Critique: "Filtering duplicate reads from 454 pyrosequencing"

Susanne Balzer, Ketil Malde1, Markus A. Grohme3 and Inge Jonassen

The paper describes a novel approach for filtering duplicate reads from 454 pyrosequencing data. This problem is motivated by the need of reduce sequencing errors and artifically duplicated reads in some applications such as de-novo whole genome sequencing or metagenomics. Existing solutions are often based on nucleotide sequences, while raw flowgram values, which contain additional information, are unsused.

Authors present a new software tool JATAC, which can be used for accurate duplicates filtering and accepts 454 flowgrams as input. Approach is based on reads clustering, performed by calculating all pairwise distances between reads. For distances calculation, probability of homopolymers having same length when observing corresponding flowgram is being used. The method was benchmarked on 3 different bacterial datasets and it showed better results, compaired to existing solutions.

Advantages of this approach are clear: usage of raw flowgram data gives more accurate estimation of reads similarity and, as a result, more precise duplicates filtering. While it looks quite sexy, there are some weak points in presented paper.

First of all, hierarchial clustering is performed according to some threshold constant, which was chosen empirically. It is unclear, why same constant should be used for clusters with different size and different read diversity. There are already some existing approaches (which are more accurate) with probabilistic models for reads clustering (see [1]).

Secondly, there are some more unmotivated empirical constants in distance calculating algorithm, which corresponds to maximum size of homopolymer, involved into flowgrams comparison. Thirdly, no running time and memory usage benchmarking was performed.

To sum up, despite of not very accurate reads clustering, presented software seems to be more accurate, comparing with existing solutions. Software could be used for some general filtering 454 sequencing data. Authors do not recommend using JATAC for IonTorrent data in present version.

1 References

1. Quince et al. Removing Noise From Pyrosequenced Amplicons. BMC Bioinformatics 2011, 12:38