VARIANT DETECTION MODEL WITH IMPROVED ROBUSTNESS AND ACCURACY FOR LOW-DEPTH TARGETED NEXT-GENERATION SEQUENCING DATA

Abstract. Massively parallel sequencing data has been generated by next-generation sequencing (NGS) technology for single nucleotide variants (SNVs) identification among related populations. To address the detection of SNVs, a variety of methods are being under-represented. However, by the reason of the error rate of the clonal heterogeneity and the limitation of the algorithms, identifying the true variants at minor allele frequencies for the very low sequencing read depth remains challenging.

We propose a novel empirical Bayesian model to call variants accurately. And the sensitivity of this hierarchical model is analyzed by the improper prior, information prior and non-information prior on the synthetic sequence data with varying read depth and a range of allele frequencies. Our model with information prior (log-normal prior) and non-information prior (Jeffreys prior) both performs a high accuracy (96%) when applied to a known 0.1% minor allele frequency within very low read depth (39). Moreover, the model with Jeffreys prior presents much lower false discovery rate (FDR) to a known 0.1% minor allele frequency event, compared with using improper prior. For further validation, our statistical model was applied on the yeast sequence data which also shows a high specificity and sensitivity over a wide range of read depth. Thus our Bayesian model achieves a enhanced robustness and accuracy when calling variants for the low read depth and minor allele frequencies.