De novo assembly combining SOLiD mate-pair and 454 data

Caniato E, Vezzi A, Albiero A, Campagna D, Schiavon R, D'Angelo M, Zamperin G, Forcato C, Vitulo N, Valle G

Motivation

The advancements in DNA sequencing have opened the possibility to carry out de novo sequencing projects at a reasonable cost and in a short time. The technology is still evolving very rapidly as several "third generation" instruments have been announced, however at the moment two main types of DNA sequencers are available: 1) the Roche-454 that is based on pyrosequencing and produces about 1 million reads of "long" reads of 450 bases per run; 2) the Illumina-GA and the Applied Biosystems SOLiD that are based on different technologies and produce 150-300 million "short" reads of 50-75 bases per run. De novo assembly of genomic sequences have been successfully achieved using either long or short reads; long reads are much easier to assemble and a relatively low coverage (such as 15x) is sometime enough to succeed. Unfortunately, the cost per base of long reads is much higher than short reads. On the other hand, the assembly of genomic sequences from short reads requires a very high coverage (such as 70x) and a very powerful computing facility. Here we propose a mixed approach that makes use 454 shotgun reads coupled with SOLiD "mate-pairs". The combination of the two types of data allows a considerable reduction in sequencing costs. Moreover, the assembling algorithm that we propose requires a relatively small computing facility.

Methods

Most shotgun assembly methods are based on two main steps. Firstly, overlapping sequences are assembled into contigs. Secondly, mate pairs (i.e. the pair of reads obtained at the end of the same library insert) are used to sort the contigs, giving them an order and direction into the "scaffold". This second step is not based on sequence alignment, but only on the relative position of the mate pairs within each other an within their respective contigs. We use Newbler (V.2.3 - Roche) to obtain contigs from the 454 reads while for the scaffolding we have developed a new algorithm that uses PASS (Campagna et al., 2009) to align the SOLiD reads on the 454 contigs and CONSORT, a new bioinformatic tool that we designed to sort contigs into scaffolds. The main steps of CONSORT are shown in the figure.

CRIBI, Università di Padova, via Ugo Bassi 58/b, Padova, Italy, Italy

Results

We applied CONSORT to integrate SOLiD mate pair data and 454 data for the assembly of the Tomato genome. The results are very satisfactory, showing a considerable improvement in the estimate of gap size as well as in the scaffolding of the contigs.

Availability

http://pass.cribi.unipd.it/consort/

Contact e-mail

giorgio.valle@unipd.it

Image