April 29, 2006

The Genome Network and FANTOM Consortium
Genomic Sciences Center
Genome Exploration Research Group
Project Director: Dr. Yoshihide Hayashizaki

Discovery of novel mechanisms to control gene expression (Structure of promoters implicate function)

- The majority of mammalian promoters are "broad" type and evolve rapidly -

- We have identified the majority of gene starts in human and mouse.
- Contrary to the beliefs that RNA transcription starts from defined positions, the majority of mammalian promoters are of a "broad" type, having many closely located transcriptional starting sites, and only a minority of the promoters are "narrow" type.
- The evolutionary speed of the "broad" type promoter is faster than the classical narrow type, and it can be a clue to explain the complexity and evolution of higher organisms.

The Genome Network and FANTOM Consortium of scientists (organized by RIKEN with 45 institutions of 11 countries. Dr. Yoshihide Hayashizaki; General Organizer, Dr. Piero Carninci; Scientific coordinator) achieved the comprehensive analysis of transcriptome and published a series of milestone articles (thirteen) in two journals, one article in "Nature Genetics" (April 28th) and a 12-article collection in the open access journal, "PLoS Genetics" (April 28th). The full text of all articles in this special collection in PLoS Genetics are freely available immediately upon publication.

We identified 5-10 times more promoters in mouse (236,498) and human (190,513) than what was known before and found that a "broad" promoter initiates transcription at many closely located sites, in contrast to the "single peak" textbook example. The majority of the mammalian promoters are of the "broad" structural type, where the "broad" promoter type regulates more widely expressed genes, while the narrow promoter type tends to regulate tissue/context-specific genes.
From comparative analysis of human, mouse, rat, dog, and chimpanzee, we found that the evolutionary speed of broad type promoter is faster, and this can be a clue to explain the complexity of higher organisms. The promoters are not only located at upstream of the first exon but also at internal regions, and especially at the last exon.

Pseudogenes are thought to be remnants of ancient genes, but 9583 of those genes has turned out to have active promoters and are actually transcribed to RNA. The database of these promoters and RNAs will provide important insights to life science and can be accessed from the http://gerg01.gsc.riken.jp/cage/ and http://www.ddbj.nig.ac.jp/Welcome-e.html (On the servers of DDBJ, National Institute of Genetics) While all these 13 papers are addressing different aspects of mammalian transcription control, a unifying finding is that the transcription landscape is much more complex than previously appreciated.

Background: Charting the location of all promoters in human and mouse
A large part of the complexity of mammals is due to their intricate system of controlling which genes are active in each cell of the body. The current view of gene regulation is that each gene has a DNA region just before the start of the gene that contains signals necessary for activating the gene - this region is called the promoter. The finding and characterizing of promoters is one of the most important fields of biological sciences. Previously, scientists in the Genome Network and FANTOM Consortium reported that the majority of mammalian genomes are transcribed, and that at least 180,000 different forms of transcripts exist (Science, Sept 2). Now, as reported in Nature Genetics this week, they have identified the core promoters of the large majority of genes in the mouse and human genome by sequencing more than 12 million sequence tags corresponding to starts of RNA molecules. Their data expands the number of characterized "core promoters" by about 5-10 fold, which provides a major resource to the biomedical research community.

Identifying the regulatory signals in the genome is essential to reconstruct the logic of control in biological systems. In essence, we need to understand the language cells use to "read" DNA. The novel dataset presented is an essential part of this undertaking and has brought us several surprises.

Methods and Achievements;

The architecture of mammalian promoters
The traditional view of gene regulation is that most genes have a single promoter and a
single start position, which is determined by a specific sequence (the TATA box) located nearby. Unexpectedly, this study showed that the large majority of mammalian genes have fuzzy boundaries: the start can be described as a distribution of different start sites, which can have different shapes. The distribution shape is strongly linked to both the type of signals found in the promoter and the function of the genes regulated by the promoter: genes used in many tissues have broad distributions of start sites, while genes which are only turned on in specific tissues or time-points use only one start site. The concept of promoters with single gene start sites has been the dominating promoter model in biology, but it is now clear that this is not representative for most promoters. Instead of having a TATA-box sequence, promoters with broad distributions of start sites are enriched CpG island sequence patterns. This architecture permits gradual changes during evolution, so that individual species can differ subtly in the way they control individual genes. Therefore, broad type promoters evolve faster than promoters with single start sites. This was confirmed by a comparative analysis of start sites in mouse and human: this being the first time such a comparison has been possible to make.

The analysis has revealed new classes of start sites, within and at the end of genes, producing transcripts that read through into neighboring genes, producing different forms of RNA that encode different proteins, or do not produce a protein but act solely as regulators. The transcriptome (the whole content of information read from the genome by the cell) has become a great deal more complicated, but at the same time we now have the tools to begin to analyze that complexity. A greater understanding on how and when genes are active is necessary for the next generation of clinical treatments to combat disease.

**Mammals have a staggeringly complex DNA and RNA universe**

The May issue of PLoS Genetics contains 12 articles deriving from the activity of the Genome Network and the FANTOM Consortium project, which further expand on the biological findings derived from the analysis of full-length cDNAs and cap-analysis gene expression (CAGE) technologies. These papers describe analysis results using novel datasets and computer algorithms for improved analysis of this type of data. Although the papers address a wide array of different biological questions, they have an underlying common theme - the revelation of a staggeringly complex transcriptome. For instance, there are more than 1000 regions that are organized as "chains" of multiple transcripts, which share expressed overlapping RNAs and/or controlling elements,
implicating a complex chain of regulating events. These chains were shown to be conserved between mouse and human during the evolution.

The conservation suggests that the regulation of these genes is achieved not by regulating the genes locally, but by regulatory mechanisms that control much broader genomic regions (loci). Such kind of regulations may also include non-coding transcripts (RNA transcripts not producing proteins) and sense-antisense RNAs (mirrored sequences that are able to bind to each other). A related finding is a novel class of very long RNA transcripts, which are transcribed through very large regions covering many different genes.

One likely function of these long RNA molecules is to change the status of the genome regions through which they are transcribed, either by activating or inactivating them. Therefore, these RNAs can be considered as “regional genomic switches”. Linked to this, a large number of active "pseudogenes" were found. This is surprising, as pseudogenes are commonly thought to be fossil testimonials of ancient genes. Since they are still used by the cell, they are likely to have some yet unknown function - they might have become regulatory non-coding RNA.

On the other end of the scale, at least 1100 new very short proteins (less than 100 amino acids) were identified. Alternative splicing adds variability to proteins sequences encoded from the same gene,. More than 8% of the transcriptional units can produce two or more proteins that are so profoundly different that they localize into different cellular compartments.

The non-coding RNA world has been further analyzed using the CAGE data; in particular sense-antisense regulation of gene expression. These sense-antisense RNAs tend to occur on promoter regions of genes active in the cell nucleus (such as transcription factors), implying specific regulatory systems for these genes.

The dataset has also enabled deeper studies of promoter structure and evolution. The majority of promoters have CpG island patterns. These types of promoters were found to be under positive selection in primates: in other words, they tend to mutate more rapidly than other sequences in the recent human evolution, thus highlighting that the recent evolution has been mainly related to changes of expression of the genes, rather than changes in the gene product. This is correlated with the finding that broad type promoters evolve fast, since many of these promoters have the CpG patterns.

**Future perspectives: redefining regulatory mechanisms**

The results presented in Nature Genetics and PLoS genetics this week represent
significant advances in mammalian biology. A large part of the findings are challenging established paradigms of modern biology, urging scientists to reconsider several generally accepted models.

In the last few years, a wide variety of large-scale techniques became available for scientist to study the whole genome at the same time. These genome-wide studies have given insights that were not possible to make when studying single genes. This observation is true for the study above, and we can expect other significant advances in the future based on this and future large-scale datasets.

The next challenge is to understand the actual signals that regulate genes, and how genes interact. This challenge motivates the next phase of the Genome Network Project, which aims to identify the interacting genes and their modality of interaction. This kind of holistic knowledge of biological phenomena, which are based on the understanding of the whole biological system rather than the fragmented function of individual or restricted number of genes, will promote novel developments in biomedical sciences including better diagnosis and treatments of patients.

For more information, please contact:

RIKEN Public Relations Office
Email: koho@riken.jp