To the Editors,

We are submitting our manuscript “Variational inference for rare variant detection in deep, heterogeneous next-generation sequencing data” as an original methodology article to BMC Bioinformatics. We developed a variational expectation maximization (EM) inference algorithm to detect rare single nucleotide variants and estimate non-reference allele frequencies in heterogeneous samples.

We have previously developed a variant detection model and published “RVD2: an ultra-sensitive variant detection model for low-depth heterogeneous next-generation sequencing data” on Bioinformatics, 31(17), 2015. However, the problem of RVD2 is that it is computationally slow for large data set. To address this problem, we developed a variational EM algorithm that is more computationally efficient than RVD2 on tests of low coverage sequencing data. Furthermore, we applied our variational EM algorithm on a novel data set that is a directed evolution longitudinal yeast data. We are able to detect novel variants that were not detected by the original publication.

We proved the accuracy of our variational EM algorithm compared with many state-of-the-art algorithms on a synthetic data set. We also detected rare variants with low non-reference allele frequencies (<1.0%) early in the evolutionary time course in the longitudinal data set.

The synthetic data set is available on <http://dna-discovery.stanford.edu/software/rvd/> under the heading of “Synthetic DNA Data”. The longitudinal data set is available on the NCBI Sequence Read Archive.

Thank you for your consideration.

Fan Zhang or Patrick Flaherty ?

May 22, 2016