Detection of alternative splicing in Vitis vinifera and validation by short sequencing reads

Picardi E[1], Horner D[2], Chiara M[2], Pesole G[1,3]

[1] Dipartimento di Biochimica e Biologia Molecolare “E.Quagliariello”, Università di Bari, Via Orabona 4, 70126 Bari, Italy
[2] Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, Via Celoria 26, 20133 Milano, Italy
[3] Istituto Tecnologie Biomediche-Sede di Bari, Consiglio Nazionale delle Ricerche, Via Amendola 122/B, 70126 Bari, Italy

Motivation

Alternative splicing enables the production of functionally different transcripts and proteins from the same gene locus and represents a key molecular mechanism to optimize and expand the coding capacity of eukaryotic genomes. In plants, alternative splicing has been considered to be much less frequent than in animals such as humans where over 95% of genes are alternatively spliced. However, recent genome-wide computational analyses have revealed that in flowering plants alternative splicing is more prevalent than previously thought. In Arabidopsis and Oryza, for example, 32.5% and 23.5% of annotated genes, respectively, are alternatively spliced. The differences between plants and animals concern not only the relative abundance of alternatively spliced genes and the average number of isoforms/gene but also the type of events preferred. Exon skipping is frequent in animals (about half of total events) but rare in plants; while intron retention is frequent in plants (in the range of 33-41% of total events) but rare in animals (Barbazuk et al 2008). These findings suggest that a thorough investigation of alternative splicing in plants is needed. For this reason, we have developed an ad hoc computational framework to detect and classify splicing events in plants given the complete genome sequence and a related pool of ESTs/FLcDNAs. Our framework has been applied to the complete genome of Vitis vinifera. It uses the EasyCluster program to build gene-oriented clusters of ESTs/FL-cDNAs and the ASPic software to detect alternative splicing events in given clusters. Finally, all predicted splice sites have been validated by mapping Illumina short reads onto reconstructed grapevine transcripts.

Methods

ASPic is a well-established software to predict alternatively splicing transcripts and events given a genomic region and a set of related EST/FL-cDNA sequences (Castrignanò et al. 2008; Castrignanò et al. 2006). Currently, genome-wide detection of alternative splicing by ASPic can be performed employing RefSeq transcripts and ESTs/FL-cDNAs from the Unigene database that map to specific genome locations. However, reliable gene annotations and EST clusters are available for a limited number of organisms and very few plants. In order to apply ASPic to organisms with incomplete gene annotations or unsatisfactory Unigene entries, gene-oriented clusters of ESTs/FL-cDNAs have been generated by EasyCluster. This program has been developed for a quick and reliable genome-based clustering of EST/FL-cDNA sequences. Each cluster and the corresponding genomic region are, then, used as input for ASPic enabling the prediction of alternative FL-transcripts and splicing events. Finally, all predicted transcripts are aligned to millions of Illumina Whole Transcriptome reads using Blat, in order to experimentally validate computationally detected splice sites.

Results

More than 550,000 sequences, including both ESTs and FL-cDNAs from VIGNA consortium and public ESTs from dbEST, have been mapped onto the 8x assembly of the grape genome and resulting alignments have been used to build gene-oriented clusters of ESTs using EasyCluster. In less than one hour, we obtained 15,536 clusters with an EST content ranging from 1 to 1675. All generated clusters have been analysed by ASPic to detect alternatively spliced transcripts and splicing events. On average, we recovered 1.5 transcripts per gene and we found that 30% of genes are subjected to alternative splicing. This last finding is in line with previous estimation of alternative splicing (32.5%) in Arabidopsis thaliana using EST/FL-cDNA sequences. Transcript comparisons per gene revealed, moreover, that also in grape the intron retention event is the most frequent (36%). Exon skipping events and alternative donor/acceptor events are equally frequent (16%) and, thus, more abundant than expected. Finally, we validated ASPic predicted splice sites by mapping more than 170,000,000 short Illumina reads onto computa-

tionally assembled grape transcripts. About 90% of all detected splice sites were supported by at least three independent reads. Moreover, about 75% of all reconstructed transcripts were fully validated by short reads, demonstrating the reliability of our framework in detecting correct alternative splicing events in grapevine.

Contact : e.picardi@biologia.uniba.it