

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.

SIGNATURE OF CANDIDATE: \_\_\_\_\_  
[Xuan Zhang]

Production Note:  
Signature removed prior to publication.

DATE: 5<sup>th</sup> 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

Certificate of Authorship/Originality . . . . .	<b>i</b>
Acknowledgment . . . . .	<b>iii</b>
List of Figures . . . . .	<b>ix</b>
List of Tables . . . . .	<b>xi</b>
List of Publications . . . . .	<b>xiii</b>
Nomenclature . . . . .	<b>xv</b>
Abstract . . . . .	<b>xv</b>
<b>Chapter 1 Introduction . . . . .</b>	<b>1</b>
1.1 Background . . . . .	<b>1</b>
1.1.1 Genomic Sequencing Data . . . . .	<b>2</b>
1.1.2 Error Correction Strategy . . . . .	<b>3</b>
1.1.3 Applications of Sequencing Data . . . . .	<b>6</b>
1.2 Research Motivations . . . . .	<b>7</b>
1.3 Research Objectives and Contributions . . . . .	<b>9</b>
1.4 Thesis Structure . . . . .	<b>11</b>
<b>Chapter 2 Related Work and Literature Review . . . . .</b>	<b>14</b>
2.1 Error Correction Classification . . . . .	<b>14</b>
2.1.1 Classification According to Data characteristics . . . . .	<b>14</b>
2.1.2 Classification According to Correction Strategies . . . . .	<b>15</b>
2.2 The Existing Error Correction Methods . . . . .	<b>16</b>
2.2.1 The Kmer Based Methods . . . . .	<b>16</b>
2.2.2 The Alignment Based Methods . . . . .	<b>17</b>

2.2.3	The suffix array Based Methods . . . . .	19
2.3	Error Correction Evaluation metrics . . . . .	21
2.3.1	Statistical Error Correction Performance . . . . .	21
2.3.2	Statistical Assembly Performance . . . . .	21
2.4	Summary . . . . .	22
 <b>Chapter 3 Instance-based Error Correction for Short Sequencing reads . . . . . 23</b>		
3.1	Introduction . . . . .	23
3.2	Methods . . . . .	27
3.2.1	Reads extraction . . . . .	27
3.2.2	Error correction step . . . . .	29
3.3	Experiments and Results . . . . .	31
3.3.1	Sequencing Read Datasets . . . . .	34
3.3.2	Evaluation Metrics . . . . .	35
3.3.3	Performance Evaluation . . . . .	37
3.4	Summary . . . . .	44
 <b>Chapter 4 Error Correction Method for MicroRNA Sequencing Reads . . . . . 46</b>		
4.1	Introduction . . . . .	46
4.2	Methods . . . . .	49
4.2.1	A 3-layer Kmer Lattice Structure . . . . .	50
4.2.2	Error Correction Steps . . . . .	51
4.3	Experiments and Results . . . . .	54
4.3.1	Sequencing Read Datasets . . . . .	55
4.3.2	Evaluation Metrics . . . . .	58
4.3.3	Performance Evaluation . . . . .	60
4.3.4	Case Studies . . . . .	68
4.4	Summary . . . . .	83



<b>Chapter 5 Extended Error Correction Method for Small RNA Sequencing Reads . . . . .</b>	<b>85</b>
5.1 Introduction . . . . .	85
5.2 Methods . . . . .	87
5.2.1 A 3-layer Read Lattice Structure . . . . .	87
5.2.2 Proportional Correction . . . . .	88
5.3 Experiments and Results . . . . .	91
5.3.1 Read Datasets . . . . .	91
5.3.2 Evaluation Metrics . . . . .	94
5.3.3 Performance Evaluation . . . . .	95
5.4 Summary . . . . .	100
<b>Chapter 6 Conclusion and Future Work . . . . .</b>	<b>102</b>
6.1 Conclusion . . . . .	102
6.2 Summary of Important Results . . . . .	104
6.3 Perspectives and Future Research . . . . .	105
<b>Bibliography . . . . .</b>	<b>107</b>



# List of Figures

1.1	Thesis structure. . . . .	12
3.1	An example of read correction on eight reads. The base-type frequency $f(r)$ and dominance value $p(r)$ of a base are shown below that base. For the columns of bases, the dominant nucleotide types are in bold and the erroneous bases are in the red color. The updated sequence is on the top and the correction details are listed in the right. For example, in column 26, there are two dominant nucleotides ( <i>G and C</i> ). $f(C)$ is larger than $f(G)$ , so the nucleotide C is used to update the sequence, and the erroneous base A (position[26]) in read4 is corrected to C. For read3, the third base after ranking is below the threshold, so read3 is labeled ‘out’ and deleted from the extracted subset. . . . .	32
3.2	Two examples of point mutations in case studies. . . . .	44
4.1	A 3-layer kmer lattice structure. . . . .	51
4.2	The numbers of corrected bases vary at different lengths of kmers. . . . .	54
4.3	Proportions of unique-read count are changed compared with uncorrected data in average of 12 salmon datasets. (a) The miREC runs with continuous k value from 8 to 20. (b) The miREC runs with continuous k value from 8 to 25. . . . .	66

*List of Figures*

---

4.4	Dataset entropy changes before and after the error correction by miREC on the 12 salmon miRNA datasets. (a) when the continuous k settings from 8 to 20; (b) when the continuous k settings from 8 to 25. . . . .	67
4.5	The distributions of correction positions. . . . .	67
4.6	A rare miRNA in the Alzheimer’s disease patient aged 94 showing significant copy count change after error correction. . . . .	76
5.1	A 3-layer read lattice structure. . . . .	88

# List of Tables

1.1	Comparison with different sequencing techniques . . . . .	4
2.1	Comparison with the existing error correction methods . . . . .	20
3.1	Description of the datasets . . . . .	35
3.2	Genes related to lung cancer on human chromosome one . . . . .	36
3.3	Performance comparison of instance-based error corrections . . . . .	38
3.4	Performance comparison. Instance-based approach vs global approach . . . . .	40
3.5	Assembly results compared with the ground truth . . . . .	41
3.6	The contigs from corrected reads vs the reference sequence . . . . .	42
4.1	Description of our simulated datasets . . . . .	56
4.2	Description of twelve wet-lab salmon miRNA sequencing datasets and four human miRNA datasets. . . . .	57
4.3	Outstanding error correction performance by our miREC in comparison with the best available tools . . . . .	62
4.4	Changes in the read counts of some miRNAs . . . . .	70
4.5	isomiRNA detection . . . . .	72
4.6	Rank changes of the top-10 common miRNAs in salmon heart and brain tissues after error correction . . . . .	73
4.7	Ranking position change of tissue-specific miRNAs in the heart tissue (vs the liver tissue) before and after error correction . . . . .	75

4.8	Copy count change and ranking position change of top-10 differentially expressed miRNAs in the high glucose incubated human beta cell after error correction, and those in the low glucose incubated human beta cell after error correction. . . .	77
4.9	Ranking position changes of age-specific miRNAs in brain tissue from an Alzheimer male patient aged 94 (vs a patient aged 75) before and after error correction . . . . .	78
4.10	Copy count enhancement of 10 rare miRNAs after error correction in the human brain dataset related to an Alzheimer’s disease patient aged 94 . . . . .	80
4.11	Correction performance by miREC in comparison with Karect on the sequencing reads of the 963 miRXplore Universal Reference miRNAs (pure control and spike-in) . . . . .	81
5.1	Description of used datasets . . . . .	92
5.2	Details of wet-lab sequencing datasets with a few known errors	93
5.3	Outstanding error correction performance by our SRNAEC in comparison with the best available tools . . . . .	97
5.4	Error correction results on raw sequencing datasets with injected errors . . . . .	98
5.5	Changes of unique small RNA reads and entropy . . . . .	99

# 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

*List of Publications*

---

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.

