

Transcriptome annotation using tandem SAGE tags

E. Rivals⁽¹⁾, A. Boureux⁽²⁾, M. Lejeune⁽²⁾, F. Ottonnes⁽²⁾, O.P. Pérez⁽³⁾, J. Tarhio⁽³⁾, F. Pierrat⁽⁴⁾,
F. Ruffe⁽²⁾, T. Commes⁽²⁾ and J. Marti⁽²⁾

⁽¹⁾Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR CNRS 5506, Université Montpellier 2,
161 rue Ada, 34392 Montpellier 05, France

⁽²⁾Institut de Génétique Humaine, UPR CNRS 1142, 141 rue de la Cardonille, 34396 Montpellier 05, France

⁽³⁾Helsinki University of Technology, P.O. Box 5400, FI-02015 HUT, Finland.

⁽⁴⁾Skuld-Tech, 134, rue du Curat – Bât. Amarante, 34090 Montpellier, France
Anthony.Boureux@univ-montp2.fr

Analysis of several million expressed gene signatures (tags) revealed an increasing number of different sequences, largely exceeding that of annotated genes in mammalian genomes. Serial Analysis of Gene Expression (SAGE) can reveal new polyadenylated RNAs transcribed from previously unrecognized chromosomal regions. However, conventional SAGE tags are too short to identify unambiguously unique sites in large genomes. Here, we design a novel strategy with tags anchored on two different restrictions sites of cDNAs. New transcripts are then tentatively defined by the two SAGE tags in tandem and by the spanning sequence read on the genome between these tagged sites. Having developed a new algorithm to locate these tag-delimited genomic sequences, we first validated its capacity to recognize known genes and its ability to reveal new transcripts with two SAGE libraries built in parallel from a single RNA sample. Our algorithm proves fast enough to experiment this strategy at a large scale. We then collected and processed the complete sets of human SAGE tags to predict yet unknown transcripts. A cross-validation with tiling arrays data shows that 47% of these tag-delimited genomic sequences overlap transcriptional active regions. Our method provides a new and complementary approach for complex transcriptome annotation.