BIAS CORRECTION IN FINDING COPY NUMBER VARIATION WITH USING READ DEPTH-BASED METHODS IN EXOME SEQUENCING DATA

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Fatma Balcı

August, 2014

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

Assist. Prof. Can Alkan(Advisor)

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

Assist. Prof. Buğra Gedik

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

Assist. Prof. Özlem Çavuş

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

BIAS CORRECTION IN FINDING COPY NUMBER VARIATION WITH USING READ DEPTH-BASED METHODS IN EXOME SEQUENCING DATA

Fatma Balcı
M.S. in Computer Engineering
Supervisor: Assist. Prof. Can Alkan
August, 2014

Medical research has striven for identifying the causes of disorders with the ultimate goal of establishing therapeutic treatments and finding cures since its early years. This aim is now becoming a reality thanks to recent developments in whole genome (WGS) and whole exome sequencing (WES). Despite the decrease in the cost of sequencing, WGS is still a very costly approach because of the need to evaluate large number of populations for more concise results. Therefore, sequencing only the protein coding regions (WES) is a more cost effective alternative. With the help of WES approach, most of the functionally important variants can be detected. Additionally, single nucleotide polymorphisms (SNPs) that are located within coding regions are the most common causes for Mendelian diseases (i.e. diseases caused by a single mutation). Moreover, WES approaches require less analysis effort compared to whole genome sequencing approaches since only 1% of whole genome is sequenced. Besides the advantages, there are also some shortcomings that need to be addressed such as biases in GC−content and probe efficiency. Although there are some previous studies on correcting GC−content related issues, there are no studies on correcting probe efficiency effect. In this thesis, we provide a formal study on the effects of both GC−content and probe efficiency on the distribution of read depth in exome sequencing data. The correction of probe efficiency will make it possible to develop new CNV discovery methods using exome sequencing data.

Keywords: Copy number variations, read depth, bias correction, GC content, exome sequencing, next-generation sequencing, probe efficiency, DNA sequencing.

ÖZET

DİZİ DERİNLİĞİ YÖNTEMİ KULLANILARAK KOPYA SAYISI FARKLILIKLARINI TESPİT ETMEDE EKZOM DİZİLEME DATALARINDA VAROLAN ETKİLERİN DÜZELTİLMESİ

Fatma Balcı
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Yrd. Doç. Can Alkan
Ağustos, 2014

İnsanlığın varoluşundan bu yana tıptaki araştırmalar, hastaları tedavi edebilmek ve hastalıkların çaresini bulabilmek adına bunların altında yatan sebepleri bulmak için yapılmıştır. Bu amaç, son zamanlarda tüm genom ve tüm ekzom dizilemede yaşanan gelişmeler sayesinde gerçekleştirilebilmektedir. Dizileme maliyetlerinde yaşanan azalmalara rağmen, daha doğru sonuçlar elde edebilmek adına çok sayıda insan genomunun dizilenme ihtiyacı olduğundan tüm genom dizileme halen yüksek maliyetli bir yöntemdir. Bu sebeple, sadece protein kodlayan bölgeleri dizileyen tüm ekzom dizileme yöntemi nispeten daha az maliyetli bir alternatiftir. Tüm ekzom dizileme yaklaşımlarının yardımıyla, fonksiyonel önem taşıyan varyantların çoğu bulunabilmektedir. Buna ek olarak, Mendeliyen (tek mutasyon kaynaklı) hastalıkların en büyük sebebi olan tek nükleotid polimorfizmlerinden, ekzon bölgelerinde yer alanlar da bulunabilmektedir. Ayrıca tüm ekzom dizilemeye dayalı yaklaşımlar, insan genomun sadece %1'lik kısmını kapsadığından diğer yaklaşıma göre analiz yaparken daha az çaba gerektirmektedir. Ancak doğru sonuçlar elde edebilmek için ekzom dizileme datasında varolan prob etkinliği ve GC içeriği gibi sapma etkilerinin duzeltilmesi gerekmektedir. Bunlardan GC içeriği sapmasını düzeltmek icin yapılmış bazı çalışmalar bulunmaktadır. Ancak literatürde, prob etkinliği sapmasını düzeltmek amacıyla yapılan bir çalışma bulunmamaktadır. Bu tezde ekzom dizileme datasına ait dizi derinlemesi dağılımında varolan prob etkinliliği ve GC içeriği sapmaları üzerinde çalışılmıştır. Prob etkinliği sapmasının düzeltilmesiyle birlikte, ekzom dizileme datası kullanan yeni kopya sayısı varyantı bulma metotları geliştirilmek mümkün olacaktır.

Anahtar sözcükler: Kopya sayısı farklılıkları, dizi derinliği, etki düzeltme, GC içeriği, ekzom dizileme, yeni nesil dizileme, prob verimliliği, DNA dizileme.

Acknowledgement

Foremost, I would like to express my sincere gratitude to my advisor Assist. Prof. Can Alkan for the continuous support of my research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor.

Besides my advisor, I would like to thank the rest of my thesis committee: Assist. Prof. Buğra Gedik and Assist. Prof. Özlem Çavuş for their support.

My special thanks goes to Basri Kahveci for his endless patience and faith. I couldn't be finished with this work without his support.

I would like to thank TUBITAK for offering me the scholarship opportunity, through grant 112E135.

I thank my hardworking friend Elif Dal in Alkan Lab. I also thank my friend Havva Gülay Gürbüz for all the fun we have had in the last two years.

Last but not the least, I would like to thank my brother Ahmet Balcı for his endless support. I wouldn't be who I am without him. I would like to thank my parents Makbule Balcı and Zeki Balcı for supporting me spiritually throughout my life.

Contents

1 Introduction **1**

1.1 Motivation 2

1.2 Problem Statement 3

1.3 Contributions 4

2 Background **6**

2.1 DNA Sequencing 6

2.1.1 DNA 6

2.1.2 Gene 7

2.1.3 Chromosome 8

2.2 DNA Sequencing Technologies 9

2.2.1 Sanger Sequencing (First-Generation Sequencing Technology) 9

2.2.2 Next-Generation Sequencing (Second-Generation Sequencing Technology) 10

2.2.3 Next Next Generation Sequencing (Third - Generation Sequencing Technology) (Single Molecule Sequencing) 15

2.2.4 Fourth-Generation Sequencing Technology (Nanopore Sequencing) . . . 16

2.3 Genome Sequencing . . . 17

2.3.1 Whole Genome Sequencing . . . 18

2.3.2 Whole-Exome Sequencing . . . 18

2.4 Genomic Variations . . . 20

2.4.1 Single Nucleotide . . . 20

2.4.2 2 bp to 1,000 bp . . . 20

2.4.3 1 kb to Submicroscopic . . . 21

2.4.4 Microscopic to Subchromosomal . . . 22

2.4.5 Whole Chromosomal to Whole Genome . . . 23

2.5 The Effects of Copy Number Variations on Human Health and Phenotype . . . 23

3 Finding Copy Number Variations in Exome Sequencing Data 26

3.1 Four-Step Procedure . . . 26

3.1.1 Mapping . . . 27

3.1.2 Correcting Biases and Normalization . . . 27

3.1.3 Estimation of Copy Number . . . 30

3.1.4 Segmentation . . . 30

3.2 Methods . . . 30

3.2.1 Paired-end mapping (PEM)-based methods . . . 32

3.2.2 Split read-based methods 33

3.2.3 Read depth-based methods 34

3.2.4 Assembly-based methods 36

3.2.5 Hybrid approaches 37

4 Related Works **38**

4.1 Whole Genome Sequencing 38

4.1.1 Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing 38

4.2 Whole Exome Sequencing 41

4.2.1 Copy Number Variation Detection and Genotyping from Exome Sequencing Data 41

4.2.2 Discovery and Statistical Genotyping of Copy Number Variation from Whole Exome Sequencing Depth 41

5 Description of the Experiments **44**

5.1 Data 44

5.2 Mapping 45

5.2.1 Mapping of Reads to the Reference: MrsFAST-Ultra . . . 45

5.2.2 Calculation of read depth: Bedtools 46

5.3 Correcting Biases and Normalization 47

5.3.1 Calculation of Correlation Coefficients for Each Exon Region 51

5.3.2 Calculation of Correlation Coefficients for Each Gene Region 54

5.3.3 Finding optimum span parameter of LOESS method . . . 61

6 Conclusion **68**

6.1 Future Work 69

A Glossary **76**

B Length measurements **78**

C Timeline of DNA **79**

List of Figures

2.1 DNA structure 7

2.2 Human chromosomes 8

2.3 Workflow of the Sanger Sequencing Method 10

2.4 Workflow of the next-generation sequencing 11

2.5 454/Roche Machines 12

2.6 ABI Solid Machine and its procedure 13

2.7 Illumina Machines 14

2.8 IonTorrent Machines 15

2.9 Pacific Biosciences Machine (SMRT) 16

2.10 MinION and GridION System 17

2.11 Comparison of the size of whole genome and whole exome that are found on human genome 17

2.12 Sequencing 18

2.13 Workflow of whole exome sequencing 19

2.14 Classes of structural variation 21

2.15 Tiger with down syndrome 24

3.1 Four-step procedure to find CNVs in WES data 26

3.2 High and low coverage 29

3.3 Multiple vs. unique mapping 29

3.4 Classification of CNV detection methods 31

3.5 Paired-end mapping-based methods 33

3.6 Split read-based methods 33

3.7 Calculation depth of coverage 35

3.8 Read depth-based method 35

3.9 Assembly-based methods 36

5.1 Read depth and probe efficiency for each exon 52

5.2 Read depth and probe efficiency for each exon (0.24<GC Content<0.47) 53

5.3 Read depth and GC content for each exon 54

5.4 Read depth and probe efficiency for each gene 56

5.5 Read depth and probe efficiency for each gene (0.24<GC Content<0.47) 57

5.6 HG00629 58

5.7 HG01191 58

5.8 HG01437 59

5.9 NA19664 59

5.10 NA19707 60

5.11 NA19723 60

5.12 NA20766 61

5.13 Smoothed read depth and probe efficiency by LOESS method for each gene (HG00629 (Span=0.001)) 62

5.14 Smoothed read depth and probe efficiency by LOESS method for each gene (HG00629 (Span=0.05)) 62

5.15 Smoothed read depth and probe efficiency by LOESS method for each gene (HG00629 (Span=0.9)) 63

5.16 Smoothed read depth and probe efficiency by LOESS method for each gene (HG00629 (Span=0.005)) 63

5.17 Smoothed read depth and probe efficiency by Robust LOESS method for each gene (HG00629 (Span=0.005)) 64

5.18 Smoothed read depth and probe efficiency by LOESS method for each gene 66

5.19 Smoothed read depth and probe efficiency by LOESS method for each gene 67

List of Tables

2.1 Comparison of the DNA sequencers 10

2.2 Comparison of WES and WGS 20

3.1 Applicability of the tools to the methods 32

4.1 Summary of bioinformatics tools for CNV detection using WGS data. This table is adapted from [1]. 40

4.2 Summary of bioinformatics tools for CNV detection using WES data. This table is adapted from [1]. 43

5.1 Correlation between read depth, probe efficiency and GC content for each exon 51

5.2 Correlation between read depth, probe efficiency and GC content for each exon (0.24<GC Content<0.47) 52

5.3 Correlation between read depth, probe efficiency and GC content for each gene 55

5.4 Correlation between read depth, probe efficiency and GC content for the genes (0.24<GC Content<0.47) 55

5.5 Correlation between read depth, probe efficiency and GC content for each smoothed gene data 64

5.6 Correlation between read depth, probe efficiency and GC content for each smoothed gene data ($0.24 <$ GC Content < 0.47) 65

Chapter 1

Introduction

Although giant strides have been made in recent years in the field of bioinformatics, there remains an open question as to find copy number variation (CNV) more accurately to better understand the underlying genetic causes of several diseases, such as autism and schizophrenia. There are four basic sequence signatures that can be used to identify CNV (see Section 3.2); but the read depth-based method is the most reliable with whole exome sequencing (WES) data. However, there are several errors in coverage, named biases, introduced in exome capture, which prevent this method to work accurately, as they dramatically alter the read depth distribution properties, and they fail to provide accurate results data because of these biases.

GC-content and probe capture efficiency are two causes of the biases in exome sequencing data. Some studies exist in the literature to correct the GC-content bias, although most have limitations. Moreover, there isn't any work about probe capture efficiency. This thesis characterizes the effects of GC-content affecting exome sequencing read depth distribution; and demonstrates that the correction of GC-content and probe capture efficiency simultaneously works better in smoothing the depth distribution. Our study described in the following chapters attempts to decrease the biases in WES data through identifying the effects of both GC-content and probe capture efficiency in the read depth distribution for accurately characterizing CNV. We also provide an insight into how to correct

for these biases using a well-known statistical smoothing technique, called *locally weighted scatterplot smoothing* (LOESS). After this error correction step, it will be easier to apply more standard CNV identification algorithms to better discover CNV using WES data.

In chapter 2, we present background information to help better understand the biological and technical concepts. We define the biases, and provide comparisons of CNV discovery algorithms, and provide the basic steps of these methods in Chapter 3. In Chapter 4, we discuss the related previous studies. The description of the experiment is given in detail in Chapter 5. Finally, we evaluate our formulations and source of bias characterizations in Chapter 6.

1.1 Motivation

Obtaining accurate knowledge of nucleic acid composition is crucial to all life sciences. Deciphering DNA sequences started to shed light on novel biological functions and phenotypic differences, which increased the demand for highly efficient sequencing technologies. As a consequence, an era of synthetic genomics and personalized medicine is expected to start within the next few years.

The aim of this thesis is to help to find copy number variation (CNV) accurately in whole exome sequencing (WES) data by calculating read depth (RD) to be able to get rid of the analysis burden of whole genome sequencing (WGS) data.

The new cost-efficient and high throughput strategies for DNA sequencing are now the leading power house of discoveries in life sciences. The new sequencing technologies, commonly referred to as *high throughput sequencing* (HTS), or *next-generation sequencing* (NGS) started to appear in 2007. The advantages of HTS platforms are many: Cost of sequencing is reduced 10,000-fold, while data throughput is increased 30-fold per base per day. Despite the improvements, data generated with HTS platforms are more difficult to analyze without a *priori* information, however the availability of whole genome assemblies for humans and

all major model organisms has strengthened the potential utility of HTS.

In addition, general progress in technology across other related fields, including chemistry, nucleotide biochemistry, computation, data storage, and others, helped make better use of the HTS data. However, we still need to improve algorithms to better characterize genomic variation. Although the methods to discover and genotype single nucleotide polymorphisms (SNPs) are maturing, accurate detection of copy number variation (CNV) is still lacking, but there exist some algorithms with different strengths and weaknesses in the literature. [2]

1.2 Problem Statement

Depth of coverage is the average number of reads representing a given nucleotide in the reconstructed sequence. Most of the CNV discovery methods using WGS data perform statistical tests based on a Poisson model, in which reads are assumed to be distributed uniformly across the genome, since the sequence reads are assumed to be chosen randomly from the genome. It means RD in a region should follow a Poisson distribution with mean directly proportional to the size of the region and to the copy number. However, this assumption hardly holds even for normal genomes due to the biases mentioned in Chapter 3.

GC-content and probe capture efficiency are two sources of bias in data. GC-content is the percentage of guanine (G) and cytosine (C) bases in a genomic region. GC-abundance is heterogeneous across the genome and often correlated with functionality so it affects each region differently and needs to be corrected. Exome sequencing involves exon-capture step by which the coding regions are selected from the total genome DNA by means of hybridization. The characteristics of the probe, such as length, conformation and abundance on the solid phase, are of relevance in determining the capture efficiency. This capture efficiency is different for each probe due to the characteristics. Probe capture efficiency determines whether if an exon will be captured or not and the length of the region captured. Therefore, it also needs to be corrected.

CNV discovery methods with using WES data is mostly affected by GC-content and probe capture efficiency biases that can only be corrected locally with using smoothing methods to be able to reveal data points including CNV in the graph.

Although there are lots of computational methods and tools in the field to find CNV by using read depth-based (RD-based) methods, all of them suffer from various types of biases. After the work completed in this thesis, almost all of the tools will work more reliably by enhancing data.

1.3 Contributions

Learning more about the relationship between copy number variation (CNV) and exome sequencing, could help understanding the effects of CNV on humans and lead to huge improvements on comprehending of the underlying causes of some important diseases and phenotypic changes of humans. Although CNV can affect the other species, we are only interested in human genomes, since most available whole exome sequencing (WES) data are generated from human samples. There is a problem in finding the underlying causes of some important diseases such as schizophrenia and cancer.

Despite the great number of researches on finding CNV, most of them is based on finding CNV in whole genome sequencing (WGS) data. This problem has negatively impacted by the magnitude of the WGS data because a human genome is approximately comprised of 3 billion nucleotides. There are some powerful tools available to find CNV in WGS data. It seems like that they can be used for whole exome sequencing (WES) data, but the usage of these tools is not possible due to some biases. A possible cause of this problem is the difference between probe efficiencies. A study which investigates to understand the relationship between probes and read depth by improving an algorithm could remedy to use WES data with some healing.

Firstly, we need to find the read depth and GC-content of the exon regions

in the samples' genomes. We benefit from the 1000 Genomes data and Agilent Sure Select Capture Kit data with this aim. After reducing some noises in data, such as sex chromosomes, we calculate the correlation coefficient to understand the relationship between read depth and probe efficiency. We demonstrated that there is an important relationship between read depth and probe efficiency.

At the end of this process, we should normalize the data with the help of *locally weighted scatterplot smoothing* (LOESS) method because each bases in WES data are generally affected by the regions that are close to them. Therefore, we need to evaluate each DNA regions locally. We need to correct data as much as possible to be able to develop statistical approaches with using WES data. The hardest part of this process is to separate bias-based and CNV-based deviations in the data.

People who are working in the field of bioinformatics, patients with genetic diseases, doctors who want to understand the underlying causes of genetic diseases, and scientists who work in the related fields of science may benefit from this thesis. If successful tools we plan to develop using the methods presented in this thesis may also be used in clinical sequencing tests that we expected to be used in all hospitals within the next few years.

Chapter 2

Background

2.1 DNA Sequencing

2.1.1 DNA

The hereditary material in humans and almost all other organisms is called as deoxyribonucleic acid (DNA). Human DNA is mostly found in the cell nucleus, but it can also be found in the mitochondria in a small amount. Adenine (A), cytosine (C), guanine (G), and thymine (T) are the coding elements of DNA, nucleotides.

Adenine and thymine are called as purines, whereas cytosine and guanine are called as pyrimidines. Each nucleotide consists of a phosphate group, a 5-carbon sugar (deoxyribose), and a nitrogen containing base attached to the sugar. These four types of nucleotides differ only in the nitrogenous base. The order of these nucleotides determines the information to build and maintain an organism. Nucleotides are located in two long strands that form a spiral, double helix. Almost each cells in our body has the same DNA. Approximately, 3 billion bases exist in human DNA.

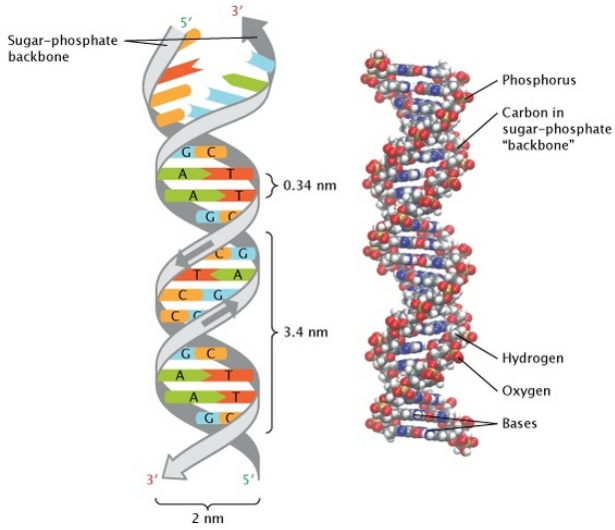

Figure 2.1: The grey ribbons that represent the sugar-phosphate backbone have arrows that run in opposite directions to indicate that the two strands of the helix are arranged in an anti-parallel manner. The upper end of one strand is labeled five prime (5'), and the lower end of the same strand is labeled three prime (3'). The nucleotide bases are shown as differently colored rectangles. The nucleotide guanine (G), shown in blue, binds with the nucleotide cytosine (C), shown in orange and the nucleotide adenine (A), shown in green, binds with the nucleotide thymine (T), shown in red. Gold spheres represent phosphorus atoms, grey spheres represent carbon atoms, white spheres represent hydrogen atoms, red spheres represent oxygen atoms, and blue spheres represent nitrogen atoms. This figure is adapted from [3].

2.1.2 Gene

The principle physical and functional unit of heredity is called as a gene. A gene is used as a template to make protein molecules and they are made up from DNA. The number of genes found in a human body has been estimated between 20.000 and 25,000 by the Human Genome Project. There are two copies of genes in human genomes. One of them is inherited from mother, whereas another is from father. Almost all genes are common between humans except 1% of the

genes.

2.1.3 Chromosome

The thread-like structures in which DNA located in human body are called as chromosomes. 46 chromosomes are found in each human cell. 23 of them are inherited from mother, whereas the remaining 23 of them are inherited from father. Humans have 22 pairs of autosomes and one pair of sex chromosomes, the X and Y.

Autosomes are roughly ordered due to their size. The largest chromosome is Chromosome 1 which has approximately 2,800 genes. Moreover, the smallest chromosome is Chromosome 22 which has approximately 750 genes. These genes are providing instructions for making proteins. Changes in the structure or number of copies of a chromosome can cause problems with health and development, but it doesn't have to cause any problems.

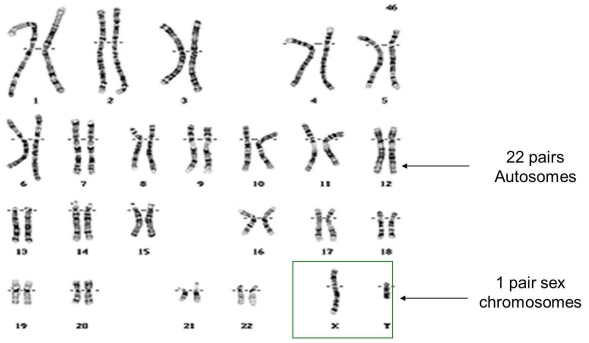

Figure 2.2: Human chromosomes. This figure is adapted from [4].

A pair of sex chromosomes are found in each human normally. Females have two X chromosomes, whereas male have one X and one Y chromosome.

2.2 DNA Sequencing Technologies

Nucleic acid sequencing is a way to determine the exact order of the DNA. The usage of nucleic acid sequencing has become accessible for researchers in the past decade. These sequencing techniques are key tools in many fields ranging from archeology, genetics, anthropology, biotechnology, forensic sciences to molecular biology. As the first major foray into DNA sequencing, The Human Genome Project is completed in 2004 at the cost of approximately $3 billion. It was a 14-year-long endeavor. The project is completed with using Sanger sequencing which is developed in 1975 by Frederick Sanger.

There are different kinds of platforms for DNA sequencing in the market. Four generations of DNA sequencing technologies can be distinguished by their nature and the kind of output they provide. The field of DNA sequencing technology development has a rich and diverse history.

2.2.1 Sanger Sequencing (First-Generation Sequencing Technology)

Sanger sequencing method had become the gold standard for 30 years after its discovery in 1977. This method uses DNA polymerase which makes use of inhibitors that terminate the newly synthesized chains at specific residues.

DNA to be sequenced can be prepared in two different ways, shotgun de novo sequencing or targeted resequencing. The output of both methods is an amplified template. Then, template denaturation, primer annealing, and primer extension are performed in a cycle sequencing. Primer is an oligonucleotide complementary to target DNA and leads the DNA extension. With the help of fluorescently labeled ddNTPS, each round of primer extension is halted. Labeled ddNTPs in its current form are mixed with regular, non-labeled, and non-terminating nucleotides in a cycle sequencing reaction. The label on the terminating ddNTP of any fragment corresponds to the nucleotide identifying its terminal position. To separate sequences by length and to provide subsequent interrogation of the

terminating base capillary electrophoresis is applied. Software provides DNA sequences and also their error probabilities for each base-call. [5] [6]

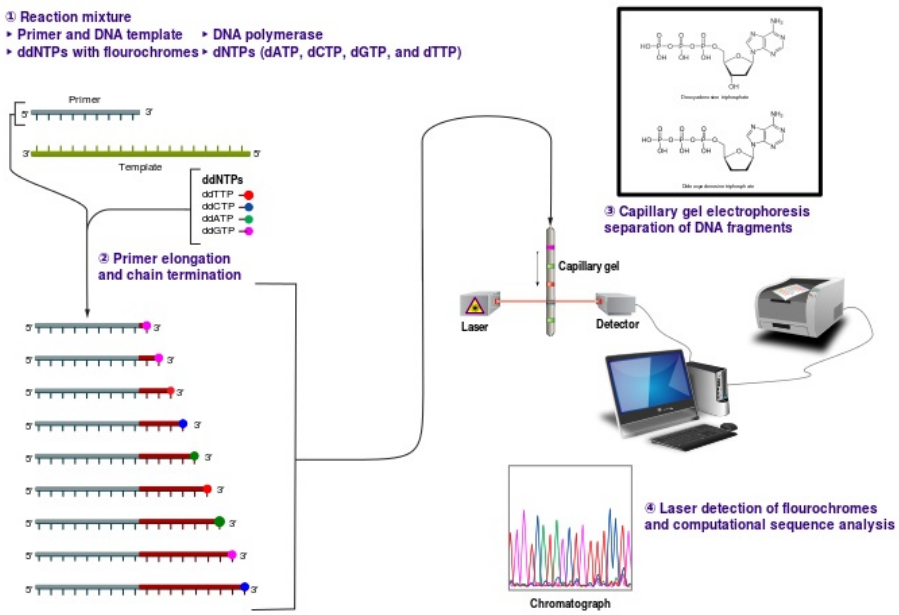

Figure 2.3: Workflow of the Sanger Sequencing Method. This figure is adapted from [7].

2.2.2 Next-Generation Sequencing (Second-Generation Sequencing Technology)

Machine	Capacity	Speed	Read Length	Cost Per Base
454 Roche	35-700 Mb	10-23 hours	400-700 bp	$714/14285 \times 10^{-8}$ €
SOLiD	90-180 Gb	7-12 days	75 bp	$3/5 \times 10^{-8}$ €
Illumina	6-600 Gb	2-14 days	100-250 bp	$2/333 \times 10^{-8}$ €
Ion Torrent	20 Mb-1 Gb	4-5 hours	200 bp	$100/10000 \times 10^{-8}$ €
PacBio	1 Gb	30 minutes	3,000 bp	$60/80 \times 10^{-8}$ €

Table 2.1: Comparison of the DNA sequencers. This table is adapted from [8].

After the completion of the Human Genome Project, cheaper and faster sequencing methods are demanded in the market. This demand has revealed the development of next-generation sequencing (NGS) methods. NGS is used for a fast, affordable, and through way to determine the underlying genetic causes of diseases. Millions of fragments of DNA from a single sample can be sequenced in unison with NGS. Massively parallel sequencing technology performed in NGS facilitates high-throughput sequencing. With this technology an entire genome are sequenced in less than ten days. In addition to these advantages of NGS, the cost required for a whole human genome has decreased with this technology. It is also minimizing the need for the fragment-cloning methods which are frequently used in Sanger sequencing.

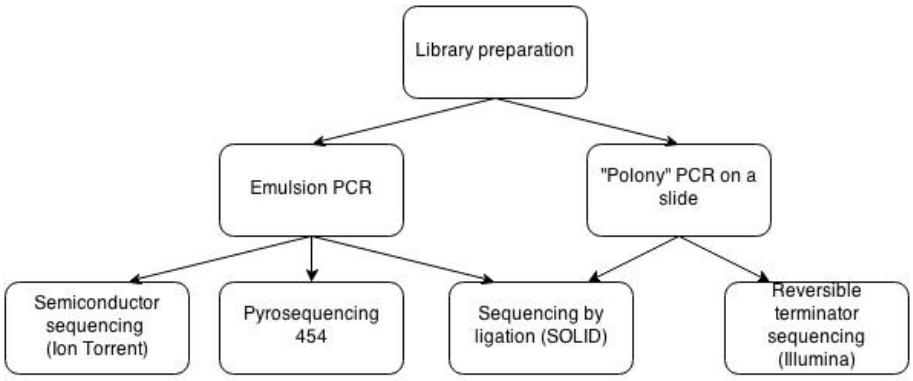

Figure 2.4: Workflow of the next-generation sequencing. This figure is adapted from [8].

After the appearance of the first 2nd generation sequencer in 2005, several second generation sequencers followed this emergence in the market. They are working conceptually similar although they have differences in sequencing biochemistry as well as in how the array is sequenced.

2.2.2.1 454/Roche

The first next-generation sequencing technology was 454 pyrosequencing. 454/Roche sequencing method consists of library preparation, emulsion PCR,

and pyrosequencing.

454 is based on the "sequencing by synthesis principle" which means taking the single stranded DNA to be sequenced and sequencing its complementary in an enzymatic way. In this method the activity of DNA polymerase is monitored by another enzyme, chemiluminescene. When the complementary is bound by the single-stranded sequenced DNA, light is produced. Sequencing is completed by the produced chemiluminescent signals [6].

Figure 2.5: 454/Roche Machines. This figure is adapted from [8].

2.2.2.2 ABI Solid

Emulsion PCR is used to generate the clonal sequencing features in the sequencing process of ABI Solid sequencing technology. Di-base sequencing technique in which two nucleotides are read via sequencing by ligation is used at each step of the sequencing process.

Although there are 16 base combinations of di-bases are possible, 4 dyes are used by the system. Therefore, 4 di-bases are represented by a single color. All bases are interrogated twice by the sequencing machine. Each following base can be derived in this way if the previous base is known. Moreover, a misidentified color can change all of the following bases in the translation. ABI Solid is rarely preferred nowadays. [9]

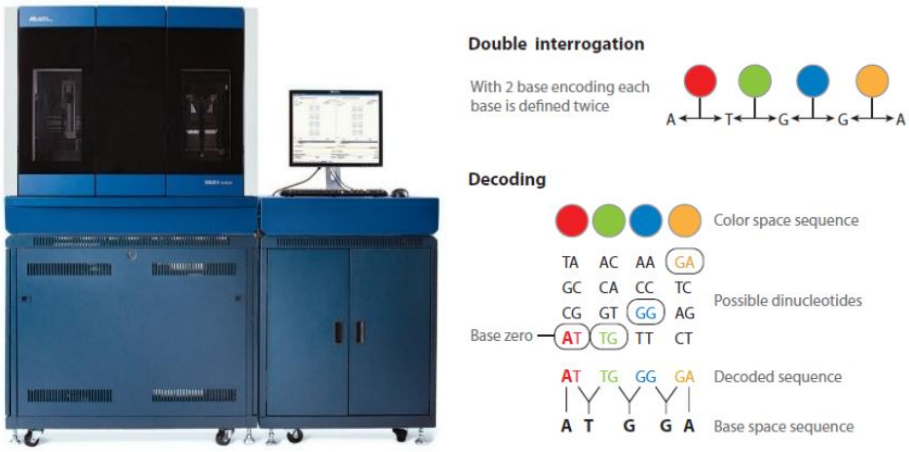

Figure 2.6: ABI Solid Machine and its procedure. This figure is adapted from [8].

2.2.2.3 Illumina

Although better platforms exist in the market, Illumina is the market leader due to the lower prices. Data produced by Illumina's machine is also used in this project. Illumina sequencing is comprised of library preparation, clustering, and sequencing processes.

In the first part, DNA simultaneously fragment and tag the extracted and purified DNA with adapters. Reduced cycle amplification adds the sequencing primary binding sites, indices, and regions that are complementary to the flow cell oligos after the ligation of adapters.

Each fragment molecule is amplified isothermally in the clustering part. The flow cell is a glass line with lanes. Each lanes is a channel coded with the composed of two types of oligos. Complementary oligo to the adapter region on one of the fragment strands, enables hybridization. A polymerase creates a complement of the hybridized fragment. The double stranded molecule is denatured and the original template is washed away. The strands are amplified the bridge amplification clonally. In this process the strands pulls over and the adapter

region hybridizes to the second type of the oligo on the flow cell. The complementary strand is generated by polymerases forming a double stranded bridge which is then denatured resulting in two single stranded copies of the molecule. The process is then simultaneously repeated for millions of clusters resulting for all fragments. Then, the reverse strands are washed away.

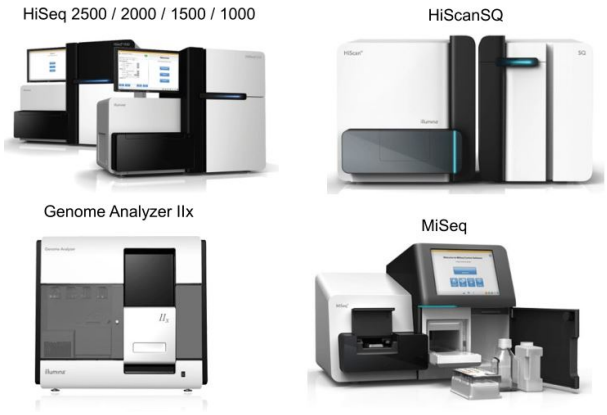

Figure 2.7: Illumina Machines. This figure is adapted from [8].

In the last part, sequencing, first sequencing primer is extended to produce the first read. One of the 4 fluorescently tagged nucleotide is incorporated based on the sequence of the template. After each incorporation, a light source is used and a characteristic fluorescent signal is emitted. This process is known as "sequencing by synthesis". The length of a read is determined by the cycle number and the base call is determined by the emission wave length. Hundreds of millions of clusters are sequenced in a massively parallel process. This entire process generates billions of reads representing all the fragments. [10] [11]

2.2.2.4 IonTorrent

Workflow of Ion Torrent is comprised of library preparation, emulsion PCR, and semiconductor sequencing processes. A hydrogen ion is naturally released as a by-product, when a nucleotide is incorporated into a strand of DNA by a polymerase.

Ion Torrent works on the principle of detection of these hydrogen ion releases in a massively parallel manner. These ions are detected on ion-semiconductor sequencing chips. Ion Torrent technology creates a direct connection between the chemical and digital events. [12]

Figure 2.8: IonTorrent Machines. This figure is adapted from [8]

2.2.3 Next Next Generation Sequencing (Third - Generation Sequencing Technology) (Single Molecule Sequencing)

Single molecule sequencing has ability to resequence the same molecule multiple times for improved accuracy and the ability to sequence molecules that cannot be readily amplified because of extremes of GC content, secondary structures, and other reasons. The main focus of the molecule sequencing technology is generally on read length, error rate, and throughput. Potential for lower cost, higher throughput, improved quantitive accuracy, and increased read lengths are offered by single molecule sequencing. [13]

2.2.3.1 Pacific Biosciences (SMRT)

Pacific Biosciences (SMRT) that comprised of library preparation and sequencing processes is an example of single molecule sequencing. Pacific Biosciences also developed a "sequencing by synthesis" approach using fluorescently labeled

nucleotides.

Figure 2.9: Pacific Biosciences Machine (SMRT). This figure is adapted from [8].

DNA is constrained to a small volume in a zero-mode wave guide and fluorescently labeled nucleotide near the DNA polymerase is measured because of low capability of light penetration. Each nucleotide has a characteristic incorporation time which helps improving base calls. However, the raw error rate is significantly higher than any other current sequencing technology. Therefore, it creates challenges for variation detection. [13]

2.2.4 Fourth-Generation Sequencing Technology (Nanopore Sequencing)

The results of fourth-generation sequencing technologies have not been sufficiently evaluted yet because this is the newest sequencing technology and it requires time. There is limited information about this technology for now.

2.2.4.1 Oxford Nanopore

Oxford Nanopore follows two parallel strategies. "Exonuclease sequencing" is the first one which based on exonuclease digestion of a single-stranded template into nucleotides that are fed into a nearby protein nanopore in a lipid membrane. "Strand sequencing" is the second strategy which is like feeding thread through the eye of a needle.

Figure 2.10: MinION and GridION System.This figure is adapted from [8].

Detailed information about nanopore sequencing is not given here because it requires time to be evaluated sufficiently. [8] [14]

2.3 Genome Sequencing

Genome sequencing can be divided into two categories that is defined below as whole genome sequencing (WGS) and whole exome sequencing (WES).

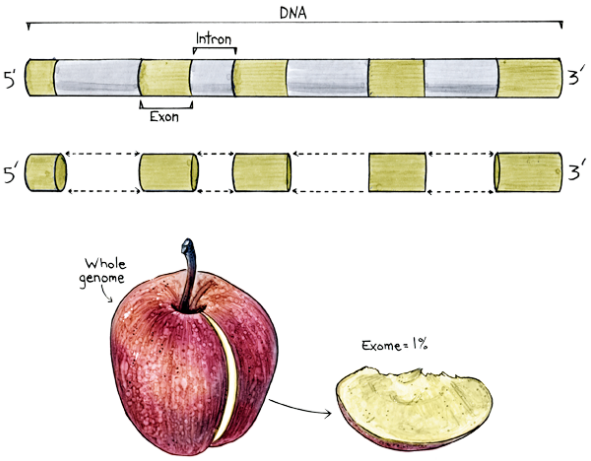

Figure 2.11: Comparison of the size of whole genome and whole exome that are found on human genome. This figure is adapted from [15].

2.3.1 Whole Genome Sequencing

Whole genome sequencing (WGS) is a laboratory method that reads the exact sequence of all DNA bases in an entire genome. The genome contains an individual's entire genetic code which means all of their genetic information. This entire genetic code includes protein coding regions (exons as well as the areas of the genetic code that do not give instructions for making proteins which are non-coding DNA sequences (introns).

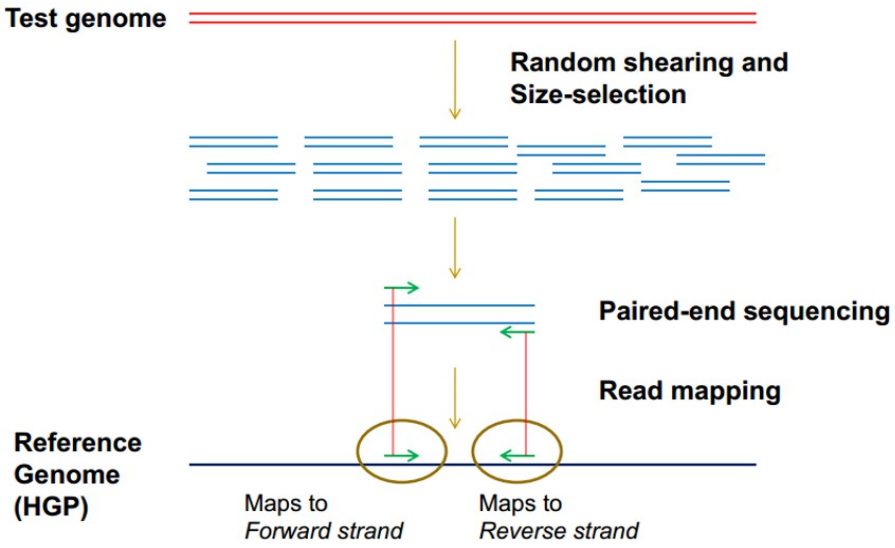

Figure 2.12: Sequencing

2.3.2 Whole-Exome Sequencing

All genomic regions coding for proteins and untranslated regions flanking them form the exome of an organism. The exome comprises just over 1% of the genome and 230,000 exons. It provides sequence information for protein-coding regions.

Most of the inherited disorders are believed to reside on the coding regions so laboratories can focus exclusively on exon regions. It gives an opportunity to eliminate the tremendous mass of non-coding DNA in the genome. The major advantage of using WES instead of WGS data is using only 1-2% of human

genome. The cost for sequencing is lowered significantly in this way. [1] [16]

The possibilities for analyzing exome has changed with NGS. In the past years exome sequencing is widely used in gene discovery and the identification of disease-causing mutations in pathogenic presentations in that the exact genetic cause is not known. Rare mutations that change the function of a protein which is the cause of most Mendelian and non-Mendelian diseases.

Almost half of the reported CNVs overlap with exon regions in DNA. Therefore, usage of exome sequencing data is more sensible in terms of time and cost efficiency. Most of the essential information about life mechanism of a human body can be obtained with exome sequencing data so redundant data are not necessarily processed.

Even though the reads have a non-uniform distribution, data can be improved by the work discussed in this thesis and some related works [10] [17].

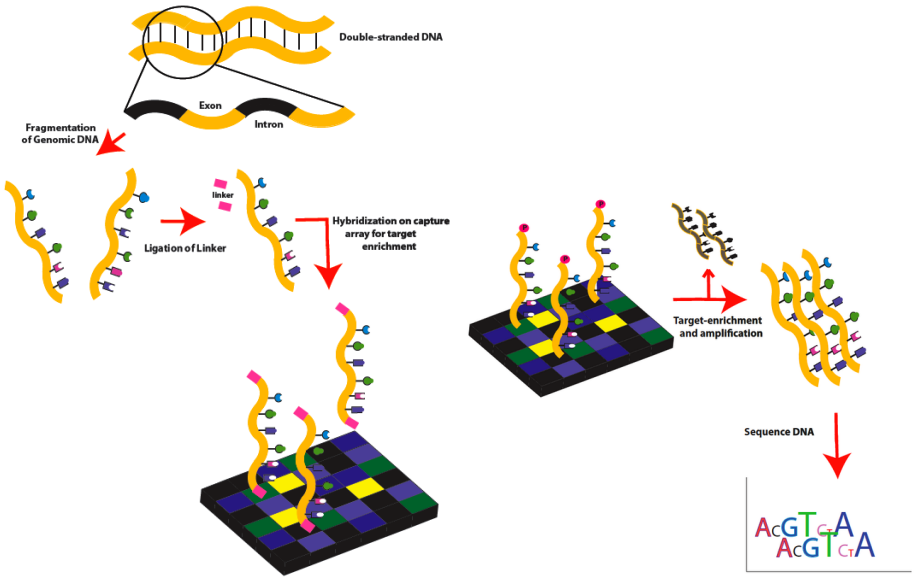

Figure 2.13: Workflow of whole exome sequencing. This figure is adapted from [18].

	WES	WGS
Target	50 Mb	3 Gb
Depth of Coverage	Biased	Poisson
Mapping	Fast	Slow
Analysis time	Short	Long
RD-based CNV discovery tools [1]	12 tools	15 tools
Cost	$900	$5,000
Usage in CNV Discovery	Less common due to biases	Common

Table 2.2: Comparison of WES and WGS

2.4 Genomic Variations

Single nucleotide variants (SNPs), small insertions or deletions (indels), copy number variations (CNVs), and large structural variants are called together as genomic variation. [1]

2.4.1 Single Nucleotide

- **Base change - substitution - point mutation:** Substitution is a type of mutation where one base pair is replaced by a different base pair.
- **Insertion - deletions (indels):** Indel describes relative gain or loss of a segment in a genomic sequence.

2.4.2 2 bp to 1,000 bp

- **Microsatellites:** These sequences are composed of non-coding DNA and they are not parts of genes. These are used as genetic markers to follow the inheritance of genes in families.

- **Minisatellites:** These are generally situated near genes and they are repeated segments of the same sequence of multiple triplet codons. Moreover, minisatellites are useful as linkage markers due to their highly polymorphic nature.
- **Indels:** Indel describes gain or loss of a segment in a genomic sequence.
- **SNP/SNV:** SNPs are a type of polymorphism which involves variation of a single base pair.
- **Variable Number Tandem Repeats - VNTRs:** Linear arrangement of multiple copies of short repeated DNA sequences that vary in length and are highly polymorphic, making them useful as markers in linkage analysis.

2.4.3 1 kb to Submicroscopic

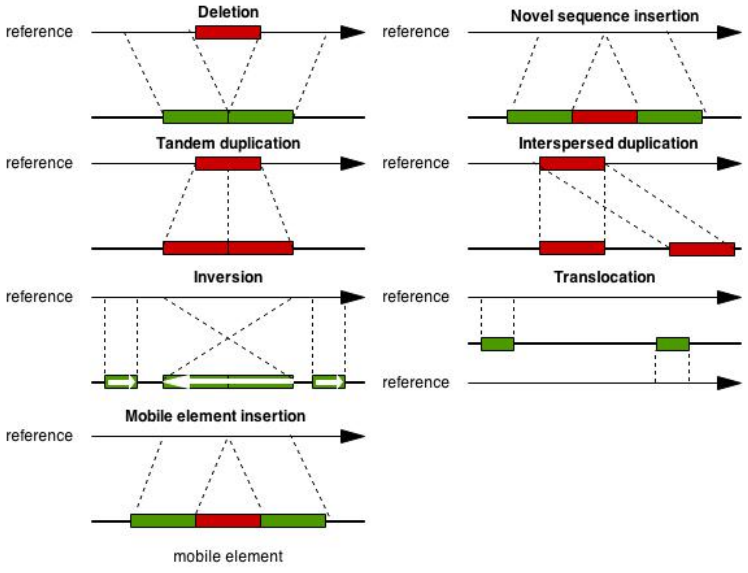

Figure 2.14: Classes of structural variation. This figure is adapted from [8].

- **Copy number variants (CNVs):** Copy number variants comprise of deletions, insertions, and duplications. There is another term that is called as CNP creating confusion. CNP is a CNV that occurs in more than 1%

of the population. CNVs have been observed in the comparison of two or more genomes.

  - **Copy number gain (insertions or duplications):** A sequence alteration whereby the copy number of a given region is greater than the reference sequence.
  - **Copy number loss (deletions):** A sequence alteration whereby the copy number of a given region is less than the reference sequence.

- **Segmental duplications:** A segment of DNA >1 kb in size that occurs in two or more copies per haploid genome, with the different copies sharing >90% sequence identity.
- **Translocation:** A region of nucleotide sequence that has translocated to a new position.
- **Inversion:** A continuous nucleotide sequence is inverted in the same position.
- **CNV regions (CNVRs):** Merging of independently ascertained, but overlapping, genomic segments creates the representation of a CNV locus.

2.4.4 Microscopic to Subchromosomal

- **Segmental aneusomy:** Disorder that results from the inappropriate dosage of crucial genes in a genomic segment.
- **Chromosomal deletions - losses:** In this disorder, entire chromosomes, or large segments of them, are missing.
- **Chromosomal insertions - gains:** In this disorder, entire chromosomes, or large segments of them, are duplicated.
- **Chromosomal inversions:** In this disorder, entire chromosomes, or large segments of them, are altered.

- **Intrachromosomal translocations:** A segment breaks off the chromosome and rejoins it at a different location.
- **Heteromorphisms:** A chromosome pair with some homology but differing in size, shape, or staining properties. Homologous chromosome pair which are not morphologically identical (eg the sex chromosomes).
- **Fragile sites:** A chromosomal region that has a tendency to break.

2.4.5 Whole Chromosomal to Whole Genome

- **Interchromosomal translocation:** Recombination resulting from independent assortment.
- **Isochromosome:** A chromosome with two genetically and morphologically identical arms.
- **Marker chromosomes:** A structurally abnormal chromosome in which no part can be identified.
- **Aneuploidy :** This is the state of having an abnormal number of chromosomes. Down syndrome is an example of aneuploidy.
- **Aneusomy:** is the condition in which an organism is made up of cells that contain different numbers of chromosomes. [19] [2] [20]

2.5 The Effects of Copy Number Variations on Human Health and Phenotype

A human carries two copies of most genes. One copy comes from mother genome and one copy comes from father genome. Occasionally alterations in a chromosome can lead to the gain or a loss of one copy. A deletion can occur when a fragment of DNA is lost. It can occur either during copying or when the genes are shuffles during meiosis. A duplication can occur whereby which we gain an

additional copy of a gene by the same mechanisms. Deletions and duplications of greater than 1,000 nucleotides are called copy number variants.

Figure 2.15: Tiger with down syndrome

A difference in the copy number of a gene can increase or decrease the level of that genes activity so it may cause diseases, phenotypic changes or nothing. For instance, when a copy of a gene is deleted, the cell may produce half as much protein as compared to a normal cell. There are many diseases that are cause by changes in gene copy number.

Next-generation sequencing technology is used to detect copy number variations in both healthy and diseased people. Although copy number variations do not necessarily have a negative effect on human health, large number of CNVs have an association with a disease or directly involve in. The most well-known health problem because of CNV is Down Syndrome, which is caused by having an extra copy of chromosome 21.

Some of the most known health problems due to CNVs are autism, schizophrenia, Turner syndrome, cancer, neuropsychiatric disorders, and obesity. Rare CNVs may account for 15% of cases of pediatric neurodevelopmental diseases Backenroth et al [21]. Although both rare and common CNVs are thought to carry substantial risk for disease, much recent activity has focused on the role played in disease by rare CNVs, given the smaller cohort sizes required to attain statistical significance for identifying highly penetrant risk-associated rare CNVs.

As another example, a recent study found that severe obesity is often associated with a significant burden of large rare CNVs.

To comprehend the underlying causes of these in a less costly way, an efficient sequencing approach is needed. Whole exome sequencing, the most cost-effective way which is known, has the potential to rapidly detect copy number variations that cause these things mentioned above in human coding regions.

WES has been commonly used in not only the detection of pathogenic variants for Mendelian diseases, but also discovery of susceptible loci for complex diseases. By using these kinds of approaches mentioned in this thesis, underlying causes of some diseases and phenotypes and also the treatments of CNV-based diseases can be found. [22] [19]

Chapter 3

Finding Copy Number Variations in Exome Sequencing Data

The steps required to find copy number variations in exome sequencing data consist of mapping, correcting biases and normalization, estimation of copy number and segmentation parts. The main aim of the thesis is correcting some of the biases in exome sequencing data, especially GC-content and probe efficiency.

3.1 Four-Step Procedure

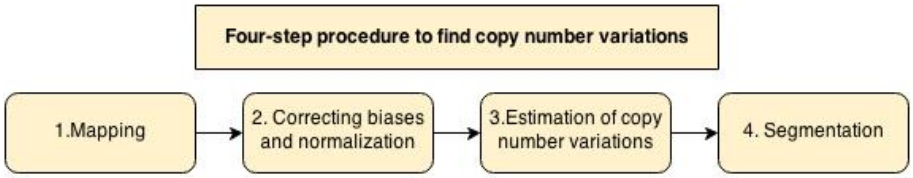

Figure 3.1: Four-step procedure to find CNVs in WES data

3.1.1 Mapping

Although human genome has a continuous structure to read, sequencing machines can only read a few hundred DNA letters at a time. These short DNA letter sets are aligned to the reference genome. This causes some mapping problems when these small reads are used for constructing whole part wanted. Mapping is the first step to find CNVs so the problems in mapping affect the reliability of the results. The choice of single or multiple mapping affects is one of these effects. This choice can change copy number in some regions. It can either cause unexpected increases or decreases in copy numbers.

Mapping are also affected by repeats in mapping part. Approximately 45% of a human genome is repeats such as LINEs, SINEs, retrovirus-like elements, and DNA transposon fosils. RepeatMasker program is available to identify repeats for known ones, but this cannot identify all of these repeats.

3.1.2 Correcting Biases and Normalization

Systematic false positive and false negative results of exome sequencing are identified. Mapping and systematic sequencing errors cause false positives. These false positives are removed by comparing each sample against previously sequenced exomes so using more samples help to improve data. On the other hand, low overall coverage, poor capture efficiency, and difficulty in unambiguously aligning repetitive regions cause false negative results.

Other challenges also exist in this field. Reads have non-uniform distribution and exome is not a continuous search space so this does not allow researchers to use statistical approaches easily due to its sparse structure. This is generally caused by different efficiency of exon capture probes. Another challenge is resolution limited by distance between exons. Some other challenges will be explained below.

Probe capture efficiency

Exome sequencing involves exon-capture step by which the coding regions are selected from the total genome DNA by means of hybridization, either to a microarray or in solution. The characteristics of the probe, such as length, conformation and abundance on the solid phase, are of relevance in determining the capture efficiency. Exon capture techniques are not efficient exactly because 3%-5% of the exons can't be captured and sequenced. [16]

GC-content

GC-content is the percentage of guanine and cytosine bases in a genomic region. It is higher in protein coding regions than intron regions of a genome. GC abundance is heterogeneous across the genome and often correlated with functionality.

GC content bias correction is a necessity for read-depth based copy number detection tools. It was previously shown that there is a positive correlation between read depth and GC-content of a region. Average read depth of a region has a unimodal relationship with its GC-content. Regions that have extremely high or low GC-content might be excluded from the analysis. Bins with high or low GC-content have lower mean read depth than bins with medium GC-content due to the difference in probe efficiency and sequencing.

Most of the sequencing tools are affected by this bias so there are few tools to correct GC-content bias.

Coverage

Multiple copies of a genome are randomly broken into small fragments. Chunks of these fragments are sequenced, generating reads. Reads are merged in regions that they overlap. Coverage is defined as the average number of reads overlapping each base.

The majority of the reads form the final consensus sequence. The higher the coverage of a consensus sequence segment, the more confident you can be in the

accuracy of that segment.

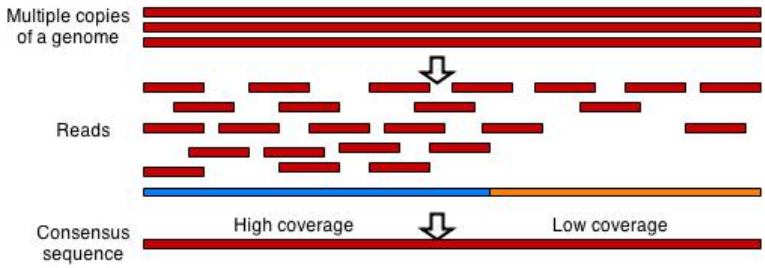

Figure 3.2: High and low coverage

Coverage is calculated by the formula given below:

$$average\ coverage = \frac{number\ of\ reads \times read\ length}{exon}$$

Mapping

Next-generation sequencing generates short reads that are mapped to a reference genome and some of these reads, multi-reads, are not uniquely mapped to the reference. The number of multi-reads depends on read lengths, allowed number of mismatches, and choosing paired-end or single-end sequencing. On the other hand, the presence of repetitive regions has also an effect on multi-mapping.

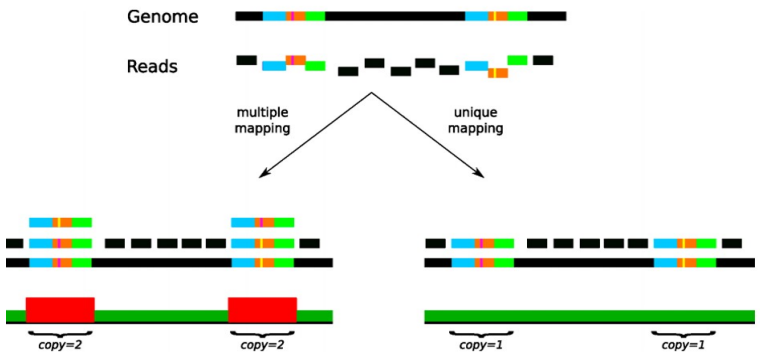

Figure 3.3: Multiple vs. unique mapping

Polymerase Chain Reaction (PCR) Process

The PCR is an *in vitro* method for the enzymatic synthesis of specific DNA sequences, using two oligonucleotide primers that hybridize to opposite strands and flank the region of interest in the target DNA. In short, it represents a form of "in vitro cloning" that can generate, as well as modify, DNA fragments of defined length and sequence in a simple automated and cyclic reaction.

One of the major cause of distortion in WES data is the PCR process. Less reads are created when genomic fragments have lower PCR rates, which is also affected by GC%.

3.1.3 Estimation of Copy Number

In the third step of read depth-based methods, the aim is to estimate accurate copy numbers along the chromosome to determine gain or loss with the normalized read depths. This step changes due to the sample size and selected path.

3.1.4 Segmentation

Segmentation is the process that combines all the reads from same continuous region into a segment with determined boundaries. An ideal segmentation approach will merge adjacent data points with same copy number into one segment and divide regions with different copy numbers into different segments. The copy number state should also be evaluated for each region. The challenging part of the segmentation is to separate random effects from copy number variations. [1]

3.2 Methods

Due to the increasing demand for copy number variation (CNV) detection, a lot of algorithms have been appeared in the field. Choice of using whole genome

sequencing-based (WGS-based) tools or whole exome sequencing-based (WES-based) tools are mostly based on need in a research. However, WES-based tools are getting more popular due to its efficiency in both cost and time.

Here is the table depicting types of variations that can be found by specific tools. Read depth-based (RD-based) method is the main concern of this thesis and duplications and deletions are the variation types that we want to find.

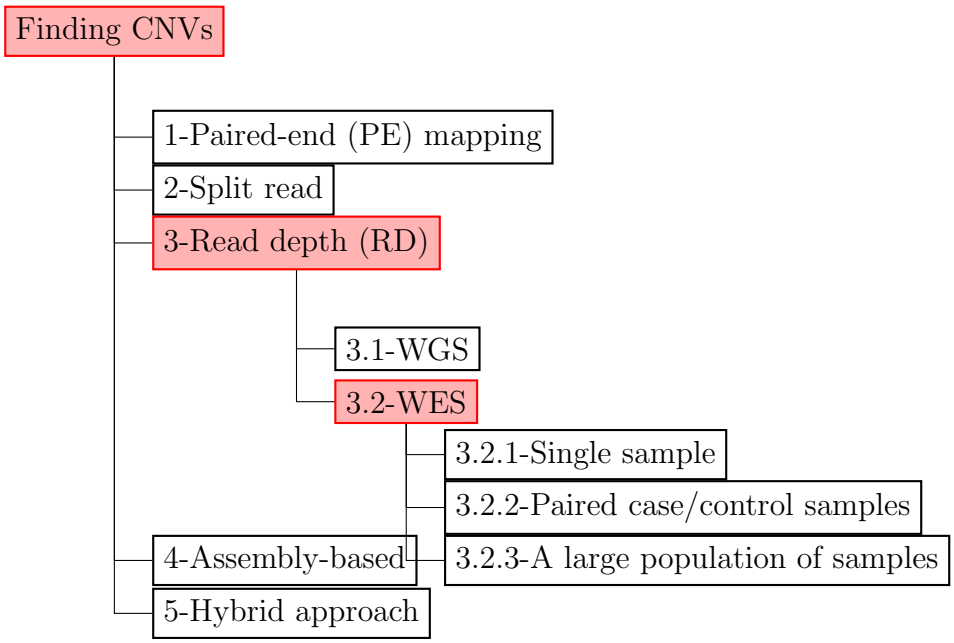

Figure 3.4: Classification of CNV detection methods

Some of the popular tools for CNV detection with using WES data are CONDEX [23], CONIFER [24], CONTRA, Control-FREEC, ExoCNVTest [25], ExomeCNV, ExomeDepth [26], PropSeq [27], SeqGene, VarScan2, and XHMM [28].

SV classes	Read pair	Read depth	Split read
deletion	yes	yes	yes
novel sequence insertion	yes	*no*	yes
mobile element insertion	yes	*no*	yes
inversion	yes	*no*	yes
interspersed duplication	yes	yes	yes
tandem duplication	yes	yes	yes

Table 3.1: Applicability of the tools to the methods

The quantitive relationship between true copy number and depth is distorted by target-specific and sample-specific biases in capturing, PCR amplification, sequencing efficiency, and *in silico* read mapping, GC-content of the targets, target size, sequence complexity, proximity to segmental duplications, single nucleotide polymorphisms(SNPs), DNA concentration, hybridization temperature, experimental sample batching, and various indeterminate factors. Hence, RD-based methods using WGS data are not applicable to WES data if the extra biases are not accounted for. Moreover, the assumption of normal distribution may no longer be valid due to the biases regarding read depth distribution and most of the CNV breakpoints could not be detected due to the discontinuation of genomic regions. Lastly, the widely applied segmentation algorithms to merge windows in WGS may not be applicable due to the non-continuous distributions of the reads in WES data.

3.2.1 Paired-end mapping (PEM)-based methods

Finding CNV using NGS data was first made by PEM methods that can only applied to paired-end reads. PEM-based method is based on identification of CNVs from discordantly mapped paired-reads whose distances are significantly different from the predetermined average insert size.

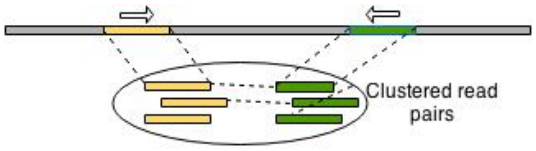

Figure 3.5: Paired-end mapping-based methods

In addition to insertions and deletions, this is also used for identification of mobile element insertions, inversions, and tandem duplications. However, this is not applicable for insertion events which are larger than the average insert size.

There are two different approaches in PEM methods which are the clustering approach in which predefined distance is provided and the model-based approach in which statistics is used to define a distance.

Unlike WGS data, the non-random nature of reads from WES limit the applicability of PEM-based methods for CNV discovery. For instance, the insert size for paired-end reads in WES data may not be long enough to detect CNV. The PE reads should span the CNV breakpoints, but they may not be within exons, thus not captured.

3.2.2 Split read-based methods

Split read-based (SR-based) methods are conceptually used to find insertions and deletions (indels). These methods use read pairs. First read is aligned to the reference genome uniquely while the other read fails to map or maps partially to the reference genome.

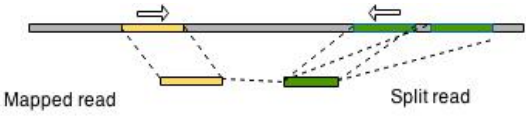

Figure 3.6: Split read-based methods

The breakpoints of structural variations (SVs) are provided due to these unmapped or partially mapped reads. The incompletely mapped reads are splitted into fragments in SR-based methods. First and last parts of these fragments are aligned to the reference genome and the exact start and end positions are found. SR-based methods can only be applied to the unique regions based on read length.

In contrast to whole genome sequencing (WGS) data, whole exome sequencing (WES) results in nonuniform read depth (RD) between the captured regions and systematic biases that affect data strongly between batch of samples. The split reads should also span the CNV breakpoints, but they may not be within exons, thus not captured. These biases and the sparse nature of the capture make WES unsuitable for well-known CNV detection algorithms [29] [30] [24].

3.2.3 Read depth-based methods

This is the most common approach to estimate copy number due to the accumulation of next generation sequencing (NGS) data. The most appropriate method between these 5 categories to find copy number variations (CNVs) with the help of whole exome sequencing (WES) data is the based on read depth (RD).

Depth of coverage, also known as coverage, is calculated by counting the number of reads which cover each base and then calculating their average for each target. RD-based methods are rely on the correlation between the depth of coverage in a genomic region and the copy number of that region. If there is higher count than expected, then there is a duplication in a region. On the contrary, there is a deletion if there is lower count than expected in a region. Exact copy numbers, CNV in complex genomic region classes, and large insertions can be found with the usage of RD-based methods.

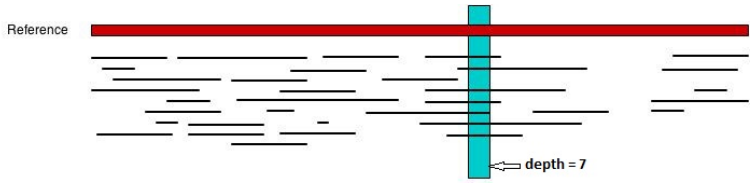

Figure 3.7: Calculation depth of coverage

Two major next-generation sequencing approaches, WES and WGS, are used to detect CNVs. If WGS-based tools are used theoretically, the full spectrum of variant can be detected. However, WES-based tools are more effective in terms of time and cost efficiency.

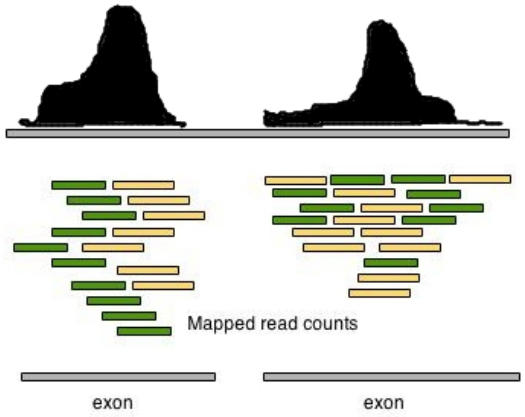

Figure 3.8: Read depth-based method

RD-based methods are classified into 3 categories in terms of the sample size: single sample, paired case/control samples, and a large population of samples. In single sampled-studies the aim is to estimate the read depth distribution using mathematical models and detect the regions with abnormal depth different from the overall distribution. In the second type of method, paired case/control samples, control sample are thought as a reference genome and copy numbers in case sample are reported as relative copies compared to the control sample. These copy numbers are not exact copy numbers. For large population of samples cases, the overall mean of the read depth from multiple samples are used to detect the discordant copy numbers in each sample. This method generally reflects the

exact copy numbers.

CNV discovery from WES data is challenging because of the non-contiguous nature of captured exons (All exons cannot be detected by current technologies). This challenge is compounded by the complex relationship between read depth and copy number which is affected by biases in targeted genomic hybridization, sequence factors such as GC-content, and batching of samples during collection and sequencing. Approaches based on Gaussian and other popular distributions used in CNV detection with using WGS data are not working on CNV detection tools with using WES data due to biases and indeterminate factors.

3.2.4 Assembly-based methods

Firstly, contigs are constructed by using DNA fragments and compared them to the reference genome as a guide. The genomic regions with discordant copy numbers are determined in this way.

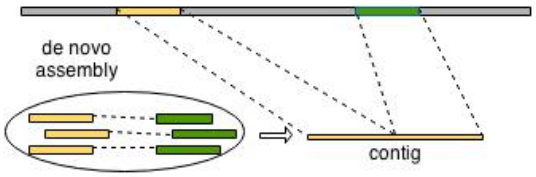

Figure 3.9: Assembly-based methods

These kinds of methods propose an unbiased approach to detect novel variants from single base to large structural variations, but these are generally used for the other small-sized genomes due to their huge demand on computational resources.

In AS-based methods, reads should also span the CNV breakpoints, but they may not be within exons, thus not captured. These biases mentioned above and the sparse nature of the capture make WES unsuitable for AS-based methods.

3.2.5 Hybrid approaches

Two or more approaches mentioned above are used together in this approach. Though there has been a great progress in each of these approaches mentioned above, none of them is able to detect all variants in a genome precisely.

Each of the methods above has its own advantages and disadvantages so taking advantages creates need to combine some of these methods to increase the performance in detecting variants and reduce false positive rates. [1]

Chapter 4

Related Works

4.1 Whole Genome Sequencing

4.1.1 Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing

GC content bias is explained as the dependence between read coverage and GC content found on Illumina sequencing data. This bias is not consistent between samples and there is no consensus to find the best method to remove this bias in a single sample. In this paper, a model that produces predictions at base pair level, allow strand-specific GC-effect correction regardless of the downstream smoothing or binning is presented.

Most current correction methods follow a common path. Fragment counts and GC counts are binned to a chosen bin-size. A curve that describes the conditional mean fragment count per GC value is estimated. This curve determines a predicted count for each bin based on the bin's GC. These obtained predictions can be used to normalize the original signal directly. These methods do not use any prior knowledge about the effect whereas they remove most of the GC effect.

A descriptive approach to investigate the common structures that is found in

GC curves of DNA sequencing data is described in this work. The effect of GC on fragment count for many high coverage human genomes which are from different labs is studied. Copy numbers for normal genomes change rarely so observed variability in fragment count may always be attributed to technical effects rather than biological effects. A single position model is used to estimate the effect of GC on the fragment counts.

LOESS method is used by them. First, read depth and GC content are calculated for each bin. The GC bias curve is determined by loess regression of count by GC on a random sample of 10000 high mappability (> 0.9) bins. The smoothness parameter, also known as span, for the *LOESS* should be tuned to produce curves that are smoothing but still capture the main trend in data. This parameter is estimated as 0.3 in this work.

Unlike the other bias correction methods, such as BEADS, predicted fragment rates for the genomic location rather than for the observed reads are generated. [31]

Tools	Description
SeqSeq	Detecting CNV breakpoints using massively parallel sequence data
CNV-seq	Identifying CNVs using the difference of observed copy number ratios
RDXplorer	Detecting CNVs through event-wise testing algorithm on normalized read depth of coverage
BIC-seq	Using the Bayesian information criterion to detect CNVs based on uniquely mapped reads
CNAseg	Using flowcell-to-flowcell variability in cancer and control samples to reduce false positives
cn.MOPS	Modeling of read depths across samples at each genomic position using mixture Poisson model
JointSLM	Population-based approach to detect common CNVs using read depth data
ReadDepth	Using breakpoints to increase the resolution of CNV detection from low-coverage reads
rSW-seq	Identifying CNVs by comparing matched tumor and control sample
CNVnator	Using mean-shift approach and performing multiple-bandwidth partitioning and GC correction
CNVnorm	Identifying contamination level with normal cells
CMDS	Discovering CNVs from multiple samples
mrCaNaVar	A tool to detect large segmental duplications and insertions
CNVeM	Predicting CNV breakpoints in base-pair resolution
cnvHMM	Using HMM to detect CNV

Table 4.1: Summary of bioinformatics tools for CNV detection using WGS data. This table is adapted from [1].

4.2 Whole Exome Sequencing

4.2.1 Copy Number Variation Detection and Genotyping from Exome Sequencing Data

Copy number inference from exome reads (CoNIFER) is a novel method that uses singular value decomposition (SVD) normalization to discover rare genic copy number variants (CNVs) as well as genotype copy number polymorphic (CNP) loci with high sensitivity and specificity from whole exome sequencing (WES) data. It can be used to discover disruptive genic CNVs that are missed by standard approaches reliably. In this study both read depth (RD) data from WES data with SVD methods to discover rare CNVs and genotype known CNP regions from HapMap samples are combined.

The workflow of CoNIFER starts with dividing fastq-formatted WES reads into non-overlapping 36 base paired-sets and aligning them to the targeted regions. These reads are aligned by allowing up to two mismatches per each read set. Reads per thousand bases per million reads sequenced (RPKM) values are calculated and then these are transformed into ZRPKM values, RPKM values transformed into standardized z-scores, based on the median and standard deviation of each exon across all samples. ZRPKM values are used as input for the SVD transformation. The strongest k singular values that depends on data are removed. SVD normalization is used to overcome coverage biases introduced by the capture and sequencing of exomes. [24]

4.2.2 Discovery and Statistical Genotyping of Copy Number Variation from Whole Exome Sequencing Depth

Exome hidden Markov model (XHMM) is a statistical tool that uses principal component analysis (PCA) to normalize exome read depth and uses hidden Markov model (HMM) to discover exon-resolution CNV and genotype variation across samples.

The workflow of XHMM starts with aligned exome read BAM-formatted files. Firstly, depth of coverage is calculated. PCA is run on the sample-by-target-depth matrix by rotating the high dimensional data to find the main components in which depth varies across multiple samples and targets and some of the largest effects that depends on data are removed. Normalized data is trained and HMM is run to discover CNVs spanning adjacent targets. At the end of the process, CNV calls and genotype qualities for all samples is outputted.

The main limitation of XHMM tool is the requirement of large number of samples because of the PCA normalization step. The efficiency of PCA depends on data size. [28]

Tools	Description
Control-FREEC	Correcting copy number using matched case-control samples or GC contents
CoNIFER	Using SVD to normalize copy number and avoiding batch bias by integrating multiple samples
XHMM	Using PCA to normalize copy number and HMM to detect CNVs
ExomeCNV	Using read depth and B-allele frequencies from exome sequencing data to detect CNVs and LOHs
CONTRA	Comparing base-level log-ratios calculated from read depth between case and control samples
CONDEX	Using HMM to identify CNVs
SeqGene	Calling variants, including CNVs, from exome sequencing data
PropSeq	Using the read depth of the case sample as a linear function of that of control sample to detect CNVs
VarScan2	Using pairwise comparisons of the normalized read depth at each position to estimate CNV
ExoCNVTest	Identifying and genotyping common CNVs associated with complex disease
ExomeDepth	Using beta-binomial model to fit read depth of WES data

Table 4.2: Summary of bioinformatics tools for CNV detection using WES data. This table is adapted from [1].

Chapter 5

Description of the Experiments

5.1 Data

In this thesis, the correlation between read depth, GC content, and probe efficiency is systematically evaluated using 1000 genomes data. We tested our methods on 7 samples from different populations around the world. Samples, HG00629, HG01191, HG01437, NA19664, NA19707, NA19723, and NA20766, are chosen randomly between the other samples of 1000 genomes project. These samples data were created by using the Illumina HiSeq2000 sequencing technology. Whole genome sequencing data is mapped to the exon regions in human genome (version hg19) so the data are used as whole exome sequencing data with this way. The Agilent Sure Select Capture Kit annotation was used to capture the exomes in the data we analyzed. There are also different whole exome capturing tools like Illumina TrueSeq Capture Kit. Although other capture kit annotations haven't been tested by using our method, our method can also be used for them theoretically.

The gene positions are token from UCSC Genome Browser refFlat (hg19) data. Probe efficiency, average of GC-content, and read depth are calculated for each gene by using these positions and exon information included by these genes. Known common deletions, duplications, and low-coverage regions of 1000

Genomes data found in this work are used to remove data noise as much as possible.

We only analyzed the autosomal chromosomes (i.e. no sex chromosomes).

5.2 Mapping

5.2.1 Mapping of Reads to the Reference: MrsFAST-Ultra

Although increasing read lengths, the mapping step remains as an important problem. The accuracy of found structural variants are partially related to this step. Tools that report the best mapping location for each read are not appropriate for structural variation detection where it is important to report multiple mapping loci for each read. MrsFast fills this gap as a fast, cache oblivious, and SNP-aware aligner that can handle the multi-mapping of next-generation sequencing reads efficiently.

MrsFAST is a mapping algorithm that rapidly finds all mapping locations of a collection of short reads from a donor genome in the reference genome within a user-specified number of mismatches. It is specifically designed for reads generated by Illumina sequencing machines.

Two main steps are included in the tool. The first step consists of building an index from the reference genome for exact anchor matching. The second step consists of computing all anchor matching for each of the reads in the reference genome through the index, extends each match to both left and right and checks if the overall alignment is within the user defined error threshold.

The mapping task is simply partitioned into independent threads executed by a single-core which is defined by user. Multi-thread option is used as using 8 threads in this work. Moreover, paired-end mode is preferred because of the reads used in this work. Owing to the repetitive content of human genome sequence,

the most comprehensive assemblies are derived from paired-end reads, where the sequence reads are obtained from both ends of each DNA fragment. Disable-nohits option is selected to speed up the process to prevent the accumulation of useless data. As a reference the hg19 version of human genome is selected. [32]

5.2.2 Calculation of read depth: Bedtools

Bedtools is introduced for comparison, manipulation, and annotation of genomic features in different file formats. Bedtools is used frequently in different steps of this work. Firstly, the output of MrsFast is sam-formatted and the input required to use count option of bedtools is bed-formatted. Therefore, sam-formatted files are converted into the bed-formatted files.

Common low-coverage regions, deletions, and duplications of the 1000 genome samples are removed and the remaining parts are assumed as normal. The exon coordinates of the Agilent Sure Select Capture Kit annotation is used as the reference to find read-depths of each region. Read counts for each exon is calculated to find read depths. This process is done by using count option of bedtools intersect.

5.2.2.1 Common Data File Formats

Sequence Alignment/Map Format (SAM): SAM format is used to store read alignments against reference sequences. It supports short and long reads which are produced by different sequencing platforms including Illumina. It also supports single-end and paired-end reads and combining reads of different types, including color space reads from ABI/SOLiD.

Binary Alignment/Map Format (BAM): BAM format is the binary representation of SAM format and keeps exactly the same information as SAM. It is designed to improve the performance. It is the compact version of SAM format. It is used in a common way.

Browser Extensible Data Format (BED): BED format is an exact and flexible way to represent genomic features and annotations. With the help of the bedtools and some other similar tools, this format ease bioinformaticians' work.

Variant Call Format (VCF): VCF is a generic format which is used to store most prevalent forms of DNA polymorphisms including SNPs, insertions, deletions, and large structural variants together with rich annotations. This format is provided to use in 1000 genomes data processes. VCF format is also used to comprehend 1000 genomes data in this work. [33] [34] [35]

5.3 Correcting Biases and Normalization

Systematic errors are platform-dependent. In the context of this work, we focus on Illumina data. Current studies about Illumina data evaluation have revealed several biases, a non-random distribution of reads in the sequenced sample over the reference and a non-random distribution of errors. Although there are some popular works to correct GC-content bias, there is no work to correct probe efficiency bias of exome sequencing data.

The aim of this work is to solve the probe efficiency bias in exome sequencing data and make the popular finding copy number variation tools work with correct results. Understanding of the relationship between GC-content, probe efficiency biases, and read depth is required to accomplish this aim. In this part, the effect of both GC-content and probe efficiency will be evaluated together.

Correlation

Correlation coefficient is used to understand the relationships between different data sets. These data sets can be composed of two or more. In this part, correlation between two data sets will be discussed.

Simple correlation coefficient is represented by r. Positive correlation coefficient means that both values increase together and negative correlation coefficient

means that one of these values increases while the other value decreases. Additionally, there is no relationship between them if correlation coefficient equals zero. [36]

In this work, correlation coefficient is used to understand the relationship between read depth and GC-content and the relationship between read depth and probe efficiency. Correlation coefficient is calculated by the formula below:

$$r_{rd,pr} = \frac{\sum_{i=1}^{n}(rd_i - \overline{rd})(pr_i - \overline{pr})}{\sqrt{\sum_{i=1}^{n}(rd_i - \overline{rd})^2 \sum_{i=1}^{n}(pr_i - \overline{pr})^2}}$$

$$r_{rd,gc} = \frac{\sum_{i=1}^{n}(rd_i - \overline{rd})(gc_i - \overline{gc})}{\sqrt{\sum_{i=1}^{n}(rd_i - \overline{rd})^2 \sum_{i=1}^{n}(gc_i - \overline{gc})^2}}$$

Multiple Correlation

Multiple Correlation measures the amount of linear association between one dependent variable and more than one independent variables. It is an extension of simple correlation (frequently just called as correlation). It helps to determine whether if more than one independent variable should be included in the model. Multiple Correlation Coefficient is represented by R and calculated by the formula below:

$$R_{y,x_1,x_2} = \frac{\sqrt{{r_{yx_1}}^2 + {r_{yx_2}}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}}{\sqrt{1 - r_{x_1x_2}^2}}$$

where y is the dependent variable, x_1 and x_2 are the independent variables, r_{yx_1} is the correlation coefficient between y and x_1, r_{yx_2} is the correlation coefficient between y and x_2, and $r_{x_1x_2}$ correlation coefficient between x_1 and x_2.

In this work, independent variables are GC-content and probe efficiency while dependent variable is read depth. Adapted version of multiple correlation formula to our work is demonstrated below:

$$R_{rd,pr,gc} = \frac{\sqrt{{r_{rd,gc}}^2 + {r_{rd,pr}}^2 - 2r_{rd,gc}r_{rd,pr}r_{gc,pr}}}{\sqrt{1 - r_{gc,pr}^2}}$$

where rd is the dependent variable, pr and gc are the independent variables, $r_{rd,gc}$ is the correlation coefficient between read depth and GC content, $r_{rd,pr}$ is the correlation coefficient between read depth and probe efficiency, and $r_{gc,pr}$ is the correlation coefficient between GC content and probe efficiency.

Smoothing data (LOESS)

LOESS method uses locally weighted linear regression to smooth data. It is a regression model, no parameter needs to be estimated except the smoothness. The smoothness parameter, also known as span, for the *LOESS* should be tuned to produce curves that are smooth but still capture the main trend in data.

A span represents the percentage of the data in which we are interested. This span value determines the smoothness of the curve and is chosen between 0 and 1. Each value is smoothened by using neighboring data within the span so this method is local.

The size of span has an important effect on the curve. If a span is too small, it produces a curve characterized by a lot of noise. If a span is too large, the regression will be over-smoothened and thus the local polynomial may not fit the data well. This will result in high variance. The weight function is also defined for the data within this span. This will result in loss of information and the fit will have large bias. Therefore, the tradeoff between bias and variance should be well-evaluated.

The degree of polynomial has also effects on the curve. A higher degree provides a better approximation and less bias. On the other hand, it requires more coefficients to estimate. The best strategy is to choose lower degree of polynomial and concentrate on choosing best bandwidth.

The weight function to be chosen has less effect than the other things mentioned above. The most chosen weight function is the tricube weight function.

LOESS procedure is similar to what is commonly used in local regression. It is assumed that the data are generated by the function below:

$$y_i = g(x_i) + \epsilon_i$$

where g is a smoothing function of the independent variables. The aim is to find smoothing function.

Constructing LOESS curve procedure:

1. Let x_i denote a set of n values for a predictor variable and let y_i represent the corresponding response.

2. Choose a span value (α) between 0 and 1. Let k be the greatest integer less than or equal to $\alpha \times n$.

3. Find the k points for each x_0 in the data set that are closest to x_0. These x_i comprise a neighborhood of x_0, and this set is denoted by $N(x_0)$.

4. Compute the distance of the x_i in $N(x_0)$ that is the furthest away from x_0 using

$$\Delta_k(x_0) = max_{x_i \in N_0} |x_0 - x_i|$$

5. Assign a weight function to each point (x_i, y_i), $x_i \in N_0$, using the tri-cube weight function:

$$w(x) = \begin{cases} (1 - |x|^3)^3 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases}$$

6. Obtain the value y_0 of the curve at the point x_0 for the given α using a weight least squares fit of the points x_i in the neighborhood $N(x_0)$.

7. Repeat steps 3 through 6 for all x_0 of interest. [37]

After smoothing the data, correlation improves. However, calculating correlation after smoothing gives false positive results in this way. If smoothing is done first, the burden of carrying through the uncertainty of that smoothing to

the estimated correlations will appear, which will be far less certain than when computed for unsmoothed data. Smoothing induces spurious correlations.

5.3.1 Calculation of Correlation Coefficients for Each Exon Region

To understand the relationship between read depth, GC-content and probe efficiency correlation coefficients for each exon are calculated.

Samples	Rd and Pr	Rd, Pr, and GC
HG00629	0.6478	0.7118
HG01191	0.6375	0.7095
HG01437	0.6483	0.6708
NA19664	0.6383	0.6640
NA19707	0.6484	0.6733
NA19723	0.6508	0.6751
NA20766	0.6509	0.6768

Table 5.1: Correlation between read depth, probe efficiency and GC content for each exon

In a recent study of the Illumina HiSeq and Genome Analyzer systems, [36] a positive correlation between read depth and GC-content was observed when GC percentage is within the spectrum of 24% to 47%.

Samples	Rd and Pr	Rd, Pr, and GC
HG00629	0.79404	0.7959
HG01191	0.78969	0.7899
HG01437	0.76086	0.7644
NA19664	0.75367	0.7568
NA19707	0.76771	0.7713
NA19723	0.76126	0.7650
NA20766	0.77169	0.7752

Table 5.2: Correlation between read depth, probe efficiency and GC content for each exon (0.24<GC Content<0.47)

Starting from this point of view, the data were also evaluated within this interval and the increase in the positive correlation coefficient was observed.

It is expected that the values of probe efficiency and read depth are directly proportional and almost on a linear fit line and the graphics almost meets the expectation. However, there are lots of outliers seen on the graphs.

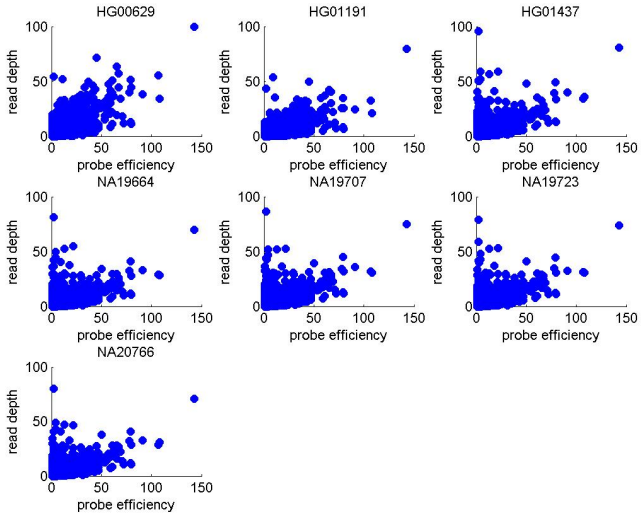

Figure 5.1: Read depth and probe efficiency for each exon

Data points are not on a linear line in the graphics due to the other biases, such as mapping errors, and the possibility of being copy number variations.

First of all, we evaluate the effect of the GC-content. We then re-calculate exons of which GC-content is within the spectrum of 24 to 47%.

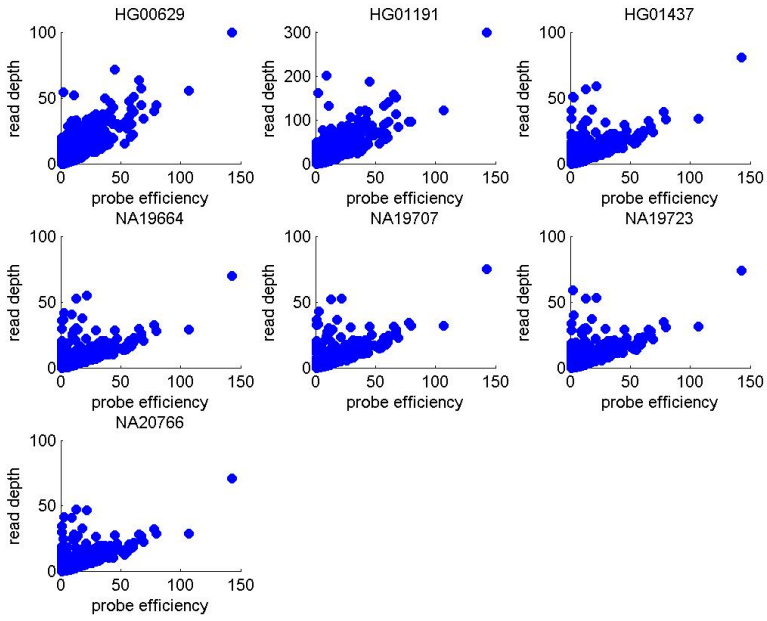

Figure 5.2: Read depth and probe efficiency for each exon (0.24<GC Content<0.47)

As it is expected statistically, read depth and GC-content data points have graphics like below.

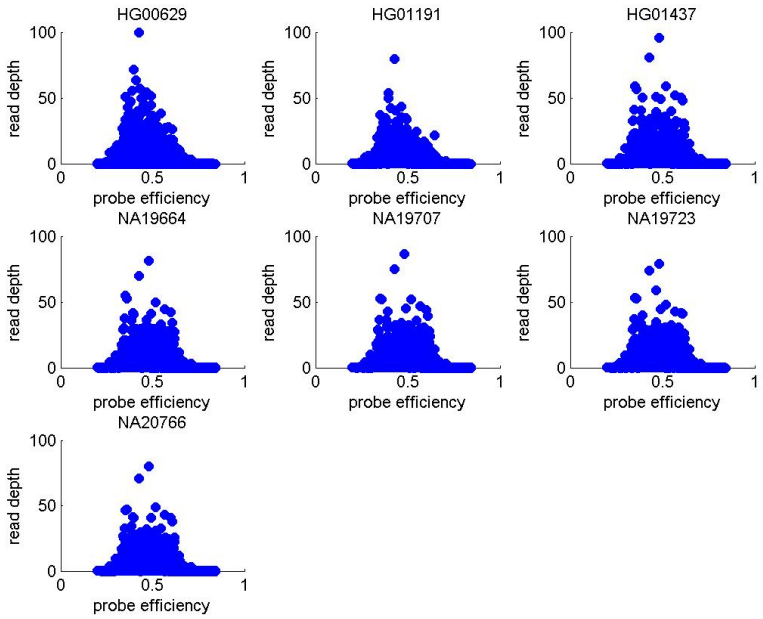

Figure 5.3: Read depth and GC content for each exon

5.3.2 Calculation of Correlation Coefficients for Each Gene Region

The possibility of reducing noises in data by using genes regions of genome is evaluated. To understand the relationship between read depth, GC-content and probe efficiency correlation coefficients for each genes are also calculated. All exons in each gene are expected to have almost the same characteristics.

Starting from this expectation, the usage of genes' and exons' information are evaluated separately. As a result, using gene information increases the correlation coefficient positively.

Samples	Rd and Pr	Rd,Pr, and GC
HG00629	0.9011	0.9134
HG01191	0.8885	0.9037
HG01437	0.9170	0.9274
NA19664	0.9140	0.9248
NA19707	0.9146	0.9256
NA19723	0.9151	0.9262
NA20766	0.9163	0.9273

Table 5.3: Correlation between read depth, probe efficiency and GC content for each gene

Using the idea of the recent study of the Illumina HiSeq and Genome Analyzer systems mentioned above, [36] the correlation between read depth, probe efficiency, and GC-content were observed when GC percentage is within the spectrum of 24% to 47%. It was concluded with the same thing that the increase in the positive correlation coefficient is observed.

Samples	Rd and Pr	Rd,Pr, and GC
HG00629	0.9809	0.9814
HG01191	0.9787	0.9804
HG01437	0.9698	0.9732
NA19664	0.9709	0.9738
NA19707	0.9704	0.9737
NA19723	0.9697	0.9732
NA20766	0.9715	0.9748

Table 5.4: Correlation between read depth, probe efficiency and GC content for the genes (0.24<GC Content<0.47)

The evaluation of each exon separately affects the relationship of GC-content and read depth. Due to the increase in correlation coefficients when using gene information, it is expected that most of the data should be on a linear line.

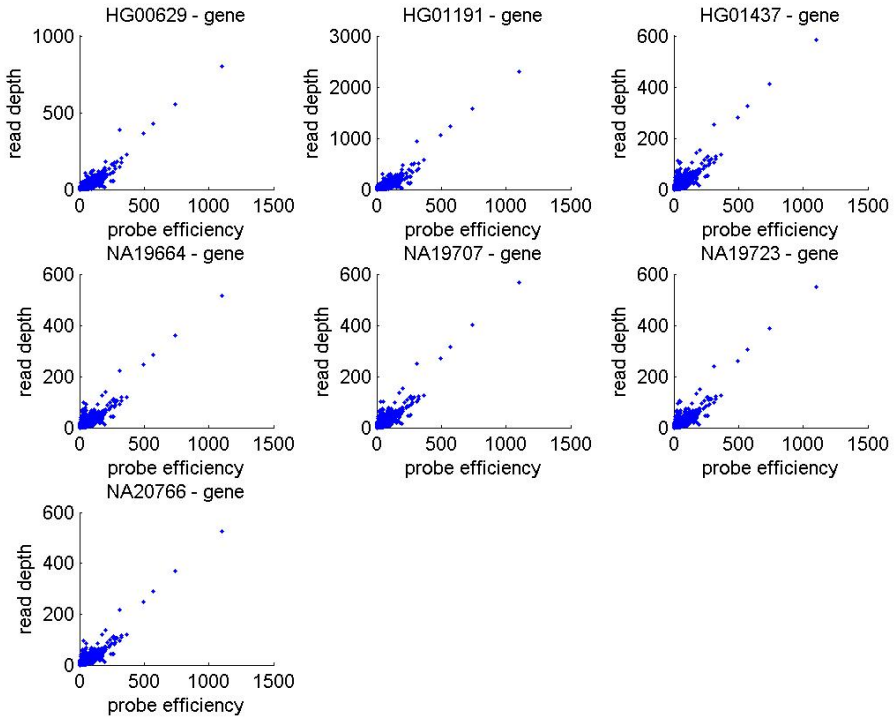

Figure 5.4: Read depth and probe efficiency for each gene

Using the same idea of the study mentioned above, [36] the correlation between read depth, probe efficiency, and GC-content were observed when GC percentage is within the spectrum of 24 to 47%. It was concluded with the same thing that the increase in the positive correlation coefficient is observed for genes' information.

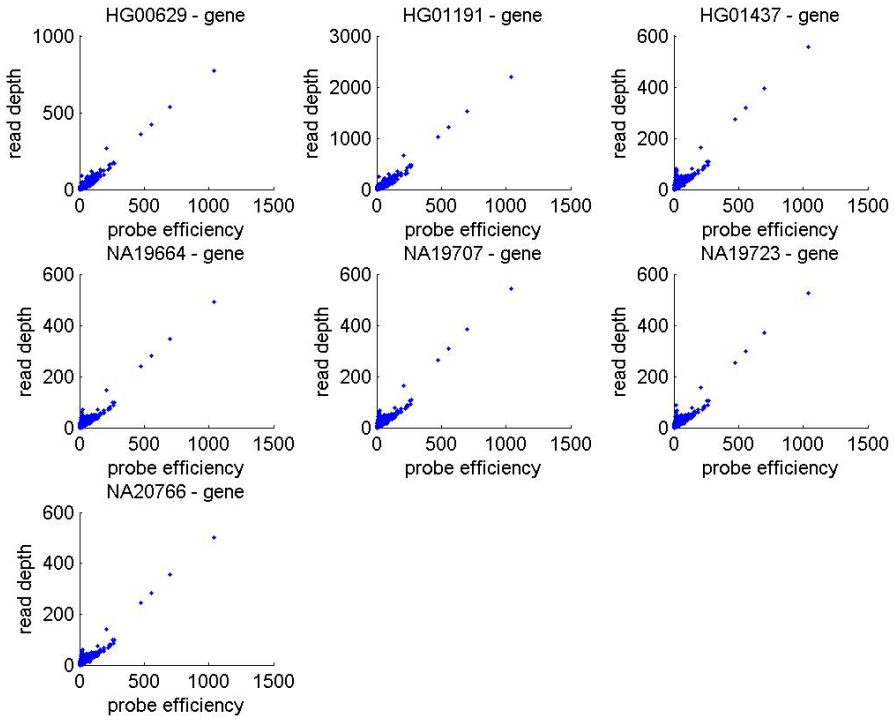

Figure 5.5: Read depth and probe efficiency for each gene (0.24<GC Content<0.47)

After these corrections, there are also some outliers on the graphs. Most of these outliers probably belong to copy number variations and some other biases.

To see the big picture, all processes for each sample are showed in the figures below. As seen, all of them are almost in the same manners for each process. Some differences may be seen because of the process in the sequencing part, different diseases and phenotypes that they have, and some other things like these.

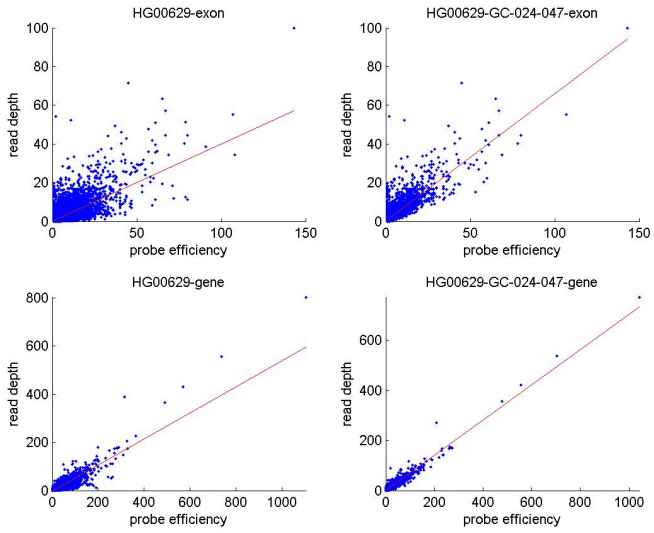

Figure 5.6: **a.** Read depth and probe efficiency of HG00629 for each exon **b.**Read depth and probe efficiency of HG00629 for each exon ($0.24 <$ GC-Content < 0.47) **c.** Read depth and probe efficiency of HG00629 for each gene **d.** Read depth and probe efficiency of HG00629 for each gene ($0.24 <$ GC-Content < 0.47)

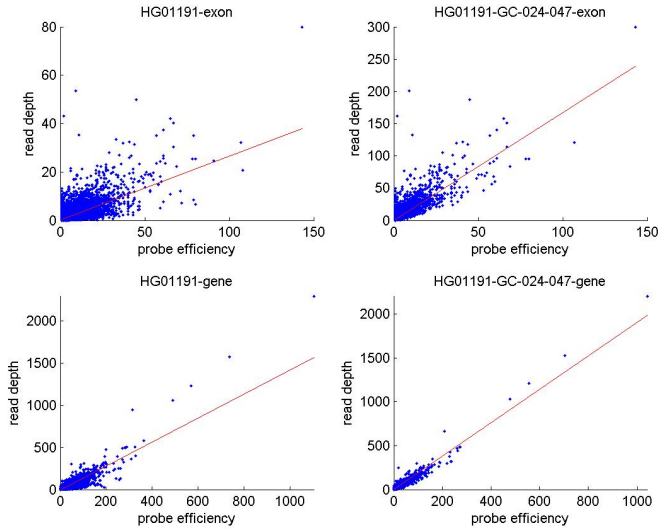

Figure 5.7: **a.** Read depth and probe efficiency of HG01191 for each exon **b.**Read depth and probe efficiency of HG01191 for each exon ($0.24 <$ GC-Content < 0.47) **c.** Read depth and probe efficiency of HG01191 for each gene **d.** Read depth and probe efficiency of HG01191 for each gene ($0.24 <$ GC-Content < 0.47)

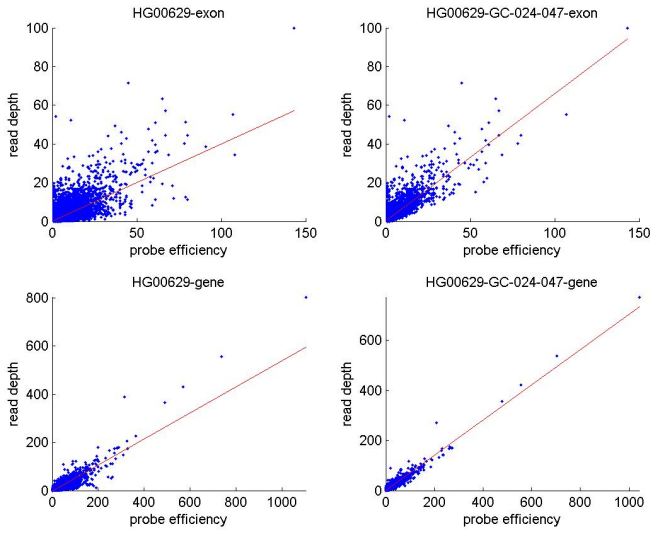

Figure 5.8: **a.** Read depth and probe efficiency of HG01437 for each exon **b.**Read depth and probe efficiency of HG01437 for each exon ($0.24 <$ GC-Content < 0.47) **c.** Read depth and probe efficiency of HG01437 for each gene **d.** Read depth and probe efficiency of HG01437 for each gene ($0.24 <$ GC-Content < 0.47)

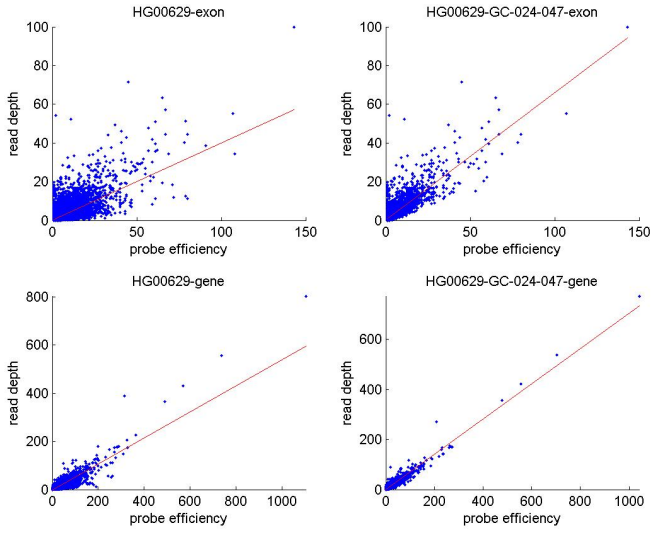

Figure 5.9: **a.** Read depth and probe efficiency of NA19664 for each exon **b.** Read depth and probe efficiency of NA19664 for each exon ($0.24 <$ GC-Content < 0.47) **c.** Read depth and probe efficiency of NA19664 for each gene **d.** Read depth and probe efficiency of NA19664 for each gene ($0.24 <$ GC-Content < 0.47)

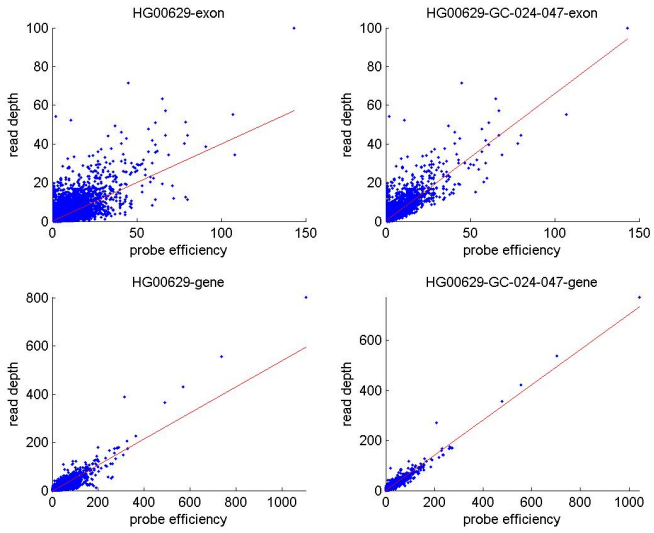

Figure 5.10: **a.** Read depth and probe efficiency of NA19707 for each exon **b.**Read depth and probe efficiency of NA19707 for each exon ($0.24 <$ GC-Content < 0.47) **c.** Read depth and probe efficiency of NA19707 for each gene **d.** Read depth and probe efficiency of NA19707 for each gene ($0.24 <$ GC-Content < 0.47)

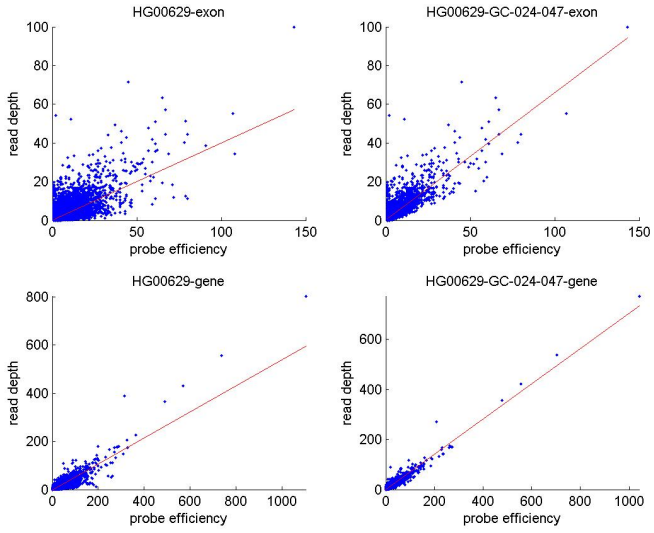

Figure 5.11: **a.** Read depth and probe efficiency of NA19723 for each exon **b.**Read depth and probe efficiency of NA19723 for each exon ($0.24 <$ GC-Content < 0.47) **c.** Read depth and probe efficiency of NA19723 for each gene **d.** Read depth and probe efficiency of NA19723 for each gene ($0.24 <$ GC-Content < 0.47)

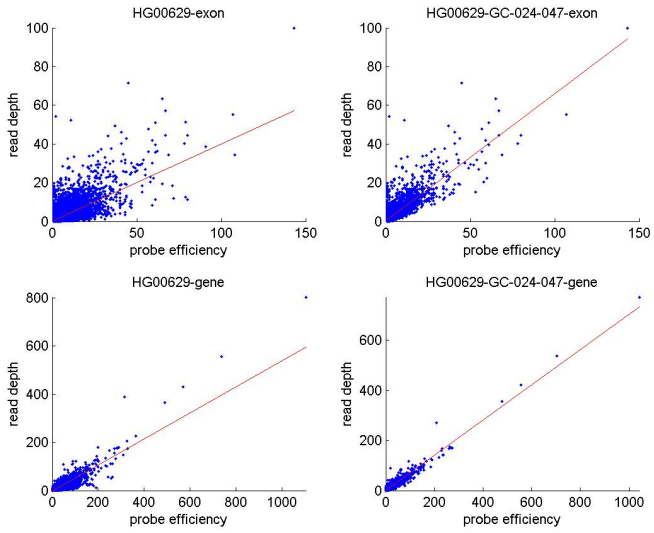

Figure 5.12: **a.** Read depth and probe efficiency of NA20766 for each exon **b.**Read depth and probe efficiency of NA20766 for each exon ($0.24 <$ GC-Content < 0.47) **c.** Read depth and probe efficiency of NA20766 for each gene **d.** Read depth and probe efficiency of NA20766 for each gene ($0.24 <$ GC-Content < 0.47)

5.3.3 Finding optimum span parameter of LOESS method

Span parameter is represented by α and it defines the interval to be contained in the calculation for each data point. Choice of span value depends on data size and distribution of data. α value is chosen between 0 and 1. For example, 10% of the data is used when $\alpha = 0.1$. Starting from this point, the smoothened data will be overfitted when α goes to 1. In contrast, the smoothed data will be underfitted when α goes to 0.

Finding the optimal value is the most difficult part of using the *LOESS* method. For the case of this work, data for all genes don't require a great changes because most of the data is on the fit line. Therefore, the span value α was chosen as 0.05 after trying different span values.

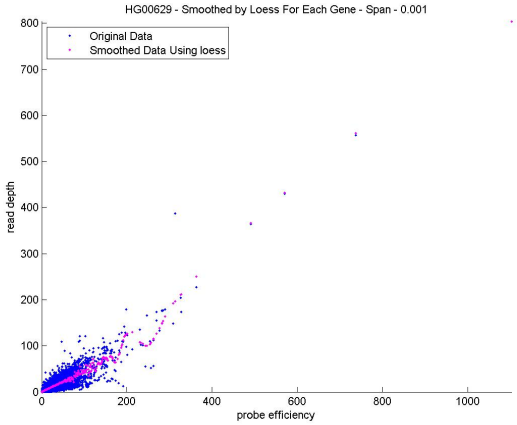

Figure 5.13: Smoothed read depth and probe efficiency by LOESS method for each gene (HG00629 (Span=0.001))

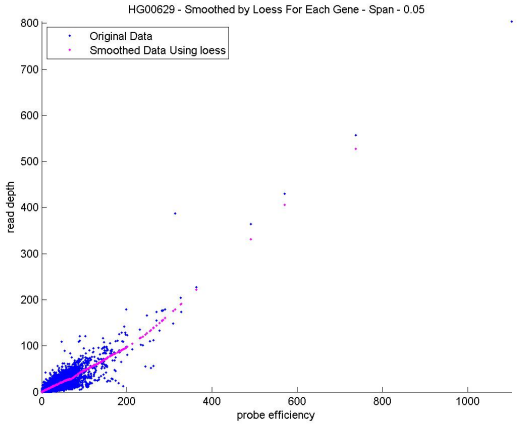

Figure 5.14: Smoothed read depth and probe efficiency by LOESS method for each gene (HG00629 (Span=0.05))

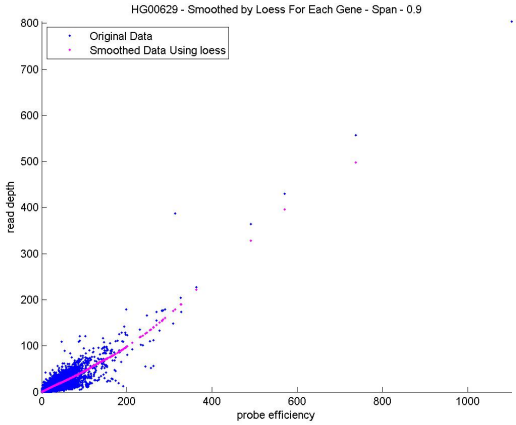

Figure 5.15: Smoothed read depth and probe efficiency by LOESS method for each gene (HG00629 (Span=0.9))

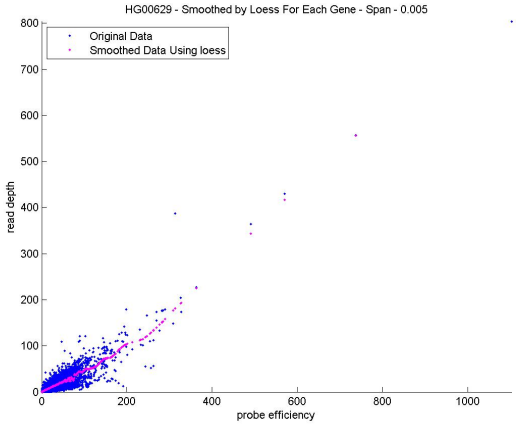

Figure 5.16: Smoothed read depth and probe efficiency by LOESS method for each gene (HG00629 (Span=0.005))

There is another *LOESS* method that is known as *Robust LOESS*. This method is also tried, but it doesn't give results better than *LOESS* method. This method is more appropriate when there are many outliers.

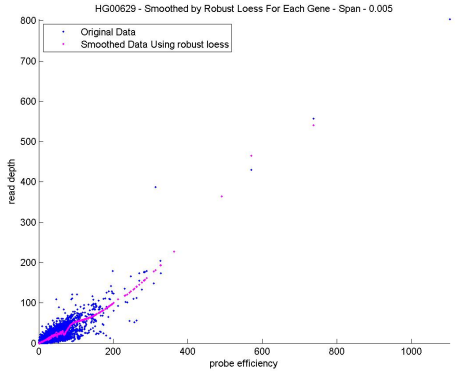

Figure 5.17: Smoothed read depth and probe efficiency by Robust LOESS method for each gene (HG00629 (Span=0.005))

Correlation coefficients are calculated for smoothed samples' data and coefficient values are increased so there is a strong relationship between read depth and probe efficiency like read depth, probe efficiency, and GC content. However, increases of multiple correlation coefficients for all genes and genes having specified GC-percentage between 24% and 47% are little or nothing. Therefore, they don't require special interest. That's why we just correct read depth and probe efficiency with using *LOESS* method for genes at the end.

Samples	Rd and Pr	Rd, Pr, and GC
HG00629	0.9806	0.9865
HG01191	0.9721	0.9802
HG01437	0.9756	0.9779
NA19664	0.9754	0.9782
NA19707	0.9741	0.9768
NA19723	0.9746	0.9773
NA20766	0.9758	0.9784

Table 5.5: Correlation between read depth, probe efficiency and GC content for each smoothed gene data

Using the idea mentioned above, [36] the correlation between read depth, probe efficiency, and GC-content were observed when GC percentage is within the spectrum of 24% to 47% for gene. It was concluded with the increase in the positive correlation coefficient after smoothing data.

Samples	Rd and Pr	Rd, Pr, and GC
HG00629	0.9921	0.9989
HG01191	0.9960	0.9968
HG01437	0.9915	0.9942
NA19664	0.9921	0.9944
NA19707	0.9907	0.9931
NA19723	0.9907	0.9934
NA20766	0.9911	0.9934

Table 5.6: Correlation between read depth, probe efficiency and GC content for each smoothed gene data ($0.24 <$ GC Content < 0.47)

LOESS method was applied to all samples demonstrated in the next page and it improved the data.

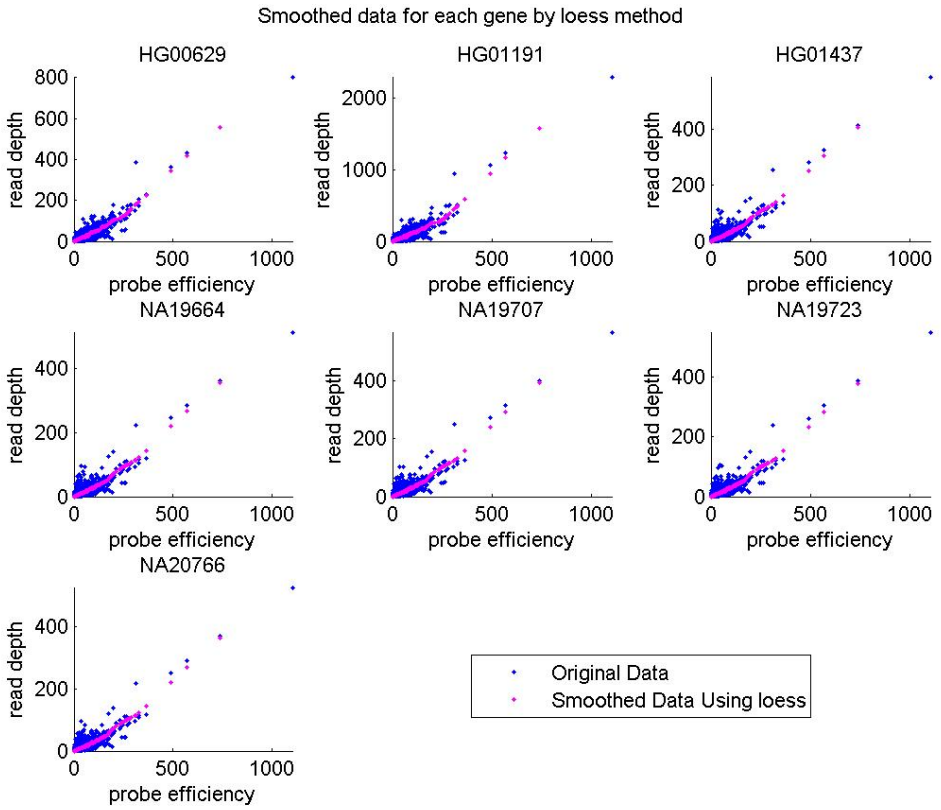

Figure 5.18: Smoothed read depth and probe efficiency by LOESS method for each gene

LOESS method was applied to samples whose GC-percentage is within the spectrum of 24% to 47% demonstrated below and it gives better results.

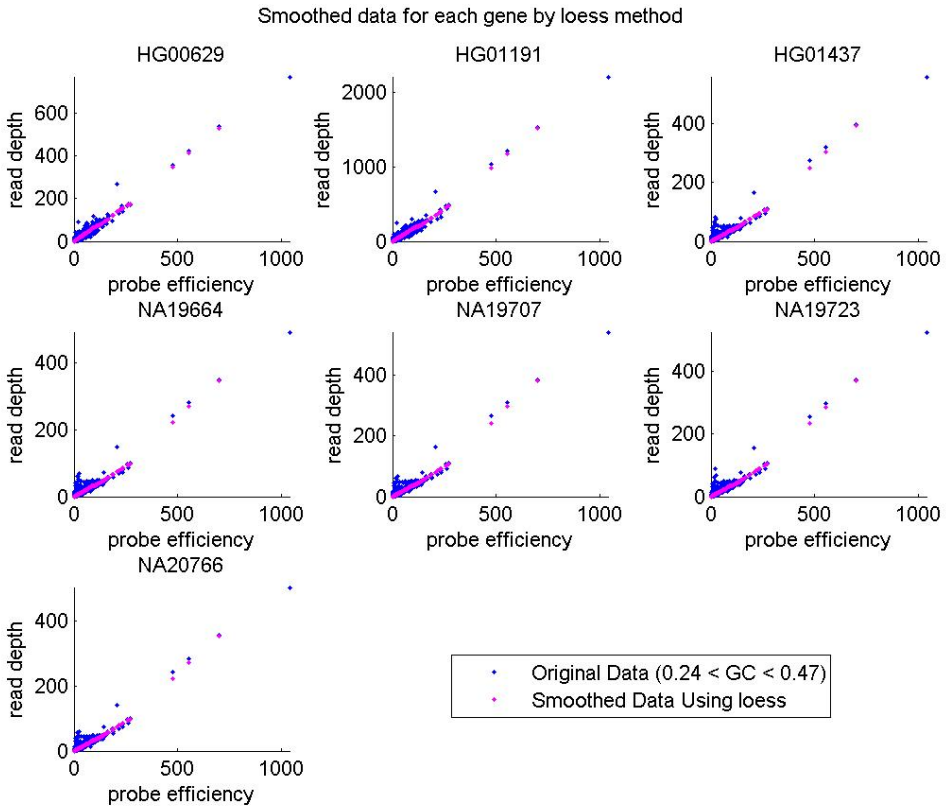

Figure 5.19: Smoothed read depth and probe efficiency by LOESS method for each gene

Chapter 6

Conclusion

Due to its lower cost and higher throughput, exome sequencing is to find genetic causes for common diseases. Whole exome sequencing has the potential to detect copy number variations rapidly. However, whole exome sequencing has some limitations because it covers approximately 1% of whole genome.

As a result of covering limitations, the full spectrum of CNVs and breakpoints cannot be completely characterized and large CNVs cannot be detected precisely. On the other hand, whole exome sequencing data gives a quick insight into copy number variation patterns for a specific disease and phenotype.

Whole exome sequencing data have higher depth for targeted regions in contrast to whole genome sequencing data. Higher read depth is ideal for more accurate copy number variations using read depth-based methods which is mentioned in this work.

Due to different capture efficiency, the depth from different genomic regions may vary substantially. These different capture efficiencies should be normalized. Our aim is to minimize this capture efficiency bias as much as possible. There are no methods that claims that can correct all biases in sequencing data. All methods (or algorithms) about finding copy number variations have their own advantages and disadvantages. The choice of tool is dependent on the aim of

research.

6.1 Future Work

There have been substantial improvements on both sequencing and calling variations. New sequencing technologies have appeared in the field recently. There are some studies that use third generation sequencing and they offer longer reads that will greatly ease read alignment and CNV detection in repetitive regions of genome and significantly reduce mapping errors due to incorrect sequencing. The increased size of short read will also strengthen the statistical power of read depth-based (RD-based) methods. Although fourth generation sequencing technology, such as Oxford Nanopore, has also appeared, the reliability of this new technology is not yet known. More research on this technology is needed to decide whether if it is reliable or not.

In this thesis, only autosomal chromosomes are investigated because sex chromosomes and the other chromosomes are needed a special interest due to their distinctive structures. As an improvement, other kinds of chromosomes can also be investigated.

As a future work of this thesis, a tool will be developed using corrections mentioned in this work. The improvements, discussed in the thesis, on exome sequencing data will be used in the tool that we have developed. Only two steps of whole procedures of finding copy number variations using whole exome sequencing data are worked on and the remaining two steps should be done. Tools used in detecting CNVs are really hard to use even the popular ones so there is a need for new tools in this field. By using the corrected and normalized exome sequencing data in one of the appropriate semi-supervised or supervised learning algorithms chosen due to labeled data size copy numbers can be found. In the last step of the procedure, one of the standard segmentation algorithms can be applied or a new segmentation algorithm can also be developed. Exome sequencing is a special form of targeted sequencing so this new developed tool

can be adapted to any type of targeted sequencing, such as molecular inversion probe (MIP) based targeting. If successful tools we plan to develop using the methods presented in this thesis may also be used in clinical sequencing tests that we expected to be used in all hospitals within the next few years. [38]

Bibliography

[1] M. Zhao, Q. Wang, Q. Wang, P. Jia, and Z. Zhao, "Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives," *BMC Bioinformatics*, vol. 14, no. Suppl 11, p. S1, 2013.

[2] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping," *Nature Reviews Genetics*, vol. 12, no. 5, pp. 363–376, 2011.

[3] L. Pray, "Discovery of dna structure and function: Watson and crick," *Nature Education*, vol. 1, no. 1, 2008.

[4] "Copy number imbalance." `http://www.cambridgebluegnome.com/applications/copy-number-imbalance/`. Accessed May 18, 2014.

[5] E. Pettersson, J. Lundeberg, and A. Ahmadian, "Generations of sequencing technologies," *Genomics*, vol. 93, no. 2, pp. 105–111, 2009.

[6] J. Shendure and H. Ji, "Next-generation dna sequencing," *Nature biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[7] "Sanger sequencing." `http://commons.wikimedia.org/wiki/File:Sanger-sequencing.svg`. Accessed May 17, 2014.

[8] A. Vierstraete, "Next generation sequencing for dummies." University Lecture, june 2012.

[9] A. Magi, M. Benelli, A. Gozzini, F. Girolami, F. Torricelli, and M. L. Brandi, "Bioinformatics for next generation sequencing data," *Genes*, vol. 1, no. 2, pp. 294–307, 2010.

[10] A. Grada and K. Weinbrecht, "Next-generation sequencing: methodology and application," *Journal of Investigative Dermatology*, vol. 133, no. 8, p. e11, 2013.

[11] "Illumina sequencing." `http://www.illumina.com/`. Accessed May 17, 2014.

[12] A. Verma and A. Singh, *Animal Biotechnology: Models in Discovery and Translation*. Academic Press, 2013.

[13] J. F. Thompson and P. M. Milos, "The properties and applications of single-molecule dna sequencing," *Genome Biol*, vol. 12, no. 2, p. 217, 2011.

[14] J. Perkel, "Making contact with sequencing's fourth generation," *BioTechniques*, vol. 50, no. 2, pp. 93–95, 2011.

[15] "Whole genome and exome sequencing." `https://www.my46.org/intro/whole-genome-and-exome-sequencing`. Accessed May 17, 2014.

[16] W. W. Grody, B. H. Thompson, and L. Hudgins, "Whole-exome/genome sequencing and genomics," *Pediatrics*, vol. 132, no. Supplement 3, pp. S211–S215, 2013.

[17] K. Stangier, F. Ernst, Y. Kumar, and T. Paprotka, "Target specific enrichments combined with next-generation sequencing revolutionizes the analysis of human exomes." `http://www.gatc-biotech.com/fileadmin/Kundendaten/bilder/landingpage/WhitePaper_GATC_AllExome.pdf`. Accessed May 15, 2014.

[18] "Exome sequencing." `http://en.wikipedia.org/wiki/Exome_sequencing`. Accessed May 19, 2014.

[19] "Overview of structural variation." `http://www.ncbi.nlm.nih.gov/dbvar/content/overview/`. Accessed May 15, 2014.

[20] S. W. Scherer, C. Lee, E. Birney, D. M. Altshuler, E. E. Eichler, N. P. Carter, M. E. Hurles, and L. Feuk, "Challenges and standards in integrating surveys of structural variation," *Nature genetics*, vol. 39, pp. S7–S15, 2007.

[21] D. Backenroth, J. Homsy, L. R. Murillo, J. Glessner, E. Lin, M. Brueckner, R. Lifton, E. Goldmuntz, W. K. Chung, and Y. Shen, "Canoes: detecting rare copy number variants from whole exome sequencing data," *Nucleic acids research*, p. gku345, 2014.

[22] J. Wu, Y. Li, and R. Jiang, "Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies," *PLoS genetics*, vol. 10, no. 3, p. e1004237, 2014.

[23] A. Ramachandran, M. Micsinai, and I. Pe'er, "Condex: Copy number detection in exome sequences," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, pp. 87–93, IEEE, 2011.

[24] N. Krumm, P. H. Sudmant, A. Ko, B. J. O'Roak, M. Malig, B. P. Coe, A. R. Quinlan, D. A. Nickerson, E. E. Eichler, *et al.*, "Copy number variation detection and genotyping from exome sequence data," *Genome research*, vol. 22, no. 8, pp. 1525–1532, 2012.

[25] L. J. Coin, D. Cao, J. Ren, X. Zuo, L. Sun, S. Yang, X. Zhang, Y. Cui, Y. Li, X. Jin, *et al.*, "An exome sequencing pipeline for identifying and genotyping common cnvs associated with disease with application to psoriasis," *Bioinformatics*, vol. 28, no. 18, pp. i370–i374, 2012.

[26] V. Plagnol, J. Curtis, M. Epstein, K. Y. Mok, E. Stebbings, S. Grigoriadou, N. W. Wood, S. Hambleton, S. O. Burns, A. J. Thrasher, *et al.*, "A robust model for read count data in exome sequencing experiments and implications for copy number variant calling," *Bioinformatics*, vol. 28, no. 21, pp. 2747–2754, 2012.

[27] G. J. Rigaill, S. Cadot, R. J. Kluin, Z. Xue, R. Bernards, I. J. Majewski, and L. F. Wessels, "A regression model for estimating dna copy number applied to capture sequencing data," *Bioinformatics*, vol. 28, no. 18, pp. 2357–2365, 2012.

[28] M. Fromer, J. L. Moran, K. Chambert, E. Banks, S. E. Bergen, D. M. Ruderfer, R. E. Handsaker, S. A. McCarroll, M. C. ODonovan, M. J. Owen, *et al.*, "Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth," *The American Journal of Human Genetics*, vol. 91, no. 4, pp. 597–607, 2012.

[29] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads," *Bioinformatics*, vol. 25, no. 21, pp. 2865–2871, 2009.

[30] E. Karakoc, C. Alkan, B. J. O'Roak, M. Y. Dennis, L. Vives, K. Mark, M. J. Rieder, D. A. Nickerson, and E. E. Eichler, "Detection of structural variants and indels within exome data," *Nature methods*, vol. 9, no. 2, pp. 176–178, 2012.

[31] Y. Benjamini and T. P. Speed, "Summarizing and correcting the gc content bias in high-throughput sequencing," *Nucleic acids research*, p. gks001, 2012.

[32] F. Hach, F. Hormozdiari, C. Alkan, F. Hormozdiari, I. Birol, E. E. Eichler, and S. C. Sahinalp, "mrsfast: a cache-oblivious algorithm for short-read mapping," *Nature methods*, vol. 7, no. 8, pp. 576–577, 2010.

[33] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, *et al.*, "The variant call format and vcftools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[34] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, *et al.*, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[35] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.

[36] A. E. Minoche, J. C. Dohm, H. Himmelbauer, *et al.*, "Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems," *Genome Biol*, vol. 12, no. 11, p. R112, 2011.

[37] W. L. Martinez, A. Martinez, and J. Solka, *Exploratory data analysis with MATLAB*. CRC Press, 2004.

[38] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, *et al.*, "Targeted capture and massively parallel sequencing of 12 human exomes," *Nature*, vol. 461, no. 7261, pp. 272–276, 2009.

[39] R. Dahm, "Friedrich miescher and the discovery of dna," *Developmental Biology*, vol. 278, no. 2, pp. 274–288, 2005.

Appendix A

Glossary

BAM: Binary Aligment/Map Format
BEADS: Bias Elimination Algorithm for Deep Sequencing
BED: Browser Extensible Data Format
CNP: Copy Number Polymorphism
CNV: Copy Number Variation
dATP: deoxyAdenosine TriPhospate
dCTP: deoxyCytidine TriPhospate
dGTP: deoxyGuanosine TriPhospate
dTTP: deoxyThymidine TriPhospate
DNA: Deoxyribonucleic Acid
Indel: Insertion and Deletion
Kb: Kilo base
LINE: Long Interspersed Elements
LOESS: Locally Weighted Regression Scatter Plot Smoothing Method
Mb: Mega base
NGS: Next-Generation Sequencing
PCA: Principal Component Analysis
PCR: Polymerase Chain Reaction
PE: Paired-End
SAM: Sequence Alignment/Map Format

SE: Single-End

SINE: Short Interspersed Elements

SNP: Single Nucleotide Polymorphism

SV: Structural Variation

VCF: Variant Call Format

VNTR: Variable Number Tandem Repeats

WES: Whole Exome Sequencing

WGS: Whole Genome Sequencing

Appendix B

Length measurements

The following abbreviations are commonly used to describe the length of a DNA molecule:

bp := One base pair corresponds to roughly 618 or 643 daltons (the standard unit that is used for indicating mass on an atomic or molecular scale) for DNA.

kb (kbp) := kilo base pairs = $1,000$ bp.

Mb := mega base pairs = $1,000,000$ bp.

Gb := giga base pairs = $1,000,000,000$ bp.

Appendix C

Timeline of DNA

1865: Gregor Mendel discovers through breeding experiments with peas that traits are inherited based on specific laws (later to be termed "Mendel's laws").
1866: Ernst Haeckel proposes that the nucleus contains the factors responsible for the transmission of hereditary traits.
1869: Friedrich Miescher isolates DNA for the first time.
1871: The first publications describing DNA ("nuclein") by Friedrich Miescher, Felix Hoppe-Seyler, and P. Plósz are printed.
1882: Walther Flemming describes chromosomes and examines their behavior during cell division.
1884 - 1885: Oscar Hertwig, Albrecht von Kölliker, Eduard Strasburger, and August Weismann independently provide evidence that the cell's nucleus contains the basis for inheritance.
1889: Richard Altmann renames "nuclein" to "nucleic acid."
1900: Carl Correns, Hugo de Vries, and Erich von Tschermak rediscover Mendel's Laws.
1902: Theodor Boveri and Walter Sutton postulate that the heredity units (called "genes" as of 1909) are located on chromosomes.
1902 - 1909: Archibald Garrod proposes that genetic defects result in the loss of enzymes and hereditary metabolic diseases.
1909: Wilhelm Johannsen uses the word "gene" to describe units of heredity.

1910: Thomas Hunt Morgan uses fruit flies (Drosophila) as a model to study heredity and finds the first mutant (white) with white eyes.

1913: Alfred Sturtevant and Thomas Hunt Morgan produce the first genetic linkage map (for the fruit fly Drosophila).

1928: Frederick Griffith postulates that a "transforming principle" permits properties from one type of bacteria (heat-inactivated virulent Streptococcus pneumoniae) to be transferred to another (live nonvirulent Streptococcus pneumoniae).

1929: Phoebus Levene identifies the building blocks of DNA, including the four bases adenine (A), cytosine (C), guanine (G), and thymine (T).

1941: George Beadle and Edward Tatum demonstrate that every gene is responsible for the production of an enzyme.

1944: Oswald T. Avery, Colin MacLeod,and Maclyn McCarty demonstrate that Griffiths "transforming principle" is not a protein, but rather DNA, suggesting that DNA may function as the genetic material.

1949: Colette and Roger Vendrely and André Boivin discover that the nuclei of germ cells contain half the amount of DNA that is found in somatic cells. This parallels the reduction in the number of chromosomes during gametogenesis and provides further evidence for the fact that DNA is the genetic material.

1949 - 1950: Erwin Chargaff finds that the DNA base composition varies between species but determines that within a species the bases in DNA are always present in fixed ratios: the same number of A's as T's and the same number of C's as G's.

1952: Alfred Hershey and Martha Chase use viruses (bacteriophage T2) to confirm DNA as the genetic material by demonstrating that during infection viral DNA enters the bacteria while the viral proteins do not and that this DNA can be found in progeny virus particles.

1953: Rosalind Franklin and Maurice Wilkins use X-ray analyses to demonstrate that DNA has a regularly repeating helical structure.

1953: James Watson and Francis Crick discover the molecular structure of DNA: a double helix in which A always pairs with T, and C always with G.

1956: Arthur Kornberg discovers DNA polymerase.

1957: Francis Crick proposes the "central dogma" (information in the DNA is translated into proteins through RNA) and speculates that three bases in the

DNA always specify one amino acid in a protein.

1958: Matthew Meselson and Franklin Stahl describe how DNA replicates (semi-conservative replication).

1961 - 1966: Robert W. Holley, Har Gobind Khorana, Heinrich Matthaei, Marshall W. Nirenberg, and colleagues crack the genetic code.

1968 - 1970: Werner Arber, Hamilton Smith, and Daniel Nathans use restriction enzymes to cut DNA in specific places for the first time.

1972: Paul Berg uses restriction enzymes to create the first piece of recombinant DNA.

1977: Frederick Sanger, Allan Maxam, and Walter Gilbert develop methods to sequence DNA.

1982: The first drug (human insulin), based on recombinant DNA, appears on the market.

1983: Kary Mullis invents PCR as a method for amplifying DNA in vitro.

1990: Sequencing of the human genome begins.

1995: First complete sequence of the genome of a free-living organism (the bacterium Haemophilus influenzae) is published.

1996: The complete genome sequence of the first eukaryotic organism - the yeast S. cerevisiae - is published.

1998: Complete genome sequence of the first multi-cellular organism - the nematode worm Caenorhabditis elegans - is published.

1999: Sequence of the first human chromosome (22) is published.

2000: The complete sequences of the genomes of the fruit fly Drosophila and the first plant - Arabidopsis - are published.

2001: The complete sequence of the human genome is published.

2002: The complete genome sequence of the first mammalian model organism - the mouse - is published.

2005: 454 pyrosequencing has appeared.

2008: Illumina DNA sequencing technology (sequencing by synthesis) has appeared.

2009: SoLiD (sequencing by ligation) has appeared.

2011: PacBio DNA sequencing technology has appeared.

2013: Oxford Nanopore DNA sequencing technology has appeared. [39]