



Cairo University

High-Performance system for detecting novel altered regions of Hepatocellular carcinoma using high-throughput sequencing

By

Esraa Mamdouh Hashim Shabeb

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
BIOMEDICAL ENGINEERING AND SYSTEMS

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2017

High-Performance system for detecting novel altered
regions of Hepatocellular carcinoma using high-throughput
sequencing

By
Esraa Mamdouh Hashim Shabeb

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
BIOMEDICAL ENGINEERING AND SYSTEMS

Under the Supervision of

Prof. Dr. Ayman M. Eldeib

Assoc.Prof. Dr. Mai Mohamed S.
Mabrouk

.....
Professor
Biomedical Engineering and Systems
Department
Faculty of Engineering, Cairo University

.....
Associate Professor and Department Head
Of Biomedical Engineering
Faculty of Engineering, Misr University for
Science and Technology

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2017

High-Performance system for detecting novel altered regions of Hepatocellular carcinoma using high-throughput sequencing

By

Esraa Mamdouh Hshim

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
BIOMEDICAL ENGINEERING AND SYSTEMS

Approved by the Examining Committee:

Prof. Dr. Ayman Mohamed Eldeib	Main Advisor
---------------------------------------	--------------

Assoc.Prof. Dr. Mai Mohamed Saied. M. Mabrouk	Member
--	--------

Associate Professor of Biomedical Engineering, Department Head of Biomedical Engineering, Misr university for science and technology

Prof. Dr. Ahmed Hesham Kandil	Internal Examiner
--------------------------------------	-------------------

Prof. Dr. Mohamed Waleed.T. Fakhr	External Member
--	-----------------

Computer Engineering Department, AAST, Cairo

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2017

Engineer: Esraa mamdouh hashim shabib
Date of Birth: 1 / 4 / 1986
Nationality: Egyptian
E-mail: bioeng_esraa@hotmail.com
Phone.: 01008785879
Address: 6-october
Registration Date: 1/ 3 / 2013
Awarding Date: / / 2017
Degree: Doctor philosophy
Department: Biomedical Engineering and Systems department
Supervisors: Prof.Dr. Ayman Mohamed Eldeib
Ass.Prof.Dr. Mai saied mabrouk



Examiners: Ass.Prof. Dr. Mai saied mabrouk
Prof. Dr. Ayman Mohamed Eldeib
Prof. Dr. Ahmed Hesham Kandeel
Prof. Dr. Mohamed waleed Fakhr, Professor

Title of Thesis: High Performance system for detecting novel altered regions of Hepatocellular carcinoma using high-throughput sequencing

Key Words: Hepatocellular carcinoma.; whole-genome sequencing; Next generation sequence; bioinformatics.

Summary:

Hepatocellular carcinoma(HCC) is a tumor of the liver which usually arises in the setting of chronic liver diseases. Fibrolamellar hepatocellular carcinoma (F-HCC) is a rare entity of HCC not yet analyzed cytogenetically. The use of high-performance parallel computing techniques can enable researchers to employ large numbers of processors to run comprehensive analyses within a reasonable period. This work provides a genomic study that focuses on using bioinformatics approaches to predict the molecular causes of HCC and F-HCC by the investigation whole genome sequence of the chromosomal aberrations using single nucleotide polymorphism (SNP)arrays and Next generation sequence (NGS) by applying four statistical techniques. The study revealed 3 distinct structural variations_related genes MDM4, PRDM5, and WHSC1, these genes are a novel target signature that can help to predict survival of patients with detecting F-HCC. A new altered chromosome region amplification(**4q22.1**), this altered chromosomal region is novel for detect HCC, this finding has not previously reported being involved in liver carcinogenesis. NGS detect **EPHA5**, **UBE1L2** tumor suppressor and UGT2B28 for both F-HCC and HCC.

Acknowledgments

First, and for most, thanks to God the most merciful and most gracious.

Second, I would like to express my special appreciation and thanks to my advisor Professor Dr. Ayman eldeab, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist.

I wish to express my sincere thanks to Assist. Prof. Dr. Mai Saeed Mabrouk, for her patience, experienced advice, great help, scientific discussions, guidance, encouragement, suggesting the point of research, fruitful discussion that helped me in achieving this thesis. I really appreciate her great effort through the duration of this work. Your advice on both research as well as on my career have been priceless. I could not have imagined having a better advisor and mentor for my Ph.D study.

A special thanks to my family. Words cannot express how grateful I am to my mother, and father for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank all of my friends who supported me in writing, and incited me to strive towards my goal. At the end I would like express appreciation to my beloved husband Mahmoud Fahmy who spent sleepless nights with and was always my support in the moments when there was no one to answer my queries. I'd like to say that really without his standing beside me this work has never been completed.

I would like to dedicate this work to my lovely kids Lamar and leen.

Abstract

The area of Bioinformatics has transformed into a consequential portion of different fields of biology. In experimental sub-atomic science, several methods of Bioinformatics such as signal processing and image processing results in sizable voluminous amounts of data of raw information. In the area of genomics and genetics, Bioinformatics helps in sequencing and analysis entire genomes and detect their observed variations.

Hepatocellular carcinoma (HCC) counted as the third prevalent reason of cancer mortality worldwide. HCC is amongst the most prevalent types of aggressive tumors; HCC patient survival is poor incompletely because of HCC recurrence, also the behavior of HCC is difficult to predict. Fibrolamellar hepatocellular carcinoma (F-HCC) is a rare primary hepatic malignancy.

The etiology of F-HCC is still **uncertain**, due to the non-appearance of defining symptoms or a specific diagnostic test, it is often detected after it has metastasized, and by then, the disease is frequently progressive and deadly, the outcome of HCC patients still remains dismal due to the difficulty in detecting the disease at its early stage.

The major form of genetic mutation is copy number aberrations (CNAs), that leads to abnormal cell development and diffusion. Different microarray approaches have been effectively emerged to characterize CNAs in all sorts of cancer, but, their resolution is constrained by the location and the number of probes on the platforms array (FISH, A-CGH). SNP array and next generation sequencing platforms guarantee to characterize the copy number profiles in malignancy tumors.

The main aim of this work is to verify , develop and test the new methodologies to quickly and precisely detect the copy number changes in the human genome ,four statistical methodologies are represented by a circular binary segmentation (CBS) algorithm, Discrete stationary wavelet transformation method (DSWT), Quantisnp,and Oncosnp-seq applied to whole genome sequence for analyzing SNP array and high-throughput sequence information that identify genome-wide CNAs from the genomic data which counting for the dependence between neighboring clones by using high performance parallel computing to accelerates the performance by separating a problem into smaller sub problems that are divided to various processors.

The other objective is to detect new altered regions which can lead to discovering new markers that may help improve diagnosis to detect HCC and represent novel targets for therapeutic agents.

Results referred to a certain number of chromosome aberrations detected, the first structural variations highlight is MDM4 gene, increased expression of epidermal growth factor receptor (EGFR) is associated with tumor progression in several carcinomas that previously were reported to be altered in F-HCC. A significant deletion of chromosome 4 has been observed and recorded in this work, especially in a **4q25-26** region and seen in almost all F-HCC samples striking **PRDM5** gene.

Loss of 4p16.3 involved in WHSC1 gene, these genes are **novel target** and the second and third variation highlight for *F-HCC*. A new altered chromosome region amplification(4q22.1) have been detected this altered chromosomal region is **novel** for detecting Hepatocellular carcinoma and may lead to discovering new biomarker help for better understanding of hepatocellular carcinoma and its diagnosis because this finding has not previously reported being involved in liver carcinogenesis. The deletion aberration at **19p13.12** is responsible for the occurrence of the highly prevalent DNAJB1–PRKACA fusion protein, which is present in F-HCC it is detected by DSWT, quantisnp and Oncosnp-seq techniques.

Recently the development of Next generation sequence empowers simultaneous estimation of copy number of hundred thousands of locales in a genome, more precise estimation of copy numbers, higher resolution and coverage, accurate identification of change points, and higher tendency to distinguish new CNAs.

Table of Contents

Chapter1: Introduction	1
1.1 Overview	1
1.2 Problem Definition.....	2
1.3 Aim of the Work.	2
1.4 Thesis Organization.....	3
Chapter 2: Medical Background and literature review	4
2.1 Introduction.....	4
2.2 Bioinformatics in the medical science	4
2.3 Hepatocellular carcinoma (HCC).....	5
2.4 Fibro lamellar Hepatocellular carcinoma (F-HCC)	13
2.5 Genome analysis	14
2.5.1 High-throughput single nucleotide polymorphisms (SNP) array,	17
2.5.2 Deep sequencing of HCC using next-generation sequencing technologies	19
2.6 High performance parallel computing.....	22
Chapter 3: Technical Background	24
3.1 Introduction	24
3.2 Detection of Copy number alteration(CNA).....	24
3.2.1 Circular binary segmentation (CBS).....	25
3.2.2 Discrete Stationary wavelet transforms (DSWT).	26
3.2.3 OuantiSNP Technique	28
3.2.4 Oncosnp-Seq technique.	30
3.3 High-performance parallel Computing Architectures.....	31
Chapter 4: Material and Methods	33
4.1 Introduction	33
4.2 Database description	34
4.2.1 30-HCC SNP array cell lines:	35
4.2.2 13-F-HCC SNP array cell line:	35
4.2.3 5-HCC Whole genome sequence cases:	35

4.3	Implementation.....	35
4.4	Circular Binary Segmentation (CBS):.....	36
4.5	Discrete Stationary wavelet transforms (DSWT).....	37
4.6	OuantiSNP Technique.....	40
4.7	Oncosnp-Seq technique.....	42
4.8	Evaluation Criteria	44
4.9	High-performance parallel Computing Architectures.....	45
Chapter 5: Results and Discussions		47
5.1	Introduction	47
5.2	Results of circular binary segmentation (CBS).....	47
5.3	Results of Discrete Stationary Wavelet Transform (DSWT).....	52
5.4	Results of QuantiSNP technique:.....	62
5.5	Results of Oncosnp-seq.....	68
5.6	Results of High performance parallel computing:	73
5.7	Discussions.....	74
Chapter 6: Conclusions and Future Work.....		82
6.1	Conclusions	82
6.2	Future work	84

List of tables

Table 2-1 common causes of HCC in Arab world.....	11
Table2-2 comparisons of imaging approaches with their limitations.....	13
Table 4-1 Control hidden states and biological interaction	40
Table 4-2 Control states and biological interaction	43
Table 5-1 Validation of CBS algorithm.....	52
Table 5-2 prediction accuracy of each chromosome for 30-HCC sample.....	59
Table 5-3 summary of abnormal chromosomal regions of f- HCC samples	61
Table 5-4 outlines irregular chromosomal points of Hepatocellular carcinoma samples.	65
Table 5-5 summery of related gens within abnormal chromosomal regions of Fibro lamellar HCC samples	67
Table 5-6 summary of abnormal chromosomal aberrations of WGS of HCC samples....	70
Table 5-7 comparison of statistical techniques	70
Table 5-8 comparison of related work for F-HCC samples.....	71
Table 5-9 related work	72
Table 5-10 results of high performance for DSWT	74
Table 5-11 results of high performance for quantisnp program	74

List of Figures

Figure 2-1 Biomarker discovery in a cross-disciplinary domain[2]	5
Figure 2-2 mortality rate and incidence of the 15 most common cancers in the world 2012[9]	7
Figure 2-3 Geographic distribution of HCC. Incidence rates (%) in total population A, female; B, male [10].	Error! Bookmark not defined.
Figure 2-4 mortality and Incidence (per 100,000) of HCC in regions around the world 2012.[9]	9
Figure 2-5 worldwide epidemiologic data estimate HCC due HBV and HCV[13]	10
Figure 2-6 the main factor causes HCC[14]	11
Figure 2-7 Scheme of the guideline of the FISH Experiment to detect CN [2].....	15
Figure 2-8 Schematic representation of CGH [2]	16
Figure 2-9.a-CGH empower genome-wide analysis of DNA sequence [2]	17
Figure 2-10 principal of SNP array [75]	18
Figure 2-11(a,b,c,d) preparation of WGS data [34].....	20
Figure 2-12 brake down the problem to small problems [83]	23
Figure 4-1 General bock diagram of overall process for our proposed approach	34
Figure 4-2 the Ideogram of homo sapiens	36
Figure 4-3 steps of CBS technique	37
Figure 4-4 procedures to use DSWT	39
Figure 4-5 procedures of QuantiSNP technique to estimate CNA	41
Figure 4-6 overall process of Oncosnp-seq technique.....	44
Figure 4-7 the process of parallel programming using SPMD	46
Figure 5-1 gain of chromosome 1q21 to 1q22,the red bold horizontal lines represent the within-segment means computed by the CBS algorithm.....	47
Figure 5-2 loss in chromosome 4q25 to 26 and gain in 4q22.1, the red bold horizontal lines represent the within-segment means computed by the CBS algorithm.....	48
Figure 5-3 loss and gain in 8q24.12to 8q24.13, the red bold horizontal lines represent the within-segment means computed by the CBS algorithm.....	48
Figure 5-4 loss in 9p21.3, the red bold horizontal lines represent the within-segment means computed by the CBS algorithm.....	49
Figure 5-5 gain in chromosome 11q13.2 to 13.3. the red bold horizontal lines represent the within-segment means computed by the CBS algorithm.....	49
Figure 5-6 loss in chromosome 13q12.11, the red bold horizontal lines represent the within-segment means computed by the CBS algorithm.....	50
Figure 5-7 loss in chromosome 16q11.2, the red bold horizontal lines represent the within-segment means computed by the CBS algorithm.....	50
Figure 5-8 loss in chromosome 17p13.1, the red bold horizontal lines represent the within-segment means computed by the CBS algorithm.....	51

Figure 5-9 gain in chromosome 20q12, the red bold horizontal lines represent the within-segment means computed by the CBS algorithm.	51
Figure 5-10 the coefficients of approximation and detail sub-band	52
Figure 5-11 the original before denoising and denoised signals of chromosome 4.....	53
Figure 5-12 gain and loss of chromosome 1q21and 1p36.32	54
Figure 5-13 Loss of chromosome 4q25-26.....	54
Figure 5-14 gain of chromosome 4q22.1.....	55
Figure 5-15 gain of chromosome 8q24.12.....	55
Figure 5-16 loss of chromosome 8p21.3.....	56
Figure 5-17 loss of chromosome 9q21.3-23	56
Figure 5-18 gain of chromosome 11q13.2-13.3.....	57
Figure 5-19 .loss of chromosome 13q12.11.....	57
Figure 5-20 loss of chromosome 16q11.2.....	58
Figure 5-21 gain and loss of chromosome 17q12,17p13.1	58
Figure 5-22 Gain in chromosome 20q12	59
Figure 5-23 loss in chromosome 1p36.32-33	60
Figure 5-24 gain of chromosome 7p11.2.....	60
Figure 5-25 gain in chromosome 20q13.3	61
Figure 5-26 Performance comparison of CNA detection methods using 4 types of data.	63
Figure 5-27 gain with green color and loss with red color in chromosome 1	64
Figure 5-28 gain in chromosome 4 at the green region	64
Figure 5-29 loss of chromosome 1	Figure 0-30 loss in
chromosome 4.....	66
Figure 5-31 gain in chromosome 6	Figure 0-32 gain in chromosome
7.....	66
Figure 5-33 loss in chromosome 8 ,Figure 5-34 gain in chromosome 17	66
Figure 5-35 genome-wide copy number profiles (red)	68
Figure 5-36 whole genome sequence data of HCC	69
Figure 5-37 comparison of CBS,DSWT,Quantisnp with related work	71
Figure 5-38 comparison of F-HCC samples	72
Figure 5-39 NGS comparison of the results of our methods with results of previous work	73

List of Abbreviation.

HCC	Hepatocellular carcinoma
CNVs	copy number variation
CNA	Copy number alteration
ALD	alcoholic liver disease
NASH	non-alcoholic steatohepatitis
AIH	auto-immune hepatitis
PBC	primary biliary cirrhosis
PSC	primary sclerosing cholangitis
SNP	single nucleotide polymorphisms
AFLP	Lens culinaris
DCP	des- carboxyprothrombin
GPC3	Glypican-3
GP73	Golgi protein 73
AFP	Alpha-fetoprotein
F-HCC	Fibro lamellar hepatocellular carcinoma
SNP	single nucleotide polymorphisms
NGS	Next generation sequence
WGS	whole genome sequence
WES	Whole exome sequencing
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
FISH	fluorescent in situ hybridization
FDR	False discovery rate

SNV	single nucleotide variants
GA	Genome analyzer
SIMD	Single Instruction Multiple Data
MIMD	Multiple Instruction Multiple Data
DSWT	Discrete stationary Wavelet transform
SPMD	Single program multiple data
SHC	Src homology 2 domain containing
CCT3	Chaperonin containing TCP1, subunit 3 gamma
COPA	Coatomer protein complex, subunit alpha
RD	Read depth
RC	Read count
LRR	Log R Ratio
HMM	hidden markov model
SCIMM	conditional mixture modelling
FDR	false discovery rate
FT	Fourier transform
SURE	Stein's unbiased risk estimate
MSE	mean-squared error
OB-HMM	Objective Bayes Hidden-Markov Model
BF	Bayes Factor
SPMD	single-program multiple-data
CDH1	Cadherin 1, type 1, E-cadherin (epithelial)
ADRBK2	adrenergic, beta, receptor kinase 2
CRYBB2P1	crystallin, beta B2 pseudogene 1

LRP5L	low density lipoprotein receptor-related protein 5
PTPRN2	protein tyrosine phosphatase, receptor type, N polypeptide 2
DPP6	dipeptidyl-peptidase 6
EM	Expectation Maximization
JTB	Jumping translocation breakpoint
SHC	Src homology 2 domain containing
CCT3	chaperonin containing TCP1, subunit 3 gamma
COPA	coatomer protein complex, subunit alpha
HSRG1	HSV-1 stimulation-related gene 1
CRYBB2P1	crystallin beta B2 pseudogene 1
ADRBK2	adrenergic beta receptor kinase 2