Faculty of Engineering and Information Technology
University of Technology Sydney

Error Correction Algorithms for Genomic Sequencing Data

Thesis submitted in fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Xuan Zhang

September 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Xuan ZHANG, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature of Candidate: Signature removed prior to publication.

[Xuan Zhang]

Date: 5th July, 2021

Place: Sydney, Australia

Acknowledgments

First of all, I would like to express my sincere thankfulness to my principal supervisor Prof. Jinyan Li for his continuous support, assistance and advice during my Ph.D. studies. I would like to thank other members of Prof. Jinyan's research team for their kind helps on my research and daily life, they are Dr. Hui Peng, Dr. Yi Zheng, Dr. Chaowang Lan, Dr. Yuansheng Liu, Dr. Zhixun Zhao, Dr. Xiaocai Zhang, Tao Tang, Tian Lan and Pengyao Ping. I would further like to thank my parents for their love and continued support. Thanks for all the memorized scenes they brought to my life.

I am also grateful to my co-supervisor Prof. Michael Blumenstein and Prof. Gyorgy Hutvagner, for their kindly helps and suggestions during my Ph.D. career.

I sincerely acknowledge the organizations that provide financial supports for this research, including the Australia Research Council and the University of Technology Sydney.

I would like to thank all the anonymous peer reviewers for their insightful comments and suggestions to significantly improve the quality of this work.

Lastly, I would like to express my deepest gratitude to my family members and friends, who were always encouraging me when I was in difficult times.

Xuan Zhang @ Sydney
July 2021

Contents

List of Figures

List of Tables

List of Publications

Journal Papers :

J-1: **Zhang Xuan**, Liu Yuansheng, Yu Zuguo, Blumenstein Michael, Hutvagner Gyorgy, & Li Jinyan. (2021) 'Instance-based error correction for short reads of disease-associated genes', *BMC Bioinformatics* 22.6, 1-18.

J-2: **Zhang Xuan**, Ping Pengyao, Hutvagner Gyorgy, Blumenstein Michael, & Li Jinyan. (2021) 'Aberration-corrected ultrafine analysis of miRNA reads at single-base resolution: a k-mer lattice approach', Accepted by *Nucleic Acids Research.* .

J-3: **Zhang Xuan**, Blumenstein Michael, & Li Jinyan. (2021) 'Small-RNA sequencing reads restoration through accurate error rectification' plan to submit to *Genome Biology.*

Conference Papers :

C-1: Zhang Xiaocai, **Zhang Xuan**, Verma Sunny, Liu Yuansheng, Blumenstein Michael. & Li Jinyan. (2019), Detection of anomalous traffic patterns and insight analysis from bus trajectory data, *in* 'Pacific Rim International Conference on Artificial Intelligence'. Springer, pp. 307-321.

C-2: Chen Hongjie, Wang Xun, **Zhang Xuan**, Zeng Xiangxiang, Song

Tao & Rodríguez-Patón Alfonso. (2018), LncRNA-disease association prediction based on neighborhood information aggregation in neural network. *in* 'IEEE International Conference on Bioinformatics and Biomedicine'. IEEE, pp. 175-178.

Abstract

The rapid development of high-throughput next-generation sequencing (NGS) platforms has produced massive sets of genomic reads under low costs for a wide range of biomedical applications (*e.g.*, *de novo* genome assembly, read alignment, resequencing, and Single-nucleotide polymorphism discovery). A serious concern over these datasets is that machine-made sequencing data suffers from lots of random errors (such as substitutions, insertions and deletions). To the best of our knowledge, all the existing methods suffer limitations. This work aims to rectify as many errors as possible by designing strategies adapted to specific cases. Three novel error correction algorithms are designed to providing high-quality sequencing data.

The first method is to use an instance-based strategy to correct errors, as described in Chapter 3. This novel instance-based error correction method is able to provide high quality reads for any given instance case and implemented as a tool named InsEC. It is designed to correct errors in reads related to instance cases (*e.g.*, a set of genes or a part of the genome sequence. The nature of data characteristics and fine-grand features are considered to gain better correction performance. In our method, the instance-based strategy makes it possible to make use of data traits only related to an instance, which guarantees that we can approach the ground truth of the instance case and then achieve better error correction performance. In the instance extraction step, all reads related to a given instance are extracted by using read mapping strategies. In the correction step, we take advantage of alignment processes and correct errors according to the alignment. Besides, statistical models are used to avoid induced errors as well. Intensive experiments are conducted with other state-of-the-art methods on both simulated and real datasets. The results demonstrate the superiority of our method, which achieves the best error correction performance (*e.g.*, precision, recall and gain rate in average) and further assembly results (*e.g.*, N50, the length of contig and contig quality).

Chapter 4 develops the first method for miRNA read error correction. Existing error correction methods do not work for miRNA sequencing data attributed to miRNAs' length and per-read-coverage properties distinct from DNA or mRNA sequencing reads. Although the error rate can be low at 0.1%, precise rectification of these errors is critically important because isoform variation analysis at single-base resolution such as novel isomiR discovery, editing events understanding, differential expression analysis, or tissue-specific isoform identification is very sensitive to base positions and copy counts of the reads. We present a novel lattice structure combining kmers, (k-1)mers and (k+1)mers to address this problem. Moreover, the method is particularly effective for correcting indel errors. Extensive tests on datasets having known ground truth of errors demonstrate that the method is able to remove almost all of the errors, without introducing any new error, to improve the data quality from every-50-reads containing one error to every-1300-reads containing one error. Studies on wet-lab miRNA sequencing datasets show that the errors are often rectified at the 5' ends and the seed regions of the reads. Note that there are remarkable changes after the correction in miRNA isoform abundance, the volume of singleton reads, overall entropy, isomiR families, tissue-specific miRNAs, and rare-miRNA quantities.

Chapter 5 introduces a novel method for small RNA error correction which supports substitution, insertion and deletion error rectification. Compared with the miRNA error correction method, this method is more robust by supporting all kinds of small RNA sequencing (read length from 20 to 200 nucleotides). Furthermore, we improve the three-layer lattice structure and combine it by reads with the same length, length plus one and length minus one, which dramatically increases the method's efficiency. Finally, we consider RNA's isoform and propose to do correction proportionally to make a fine correction. Specifically, in the correction phase, we do not correct all potential erroneous copies to the top one candidate; Instead, we divide corrections into top 3 candidates proportionally to remain all possible recovery. With this improvement, the method achieves high error correction performance, and its precision, recall and gain rate are superior to all other existing error correction methods. Extensive experiments on simulation and raw sequencing data prove our method's ability. Thus, our error correction method does help improve data quality and is necessary for all downstream analyses.