nucleic Acids Res 2006;34:721–33.

[46] Karunatilaka KS, Rueda D. Post-transcriptional modifications modulate conformational dynamics in human U2–U6 snRNA complex. RNA 2014;20:16–23.

[47] Zhang Z, Wong GK, Yu J. Protein Coding. In: eLS. Chichester: John Wiley & Sons Ltd; 2013. http://dx.doi.org/10.1002/9780470015902.a0005017.pub2.

[48] Wu J, Xiao J, Wang L, Zhong J, Yin H, Wu S, et al. Systematic analysis of intron size parameters in diverse lineages. Sci China Life Sci 2013;56:968–74.

[49] Yu J, Wong GK, Wang J, Yang H. Shotgun sequencing. In: Encyclopedia of molecular cell biology and molecular medicine. 2nd ed. Germany: Wiley–VCH; 2005, p. 71–114.

[50] Wong GK, Passey DA, Yu J. Most of the human genome is transcribed. Genome Res 2001;11:1975–7.

[51] Wong GK, Passey DA, Huang Y, Yang Z, Yu J. Is "junk" DNA mostly intron DNA? Genome Res 2000;10:1672–8.

[52] Wang D, Su Y, Wang X, Lei H, Yu J. Transposon-derived and satellite-derived repetitive sequences play distinct functional roles in Mammalian intron size expansion. Evol Bioinform Online 2012;8:301–19.

[53] Su Z, Wang J, Yu J, Huang X, Gu X. Evolution of alternative splicing after gene duplication. Genome Res 2006;16:182–9.

[54] Zhu J, He F, Wang J, Yu J. Modeling transcriptome based on transcript-sampling data. PLoS One 2008;3:e1659.

[55] Zhu J, He F, Song S, Wang J, Yu J. How many human genes can be defined as housekeeping with current expression data? BMC Genomics 2008;9:172.

[56] Zhu J, He F, Hu S, Yu J. On the nature of human housekeeping gene. Trends Genet 2008;24:481–4.

[57] Zhou Y, Gong W, Xiao J, Wu J, Pan L, Li X, et al. Transcriptomic analysis reveals key regulators of mammogenesis and the pregnancy-lactation cycle. Sci China Life Sci 2014;57:340–55.

[58] Gong W, Pan L, Lin Q, Zhou Y, Xin C, Yu X, et al. Transcriptome profiling of the developing postnatal mouse testis using next-generation sequencing. Sci China Life Sci 2013;56:1–12.

[59] Ma L, Nie L, Liu J, Zhang B, Song S, Sun M, et al. An RNA-seq-based gene expression profiling of radiation-induced tumorigenic mammary epithelial cells. Genomics Proteomics Bioinformatics 2012;10:326–35.

[60] Zhao D, Wu J, Zhou Y, Gong W, Xiao J, Yu J. WikiCell: a unified resource platform for human transcriptomics research. Omics 2012;16:357–62.

[61] Wong GK, Wang J, Tao L, Tan J, Zhang J, Passey DA, et al. Compositional gradients in Gramineae genes. Genome Res 2002;12:851–6.

[62] Chen K, Meng Q, Ma L, Liu Q, Tang P, Chiu C, et al. A novel DNA sequence periodicity decodes nucleosome positioning. Nucleic Acids Res 2008;36:6228–36.

[63] Chen K, Wang L, Yang M, Liu J, Xin C, Hu S, et al. Sequence signatures of nucleosome positioning in *Caenorhabditis elegans*. Genomics Proteomics Bioinformatics 2010;8:92–102.

[64] Cui P, Zhang L, Lin Q, Ding F, Xin C, Fang X, et al. A novel mechanism of epigenetic regulation: nucleosome-space occupancy. Biochem Biophys Res Commun 2010;391:884–9.

[65] Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu J. The next-generation sequencing technology and application. Protein Cell 2010;1:520–36.

[66] Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012;2012:251364.

Dr. Jun Yu is a professor of Beijing Institute of Genomics, Chinese Academy of Sciences. Dr. Yu's primary research interests include genome biology and bioinformatics. He has participated in many major genome projects in China, such as the Human Genome Project (the Chinese effort), the Superhybrid Rice Genome Project, Silkworm Genome Project and Chicken Genome Diversity Project, which all led to high-impact publications in international journals, including *Science, Nature* and *PLoS Biology*. He has published over 200 peer-reviewed scientific papers and over a dozen books and book chapters. Dr. Yu has won numerous academic awards, such as Award for TWAS Prize in Agricultural Sciences for 2012, Outstanding Science and Technology Achievements (Group, 2003, Chinese Academy of Sciences), Scientific Leader of the Year 2002 by Scientific American, "Qiushi" Annual Award for Scientific Achievement (Group, 2002, QiuShi Science and Technology Foundation, Hong Kong), *etc.*