Frequency and distribution of Simple Sequence Repeats (SSRs) in pineapple fruit transcriptome

Abstract

Simple Sequence Repeats (SSRs), also known as microsatellites or Short Tandem Repeats (STRs), are ubiquitous, hypervariable and abundant in many eukaryotes' genomes. They are generally thought to differ in size due to the effect of replication error and unequal recombination. Their distribution varies across genic, intergenic regions and non-coding regions of the DNA. SSRs may be predominant in certain regions which may reflect the tendency of mutational processes to create certain repeat patterns under selective pressure. The expansion of repeat motifs in the coding regions of the gene may alter gene expression. This paper reports the discovery of tandem repeat motifs and their frequency from a fully-ripe pineapple fruit transcriptome generated from contigs derived from paired-end Solexa mRNA sequencing. Paired-end sequences, with lengths of 75 bp each and with insert sizes of 200 bp, were assembled using Velvet which generated 28,728 contigs. These contigs were then used to mine for di-, tri,- and tetranucleotide SSR motifs, which are present in the fruit, by using SynaRex (Synamatix). Primers flanking the SSR loci were then designed by using the online software PRIMER 3. Mining of the regenerated contigs showed that only 3.5% of the fully-ripe pineapple fruit transcriptome contained SSRs. Dinucleotides were the most abundant and they accounted for 498 (49.6%) contigs containing SSR followed by trinucleotides and tetranucleotides with 467 (46.5%) and 40 (3.9%) contigs, respectively. Among the dinucleotide motifs, (TC)n and (GA)n were the most abundant with 45 and 41.6% occurrences, respectively. For trinucleotide motifs, (GGA)n, (CTC)n, (AGA)n and (TTC)n showed higher occurrences compared to other trinucleotide motifs with 20.3, 16.5, 13.1 and 9.2%, respectively. Of all the contigs containing SSR, only 26.6% were suitable for designing flanking primers for PCR. Of these, 66 contigs were

dinucleotide, 190 contigs were trinucleotide and 11 contigs were tetranucleotides. Further validation of the PCR primers will provide wider application in biomarker identification, genome mapping and characterization, phenotype mapping, marker assisted selection and finally in diversity studies of pineapples.