Additional material

September 1st, 2005 (Eastern Time, USA)

The discovery of a novel RNA Continent through comprehensive analyses of mammalian transcriptomes.

The FANTOM Consortium for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute and Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute (Genome Network Core Group), involving researchers from Australia, Italy, Germany, Greece, Singapore, South Africa, Switzerland, Sweden, UK and the United States, has published two milestone papers this week in the prestigious journal *Science,* which transform our understanding of the information content of the mammalian genome.

The past 5 years have seen the completion of several mammalian genome sequences, but these are of limited value unless we can decode the way that they are translated into functions required to create a mature animal. Only around 2% of the genome is translated into proteins, the building blocks of the cells that make up our bodies. But which 2%, and how is it controlled? The key intermediate is the transcriptome.

The transcriptome is made of many individual RNA molecules (transcripts), which are copied from DNA, and in many cases, translated into proteins. Transcriptome analysis is much more demanding than genome sequencing, since thousands of RNA molecules must be individually isolated and sequenced. The FANTOM Consortium analyzed over 2 million sequences of RNAs produced from the mouse genome, obtaining more than 100,000 full length copies of these RNAs, and used several new technologies (CAGE (cap analysis gene expression), GIS (gene identification signature, in collaboration with the Genome Institute of Singapore) and GSC (gene signature cloning)), to generate more than 20 million tags representing their starts and ends in a technical tour-de-force that provides an amazing new view of our genome.

Another aspect of our study is the analysis of sense and antisense transcription overlapping in the genome. In particular, the analysis of physical cDNA clones, sequence tags and Expressed Sequence Tags (EST), provides compelling evidence that overlapping sense-antisense (S/AS) from both strands is almost universal in the genome. S/AS are especially abundant in imprinted loci, keeping with the putative role of non-coding RNA in gene silencing. We also provide experimental evidence that perturbation of an antisense RNA can alter the expression of sense mRNAs, suggesting that antisense transcription contributes to the control of transcriptional output in mammals.

The data taken together provides the biomedical research community with the tools to understand the control networks that are needed to create a mammal. A genome sequence contains not only the code for making the parts (proteins) of a mammal, but also the code for making the right forms, in the right amounts, in the right place, at the right time. The FANTOM data provide the most complete overview of a transcriptome in any species. Since mammals only have slightly more conventional genes (around 22,000) than a simple worm, the results of the FANTOM Consortium study clearly indicate that while proteins comprise the essential components of our cells, the development of multicellular organisms like mammals is controlled by vast amounts of regulatory noncoding RNAs that until recently was not suspected to exist or be relevant to our biology.
Moreover, since most proteins are similar among mammals it also suggests that many of the differences between species may be embedded in the differences in the RNA regulatory control systems, which are evolving much faster than the protein components.

If correct, these findings will radically alter our understanding of genetics and how information is stored in our genome, and how this information is transacted to control the incredibly complex process of mammalian development, with enormous implications for the future of biological research, medicine and biotechnology.

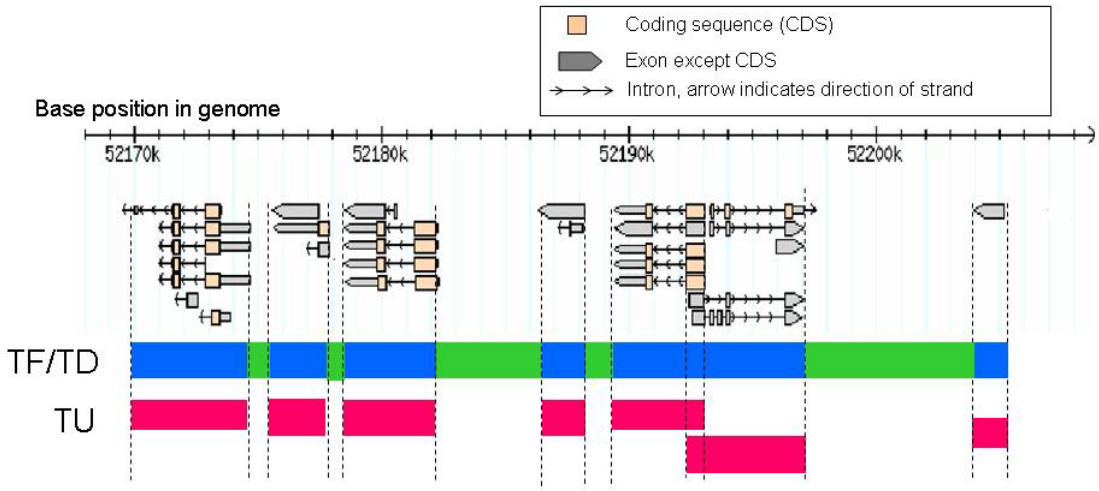
[Future directions]

With this study we challenge the concept of a gene as declared in the central dogma, a gene is no longer only the section of DNA that is transcribed to RNA, to be translated into a protein; we now declare that RNA is not simply an intermediary between DNA and functional protein, but has a multitude of intrinsic functions and activities.
This discovery of the "RNA Continent" consisting of the gigantic amount and varieties of mRNA and non-coding RNAs, is in contrast to the previous view of the genome consisting of transcriptional forests (many transcription starts) located between transcriptional deserts (no transcription starts). From this on, new studies will include ncRNA as a primary target to further unveil the transcription/regulation mechanisms in mammals.

Currently, the systematic analysis of the transcriptome is in its infancy. Comparison with independent tiling array experiments reveals that we currently target around 50% of all polyA RNAs using full-length cDNA sequencing. Other predictions indicate that the number of non polyA transcripts is close to the number of polyA transcripts. Thus, this is only the beginning systematic transcriptome analysis. Today we have two dimensions in genome analysis: time and location. In the future we will have to consider the cell age, development and tissue differentiation to be able to encompass this new, dynamic field of transcriptome analysis. Transcriptome analysis is the next new and powerful tool to apply to our understanding of gene functions, and ultimately; the secrets of life.

These studies have been carried out mainly in mice; the most widely used experimental mammalian species. Equivalent human data is not far behind, and RIKEN and many FANTOM members are actively involved in the next phase, the Genome Network consortium, which aims to use these new tools to understand human development and disease. Associated with the publication of the Science papers, all of the information will be publicly released on Internet in a unique user-friendly format. (On the servers of DDBJ, National Institute of Genetics and RIKEN; http://www.ddbj.nig.ac.jp/, http://fantom3.gsc.riken.jp/, http://www.ddbj.nig.ac.jp/whatsnew/050124-e.html)

[New concept of genes]



Transcript Desert: A genomic segment where none of the strands are transcribed

Transcript Forest: A genomic segment where RNA polymerase transcribes either strand into hnRNA or pre-mRNA.

Transcriptional Unit : A genomic segment where transcripts have exon-overlaps on the same strand (sharing direction and (partly) location)