TRANSCRIPTOMICS

Charting the functional landscape of the mammalian genome

An important goal of functional genomics is to understand how the information encoded in an organism's genome is retrieved. A range of approaches have been used to characterize functional elements in the genome that control transcription; for the human genome this is done perhaps most famously in an international collaborative project called ENCODE — Encyclopedia of DNA elements. An important study that contributes to this project has recently been published in which human and mouse transcription start sites (TSSs) are tagged and characterized on a genome-wide scale. The study provides new insights into basic promoter features, as well as the evolutionary conservation and dynamic regulation of mammalian promoters.

Wanting to refine the functional landscape of the mammalian genome, the authors turned to CAGE (cap analysis of gene expression), a technique for high-throughput analysis of TSSs and study of promoter usage. Approximately 650,000 and 730,000 TSSs were identified in the human and mouse genomes, respectively, by mapping onto unique genomic regions 20- or 21-nt long CAGE tags, which were derived from sequences (from hundreds of CAGE libraries derived from all major tissues) that lie in the proximity of the cap site.

TSSs that were defined by tag clusters of at least two tags were followed up in more detail. In fact, the distribution of tags within a cluster allowed the authors to categorize TSSs into three classes, which are conserved between mouse and human: a single peak (SP) class, indicative of a well-defined TSS; a broad shape (BR) class, indicative of multiple, weakly defined TSSs; and a bimodal/multiple (MU) class, indicative of several well-defined TSSs in one cluster. The SP class was mainly made up of TATA-box promoters, whereas the BR class of regions associated with CpG islands. The BR class predominates in mammalian genomes, whereas TATA-box promoters, mainly associated with tissue-specific and conserved gene expression, are in the minority.

The study identified TSSs in unexpected places. In some genes all of the exons showed promoter activity, whereas in other genes expressed at similar levels this activity was absent. The authors also identified a new class of promoters in 3′ UTRs. Although the function of these promoter types remains to be clarified, the authors speculate that they might have a role in enhancing RNA processing, including splicing and transcription itself.

The study provided valuable evolutionary insights. Notably, the initiator sequences — located at position −1 to +1 of TSSs — are subject to frequent changes in mammals. Moreover, pyrimidine–purine dinucleotides, which are overrepresented at the −1 to +1 sites, seem to contribute to the precise location of TSSs of the BR class. Overall, the CpG island-associated promoters seem to evolve more rapidly than the TATA-box promoters. The epigenetic control of CpG island-associated promoters and the fact that some contain bidirectional TSSs, which might regulate a locus's expression, might be important facilitators of adaptive evolution in mammals.

Having provided a wealth of information about the functional landscape of mouse and human genomes, the authors feel confident that meaningful models of transcriptional regulatory networks can be constructed on the basis of the information about core promoter sequences. And, because "technologies like CAGE are scaleable to whole organisms, these approaches pave the way for 'systematic' systems biology."

Magdalena Skipper

ORIGINAL RESEARCH PAPER Carcinci, P., Sandelin, A. & Lenhard, B. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* 28 April 2006 (doi:10.1038/ng1789)