GENOMICS

SAGE takes a look at the 5′ end

By adapting the method of long serial analysis of gene expression (LongSAGE), research groups in Singapore and Japan were able to identify the 5′ ends of new genes.

Over 98% of the human and mouse genome fall into the category of noncoding DNA. Finding sequences that do encode genes is therefore a challenge akin to finding a needle in a haystack. The method of long serial analysis of gene expression (LongSAGE) was originally developed to identify units of transcription in the genome; recently, research groups in Singapore and Japan have successfully adapted it to not only find genes but also determine their 5′ and 3′ boundaries. LongSAGE was first developed by Victor Velculescu and colleagues when the limitations of computational approaches to finding genes became apparent and an experimental rather than a bioinformatics approach to gene prediction was needed (Saha *et al.*, 2002). In LongSAGE the mRNA is trapped via its poly(A) tail and reverse transcribed, then a primer, containing a recognition site for an enzyme that cuts 21 bp upstream of this site, is annealed to the cDNA. Because the reverse transcription starts at the poly(A) tail, the cDNA consists largely of the 3′ region, so that after enzymatic digestion the small 21-bp tags comprise the 3′ region of the transcript. These tags are concatenated to long pieces of DNA, cloned, sequenced and matched to the genome. The two major advantages of LongSAGE are that it allows the identification of new genes and that it does so in a quantitative way. The number of tags obtained is a direct reflection of the number of transcripts present in the genome. What the original LongSAGE does not provide, because the tags come from the 3′ regions, is information about where the gene starts.

Two groups recently began to tackle the question of how to find the transcription start by designing a method for 5′ SAGE. The teams of Yijun Ruan at the Genome Institute of Singapore and Kouji Matsushima from the University of Tokyo used modifications of cap trapping where the 5′ end of the full-length cDNA is annealed to a biotin linker and can be retrieved via streptavidin-coated beads (Wei *et al.*, 2004; Hashimoto *et al.*, 2004). The 5′ linker also includes the recognition site for an enzyme that cleaves 20 bp away and thus generates the 5′ tag (**Fig. 1**). In the two papers, a high percentage of tags unambiguously identified distinct genes in the genome. In addition, alternative transcription start sites within the same gene were identified. Matsushima's team now plans to expand their search for new or differentially expressed genes by comparing the 5′ tags obtained from cancer cells to 5′ tags derived from normal cells.

The power of 5′ SAGE to identify transcription starts is undisputed. Velculescu commends 5′ SAGE, noting how useful it will be to pinpoint 5′ ends. However, that is only half the work that Ruan and his group want to accomplish in the long run. Their goal is to unequivocally identify gene boundaries at both ends. To achieve this they modified the original LongSAGE method so that 3′tags comprising the polyadenylation sites were obtained. This combination of 5′ and 3′ SAGE showed great potential for finding gene boundaries; however, they were able to identify both a 5′ and a 3′ tag for only a small percentage of genes. When asked about this relatively poor performance, Ruan pointed to the fact that 5′ and 3′ tags are contained in separate libraries, which decreased the likelihood that corresponding matches could be found. Also, the number of tags analyzed was too low to cover the whole mRNA pool. With a total of about 300,000 mRNA molecules per cell, the 10,000 tags these authors generated covered only a small fraction. The future technical advances that Ruan sees for 5′-3′ SAGE are higher coverage of tags and the construction of ditags, which contain the sequences of both ends in one tag. This will keep the sequence of the corresponding 5′ and 3′ ends together, making it much easier to match them to the same gene .

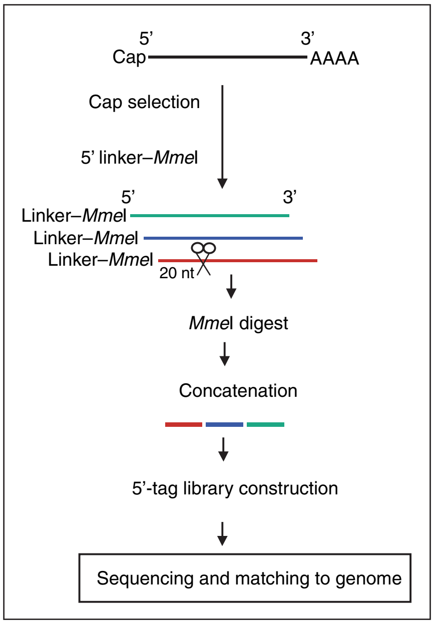

Figure 1 | Schematic of 5′ SAGE.

Methods such as 5′ and 3′ SAGE will be a tremendous methodological help in discovering new genes as well as analyzing their transcription frequency and potential alternative transcription start sites. To put it more poetically, in the words of Yijun Ruan: "With the sequence, the letters in the book of life are spelled out; now we have to learn to read it."

Nicole Rusk

RESEARCH PAPERS

Saha, S. *et al*. Using the transcriptome to annotate the genome. *Nat. Biotechnol*. **19**, 508–512 (2002).

Wei, C.L. *et al*. 5′ long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. *Proc. Natl. Acad. Sci. USA* **101**, 11701–11706 (2004).

Hashimoto, S. *et al*. 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol*. **22**, 1146–1149 (2004).