Digital gene expression data, cross-species conservation and noncoding RNA

Nicolas Philippe, Florence Ruffle, Elias Bou-Samra, Anthony Boureux, Thérèse Commes, Eric Rivals

Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS, équipe MAB, Université de Montpellier II, Montpellier, France,
CRBM, UMR 5237 CNRS, Montpellier, France
http://www.lirmm.fr/~rival

Recently developed sequencing technologies offer massively parallel production of short reads and become the technology of choice for a variety of sequencing-based assays, including gene expression. Among them, digital gene expression analysis (DGE), which combines generation of short tag signatures for cellular transcripts with massively parallel sequencing, offers a large dynamic range to detect transcripts and is limited only by sequencing depth. As recently described (Philippe et al, 2009), tag signatures can easily be mapped to a reference genome and used to perform gene discovery. This procedure distinguishes between transcripts originating from both DNA strands and categorizes tags corresponding to protein coding gene (CDS and 3'UTR), antisense, intronic or intergenic transcribed regions. Here, we have applied an integrated bioinformatics approach to investigate tags' properties, including cross-species conservation, and the ability to reveal novel transcripts located outside the boundaries of known protein or RNA coding genes. We mapped the tags from a human DGE library obtained with Solexa sequencing, against the human, chimpanzee, and mouse genomic sequences. We considered the subset of uniquely mapped tags in the human genome, and given their genomic location, determined according to Ensembl if they fall within a region annotated by a gene (CDS, UTR and intron) or an intergenic region. We found that 76.4 % of the tags located in human also matched to the chimpanzee genome. The level of conservation between human and chimpanzee varied among annotation categories: 85 % of conserved tags in the CDS, 81 % in the UTR, 76% and 73% respectively in intron and intergenic regions. With the same procedure applied to human and mouse, we obtained 11% of conserved tags in the CDS, 7% in UTR, 1% and 3% respectively in intronic and intergenic regions. We analysed in depth the common CDS and UTR tags in human and mouse for their functional relevance: 90% of them correspond to orthologous genes with a common HUGO. We used DAVID database to extract biological features, the gene clustering revealed specific molecular functions belonging to transcription cofactor and regulator activity, nucleotide binding, ligase and protein kinase, hormone receptor, histone methyltransferase or GTPase activity, and also important signaling pathways like WNT pathway. Indeed, intergenic transcription includes mainly new, non protein-coding RNAs (npcRNAs), which could represent an important class of regulatory molecules. By integrating also SAGE genie and RNA-seq expression data, we selected intergenic tags conserved across species and assayed experimentally the npcRNA transcriptome with Q-PCR. We validated 80% of the 32 tested biological cases. These results demonstrate that considering tag conservation helps to identify conserved genes and functions, which is of great relevance when investigating expressed tags located in intergenic regions.

References

1. N. Philippe, A. Boureux, L. Bréhèlin, J. Tarhio, T. Commes, E. Rivals (2009). Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. Nucleic Acids Research (NAR) Vol. 37, No. 15 e104, doi:10.1093/nar/gkp492; 2009.

Relevant Web sites

2. http://www.atgc-montpellier.fr/ngs/