

Investigating the transcriptomic repertoire based on High Throughput Sequencing data

Eric Rivals

► **To cite this version:**

Eric Rivals. Investigating the transcriptomic repertoire based on High Throughput Sequencing data. A. Denise and E. Rocha and S. Schbath. Colloque 2009 du GDR de Bioinformatique Moléculaire, Nov 2009, Paris, France. 2009, <<http://www.gdr-bim.u-psud.fr/journees-gdr.php>>. <lirmm-00833124>

HAL Id: lirmm-00833124

<https://hal-lirmm.ccsd.cnrs.fr/lirmm-00833124>

Submitted on 12 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Investigating the transcriptomic repertoire based on High Throughput Sequencing data.

Eric Rivals - CNRS, LIRMM, Montpellier, France

Keynote speech - Colloque 2009 du GdR Bioinformatique Moléculaire

Ultra-high throughput sequencing (HTS) is used to analyse the transcriptome or interactome at unprecedented depth on a genome-wide scale. These techniques yield short sequence reads that are then mapped on a genome sequence to predict putatively transcribed or protein-interacting regions. We argue that factors such as background distribution, sequence errors, and read length impact on the prediction capacity of sequence census experiments. Here we suggest a computational approach to measure these factors and analyse their influence on both transcriptomic and epigenomic assays. We developed and tuned a bioinformatic pipeline to assess the expression level of known mRNAs and predict novel splicing variants based on the transcript signatures (reads) obtained by Digital Gene Expression (DGE). However, almost 30% of the signatures map to non coding regions, suggesting the existence of unknown transcripts. To cross validate in silico those novel RNAs, we take advantage of RNA-seq, as well as other publicly available DGE data, and visualise all data in the genomic context.

Related publications :

1. Using reads to annotate the genome : influence of length, background distribution, and sequence errors on prediction capacity N. Philippe*, A. Boueux*, L. Bréhèlin, J. Tarhio, T. Commes, E. Rivals *Nucleic Acids Research (NAR)* doi:10.1093/nar/gkp492 ; 2009.
2. MPSCAN : fast localisation of multiple reads in genomes E. Rivals, L. Salmela, P. Kiiskinen, P. Kalsi, J. Tarhio *Proc. 9th Workshop on Algorithms in Bioinformatics Lecture Notes in Bioinformatics (LNBI)*, Springer-Verlag, Vol. 5724, p. 246-260, 2009.