The Scientific Legacy of the 20th Century
Pontifical Academy of Sciences, Acta 21, Vatican City 2011
www.pas.va/content/dam/accademia/pdf/acta21/acta21-gojobori.pdf

Genomic Exploration of the RNA Continent

Takashi Gojobori

Introduction

1.1. Susumu Ohno

Susumu Ohno was born in 1928 and died in the year 2000. He had spent more than 40 years in the City of Hope in Los Angeles, USA. One of his famous books is *Evolution by Gene Duplication*, which was published in 1970 (1). In this book, he pointed out that genome duplication and gene duplication are very important not only for evolution but also for function and structures of the genome. This is mainly because a duplicated copy of gene can enjoy the freedom of functional differentiation as long as the original gene can retain the original function.

1.2. Junk DNA

The term 'junk DNA' was coined by Susumu Ohno probably in 1972, a bit later than the time when the above-mentioned book was published, as long as my memory is correct. That was presented in the Brookhaven Symposium on Biology in the United States with his paper entitled '*So Much Junk DNA in our Genome'* (2). Already at that time, it was known that the human DNA genome may have had only 5% for protein-coding regions and the other 95% for non-coding or unknown functions. From this fact and also from other observations, he coined 'junk' DNA that literally represent the protein-non-coding regions of the DNA genome.

However, the term 'junk' really brought about intense controversy over its biological significance. This is because many people have believed that there are no regions of meaningless function in a human body.

According to the Merriam-Webster's online dictionary (3), the followings are given as definition of 'junk':

1) Pieces of old cable or cordage used especially to make gaskets, mats, swabs, or oakum.
1) Old iron, glass, paper, or other waste that may be used again in some form; second hand, worn, or discarded articles.
1) Something of poor quality, almost trash, something of little meaning, worth, or significance.

I do not know exactly which meaning Dr Susumu Ohno took when he coined such a term as the 'junk DNA'. According to my personal impression, however, because I had personal and intimate communication with him when he was alive, the second one of the above-mentioned definitions is probably the most appropriate. In fact, it is close to the meaning of trash, implying that human DNA contains a vast amount of trash. That is why it caused a lot of arguments.

1.3. Neo-Darwinism

The arguments on junk DNA are so important because it gives a paradigm question of whether all the genomic DNA regions are subjected to natural selection. According to Neo-Darwinism, all the traits and features of organisms are explained by natural selection and mutation with a law of inheritance. The premise is that most mutations are deleterious against the survival of organisms but all the other mutations would be very much adaptive and advantageous for survival. For the former mutations negative selection works whereas for the latter mutations positive selection operates.

1.4. Genetic drift

There is another mechanism of changing gene frequencies in a natural population of a given organism, which is called 'genetic drift'. Genetic drift is a kind of genetic phenomenon in which gene frequencies will change simply by mating. Natural selection can be a kind of second force, in particular when the population size is small and random mating dominates.

In population genetics in the 1930s, a heated controversy took place between so-called selectionists versus proponents of genetic drift (4). An American geneticist, Sewall Wright, was kind of a hero of genetic drift. On the other hand, R.A. Fisher, who was a British man, is a hero of selectionists. Somehow selectionists appeared to have won.

1.5. Motoo Kimura

A Japanese geneticist called Motoo Kimura proposed the neutral theory of molecular evolution [5,6]. He contended that at the DNA level or the genomic level, most mutations were selectively neutral. That means that mutations are not so much deleterious, not so much advantageous, either. This is now simply called as the neutral theory.

I will not go into the details of this controversy, but what I would emphasize is that this kind of controversy always exists in the background of a discussion on biological significance of transcribed RNAs in the non-protein-coding regions of the DNA genome.

1.6. TSS: Transcriptional Start Sites

So here in this paper, our specific question is how transcription start sites (TSSs) are distributed over the human genome. Of course, transcription depends heavily upon types of tissues and cells or even environmental conditions. Using over 60 different tissues and cell lines of humans, two Japanese governmental projects on transcription were conducted, in which we actively participated: One is called the 'H Invitational Human Transcript Project' (7) and the other is the 'Human Genome Network Project' (8). Using the outcome of both projects, we would discuss the genomic distribution of TSSs in conjunction with the biological significance of transcribed RNAs in the non-coding regions.

Materials and Methods

2.1. H-invitational Project

Almost eight years ago, we initiated the annotation jamboree on the human full-length cDNAs over the human genome (7). Note that cDNA is a kind of RNA information most of which are supposed to encode for proteins. Furthermore, this jamboree was conducted as an international cooperation, in which about 120 experts gathered in Tokyo, Japan and spent ten days making annotations for actual and possible human genes. This project has been called as the 'H-Invitational Project'. We still continue this endeavour, presently including not only all cDNAs but all available transcripts.

2.2. Transcription regulation network as a small universe in a cell

The transcription regulation in a given cell may be illustrated as follows. Let us suppose that there is a certain gene in a given genomic region. We now know that *prior to* the gene; there is a *cis*-acting element or a promoter region. When a certain protein such as a transcription factor binds to *cis*-acting elements in a regulatory region, it works like a switch to order the gene to be transcribed to produce messenger RNAs. Following information of messenger RNAs, a particular protein is made through splicing in the case of eukaryotes. The protein may have interaction with other proteins that are made in the same way. Then, interacted proteins may constitute a protein complex, which would bind to DNA again to switch on production of messenger RNAs of its own protein or the other proteins. This is exactly what we call network. This transcriptional network is extended over the entire human genome, which is now called the 'genome network'. It looks just like a small universe with a single cell.

2.3. Human Genome Network Project

We conducted a project called the 'Human Genome Network'. This project was carried out in collaboration with RIKEN. From messenger RNAs (mRNAs) in a given type of tissue or cell, we can obtain only the first 20-nucleotide segments from the start site of the mRNAs, using their cap sites, in the process of making cDNAs. Let us call those 20-nucleotide segments 'TSS tags'.

The TSS tags can be actually sequenced by the so-called next-generation sequencers (NGS) in an enormous amount of numbers. Therefore, once we have a sample of tissues or cells, we can sequence the TSS tags as a form of cDNAs immediately and enormously.

The TSS tags obtained are mapped, by technologies of bioinformatics, onto the human genome, meaning that transcription start sites can be identified in a genomic scale. Thus, we can now raise unique and significant questions about how much transcription is taking place over the human genome and how these transcription start sites are distributed over the human genome. The Human Genome Network Project was conducted to answer those questions as one of the purposes.

2.4. Tissue-type and cell-type dependencies

Of course, the transcription depends heavily upon the types of tissues and cells and even upon environmental conditions. From this standpoint, it will be idealistic if we are able to examine TSSs in a single cell.

For example, we are now trying to examine TSS tags from a single cell such as a monocyte, which can differentiate to a macrophage. Because it takes more time to obtain finalized results, however, we decided to examine TSS tags for a mixture of 60 different human tissues and cell lines. We believe that it will still be useful for understanding the overview of transcription activities over the entire human genome, namely for conducting *genomic exploration of the RNA continent* of humans.

2.5. Quality control of sequence data produced by NGS

In order to elucidate a distribution of TSSs over the human genome, we have made great effort to clean the data. In particular, when a single TSS tag is tried to map on the genome, it sometimes happens to be mapped in more than one location. For other TSS tags, it also sometimes happens to be no matched locations in the human genome. Those observations are apparently due to artefacts of the experimental efforts in a process of producing TSS tags.

Thus, it is very important how much the TSS tag data can be cleaned. In this case of the next generation machine called 454, we know that a specific type of sequencing errors have been expected in a certain frequency.

By making an algorithm, we can rescue a portion of sequencing errors. Conducting computer simulations, the TSS tag data obtained has been evaluated that almost 10% of sequencing errors can rescued computationally.

Results and Discussion

3.1. Distribution of TSS tags over the human genome

When a distribution of about 47 millions of TSS tags for a mixture of 60 different types of human tissues and cell lines was examined for all the chromosomes, from chromosome numbers 1 to 22 and sex chromosomes X and Y, over the human genome, it immediately became clear that transcription for producing mRNAs are taking place actively at a tremendous numbers of locations in the human genome. Taking into account the fact that the number of human genes is about 23,000~24,000 in the genome, the number of TSSs far exceeded those numbers. Thus, we assure that transcription takes place, in an enormous number, in the protein non-coding regions of human genome. We call this situation *the RNA continent* of the human genome. Of course, the information on transcription activities of genes, such as typical disease-sensitive genes, is also very useful for understanding how and when these genes are transcribed.

3.2. Distribution of TSS tags in a liver tissue

In the previous section, we discussed the distribution of TSSs for a mixture of 60 different types of human tissues and cell lines. Here, we can focus on a single tissue. Now, we can have a distribution of the transcription start sites only for the human liver.

As long as we see, the TSS tags in the human liver are very scarcely distributed over the human genome. This is apparently due to the lack of a sufficient number of TSS tags. Thus, we point out that although this kind of study is certainly feasible now, it may take a bit more time to obtain a sufficient number of TSS tags. However, the acute developments of NGS (Next-generation sequencing machines) are expected to resolve this problem because of enormous speed and capacities of sequencing capabilities.

3.3. Examination of TSSs with known distribution: Two categories of human genes

We made comparisons of the transcription start sites obtained from genomic locations of the TSS tag with the already known transcription start signs of protein-coding regions of the human genome.

From the database such as RefSeq at NCBI/NIH in the United States, or the H Invitational database that we have constructed in the H-Invitational Project, we obtained information of genomic locations for all the protein-coding regions available. Then, we made comparisons.

If the transcription site is so sharply determined, then the distribution of TSS tags should be very sharp. On the other hand, if transcription start sites are so stochastic or if they are not really sharply determined, even though they give the right direction of transcription of a given coding region, the distribution of TSS tags should manifest a broad distribution.

As a result, we observed that there were two types of coding regions, depending upon transcription start sites. One type of coding regions has very sharp transcription start sites whereas the other type of coding regions has very broad start sites.

Although we should have understood how those transcription start sites are biologically determined, we do not know how the coding regions of having broad locations of transcription start sites are regulated. Anyway, it is very interesting to know that transcription start sites are not always sharply and uniquely determined. Therefore transcription start sites seem to have a stochastic nature, which we should keep in mind.

3.4. Susumu Ohno's Junk DNA

Let me go back to a story of Susumu Ohno. When he coined 'junk DNA', he predicted that even junk DNA would be transcribed. However, transcription itself does not mean any functional significance. In this sense, Susumu Ohno was right. In 1972 he actually and clearly showed that 'junk DNA' would be transcribed (2).

Now we are confronted by a very important question to answer. Is there any functional significance for the transcription activities observed in the protein non-coding regions occupying huge portion of the human genome? Yes, partly. We have known this answer because we know there are functional non-coding RNAs such as Micro RNAs and natural antisense RNAs.

However, the problem is whether a substantial portion of non-coding regions is subjected to the so-called 'transcriptional noise'. It is just like the engine of an old car. Once you start the engine, it cannot start immediately. You need idling. Just like this, we may be observing *transcriptional idling.*

In order to make transcription possible, opening of chromosomal structures may be prerequisite. This may cause transcriptional noise or idling because of preparation for appropriate changes of chromosomal structures.

The problem was whether junk DNA is really junk or not. We do not think this may be the right question. Because we know that there must be

functional non-coding RNAs such as micro RNAs among all transcripts, the right question should be to be asked in a way is how many are functional and what percentage are not functional. We believe that the question should be changed into the new question; otherwise the RNA continent cannot be explored in an appropriate way.

Summary

We have conducted two Japanese governmental projects: the H-Invitational Human Transcript Project and the Human Genome Network Project. Using the outcome of these two projects, we examined a distribution of transcription start sites over the entire human genome. We pointed out that tremendous transcription activities are taking place in a substantial portion of protein non-coding regions that occupy a huge portion of the entire human genome. Moreover, the transcription sites for some genes are not sharply and uniquely determined. Finally, the right question to ask should be in a way how many are functional and what percentage are not functional. We believe that the question should be changed into the new question; otherwise the RNA continent cannot be explored in an appropriate way.

Acknowledgements

First of all I would like to express my special thanks to the Holy See and also to the organisers, Dr Werner Arber and Dr Jürgen Mittelstrass. In particular I would like to extend my thanks to the Chancellor, Dr Marcelo Sánchez Sorondo for his never-changing support to me.

References

[1] Ohno, S. (1970) *Evolution by Gene Duplication*, Springer Verlag, Berlin.

[2] Ohno, S. (1972) *So Much Junk DNA in our Genome*, Brookhaven Symposium, New York.

[3] Merriam-Webster's online dictionary (2011), www.merriam-webster.com

Provine, W.B. (1971) *The Origins of Theoretical Population Genetics*, With A New Afterword.

[4] Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* 217 (5129): 624-626.

[5] Kimura, M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press.

[6] Imanishi, T., other 152 authors, Ikeo, K. Gojobori, T., and Sugano S. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2, 1-21.

[7] The FANTOM Consortium: Carninci, P., Gojobori, T., Ikeo, K. and other 158 authors and Hume, D.A., and Genome Network Project Core Group: Kai, C., and other 31 authors and Hayashizaki, Y. (2005). The transcriptional landscape of the mammalian genome. *Science* 309 (5740): 1559-1563.