REMix: Base-calling using a Random Effects Mixture Model for Next Generation Sequencing Data

Xiaoxiao Sun*
Department of Statistics
University of Georgia
Athens, GA 30602, USA
xiaosun@uga.edu

Ashley Cacho
Department of Statistics
University of California
Riverside, CA 92521, USA
acach001@ucr.edu

Ping Ma
Department of Statistics
University of Georgia
Athens, GA 30602, USA
pingma@uga.edu

Xinping Cui
Department of Statistics
University of California
Riverside, CA 92521, USA
xinping.cui@ucr.edu

ABSTRACT

The remarkable technology advancement in next-generation sequencing (NGS) is revolutionizing many biological and medical applications. Base calling is a critical step in NGS procedure and has immediate impact on the accuracy and power of down-stream sequence variation analysis. It converges fluorescence intensity measurements into discretized reads which are short nucleotide sequences. However, current leading base calling methods are either too computationally expensive or not accurate to be of practical use. In this project, we directly model raw intensity signals with a random effect multivariate mixture model. By jointly estimating model parameters and posterior probability of four bases, our new base caller will avoid making too many assumptions about the type of biases and eliminate the need of data transformation that might introduce extra biases, and therefore improve the accuracy of base calling while at the same time allow all sources of error to be accounted when quantifying base calling quality and alignment quality. By incorporating open multi-processing(OpenMP) on Blacklight, Pittsburgh Supercomputing Center, our new base caller can deal with extraordinary amount of intensity data within a reasonable time period. It is also worth mentioning that our model-based base caller that avoids supervised learning has the advantage of a potentially increased yield in the case of de novo sequencing since it does not require re-sequencing of a known reference sequence for training.

Categories and Subject Descriptors

G.1 [**NUMERICAL ANALYSIS**]: Parallel algorithms; G.3 [**PROBABILITY AND STATISTICS**]: Probabilistic algorithms (including Monte Carlo)

General Terms

Algorithms

Keywords

Base Calling, EM Algorithm, OpenMP, NGS

*To whom correspondence should be addressed.