ILLUMINATING REARRANGED CANCER GENOME STRUCTURES THROUGH GENOME GRAPHS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Xiaotong Yao

May 2021

ILLUMINATING REARRANGED CANCER GENOME STRUCTURES THROUGH GENOME GRAPHS

Xiaotong Yao, Ph.D.

Cornell University 2021

Cancer genomes harbor structural variations (SV). Whole genome sequencing (WGS) characterizes SVs. They contain drivers, their patterns reflect mutational processes, and their evolution chronicles that of the cell populations. Many simple and complex patterns of SVs are discovered. Yet, one of the biggest hurdles before we wield this powerful data to achieve a more complete understanding of cancer, is that there lacks flexible and general analytical frameworks to elucidate the complexity of SVs. This is because SVs inherently alter the coordinate system of the reference genome, thus the interpretation of any junction is dependent on other overlapping junctions. Plus, due to the limited scope of short-read WGS, we cannot yet phase most junctions nor obtain long linear sequences of the rearranged chromosome. Thereby, I present genome graph framework implemented in the R packages gGnome and JaBbA, which treats rearranged genome sequences as directed graphs, where

BIOGRAPHICAL SKETCH

Xiaotong Yao is a PhD student in the Tri-institute Program in Computational Biology and Medicine at Weill Cornell Medicine.

To Rosalind Yao and Shan Huang.

ACKNOWLEDGEMENTS

Never enough gratitude to my mentor Dr. Marcin Imielinski for showing me how scientific questions are identified, defined, and can be answered through mining the data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SV – structural variation
WGS – whole genome sequencing
CN – copy number
CNA – copy number aberration
JaBbA – Junction Balance Analysis

LIST OF SYMBOLS

CHAPTER 1

INTRODUCTION

1.1 Complexity of SVs in the cancer genomes

1.2 Copy number and junctions are two facets of the same structure

1.3 Characterization of structural variations through genome graphs

1.4 From patterns to etiology

1.5 Evolution of SVs after telomere crisis

CHAPTER 2

JUNCTION-BALANCED GENOME GRAPHS REPRESENT STRUCTURALLY ALTERED GENOMES

In [2], we described

2.1 Genome graph as a general data structure to represent rearranged genome

2.2 Walks on a genome graph correspond to linear DNA sequences

Genomic sequences are strings of nucleotides and not graphs.

CHAPTER 3

DISTINCT CLASSES OF COMPLEX STRUCTURAL VARIATION UNCOVERED ACROSS THOUSANDS OF CANCER GENOME GRAPHS

CHAPTER 4

STRUCTURAL VARIANT EVOLUTION AFTER TELOMERE CRISIS

4.1

APPENDIX A

CHAPTER 1 OF APPENDIX

Appendix chapter 1 text goes here

[1]

$$k_1 = \frac{\omega}{c(1/\varepsilon_m + 1/\varepsilon_i)^{1/2}} = k_2 = \frac{\omega sin(\theta)\varepsilon_{air}^{1/2}}{c} \tag{A.1}$$

BIBLIOGRAPHY

[1] Lewis Carroll (Charles L. Dodgson). *Alice's Adventures in Wonderland*. George MacDonald, 1865.

[2] Kevin Hadi, Xiaotong Yao, Julie M Behr, Aditya Deshpande, Charalampos Xanthopoulakis, Huasong Tian, Sarah Kudman, Joel Rosiene, Madison Darmofal, Joseph DeRose, Rick Mortensen, Emily M Adney, Alon Shaiber, Zoran Gajic, Michael Sigouros, Kenneth Eng, Jeremiah A Wala, Kazimierz O Wrzeszczyński, Kanika Arora, Minita Shah, Anne-Katrin Emde, Vanessa Felice, Mayu O Frank, Robert B Darnell, Mahmoud Ghandi, Franklin Huang, Sally Dewhurst, John Maciejowski, Titia de Lange, Jeremy Setton, Nadeem Riaz, Jorge S Reis-Filho, Simon Powell, David A Knowles, Ed Reznik, Bud Mishra, Rameen Beroukhim, Michael C Zody, Nicolas Robine, Kenji M Oman, Carissa A Sanchez, Mary K Kuhner, Lucian P Smith, Patricia C Galipeau, Thomas G Paulson, Brian J Reid, Xiaohong Li, David Wilkes, Andrea Sboner, Juan Miguel Mosquera, Olivier Elemento, and Marcin Imielinski. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell*, 183(1):197–210.e32, October 2020.