SoloDel: A probabilistic model for detecting low-frequent somatic deletions from unmatched sequencing data

Junho Kim 1, Sanghyeon Kim 2, Hojung Nam 3, Sangwoo Kim 4, and Doheon Lee 1

1 Department of Bio and Brain Engineering, KAIST, Korea

2 Stanley Brain Research Laboratory, Stanley Medical Research Institute, USA.

3 School of Information and Communications, Gwangju Institute of Science and Technology, Korea

4 Severance Biomedical Science Institute, Yonsei University College of Medicine, Korea

Abstract

Motivation

Finding somatic mutations from massively parallel sequencing data is becoming a standard process in genome-based biomedical studies. There are a number of robust methods developed for detecting somatic single nucleotide variations (SNVs). However, detection algorithms for somatic copy number alteration (SCNAs) have been substantially less explored and remain vulnerable to frequently existing sampling issues: low frequency in cell population and absence of the matched control samples.

Results

We developed a novel computational method SoloDel that accurately detects low-frequent somatic deletions even without matched control samples. We first constructed a probabilistic, somatic mutation progression model that describes the occurrence and propagation of the event in the cellular lineage of the sample. We then built a Gaussian mixture model to represent the mixed population of normal and mutated cells in which distinct sets of deletions are contained. Parameters of the mixture model could be estimated using the expectation-maximization (EM) algorithm with the observed distribution of read-depth ratios at the points of discordant-read based initial deletion calls. SoloDel outperformed the existing algorithms at the most mutation frequency ranges (10~70%) in our artificially generated test data. SoloDel could also successfully recall experimentally validated somatic deletions from previously reported neuropsychiatric whole genome sequencing data.

Availability and implementation

Java-based implementation of the method is available at http://sourceforge.net/projects/solodel/