Algorithms for structural variant detection

Lin, J.

Algorithms for Structural Variant Detection

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden
op gezag van de rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college van promoties
te verdedigen op vrijdag 24 juni 2022
klokke 10.00 uur

door

Jiadong Lin

geboren te Xi'an, China

in 1991

Promotiecommissie

Promotores:	Dr. Walter A. Kosters	
	Prof.dr. Kai Ye	
Overige leden:	Prof.dr. Thomas H.W. Bäck	
	Dr. Lu Cao	
	Dr. Hailiang Mei	Leiden University Medical Center
	Prof.dr. Yianyong Sun	Xi'an Jiaotong University
	Prof.dr.ir. Fons J. Verbeek	
	Prof.dr. Gerard P. van Westen	

ISBN 978-94-6421-778-0

Het onderzoek beschreven in dit proefschrift is uitgevoerd aan het Leiden Institute of Advanced Computer Science (LIACS, Universiteit Leiden) en aan Xi'an Jiaotong University.
The research is financially supported by the Chinese Scholarship Council (CSC No. 201906280462).

Contents

1 Introduction and background **1**
- 1.1 Computational genomics . . . 1
- 1.2 Emerging DNA sequencing technologies . . . 2
- 1.3 Genome structural variations are important . . . 3
- 1.4 Detecting structural variation . . . 5
- 1.5 Pairwise sequence alignment for nucleotide sequences . . . 7
- 1.6 Usage of graphs for structural variants detection and analysis 10
- 1.7 Frequent subgraph mining . . . 13
- 1.8 Objective and outline of the thesis . . . 13
- 1.9 List of abbreviations . . . 16

2 Mako **19**
- 2.1 Introduction . . . 20
- 2.2 Materials and methods . . . 22
  - 2.2.1 Overview of Mako . . . 22
  - 2.2.2 Building signal graph . . . 24
  - 2.2.3 Detecting CSVs with pattern growth . . . 24
  - 2.2.4 Performance evaluation . . . 26
  - 2.2.5 Preparing CSV benchmarks for performance evaluation 27
  - 2.2.6 Orthogonal validation of Mako detected CSVs . . . 28
  - 2.2.7 Data availability . . . 28
- 2.3 Results . . . 29
  - 2.3.1 Mako effectively characterizes multiple breakpoints of CSV . . . 29
  - 2.3.2 Mako precisely discovers CSV unique-interval . . . 31
  - 2.3.3 Performance on real data . . . 31
  - 2.3.4 CSV subgraph illustrates breakpoints connections . . . 32
  - 2.3.5 Contribution of homology sequence in CSV formation 35
- 2.4 Conclusion . . . 37

3 SVision **39**
3.1 Introduction . . . 40
3.2 Material and methods . . . 40
3.2.1 Overview of SVision . . . 40
3.2.2 Three-channel coding of sequence . . . 41
3.2.3 Detecting CSVs from denoised images via tMOR . . . 43
3.2.4 Creating CSV graphs from denoised images . . . 44
3.2.5 Quality score of discoveries . . . 46
3.2.6 Training data and CNN model training . . . 47
3.2.7 Evaluating simple structural variants detection with real data . . . 48
3.2.8 Evaluating complex structural variant detection . . . 49
3.2.9 Analysis and validation of high-quality CSVs detected from HG00733 . . . 51
3.2.10 Data availability . . . 53
3.3 Results . . . 54
3.3.1 Evaluating simple SV detection with real data . . . 54
3.3.2 Performance of detecting complex structural variants . 56
3.3.3 CSV mediated gene structure change and genome evolution . . . 59
3.4 Conclusion . . . 62

4 SpotSV **63**
4.1 Introduction . . . 64
4.2 Material and methods . . . 66
4.2.1 Overview of SpotSV . . . 66
4.2.2 Modify reference sequence with predicted structural variants . . . 68
4.2.3 Generating denoised segments based on k-mers . . . 68
4.2.4 Calculating structural variant validation score . . . 69
4.2.5 Data availability . . . 70
4.3 Results . . . 71
4.3.1 Evaluating SpotSV with simulated data . . . 71
4.3.2 Validating structural variants in a well-characterized genome . . . 76
4.3.3 Structural variant breakpoint validation and accuracy 78
4.4 Conclusion . . . 81

5 Assessing reproducibility **83**
5.1 Introduction . . . 84
5.2 Materials and methods . . . 86
5.2.1 Read mapping and SV detection . . . 86
5.2.2 Evaluating recall and precision of each algorithm . . . 86
5.2.3 Identification and classification of PAV calls missed by each algorithm . . . 87
5.2.4 Evaluating breakpoint accuracy . . . 88
5.2.5 Examine call set overlaps between platforms and aligners 89
5.2.6 Data availability . . . 89
5.3 Results . . . 90
5.3.1 Evaluating the impact of aligners and platforms on detection algorithms . . . 90
5.3.2 Evaluation recall and precision of detection algorithms using different benchmarks . . . 92
5.3.3 Features of PAV calls missed by detection algorithms . 95
5.3.4 Examining the effects of platforms and aligners on breakpoint accuracy . . . 96
5.3.5 Effects of aligners on tumor SV detection . . . 98
5.4 Conclusion . . . 101

6 Conclusions and perspectives **105**
6.1 Conclusions . . . 105
6.2 Perspectives . . . 107

Bibliography **109**

English summary **121**

Nederlandse samenvatting **125**

Acknowledgements **129**

Curriculum vitae **131**