Bioinformatic analisys of SOLiD transcriptome data

Albiero A(1,2), Vitulo N(2), Forcato C(2), Campagna D(2), Caniato E(2), Bilardi A(2),
Schiavon R(2), D'Angelo M(2), Zimbello R(2), Valle G(2)

(1) BMR-Genomics srl, Padova
(2) CRIBI - University of Padova, Padova

Motivation

Next-generation sequencing technologies allow the production of a large amount of data holding down costs and time. SOLiD technology produces up to 5Gbp per run in one week producing about 150.000.000 of short reads of 30-35 bases. Sequences are coded in color space, a new format where all the possible combination of di-nucleotides are coded by four colors. SOLiD features coupled with ultra deep sequencing allow and improve fine SNPs/errors discrimination, alternative splicing recognition and studies of gene expression profiles. In this poster we present an investigation on SOLiD technology potential, analyzing short reads coming from Vitis vinifera transcriptome as case-study.

Methods

Two independent transcript libraries from leaf and root tissues were sequenced with SOLiD technology and aligned on grape genome with Pass software (Campagna et Al. in press). Pass performs fast alignment in color space mode allowing gaps and mismatches. We run Pass allowing up to 4 mismatches and we consider only the best hit over entire genome. Mismatches in color space have not the same mean than in base space. One mismatch in base space is coded by 2 consecutive mismatches in color space, so 4 mismatches in color space are comparable to 3 mismatches in base space (as a maximum). One mismatch in base space can indicate an error on genome sequence, an error on read sequence or a SNP but it is not always possible to discriminate among these three possibilities. On the contrary, in color space we can recognize SOLiD sequencing errors, genome errors or SNPs using read quality, transition rules, coverage and distribution of mismatches. The coverage produced by the new DNA sequencers is very high and allow a precise evaluation of the expression level of each single gene. The number of reads matching on a gene normalized for its length represents its expression level. Moreover SOLiD sequences keep the direction of original mRNA allowing identification of repeat or low complexity regions and studies on gene expression and regulation.

Results

We obtained 139.467.080 reads from leaf and 188.742.647 from root library. We aligned about 37% of total reads (35.5% of root and 39.3% of leaf library) and we obtained 151.270.460 alignments (54.9% from root and 45.1% from leaf). Alignments analysis shows that SOLiD reads present different kind of errors: perfect match are less frequent than expected and it is very important to consider mismatches. Erros could be in reads (errors of SOLiD sequencing) or in reference (errors or SNPs). The first are generally single color mismatch with low quality, the second double color (consecutive) mismatch with high quality associated. SOLiD sequencing errors could be recognized by quality analysis, and their distribution over reads length is not equal: we find that errors tend to be accumulated at the end of the read and at the first position because of the first color position refers to the last base of the adaptor. Consecutive mismatches could be read errors, SNPs or reference errors: reference errors are associated with high coverage and high color quality, SNPs are also associated with permitted color transitions while read errors are coupled with low quality and coverage values. We checked alignments position related on gene prediction coordinates: results show that most part of reads maps on gene, only 8,9% of extragenic region is covered and 60,2% of CDS and UTR are covered. Reads mapping out of genes have been considered: most part of them are grouped in low complexity regions or in regions similar to transposable elements. Simulated data on library size and gene coverage show that there is a limit on library size over that there is not a gain on gene coverage. Directionality of reads gave us a new instrument for expression analyses. In many cases we found that a gene is expressed in both strand suggesting a possible antisense-MRNA regulation.

Contact : alessandro.albiero@bmr-genomics.it