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

……..

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

Assoc.Prof. Dr. Mai Mohamed S. Mabrouk

……..

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@hotmai.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 incented 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(4**q22.1)** 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

List of tables

List of Figures

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