2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering

A Computational Approach to Detect CNVs Using High-throughput Sequencing

Myungjin Moon, Jaegyoon Ahn, Chihyun Park, Sanghyun Park
Department of Computer Science, Yonsei University
Seoul, South Korea
{psiwind, ajk , tianell, sanghyun}@cs.yonsei.ac.kr

Youngmi Yoon
Department of Information Technology, Gachon University of Medicine and Science
Seoul, South Korea
amyyoon@ cs.yonsei.ac.kr

Jeehee Yoon
Division of Information and Communication Engineering, Hallym University
Seoul, South Korea
jhyoon@hallym.ac.kr

***Abstract*—Copy-Number Variations (CNVs) can be defined as gains or losses that are greater than 1kbs of genomic DNA among phenotypically normal individuals. CNVs detected by microarray based approach are limited to medium or large sized ones because of its low resolution. Here we propose a novel approach to detect CNVs by aligning the short reads obtained by high-throughput sequencer to the previously assembled human genome sequence, and analyzing the distribution of the aligned reads. Application of our algorithm demonstrates the feasibility of detecting CNVs of arbitrary length, which include short ones that microarray based algorithms cannot detect. Also, false positive and false negative rates of the results were relatively low compared to those of microarray based algorithms.**

Keywords-Copy Number Variations; CNVs; High-throughput sequencing; Genomic variants generator;

I. INTRODUCTION

Copy-Number Variations (CNVs) can be defined as gains or losses that are greater than 1kbs of genomic DNA among phenotypically normal individuals [1, 2]. It is known that CNVs account for a significant proportion of normal phenotypic variation, including disease susceptibility [3-5]. Therefore, identifying and cataloging of CNVs are essential for the genetic and functional analysis of human genome variation.

Many algorithms were proposed to assess CNV regions of human genome using microarray, which includes Whole Genome TilePath (WGTP) array [6, 7] and SNP genotyping array [8, 9]. The popularity of these methods mainly accounts for the relatively low cost of WGTP array and SNP array. However, the resolution of these platforms limits the size of CNVs found. Generally, these methods are known to be useful to detect only medium or large sized CNVs. Moreover, high noise level of these platforms tends to result in relatively high false positive and false negative rates.

The comparison of two or more human genome sequences can detect CNVs, regardless of CNVs' sizes, more precisely [10-12]. The weakest point of the sequence comparison methods is the high cost of the human genome sequence. There has been much effort to lower the cost to get the whole human genome sequence, and high-throughput sequencing is believed to take a prime role [13]. High-throughput sequencing machine can generate enormous short reads in a short time with relatively low cost. Table 1 [13] shows the details of the reads and throughput of various high-throughput sequencing platforms.

TABLE 1. HIGH- THROUGHPUT SEQUENCING PLATFORMS [13]

Company	Format	Read Length (bases)	Expected Throughput MB(million bases) / day
454 Life Sciences	Parallel bead array	100	96
Agencourt Bioscience	Sequencing by ligation	50	200
Applied Biosystems	Capillary electrophoresis	1000	3-4
Microchip Biotechnologies	Parallel bead array	850-1000	7
NimbleGen Systems	Map and survey microarray	30	1000
Solexa (Illumina)	Parallel microchip	35	500
LI-COR	Electronic microchip	20000	14000
Network Biosystems	Biochip	800+	5
VisiGen Biotechnologies	Single molecule array	NA	1000

There have been many algorithms proposed for de novo assembly of the human genome [14-16]. What makes assembly difficult is the short length of

978-0-7695-3656-9/09 $25.00 © 2009 IEEE
DOI 10.1109/BIBE.2009.13