STRUCTURAL VARIATION DISCOVERY IN REPEAT- AND DUPLICATION-RICH REGIONS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By
Emre Dogru
February 2018

STRUCTURAL VARIATION DISCOVERY IN REPEAT- AND
DUPLICATION-RICH REGIONS

By Emre Dogru

February 2018

We certify that we have read this thesis and that in our opinion it is fully adequate,
in scope and in quality, as a thesis for the degree of Master of Science.

_____

Can Alkan(Advisor)

_____

No academic titles, only the name Y

_____

No academic titles, only the name Z

Approved for the Graduate School of Engineering and Science:

_____

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

STRUCTURAL VARIATION DISCOVERY IN REPEAT- AND DUPLICATION-RICH REGIONS

Emre Dogru

M.S. in Computer Engineering

Advisor: Can Alkan

February 2018

Write the abstract here.

Keywords: Key one, key two.

ÖZET

TEKRARLI VE DUPLİKASYON BAKIMINDAN ZENGİN BÖLGELERDE YAPISAL VARYASYON KEŞFİ

Emre Dogru
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Can Alkan
Şubat 2018

Türkçe özet buraya gelecek.

Anahtar sözcükler: Anahtar Sözlerim.

Acknowledgement

I acknowledge that . . .

Contents

List of Figures

List of Tables

Chapter 1

Introduction

- Early History (Darwin Mendel Identification of DNA etc.)

It all started with Darwin's book of The Origins of Species, which he published in 1859. Since then, the field of genomics have been paid a great deal of attention. What Darwin argues in his book is that animals that is more resilient to the environmental conditions contributed more to the genetic hereditary. His claim was based on the finches living in Galapagos Islands; those who survived are the ones that have developed some sort of adaptations according to the island they lived in. Therefore, Darwin concluded that the finches that have survived were naturally selected, what he called the process as natural selection.

In those years, there was an Augustinian monk whose name was Gregor Mendel was conducting experiments on his monastery's garden. His experiments were different than that of Darwin's, since Mendel used more scientific methods in his experiments. Mendel used pea plants to observe how the genetic hereditary passes down to younger generations. He crossbred the different pea plants each of which has different traits. Then he made two important observation. One of them is that some traits appear more than others and the second is even if a trait is no longer observed in one generation, it may reappear later in the offspring. His barely appreciated work at that time has earned him the title "Father of

Genetics".

- 1953 Watson and Crick model

- 1961 Marshall Nirenberg and Har Gobind Khorana DNA codons

- 1970 Frederick Sanger

- Human genome project

- Preparation (1987 - 1990)

- 1990 - 2001

- Going on

1.1 Sequencing technologies

1.1.1 Early Sequencing Methods

1.1.2 High-Throughput Sequencing

1.2 Genomic Variation

1.2.1 Structural Variation

1.2.1.1 Copy Number Variation

1.3 Segmental Duplication

1.3.1 Singly Unique Nucleotides

1.4 Our Contribution

Chapter 2

Gene Copy Number

2.1 Data Preprocessing

2.2 Extracting Singly Unique Nucleotides

2.3 Read Depth

2.3.1 GC Correction

2.4 Results

Chapter 3

Paralog Gene Copy Number

3.1   Motivation

3.2   Results

Chapter 4

Conclusion and Discussion

4.1 Future Work

Bibliography

Appendix A

Data

Appendix B

Code