

**T.C.
REPUBLIC OF TURKEY
HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF HEALTH SCIENCES**

**BAYESIAN NETWORKS FOR OMICS DATA ANALYSIS IN
HEPATOCELLULAR CARCINOMA SINGLE-CELL
SEQUENCING**

Muntadher Zahid JIHAD

**Program of Bioinformatics
MASTER THESIS**

ANKARA

2021

**T.C.
REPUBLIC OF TURKEY
HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF HEALTH SCIENCES**

**BAYESIAN NETWORKS FOR OMICS DATA ANALYSIS IN
HEPATOCELLULAR CARCINOMA SINGLE-CELL
SEQUENCING**

Muntadher Zahid JIHAD

**Program of Bioinformatics
MASTER THESIS**

**SUPERVISOR
Asst. Prof. Dr. İdil YET**

**ANKARA
2021**

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF HEALTH SCIENCES
BAYESIAN NETWORKS FOR OMICS DATA ANALYSIS IN
HEPATOCELLULAR CARCINOMA SINGLE-CELL SEQUENCING
Muntadher Zahid JIHAD
Supervisor: Asst. Prof. Dr. İdil YET

This thesis study has been approved and accepted as a Master dissertation in “Bioinformatics Program” by the assesment committee, whose members are listed below, on 11.02.2021

Chairman of the Committee : *Prof.Dr. Erderm KARABULUT*
Hacettepe University

Advisor of the Dissertation : *Asst. Prof. Dr.İdil YET*
Hacettepe University

Member : *Assoc. Prof. Dr. Yeşim AYDIN SON*
Middle East Technical University

This dissertation has been approved by the above committee in conformity to the relatedissues of Hacettepe University Graduate Education and Examination Regulation.

Prof. Diclehan ORHAN, MD, PhD
Director

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan **“Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge”** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. ⁽¹⁾
- Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. ⁽²⁾
- Tezimle ilgili gizlilik kararı verilmiştir. ⁽³⁾

04 / 03 /2021

Muntadher Zahid JIHAD

¹“Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge”

(1) Madde 6. 1. Lisansüstü teze ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez **danışmanın** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu** iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.

(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez **danışmanın** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulunun** gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, **tezin yapıldığı kurum** tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, **ilgili kurum ve kuruluşun önerisi** ile **enstitü** veya **fakültenin** uygun görüşü üzerine **üniversite yönetim kurulu** tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.

Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

* Tez **danışmanın** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu tarafından karar verilir.**

ETHICAL DECLARATION

In this thesis study, I declare that all the information and documents have been obtained in the base of the academic rules and all audio-visual and written information and results have been presented according to the rules of scientific ethics. I did not do any distortion in data set. In case of using other works, related studies have been fully cited in accordance with the scientific standards. I also declare that my thesis study is original except cited references. It was produced by myself in consultation with supervisor Asst. Prof. Dr. İdil YET and written according to the rules of thesis writing of Hacettepe University Institute of Health Sciences.

Muntadher Zahid JIHAD

ACKNOWLEDGMENT

First and foremost, I would like to express my gratitude and appreciation to my mentor and advisor Asst. Prof. Dr. İdil YET, for her patience, suggestions, and guidance. This research wouldn't be achieved without her persistent help.

I would like to thank my family, I am really grateful to them for what they have done for me, without them I wouldn't be here today.

Lastly, I thank my lovely wife, Nazlı for being beside me during tough times.

Part of this study has been funded by L'Oréal-UNESCO "*For Women in Science*" national award received by my supervisor in 2020.

The numerical calculations reported in this study were partially performed at TUBITAK ULAKBİM, High Performance and Grid Computing Center (TRUBA resources).

ABSTRACT

Jihad, M., Bayesian Networks for Omics Data Analysis in Hepatocellular Carcinoma Single-Cell Sequencing, Hacettepe University, Graduate School of Health Sciences, Department of Bioinformatics, Master Thesis, ANKARA, 2021. Single cell multi omics techniques have shown an advancement in unrevealing complex diseases like cancer heterogeneity by providing multi-faceted insight into their individual cellular regulations. In this study, a machine learning approach, Bayesian network (BN), has been applied to integrate genomic, epigenomic, and transcriptomic data in hepatocellular carcinoma at single cell resolution. Hepatocellular carcinoma (HCC) is the most common type of liver cancer with a high metastatic rate and reckoned for poor prognosis. Heterogeneity of tumor cells is concerned with cancer progression, metastasis, therapeutic resistance, and mortality. For this purpose, a dataset from a published study of 25 single cell sequencing of hepatocellular carcinoma were used. First, DNA methylome and transcriptome data were analyzed on their own. Copy number variations were estimated from DNA methylome data by using the Hidden Markov Model method. To reveal the causal relationship between the omics, three BN models were constructed. The models were fitted to their parameters by using maximum likelihood estimation. For model evaluation, score-based criteria, Akaike information criterion and Bayesian information criterion, were used. 207 genes with significant models have been detected. The heterogeneity of the omics and their regulation mechanisms with each other have been shown, by pointing to genes that follow different BN models that take place in major pathways in HCC.

Key words: Single cell, hepatocellular carcinoma, liver cancer, RNA sequencing, transcriptome, genome, copy number variation, DNA methylome, epigenome, multi-omics integration, Bayesian networks, machine learning.

ÖZET

Jihad, M., Hepatosellüler Karsinomun Tekil Hücre Diziliminde Omiklerin Veri Analizi İçin Bayes Ağları, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Biyoinformatik Programı, Yüksek Lisans Tezi, Ankara, 2021. Tek hücreli çoklu omik teknikleri, kendi bireysel hücresel düzenlemelerine çok yönlü bir bakış açısı sağlayarak kanser heterojenliği gibi kompleks hastalıkların ortaya çıkarmada bir ilerleme göstermiştir. Bu çalışmada, tek hücre temelli hepatosellüler karsinomda genomik, epigenomik ve transkriptomik verileri entegre etmek için bir makine öğrenimi yaklaşımı olan Bayesian ağları (BN) uygulanmıştır. Bu amaçla, yayınlanmış bir çalışmadan hepatosellüler karsinomun 25 tekil hücre dizileme veri seti kullanılmıştır. Hepatosellüler karsinom (HSK), yüksek metastatik oranla en yaygın karaciğer kanseri türüdür ve kötü prognoza sahip olduğu düşünülmektedir. Tümör hücrelerinin heterojenliği, kanserin ilerlemesi, metastaz, terapötik direnç ve mortalite ile ilgilidir. Önce, DNA metilom ve transkriptom verileri tek başlarına analiz edilmiştir. Kopya sayısı varyasyonu, Gizli Markov Modeli yöntemi kullanılarak DNA metilom verilerinden tahmin edilmiştir. Omikler arasındaki nedensel ilişkiyi incelemek için üç BN modeli oluşturulmuştur. Modeller, en çok olabilirlik kestirimi (MLE) kullanılarak parametrelerine uydurulmuştur. Model değerlendirme için puana dayalı kriterler, Akaike bilgi kriteri (AIC) ve Bayes bilgi kriteri (BIC) kullanılmıştır. Anlamlı modele sahip 207 gen tespit edilmiştir. Farklı BN model izleyen genlerin HCC'de aynı yolakta yer aldığına işaret ederek, omiklerin ve birbirleriyle regülasyon mekanizmalarının heterojenliği gösterilmiştir.

Anahtar Kelimeler: Tekil hücre, hepatosellüler karsinom (HSK), karaciğer kanseri, RNA sekanslama, Transkriptom, Genom, Kopya Sayısı Varyasyonu (KSV), DNA Metilomu, Epigenom, Çoklu omik entegrasyonu, Bayes ağları, makine öğrenmesi.

TABLE OF CONTENTS

APPROVAL	III
YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI	IV
ETHICAL DECLARATION	V
ACKNOWLEDGMENT	VI
ABSTRACT	VII
ÖZET	VIII
TABLE OF CONTENT	IX
ABBREVIATIONS	XII
LIST OF FIGURES	XIII
LIST OF TABLES	XIV
1. INTRODUCTION	1
2. LITERATURE REVIEW	4
2.1 Multi Omics Integration	4
2.2 Hepatocellular Carcinoma	8
2.2.1 Clinical Signs and Symptoms	11
2.2.2 The Molecular Landscape	11
2.3 Epigenetics	12
2.3.1 DNA Methylation	14
2.4 Gene Expression	17
2.4.1 Transcriptome	17
2.5 Copy Number Variation	18
2.6 Research Objective	22

3. MATERIAL AND METHODS	24
3.1 Downloading the Data	25
3.2 RNA Sequencing data	25
3.2.1 Data Preprocessing	25
3.2.2 Gene Expression Quantification	27
3.3 RRBS data	31
3.3.1 Data Pre-processing	31
3.3.2 Methylation Base Calling	32
3.4 CNV Estimation from RRBS Data	33
3.5 Correlation Calculations	36
3.6 Bayesian Networks	36
3.7 Gene Set Enrichment Analysis	41
3.7.1 CBioportal Database	41
4. RESULTS	42
4.1 Gene Expression Levels	42
4.2 DNA Methylation Levels	43
4.3 Copy Number Variation	46
4.4 Correlation Between Omics	48
4.5 Causality Analysis Using BN	49
4.6 Gene Set Enrichment Analysis	52
5. DISCUSSION AND CONCLUSION	61
5.1 Limitations and Possible Improvements	64
6. REFERENCES	65
7. APPENDIX	77
APPENDIX 1: Correction Step Of GC-Content and Mappability By Hmmcopy.	
APPENDIX 2: CNV Pattern of All Samples.	

APPENDIX 3: B_{IC} Scores For The Three Models With Gene Names, ΔB_{IC} , And Best Fitted Model For Each Gene.

APPENDIX 4: The Gene List Of Group 2 (Cme) And The Related Hcc Studies Reported By Cbioportal

APPENDIX 5: Digital Receipt

APPENDIX 6: Thesis Originality Report

APPENDIX 7: Girişimsel Olmayan Klinik Araştırmaları Etik Kurulu Onayı

8. CURRICULUM VITAE

101

ABBREVIATIONS

ACGH	: Array Comparative Genomic Hybridization
AIC	: Akaike Information Criterion
AS	: De-Novo Assembly
BIC	: Bayesian Information Criterion
BN	: Bayesian Networks
CA	: Combinatorial Approach
CBS	: Circular Binary Segmentation
CEM model	: CNV -> Gene Expression -> DNA Methylation
CME model	: CNV -> DNA Methylation -> Gene Expression
CNV	: Copy Number Variations
DNMT	: DNA Methyltransferase Enzyme Family
FPKM	: Fragments Per Kilobase of Transcripts Per Million Mapped Reads
HCC	: Hepatocellular carcinoma
HMM	: Hidden Markov Model
INDEP model	: INDEPENDANT model (CNV -> DNA Methylation; CNV -> Gene Expression)
LC	: Liver Cancer
MR	: Mendelian Randomization
PEM	: Paired-End Method
RD	: Read Depth
RRBS	: Reduced Representation Bisulfite Sequencing
SAN	: S-Adenyl Methionine
SNP	: Single Nucleotide Polymorphism
SR	: Split Read
SV	: Structural Variation

LIST OF FIGURES

Figure	Page
2.1. Incidence and Mortality Rates of Hcc Worldwide.	10
2.2. The Distribution of Hcc Incidence According to Age Groups in USA (2009-2013).	11
2.3. The Epigenetic Landscape Model by Waddington.	13
2.4 Epigenetic Modifications.	14
2.5. Overview of The Study.	23
3.1. The Workflow of ScTrio-Seq Method.	24
3.2. An Example of Fastq File Format.	25
3.3. Workflow of Tophat-Cufflinks Pipeline.	29
3.4. An Example of Fasta File Format.	30
3.5. An Example Graph of CNV Estimation by HMMcopy Throughout All Chromosomes.	36
3.6. The Promoter And The Gene Body Regions of Each Gene.	36
3.7. Simple Bayesian Network Structure (DAG).	37
3.8. The Three Constructed Bn: A) CEM, B) CME, C) INDEP.	39
3.9. Number of The Detected Genes From RRBS Data of Each Sample.	41
4.1. Gene Expression Distribution in All Samples.	43
4.2. Coverage of CpG Sites in CpG Regions of Each Sample.	45
4.3. Coverage of CpG Sites in Genomic Regions of Each Sample.	45
4.4. Methylation Level of Different Genomic Regions of Each Sample.	46
4.5. CNV Pattern of Some Samples.	47
4.6. Pearson Correlation Coefficient Between DNA Methylation and Gene Expression in Promoter and Gene Body Regions.	48
4.7. Pearson Correlation Coefficient Between CNV and Gene Expression	49
4.8. The Gene List of Group 1 and The Related HCC Studies Reported by Cbioportal	54
4.9. The Gene List of Group 2 (CME) and The Related HCC Studies Reported by Cbioportal	56
4.10. The Gene List of Group 2 (CEM) and The Related HCC Studies Reported by Cbioportal	58
4.11. The Gene List of Group 2 (INDEP) and The Related HCC Studies Reported by Cbioportal	60
5.1. Genes in Rtk-Ras Signaling Pathway, Highlighting FGFR3 (CME), ERRF1 (CEM), And RAC1 (CME) Genes.	63

LIST OF TABLES

Table	Page
2.1. HCC Studies in Literature That Have Used Multi-Omics Integration Methods.	7
3.1. Basic Statistics of RNA-Seq Reads Before and After Trimming	27
3.2. Basic Statistics of RRBS Reads Before and After Trimming	31
4.1. Sequence Information of RNA Sequencing Data Samples.	42
4.2. Sequence Information of DNA Methylation Data.	44
4.3. Genes With Verified BN Models and AIC and BIC Score of The Three Models and Relative Likelihood Values.	51
4.4. Number Of Genes Best Fitted to Each Model and The Evidence Strength According to BIC Scores Only	52
4.5. Genes of Group 2 (CME) That Matched to Pathways in HCC (Reported By Cbioportal).	57
4.6. Genes of Group 2 (CEM) That Matched to The Pathways in HCC (Reported By Cbioportal)	59

1. INTRODUCTION

Single-cell sequencing is an optimized next generation sequencing (NGS) method that can observe omics at the level of individual cells, unlike previous methods which analyze the omics in the level of collection of cells at tissue-level (1). Single-cell sequencing approaches help revealing the features and presence of the outlier cells in tumor tissue by defining intra-tumor heterogeneity and new cell types and states (2). Furthermore, single cell sequencing methods provide a deep understanding about the impact of cellular variability on tissue function (1). In turn, allow us to understand how these cellular dynamic changes influence the entire organism that can lead to complex diseases such as cancer, diabetes, accelerated ageing, and metabolic diseases (3-7). Cellular systems are complex networks that involve interaction of many molecules which take part in physical and chemical processes in order to carry out their biological function. High-throughput technologies such as single cell sequencing enable understanding these systems by allowing us to collect information about their molecular components (8). Many of these technologies collect a large set of specific molecular data “-omics” such as genome (9), transcriptome (10), epigenome (11), and proteome (12). Consequently, to draw a more comprehensive understanding of the biological processes, and diseases etiology, these different omics data have to be analyzed separately and then integrated (8). Multi omics integration methods are considered to be promising method to dissect the dysfunctionality in the biological system that occurs in complex diseases such as cancer, ageing, obesity, and nephrotic diseases (13-15). These integration methods mostly depend on machine learning techniques (16), and network inference in relating different omics are formulated by regression-based analysis (17) including supervised (18), unsupervised (19), mostly regression (20), factor analysis (21) and clustering (22). The omics are associated by a series of regression models that are fitted to take one of the feature as a response variable and the other feature as a predictor variable. This association then can be interpreted as a direct relationship in which, one of the feature can affect or explain the other feature (17).

Here, we propose a Bayesian network (BN) based machine learning method for multi-omics integration in single cell sequencing data. Bayesian Networks are probabilistic graphical models that represents the joint probability distributions in a factorized way (23, 24). BNs are composed of a graphical structure with a set of parameters. BNs are defined as directed acyclic graphs (DAG) consist of nodes and directed edges. Nodes represent the variables while edges represent the causal relationship between those variables. BNs are constructed of a set of conditional independence assumptions between the variables and its non-descendants given its parents (25). The parameters represent the conditional probability distributions between variables that connected directly by edges giving their causal relationship (25).

Cancer is a complex disease with increasing incidences worldwide due to the high growth of the population and environmental exposures (26). According to the World Health Organization (WHO), in 2018, cancer was the second leading cause-of-death disease with about 9.6 million mortalities (27). Although the high advances in in-vitro studies on cancer, still, the general progress of understanding cancer is slow because of complex and heterogenous characteristics of cancer cells (28). Hepatocellular Carcinoma (HCC), one of the most dangerous and deadly cancer types, is the fifth most common and the third cause-of-death cancer type worldwide (29). Because of its heterogeneity and multiple causing factors, the number of effective treatments is very low (30). Employing a variety of multi-omics integrating strategies have the potential to identify novel biomarkers that can lead to promising results in solving heterogeneity and provide a key-insight into the pathophysiology of cancer (31).

In this study, a dataset from a published study of 25 hepatocellular carcinoma single cell sequencing is used (32). Transcriptome, epigenome and genome data of these cells were analyzed on their own. Then, these datasets were integrated using a BN approach. Main aim of this study is to reveal the causal relationship between intratumor genome, DNA methylome, and transcriptome of HCC. We also aimed to identify a BN model specific to hepatocellular carcinoma tumor cells that can be prospectively used as a biomarker. By considering this aspect, we constructed three BN model alternatives of

three-way association involving copy number variation, gene expression, and DNA methylation levels in HCC single cells.

2. LITERATURE REVIEW

2.1 Multi Omics Integration

The advent of next generation sequencing (NGS) technologies led to a large amount of omics data like genomes, epigenomes, transcriptomes, proteomes, microbiome and metabolome data generation (13). Utilizing these omics data has expanded the fields in biology and advanced the understanding of the molecular biological process (33). Earlier methods have been optimized in order to examine an individual omic one at a time (34-36). In spite of the high facility of these methods, they are still unable of providing the whole picture about the characteristic's insight of complex diseases such as cancer. Recently, studies focused on linking different omics to be studied which in turn brought new challenges to the development of statistical methods for integrating multi-omics. Omics data integration have been included in a wide range of research area such as: system microbiology (37), plant system biology (38), genotype-phenotype interaction (39), and system pathology (40).

Cellular systems are complex and regulated in multiple levels, while each of these levels is a complex network that interacts with each other. As a result, when combining different omics to integrate them in order to reveal a biological signature becomes very challenging (41). Thus, many theoretical methods and novel algorithms were developed for multi-omics data integration. The major two methods were used are unsupervised and supervised data integration (8). Unsupervised data integration is a group of methods that deals with data without labeled response variables. Unsupervised methods are matrix factorization methods, network-based methods, Bayesian methods (BN), and multiple step analysis (41). In contrast to unsupervised methods, the supervised data integration methods consider the label of the data (control or disease) and call on a machine training approach in order to evaluate the introduced model. That is, the methods of supervised data integration are built upon information of data that their labels are known (42). Supervised methods are Network-based methods, Multiple kernel learning, and Multiple step analysis (42). Table 2.1 shows the recent studies in multi-omics integration that have used supervised and unsupervised data integration methods.

Bayesian networks (BN) are graphical probabilistic networks that represents the joint probability distribution in a factorized way. Bayesian networks composed of a directed acyclic graph (DAG) and a set of parameters (23, 24). DAGs consist of nodes and edges; nodes represent variables and edges represent the relationship between the variables they link. For example, when a directed edge from node A to B, A node is called the parent variable and B node the child variable. In DAG, a set of conditional independence assumptions are encoded between the variables, that is, “a variable is independent of its non-descendants given its parent” (24). A study by So-Youn Shin (43), have introduced Bayesian network method for multi-omics integration as a causal inference tool alternative to Mendelian Randomization (MR) method. Gutierrez et.al. (44), have used the Bayesian network method to integrate multi-omics data from different cell types. They revealed the causal relationship between genetic variation, DNA methylation and gene expression by proposing a different BN model. They also inferred the passive and active role of DNA methylation on gene expression (44).

Gang Liu et al (45), have used a modified cluster of cluster analysis on CNV, mRNA, DNA methylation, and miRNA data of 265 samples downloaded from TCGA. In their method, they have divided the samples into sub-clusters according to each omic and then divide the whole data into 2 groups according to the features in each sub-cluster. Their results showed that samples are classified into 5 major sub-groups (S1-S5). Each sub-group has its distinct molecular features. For example, S1 had TP53 gene mutation, and an amplification in the 8th chromosome at 8q24. While S2 and S3 had a low expression of TERT gene and telomere hypomethylation. Then, they associated these subgroups with clinical information and survival rate. They found that these subgroups are correlated with gender, alpha-fetoprotein level, alcohol intake, American Joint Committee on Cancer staging level. They found that S4 and S5 have more females than other groups, while in S1 and S5, most patients were involved in alcohol intake. They showed that this method can help in solving HCC heterogeneity by classifying it into subgroups.

Miao et al. (46), have integrated WGS and transcriptome sequencing of a set of intrahepatic HCC lesions, matched noncancerous liver tissue and blood. In the study, they

performed phylogenetic tree to differentiate between tissues according to somatic mutations, CNV and SV. After splitting the samples into groups, they performed functional enrichment analysis. They profiled tumor biomarkers that distinguish between two multifocal HCC types. They also showed that TTK protein as a prognostic marker for HCC.

A study by Yıldız et al. (47), showed that HCC distinct cell types have different responses to the same drug therapy. In this study, he analyzed 14 different HCC cell lines (7 epithelial-like and 7 mesenchymal-like cell lines) treated with 225 different small molecules downloaded from the Genomics of Drug Sensitivity in Cancer database. He performed unsupervised hierarchical clustering analysis dividing them into 2 distinct groups according to their responses to the drugs. The first group (group A) which contains early- stage epithelial like HCC cells, have shown more sensitive response to the drug. On the other hand, the second group (group B) which consisted of late-stage mesenchymal-like and epithelial-to-mesenchymal transition HCC cells, had less sensitive response. Moreover, the study showed that mTOR and P13K targeting drugs are more effective to treat epithelial-like HCC cells compared to mesenchymal-like HCC. Concluding that, more sensitive and personalized drugs are needed to be developed.

Yongmei Li et al (48), by analyzing three HCC cell lines with different metastatic potential. They used whole exome sequencing to detect somatic mutations and CNV detection, microarray for transcriptome, and a high-resolution Q Ex-active mass spectrometer for protein quantification. They performed weighted correlation network analysis (WGCNA) in order to cluster the highly associated genes or metabolites. They found that 32 metabolites were decreased, and 21 metabolites increased along with the ability of metastasis. Furthermore, they indicated that there is a relationship between metabolome and metastasis. That is, three metabolic pathways were observed to be altered in different levels such as glycolysis, that shown to have a role in premetastatic in HCC tissue.

Table 2.1. HCC studies in literature that have used multi-omics integration methods.

Study	Sample size	Omics	Integrating method	Reference
Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas	25 single HCC cells	CNV, DNA methylation, transcriptome	Unsupervised hierarchical clustering	Hou et al, 2016 (32)
A Multi-Omics Approach to Liver Diseases: Integration of Single Nuclei Transcriptomics with Proteomics and HiCap Bulk Data in Human Liver	4282 single nuclei HCC	single-nuclei RNA-seq, proteomic	k-mer clustering	Cavalli et al., 2020 (49)
Microenvironment characterization and multi-omics signatures related to prognosis and immunotherapy response of hepatocellular carcinoma	1000 HCC sample	Immune (tumor microenvironment)	unsupervised clustering	Liu et al., 2020 (50)
Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma	125 HCC sample	WES (somatic mutation), CNV,	Cox proportional hazards regression models	Guishard et al., 2012 (51)
Diverse modes of clonal evolution in HBV-related hepatocellular carcinoma revealed by single-cell genome sequencing	96 tumor cells 15 normal cells	Single cell genome wide sequencing, CNV, transcriptome	Multi-dimensional scaling analysis	Duan et al., 2018 (52)
Multi-omics Integration Reveals the Landscape of Prometastasis Metabolism in Hepatocellular Carcinoma	3 HCC cell lines	CNV, transcriptome, somatic mutation, metabolome	weighted correlation network analysis (WGCNA)	Li et al., 2018 (48)

Table 2.1 (continues)

Integrated multi-omics data analysis identifying novel drug sensitivity-associated molecular targets of hepatocellular carcinoma cells	14 different HCC cell lines	Molecular treatment values, transcriptome	unsupervised hierarchical clustering	Yildiz, 2018 (47)
Identification of prognostic biomarkers in hepatitis B virus-related hepatocellular carcinoma and stratification by integrative multi-omics analysis	174 HCC datasets (different cell lines)	Somatic mutation, CNV, transcriptome,	phylogenetic tree	Miao et al., 2014 (46)
Integrated Multiple “-omics” Data Reveal Subtypes of Hepatocellular Carcinoma	265 HCC samples (TCGA data)	CNV, transcriptome, miRNA, DNA methylation	Cluster of Clusters (COC)	Liu et al., 2016 (45)

2.2 Hepatocellular Carcinoma

Cancer is one of the most puzzling and dreaded diseases in the 21st century as it continues increasing every day without the discovery of an effective cure (53). According to Mackay J. et al. (54), one out of four people is at risk of cancer during lifetime. Among different types of cancer, liver cancer (LC) is the sixth most common cancer type and comes at the second place as the most common death causing cancer type worldwide (55). Liver cancer can be divided into two groups: primary LC and secondary LC (56). Primary LC which includes Hepatocellular carcinoma, Sarcoma and Cholangiocarcinoma, starts in the liver (56). On the other hand, secondary LC starts in another organ such as breast, colon, and pancreas in most cases and then metastasizes to the liver. Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer and fifth most common cancer type worldwide (45, 57). It is ranked as the second death-causing cancer type just after lung cancer all over the world (58). According to American Cancer Society (59), in

the USA about more than 40000 people have been diagnosed with hepatocellular carcinoma and about 30000 people died because of it in 2019.

Globally, HCC incidence rates differ from region to region, it highly occurs in eastern and south-eastern Asia (e.g., China, Vietnam, South Korea), Central and western Africa (e.g., Egypt, Senegal) (60). Most of HCC incidences are from countries with low to middle income countries (61, 62). Figure (2.1) shows the distribution of HCC cases worldwide and how the numbers are high in eastern Asia and central Africa where the medical systems are behindhand (54). According to “Cancer Statistics in Turkey” annual report of the health ministry of Turkey (63), HCC is not among the top five cancer types in Turkey which means that Turkey has a low frequency of HCC incidence . Still, a study done by Alacacioğlu et al., determined that Turkey has a high number of HBV and HCV infection rates, the major causes of HCC (64).

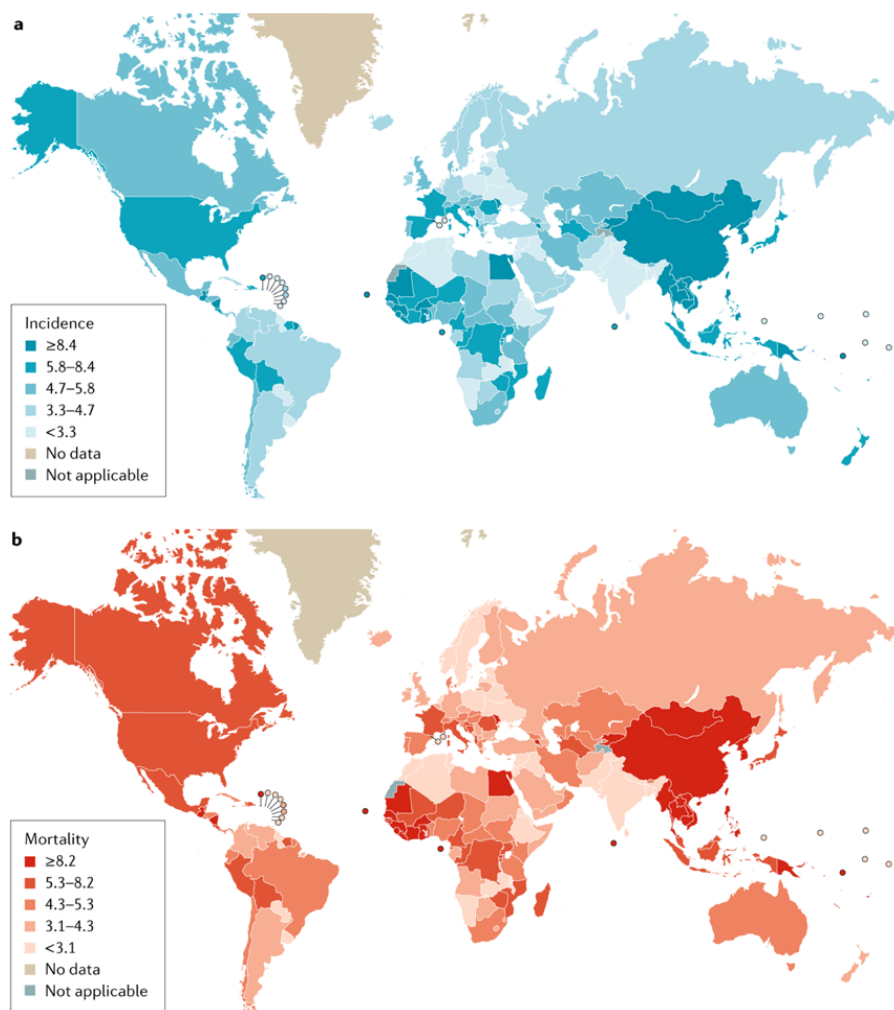


Figure 2.1. Incidence and mortality rates of HCC worldwide. Adapted from (60).

HCC occurs more in men than women (59). In a study included 963 HCC patients from 13 cities in Turkey, found that about 80% of patients were males (65). Moreover, HCC incidence shows a discrepancy with age (66). In the USA, the age group [55-64] has the highest risk of HCC (Figure 2.2), in China it is the [55-59] age group, while in North America and Europe [63-75] is the mean age to be diagnosed with HCC (62). HCC has several risk factors; cirrhosis, chronic infection with hepatitis B virus (HBV) and hepatitis C virus (HCV) are the most common ones, which occupy about three quarters of all cases (67). Besides, alcohol consumption, smoking, arsenic, non-alcoholic fatty liver diseases (NAFLD), obesity and aflatoxin contact dietary habits can also cause HCC (68).

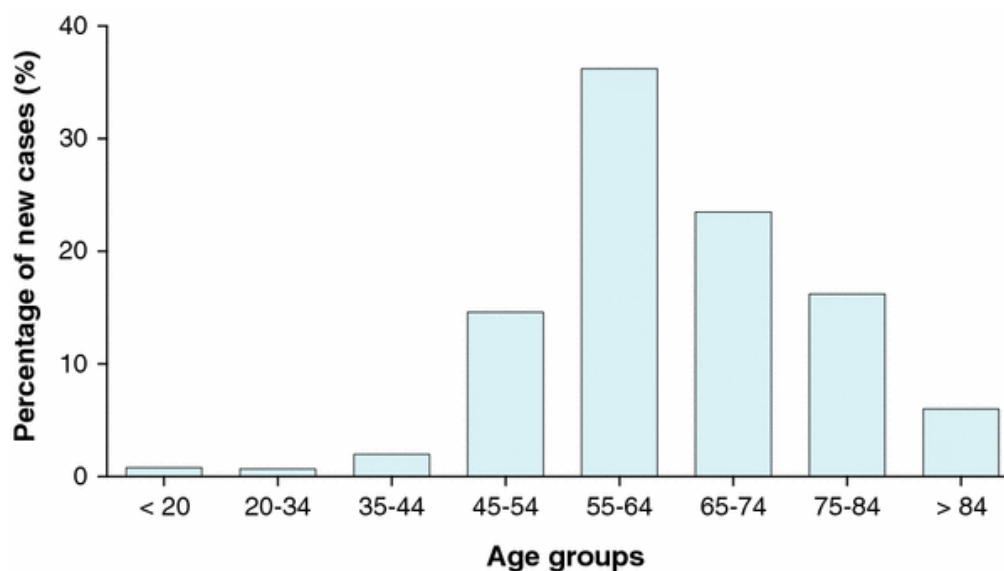


Figure 2.2. The distribution of HCC incidence according to age groups in USA (2009-2013). Adapted from (69).

2.2.1 Clinical Signs and Symptoms

In most cases, HCC is diagnosed after the tumor shows symptoms and clinical signs (29). These symptoms are usually manifested several months after tumor development (29). HCC symptoms and clinical features are similar to those in other hepatic diseases. Makes it hard for physicians to distinguish HCC from an advanced liver disease (70). Some of these symptoms are nausea, weight loss, bleeding, infections, and hepatomegaly (69). As previously mentioned, most of the time it is hard to perform an early diagnosis for HCC (70). Thus, the survival rate is short with an approximate period of [6-20] months. In the USA, more than half of HCC incidence had a survival rate less than 2 years (71).

2.2.2 The Molecular Landscape

HCC is an angiogenic tumor characterized molecularly by the dysregulation of the cell cycle and apoptosis evasion, which both play a vital role in tumor metastasis (68). Furthermore, the tumor cells of HCC go into multiple molecular disruptions such as chromosomal aberrations, genetic alterations, epigenetic changes and molecular pathway shifting (72, 73).

In a comprehensive study, as a part of the cancer genome atlas network (TCGA), performed a large-scale analysis on HCC samples from multiple platforms (74). They analyzed the CNVs and somatic mutations of 363 cases. Besides, almost for half of the samples they analyzed the DNA methylation, mRNA, miRNA, and protein expressions. In general, they observed an amplification in 1q and 8q and a deletion in 8p and 17p. Moreover, they identified a few mutated genes that are associated with HCC tissues and obtained a global hypomethylation in HCC tissues. By using unsupervised hierarchical analysis, they clustered the samples into three groups according to their genomic-epigenomic characteristics (75).

2.3 Epigenetics

By 2003, the completion of the human genome project provided us with a complete list of genes that opened the door to resolve the complexity of the human body (76). Nevertheless, the situation was more complex, that is, there is a second system in the cell with equal importance to determine which, when, and where a gene or multiple genes to be expressed during development (76). This system “Epigenetics” affects the DNA in a form of heritable marks during the division of the cell without altering the DNA sequence (77). In 1942, the term of epigenetics was introduced by Conrad Waddington (78, 79). Waddington defined it as "the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being" (80). Waddington is best known for his "epigenetic landscape model" (79, 81). During his study on embryonic development, he represented his model as a concept to illustrate the different pathway that the cell can take toward differentiation. Figure 2.3 shows “The Epigenetic Landscape Model” where the ball represents the cell, at the top of a slope. The slope has many valleys and hills which represent the genes and other regulations. As the ball rolls down the slope, these hills and valleys will direct it into different (differentiation) paths (80). Since then, with the development of genetic field, term of "epigenetics" has been modified and narrowed down gradually to become more specific (82).

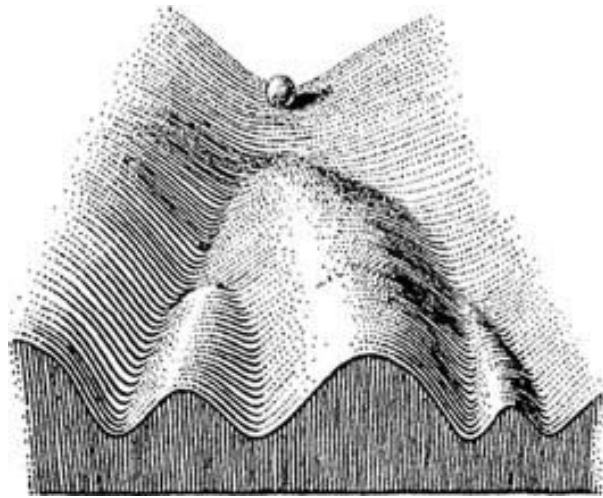


Figure 2.3. The Epigenetic Landscape Model by Waddington. Adapted from (83).

Epigenetics has been defined as “the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence” (82). The epigenetics modifications are DNA methylation, post-translational gene silencing (non-coding RNA), and histone modifications (Figure 2.4). These modifications can be affected by many factors such as age, environmental factors and lifestyle (84).

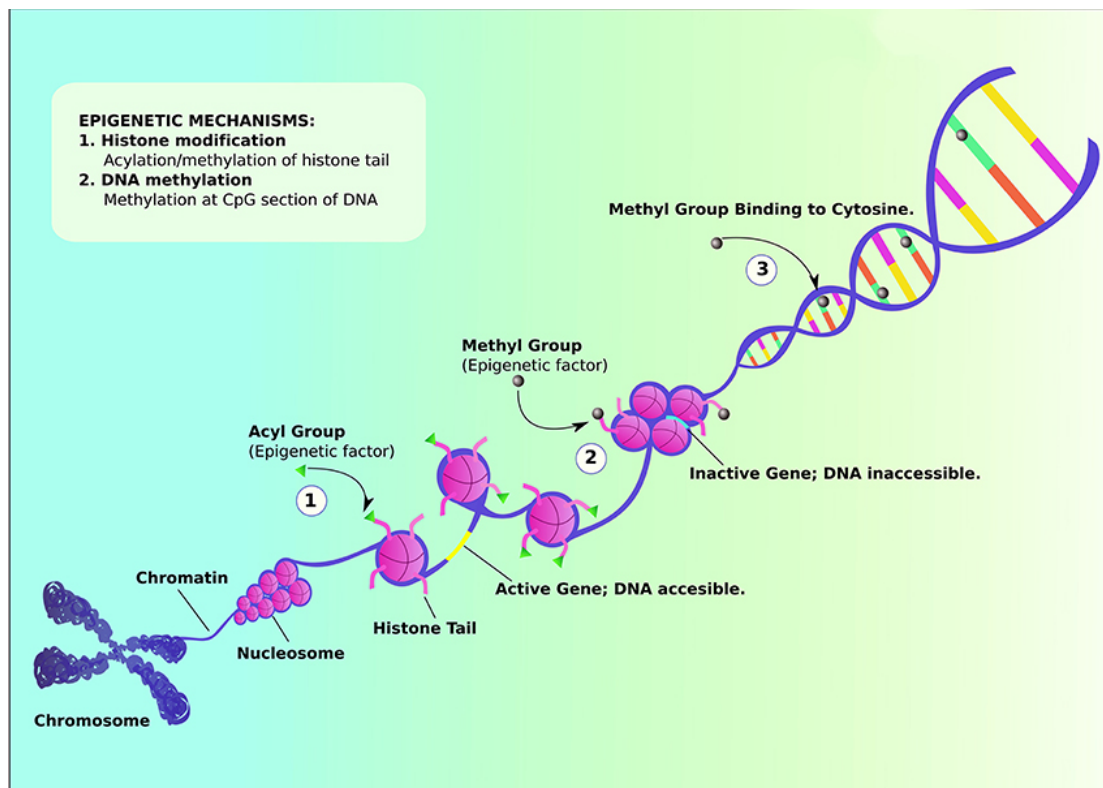


Figure 2.4 Epigenetic modifications. Adapted from (85).

2.3.1 DNA Methylation

DNA methylation is the major epigenetics mark in mammals. It is involved in gene expression regulation and cell differentiation (82, 86). It is a chemical process when a methyl group is added to cytosine nucleotide (C) residing next to guanine nucleotide (G), what is called CpG site, by one of DNA methyltransferase enzyme family (DNMT). DNMTs transfer a methyl group from S adenylyl methionine (SAM) to the fifth carbon molecule of the cytosine base from 5mC. Although DNA methylation occurs at CpG sites, most CpG rich regions “CpG islands” are unmethylated (87). CpG islands are DNA stretches about 1 kb long containing a higher density of CpG than other regions but often they are not methylated. About 70% of gene promoters lies in CpG islands. The majority of these promoters belong to housekeeping genes (88). For its importance in regulating gene expression, CpG islands are thought to be conserved during evolution (86). In a study, Illingworth applied DNA chromatography and Chip-seq techniques on human

(whole human semen, whole male blood, whole female blood), and mice (whole male blood, whole female blood) samples to identify the CpG islands in the genome of both species. He showed that the CpG islands that are linked to gene promoters are highly conserved between human and mice (89). CpG islands are related to other epigenetic modifications, that is, DNA stretches that contain histone proteins and are associated with nucleosomes are more dismissive of gene expression (7). CpG islands are characterized by having less nucleosomes than other genomic regions so it can be related to enhancing gene expression (90). Through gametogenesis and embryonic development, CpG islands go through multiple methylations. These methylations are associated with stable gene expression silencing (91). With the importance of DNA methylation in regulating gene expression during differentiation and development, this silencing has been linked to gene imprinting (92). Imprinting is the process during gametogenesis when only one of the specific inherited alleles (parental or maternal) is exclusively expressed in the offspring (78).

Despite the fact that CpG islands are associated with gene expression regulation, it is also expected to have tissue-specific patterns (93). Not only CpG islands but rather CpG island shores which have a tissue specific methylation pattern. CpG shores are regions with low CpG density flanking to CpG islands up to 2 kb in length (94). CpG shores first identified by Fienberg and his team (95). They performed genome wide bisulfite pyrosequencing on normal tissues from brain, liver, spleen and colon cancer tissue to obtain the differences among them. They find that the differentially methylated regions between the colon cancer and normal tissues are not located in CpG islands but at surrounding regions (CpG shores) addressing that these shores have a role in alternative transcript regulation (95). Furthermore, CpG island methylation has been linked with X chromosome inactivation. Gartler et al. (96), showed that the gene reactivation by 5-azaCdR treatment on the inactive X chromosome is related to the demethylation of CpG island promoter regions. Attaining the relationship between the gene silencing process on inactive X chromosome and DNA methylation. Another study by Mohandas et al., showed that an inhibitor of DNA methylation called 5-azaCdR can reactivate the inactive X chromosome (97). In addition, a study on X chromosome, Wolf and his team showed that

CpG dinucleotides clusters (CpG islands) on the inactive X chromosome are being specifically methylated (98).

Cancer was the first disease to be linked to epigenetics changes (99). As mentioned above, in normal human cells methylation happens at CpG sites that are not in CpG islands (100). But, in cancer cells the CpG islands near the promoter region become highly methylated resulting in turning off some of essential genes such as tumor-suppressor genes (101). A study by Gutierrez et al. (44), have shown that not only in the promoter region but also the methylation that occurs in the gene body region has an impact on gene expression levels. In Cancer, DNA hypomethylation in oncogenes and hypermethylation in tumor suppressor genes have a role to promote tumorigenesis (102). In HCC, Santella et al., applied genome wide methylation array on 62 HCC tumor tissues (103). Out of 2324 differentially methylated CpGs between tumor and normal tissues, about 70% were hypomethylated and 30% hypermethylated CpG sites. Likewise, another study applied by the same method on 66 pairs of HCC tumors and adjacent non-tumor tissues. They found that most of the differentially methylated CpG site are located in CpG islands while about 17% of them are located at CpG shores.

DNA methylation can be detected using different methods. Array based methods such as 450k and EPIC are widely used. Both methods provide a genome-wide screening and reports a methylation level quantification at single CpG-site level (104). On the other hand, Bisulfite genomic sequencing analysis such as whole genome bisulfite sequencing (WGBS) and reduced representations bisulfite sequencing (RRBS) (105). In these sequencing methods, DNA is treated with bisulfite before sequencing in order to determine methylation patterns. Bisulfite treatment leads to bisulfite-conversion of unmethylated C base into U while keeping the methylated C as it is. After this treatment, the DNA is sequenced by one of the NGS machine. In RRBS method, the DNA first is fragmented by using MspI restriction enzymes that have a restriction site lying in CpG islands providing a more specific coverage at high CpG regions (106).

2.4 Gene Expression

A process when the information from a gene on DNA sequence is used to synthesis a functional product. Most of the time these products are proteins, however some genes code for non-protein products such as tRNA, rRNA, miRNA and lncRNA (107, 108). Gene expression studies usually investigate the increasing or decreasing in the expression levels of a gene (or multiple genes) by measuring the abundance of its transcripts. These investigations often observe the gene responses to a drug treatment (109). mRNA is the intermediate molecule in the gene expression process that carries the needed genetic information for protein synthesis (110). Whenever a gene is expressed (or active), it produces many mRNA transcripts by transcription. So, by assessing mRNA, genetic information of a gene expression can be assessed (107). As a result, gene expression analysis methods seized mRNA as a center of interest.

2.4.1 Transcriptome

Transcriptomics is the study of the complete set of RNA -gene expression- that are produced by a specific cell or tissue (110). Many techniques that have been used for gene expression analysis, such as Northern blot, differential display, serial analysis of gene expression (SAGE) (109). However, these methods have their limitations especially when analyzing the expression of a large number of genes. For example, Northern plot have a limited number of samples to be analyzed at the same time, SAGE have a complex and laborious preparation steps with low sensitive results (111). DNA microarrays overcome these limitations and become the most abundant used technique in gene expression analysis (112). This technique is based on the hybridization of two DNA strands when the complementary sequences form hydrogen bonds between their nucleotide's base pairs (112). mRNA molecules that will be analyzed are first chopped into smaller stretches by restriction enzymes and then converted into cDNA by reverse transcriptase. Thereafter, fluorescent markers are attached to these cDNA strands. On the other hand, a large number of DNA sequences of known genes are attached on a chip. By treating these chips with fluorescent-attached cDNA, the complementary sequences of cDNA and the DNA on the chip will bind together and the fluorescent molecule will shine as a spot on the screen.

The strength of the signal from the spot shows the amount of the binding cDNA molecules giving the expression amount of that gene. Using different fluorescent colors differentiates between the studied cases. The major advantages of DNA microarray are: can be used with a large number of gene sets, can be used with a small amount, high sensitive, comparing multiple cases and conditions at the same time (113). However, microarray also has its limitations. For example, DNA microarray cannot be used to observe novel gene expression but only with genes that are known previously (114).

RNA sequencing method which is one of NGS techniques have revolutionized gene expression analysis (110). Being able not only to detect novel genes but also its ability to observe the alternative splicing, post-translational modifications, exons identifications, mutations (SNPs) and the gene expression differential in different groups and cases in a shorter time (115). RNA-seq protocol starts with cut the RNA molecules into fragments. Then, in a step called cDNA library preparation, cDNAs are generated from RNA fragments by reverse transcriptase, and finally sent for sequencing by a sequencer machine (115, 116).

2.5 Copy Number Variation

Copy Number Variation (CNV) refers to any duplication (gain) or deletion (loss) in a genomic stretch greater than 1 kilobase in size (117). Changing the copy number of a gene results in overexpression or deletion of that gene. CNVs are one of the major genomic alterations in cancer cells (118). The role of copy number variation in human diseases has gained a big interest which led to an advent in CNV recognition methods. This realization first started when CNV was subjected to the approximately 12% of the genome variation in the human population (119). In vitro methods are the most widely used methods in CNV detection studies. Array comparative genomic hybridization arrays (aCGH) have been tools of choice beside SNP microarray (120). Array CGH depends on the principle of hybridization of two samples differentially labelled to a set of targets and the signal ratio determine the CNV. SNP microarrays perform in the same principle but differ in that their probes are designed to be specific to SNPs (121).

With the development in NGS technologies, researchers have turned to statistical-based methods operating on NGS data such as whole genome, whole exome and targeted sequenced data for CNV calling. Min Zhao et al. (122), have categorized CNV detection tools into 5 approaches: paired-end (PEM), split read (SR), read depth (RD), de novo assembly (AS), and combinatorial approaches (CA).

PEM method is the first method to be used in NGS-based CNV calling (123). PEM model is only available for paired-end reads and not for single-end reads (123). In paired-end sequencing, insert size of DNA fragments have a specific distribution. PEM methods detect CNV from the inconsistently mapped reads that have significantly different distance from the predefined average insert size. It is able to detect not only CNV but also inversions, mobile element insertions and tandem repeats duplicates (122). Still, its dependency on fragments insert size makes it unable to detect CNVs larger than that insert size (124). PEM method depends on two approaches: clustering and model-based approaches. In the model-based approach a probability test is applied to discover the discordant distance between the distance distribution in the genome and the read-pairs. On the other hand, in cluster approach the predefined distance between reads is used to identify the inconsistent reads (124). Breakdancer which is one of most used CNV calling tools are based on PEM uses both model-based and clustering approaches enable it to detect small CNVs between (10-100 bps) (125).

SR methods are applied to paired end reads that detect CNVs when one of the read pairs is uniquely aligned properly to the reference genome while the other is unmapped or partially mapped (122). The discordant mapping gives a proper breaking point for CNV detection. The incompletely mapped reads are then split into fragments. The first and the last fragments are then aligned independently to the reference genome. In this remapping step the start and end position of the CNVs are precisely detected (126).

RD based methods became the major method for CNV calling because of the high read coverage that NGS data provide (127). The RD approach depends mainly on the concept of the read depth of a genomic region correlated with the copy number of that region. For example, a genomic region with a gain CNV has higher depth and the deletion has less depth than expected (128). Compared to SR and PEM methods, RD based

methods have the ability to detect the exact CNV and also it can detect large CNVs in complex regions. RD method is the major method to call CNV from whole genome sequencing (WGS) and whole exome sequencing (WES) data (129). In WGS, full variants in the whole genome can be determined, while in WES only protein-coding sequences (exons) are targeted for sequencing which enables it to result in a higher regional read coverage (130).

Depending on the study design, RD based tools generally detect CNV in three different ways: single sample, case-control samples, and cohort of samples (122). In a single sample case, CNV is estimated after applying a statistical model such as HMM and Gaussian process on the read depth distribution to detect regions with abnormal read depths. In the case-control samples studies, the control samples serve as "reference" and CNV is detected when read depth of a region in case sample is matched to that in the control. In the case of multiple samples, the CNV detection is calculated by taking the overall read depth mean from all samples which result in estimation of the inconsistent CNV in each sample (124). GATK CNV-caller provides two modes for CNV estimation: cohort and control-case mode. Control-FREEC is able to call CNVs from WGS and WES data with or without control samples (131).

The process of CNV detection using RD based methods is done through 4 steps: mapping, normalization, copy number estimation and segmentation (132). During the mapping step, after mapping short reads to the reference genome, the read depth is calculated in a defined window according to the mapped reads in that window. In the normalization step, the bias in read depth that is caused by GC content or repeat regions is corrected and normalized in order to estimate CNV. In the segmentation step, the regions with the same copy number are merged into segments and segments with discordant copy numbers are detected. HMM is one of the statistical models that are used to detect the CNV state in segmentation step (132). HMM is a probabilistic model which is used to determine the state of an unknown sequence based on the sequence of observation. The transition from a state to another is described by a matrix of probabilities (132).

From a mathematical view, after the normalization step the NGS data is similar to aCGH data. So, some algorithms that are used for CNV detection from aCGH data can be used to detect CNV from NGS data (122). For example, the statistical model Circular Binary Segmentation (CBS) was first method used to detect CNV from aCGH data by converting the noises into equal copy number segments (133). This method was then used to detect CNV from NGS data by SegSeq tool (133).

From short reads, AS based methods reconstruct DNA fragments (contigs) by assembling overlapping reads (134). Then, the assembled contigs are compared to the reference genome and the regions with inconsistent copy numbers are detected. The process of assembling the short reads without needing a reference is called de-novo assembly (134). The Cortex assembler is one of AS based method tool, that uses Bruijn graphs to collect overlapped reads from multiple samples into one graph. De Bruijn graphs present the overlapping information within a set of samples. These graphs consist of nodes, represent words of k length (k-mers), and edges join these k-mers. The variation between genomes leads to new nodes and edges to occur (135). The edges and nodes are colored differently to differentiate different samples. CNV is estimated when all nodes from different samples are collected to find a bifurcation diagram. The branches that separate different colors show the structure variation such as deletion or insertion (135).

High number of CNV caller tools that are based on the previous mentioned methods and the progress they made (122). Still, these methods failed to detect the whole types of variations in genomic copy number with both high sensitivity and specificity (124). For example, PEM based methods are able to detect many types of structural variation especially the small deletions (>1kb), but they fail to detect the precise number of the copies. While RD based methods have a high ability to detect CNV especially the large ones (>10kb), it fails to detect small CNVs (>1kb) and also perform poorly on complex regions such as translocations and inversions. In contrast, AS based methods have an advantage when a reference genome is not required. as an input and allow the detection of novel mutations, but it fails to detect CNV in repeat and duplicated regions (135). A combination of PEM and RD methods could succeed in detecting CNVs with high range in length in various genomic regions. CNVer is a tool that combines PEM and

RD based methods, detect discordant genomic fragments and breakpoints from inconsistent mapped read pairs. By this way, CNV in complex regions such as repeated and duplicated regions with different length is estimated with a high sensitivity (136). NovelSeq is another combinatorial tool that combines PEM and SR based methods to estimate novel insertions (137).

2.6 Research Objective

HCC is one of the leading death-causing diseases in the world, its incidences have shown an increase in the last decades globally (54, 138). Compared to other tumor types, HCC's detection is harder and most of the incidence can only be diagnosed in advanced stages. Furthermore, because of its heterogeneity and wide etiology, there are no effective treatments for HCC (139). Single cell sequencing technologies have been an advent in cancer studies. It helps researchers in extending their understanding about the heterogeneity of the tumor population by observing its individual cells actions (140).

This thesis consists of four parts: RNA-sequencing, DNA methylation, CNV detection, and use of BNs for integrating. Here, we integrated these omics (genome (CNV), transcriptome, and epigenome) of 25 HCC single cells by applying a machine learning approach, which is BN (Figure 2.5). Dissecting these different omics by finding the causal relationship among them at single cell resolution will extend the understanding the heterogeneity of the HCC population, and it might guide us in HCC prognosis and give more information about the disease etiology.

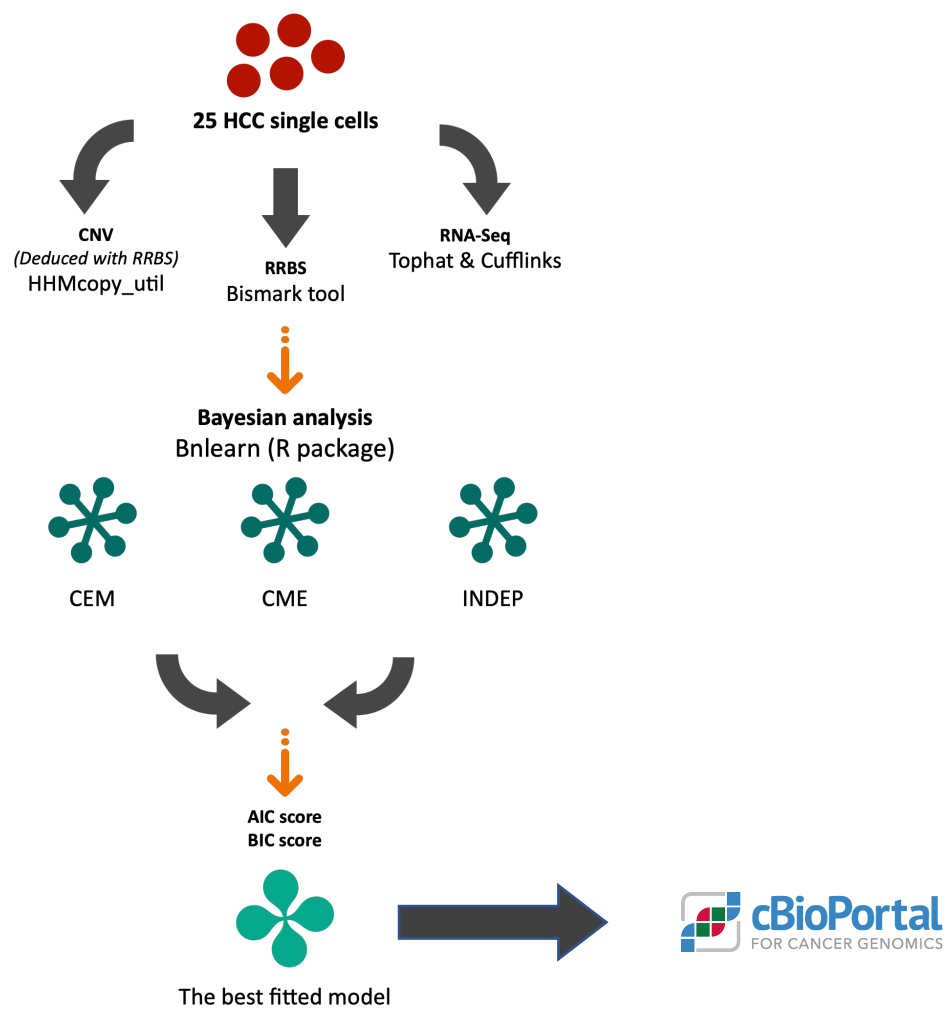


Figure 2.5. Overview of the study

3. MATERIAL AND METHODS

To explore the causal relationship among the transcriptome, genome (CNV), and epigenome in HCC single cells, a publicly available single cell sequencing data by NCBI Gene Expression Omnibus (GEO) under accession code (GSE65364) was used. The dataset contains 25 HCC single cell sequencing data generated by scTrio-seq technique (32). The pathological report showed that the tumor tissue has necrosis, extensive degeneration, and HBV related cirrhosis. ScTrio-seq is a new multi-omics single cell sequencing method developed by Yu Hou et al., (32) aims to analyze the three omics: genome (CNV), DNA methylome and transcriptome of the same cell simultaneously. After dissociating the tissue mechanically into pieces and digesting the cell suspension, the single cell is picked individually by pipetting-by-mouth. Mild lysis is done on the cytoplasm of each cell to release the mRNA only and keeps the nucleus intact. After centrifugation, the precipitate that contains nucleus is sent for DNA methylome sequencing using RRBS method, and the separated mRNA is sent to scRNA-seq after cDNA library construction. By this way, the transcriptome, DNA methylome, and later bioinformatically genome (CNV) data are yielded from the same single cell at the same time. The paired-end sequencing was done using Illumina HiSeq 2000 and Illumina HiSeq 2500 sequencers.

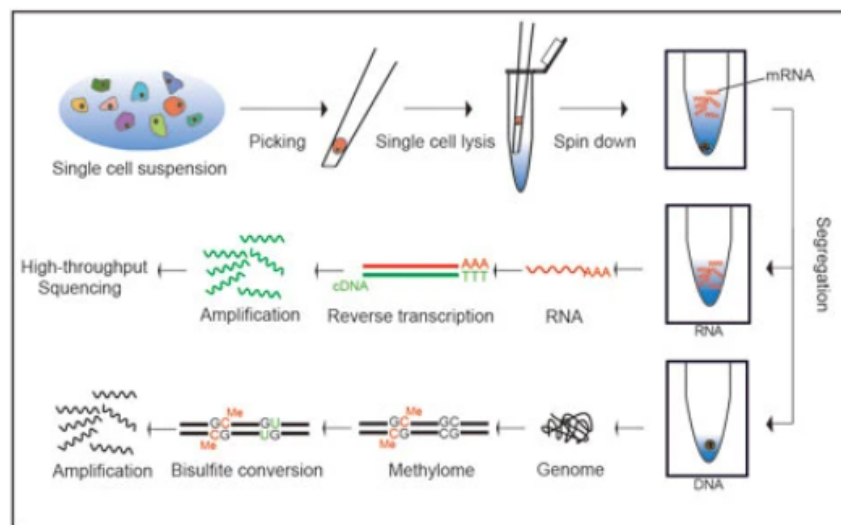


Figure 3.1. The workflow of scTrio-seq method. Adapted from (32)

3.1 Downloading the Data

The data were downloaded from the GEO database using NCBI SRA toolkit (141). SRA toolkit is a collection of tools by NCBI used for downloading and using data from the GEO database. The data were downloaded from GEO database by the command from SRA toolkit:

```
>Prefetch SRRXXX
```

This command will download the samples as SRA formatted files. SRA formatted file is a binary archive file store the raw sequencing reads to the SRA database and it can be converted to fastq by the command:

```
>Fastq-dump --outdir ~/raw_fastq/ --split-files /home/ncbi/puplic/sra/SRRXXXX.sra
```

The command will convert SRA files into fastq and split it into two reads as it is paired-end sample.

3.2 RNA Sequencing data

3.2.1 Data Preprocessing

The tools that are used in data preprocessing and quality control are Linux based tools. The analysis was performed at TUBITAK ULKABIM, high performance and grid computing center (TRUBA Resources).

Fastqc, a popular tool used for check quality control for sequencing reads from Illumina (142), was used to check the quality of raw reads fastq files. Fastq file is a text-based file that contains the short-read sequences and its corresponding quality scores. The sequence letters and quality scores are both coded with ASCII characters. Every read sequence consists of multiple lines: the “identifier” line starts with “@” symbol contains information about the sequence run, the sequence, the separator line with “+” sign, and the line of Phred score for each base (Figure 3.2).

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEEE
```

Figure 3.2. An example of fastq file format

Fastqc can be run for single or multiple samples by the command:

```
> Fastqc sample1.fastq sample2.fastq
```

Fastqc gives a HTML report that contains multiple sections:

- Basic statistics section contains the major information about the sample: the file name, file type, sequencer machine, total sequences, sequence length and GC percentage.
- Per base sequence quality, a graph shows the Phred (quality) score of each base of the read sequence. Phred score is a widely accepted score to measure the probability of a base whether to be called incorrectly (143). Phred score can be represented by the equation (3.1):

$$Q = -10 \log_{10} Pr (\text{observed allele} \neq \text{true allele}) \quad (3.1.)$$

So, for example, Q=20 shows 1% error rate meaning that the base is 99% have been called correctly.

- Per sequence quality score section shows a graph of the mean quality values per reads, allowing you to see if there are universally low-quality values over the sequence.
- Per base sequence content section shows the distribution of the four bases (A, C, G and T) over the read bases which should be equally distributed across the plot.
- Per base GC content section and per sequence GC content section shows the distribution of the GC bases over the read bases.
- Per base N content section shows if there are uncalled bases (N) in the read sequence.
- Sequence duplication level sections provide an informative plot showing the number of sequences that are duplicated in the read sequences.
- Overrepresented sequences and adaptor contamination sections shows the sequences that are highly presented in the reads and the count of them. Also, it gives the possible source of the overrepresented sequences which mostly shows the adaptor contamination in the reads.

The reports showed that there are low quality reads near the 5' end of the sequence. The raw sequencing reads were trimmed to remove the low-quality reads (<Q30) using Trimmomatic (version 0.39) (144). Trimmomatic is a trimming tool including options to trim and filter raw sequencing reads. It is able to identify the adapter sequences and read quality filter. For sequencing data, Q30 is commonly acceptable phred score to keep (145).

```
> java -jar trimmomatic-0.39.jar PE sample_1.fq sample_2.fq trimmed_sample_1.fq
trimmed_sample_2.fq HEADCROP:3
```

Table 3.1. Basic statistics of RNA-seq reads before and after trimming

Sample	SRR No.	Before Trimming			After Trimming		
		Total Sequence	Sequence length	%GC	Total Sequence	Sequence length	%GC
1	SRR1777087	3956277	101	45	3956277	98	45
2	SRR1777089	4005566	101	47	4005566	98	47
3	SRR1777091	3751180	101	46	3751180	98	46
4	SRR1777093	3612890	101	45	3612890	98	47
5	SRR1777095	4074177	101	46	4074177	98	46
6	SRR1777097	7133651	101	44	7133651	98	44
7	SRR1777099	5750626	101	41	5750626	98	41
8	SRR1777101	3445214	101	45	3445214	98	45
9	SRR1777103	3567104	101	44	3567104	98	44
10	SRR1777105	3706688	101	46	3706688	98	46
11	SRR1777107	3864532	101	42	3864532	98	42
12	SRR1777109	6636352	101	46	6636352	98	46
13	SRR1777112	6015670	101	40	6015670	98	40
14	SRR1777114	6365473	101	44	6365473	98	44
15	SRR1777116	7018374	101	44	7018374	98	44
16	SRR1777118	11163622	101	47	11163622	98	47
17	SRR1777120	9983774	101	38	9983774	98	38
18	SRR1777122	6186861	101	46	6186861	98	46
19	SRR1777124	6986571	101	45	6986571	98	45
20	SRR1777126	6607400	101	42	6607400	98	42
21	SRR1777128	5470092	101	40	5470092	98	40
22	SRR1777130	7102096	101	43	7102096	98	43
23	SRR1777132	9066635	101	46	9066635	98	46
24	SRR1777134	1908044	101	43	1908044	98	43
25	SRR1777136	3088062	101	39	3088062	98	39

3.2.2 Gene Expression Quantification

For gene expression quantification we used Tophat-Cufflinks pipeline, which is a combination of software tools used for aligning and comprehensively analyzing the gene

expression from RNA sequencing data (146). Moreover, it performs a comprehensive expression analysis, and it is a widely used tool (147). The Tophat-Cufflinks pipeline is preferred because it provides more accurate expression values, detect a greater number of genes compared to other tools such as HTseq (148), RSEM (149), and STAR (150). Tophat is a tool designed with an efficient read mapping alignment algorithm to align RNA-seq reads to reference genome without relying on a known splice site. Tophat performs the alignment process in two phases: in phase I it uses Bowtie aligner to map all the reads to the reference genome; in phase II, it assembles the mapped reads from Bowtie by using an assembly module, inferring reads into exons and transcripts (146). The pipeline recommended the use of GTF annotation file in order to annotate the results from the pipeline with their gene names (146). GTF (Gene Transfer Format) file is a tab-delimited file that contains information about genome coordinates for the use of genes annotation. There are eight columns in the GTF file: seq-name (chromosome number), source, feature, start site, end site, score, strand, frame and attribute.

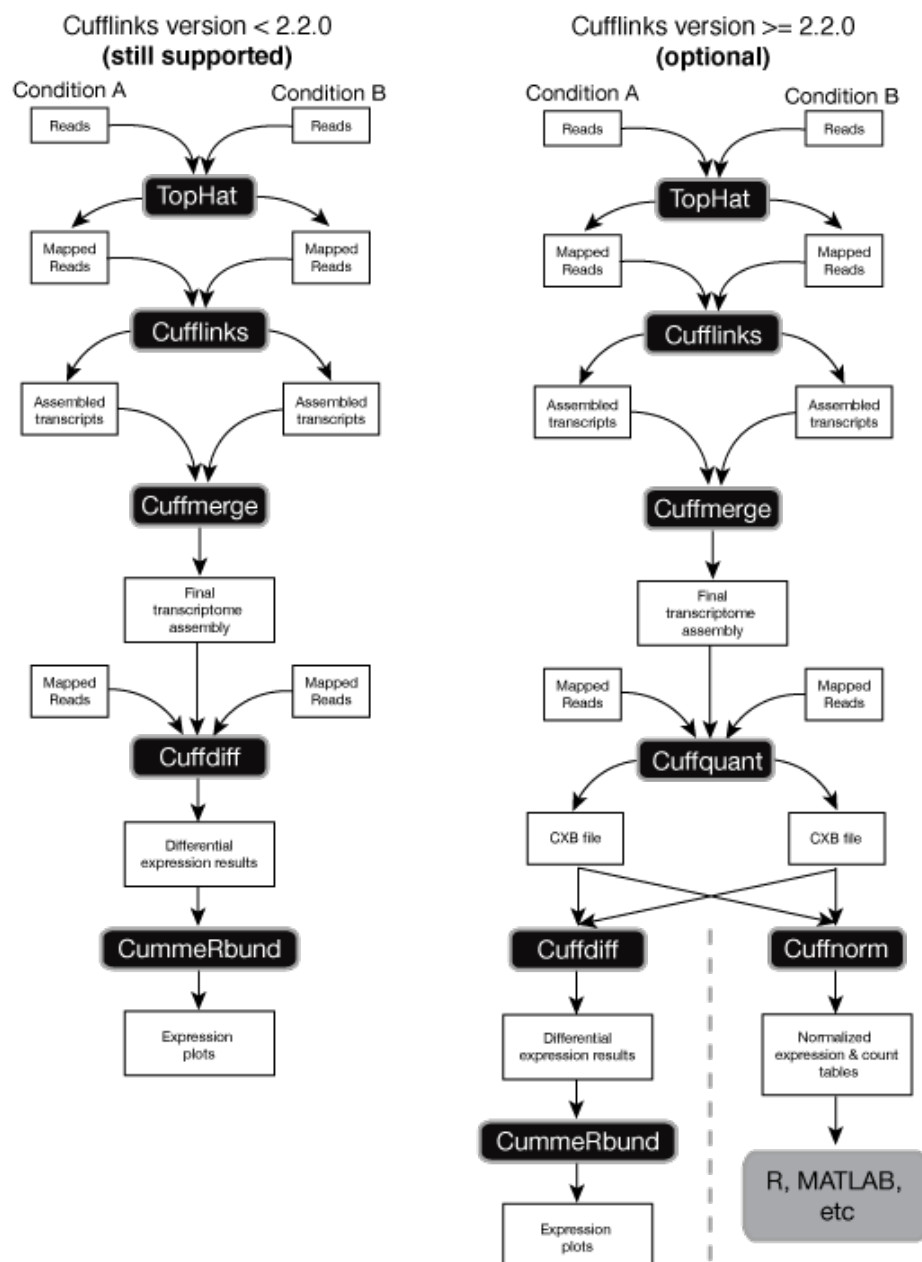


Figure 3.3. Workflow of Tophat-Cufflinks pipeline. Adopted from (146)

The trimmed reads were aligned to the human genome reference (hg 19) UCSC release fasta file (<https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/>) using Tophat (version 2.1.1) (151). Fasta file is a text-based file used to specify the reference sequence of the genome. Each chromosome is represented in two rows, the first one is the identifier

3.3 RRBS data

3.3.1 Data Pre-processing

Firstly, we checked the quality of raw RRBS sequence data (fastq) by using Fastqc following the protocol in section 3.2.1. In this data we detected an adaptor contamination in all samples. Besides other tools such as Trimmomatic (144), Cutadapt (154), and FastX (155), we preferred to use Trim-Galore tool because it has an option that is designed especially for trimming RRBS reads (156).

```
> trimgalore -illumina -rrbs sample.fastq
```

Table 3.2. Basic statistics of RRBS reads before and after trimming

Sample	SRR No.	Before Trimming			After Trimming		
		Total Sequence	Sequence length	%GC	Total Sequence	Sequence length	%GC
1	SRR1777086	5872242	101	36	5682216	101	33
2	SRR1777088	7234041	101	36	6821806	101	33
3	SRR1777090	8233353	101	36	7462253	101	33
4	SRR1777092	3914470	101	38	3195983	101	34
5	SRR1777094	8384707	101	36	8115618	101	33
6	SRR1777096	6205049	101	35	5939977	101	33
7	SRR1777098	6345297	101	35	5951204	101	33
8	SRR1777100	6497647	101	36	6193519	101	33
9	SRR1777102	5692258	101	35	5472350	101	33
10	SRR1777104	5829994	101	35	5664762	101	33
11	SRR1777106	8537614	101	36	7887331	101	33
12	SRR1777108	9067477	101	37	8109880	101	33
13	SRR1777110	6600208	101	37	5983976	101	33
14	SRR1777113	11168512	101	37	10215873	101	33
15	SRR1777115	8893147	101	37	8258172	101	33
16	SRR1777117	10583467	101	38	9905917	101	33
17	SRR1777119	7209732	101	36	6193280	101	33
18	SRR1777121	8122832	101	36	7432772	101	33
19	SRR1777123	7619708	101	36	6714080	101	33
20	SRR1777125	7467918	101	37	6855087	101	33
21	SRR1777127	9687349	101	37	7519129	101	33
22	SRR1777129	7928567	101	38	7388531	101	33
23	SRR1777131	9569206	101	37	8689899	101	33
24	SRR1777133	10643620	101	36	9126636	101	33
25	SRR1777135	8352859	101	38	7711035	101	33

3.3.2 Methylation Base Calling

Bismark (156), BS-Seekers2 (157), GSNAP (158), and BSMAP (159) are the most commonly used tools to align and call methylated CpG sites from RRBS data. We preferred to use the Bismark tool for its high accuracy and high performance in CpG calling (160). Bismark software is a program that is used for aligning bisulfite treated sequencing reads and methylation base calling. Before aligning, the human reference genome (hg 19) was prepared for the alignment by using the tool “Bismark genome preparation”. This tool indexes the reference genome and generates two [C to T] and [G to A] converted genome files. This step is necessary because of the nature of the RRBS method where the unmethylated C bases are converted into Uracil, so converting reference genome makes it compatible to align those reads (156).

Preparing reference genome command:

```
> Bismark_genome preparation -path_to_aligner ~/bowtie2 -verbose ~/hg19/
```

Then, the alignment was done by aligning the trimmed reads (fastq) to the prepared (converted) reference genome, where bam files were generated at the end. During alignment, Bismark aims to find a unique aligned read by running many alignment processes at the same time. First, Bismark transforms bisulfite reads into C to T and G to A. Then, these converted reads are aligned to the pre-converted reference genome. This mapping enables Bismark to identify the strand origin of bisulfite read and to define the methylation accurately in an unbiased way (80).

Aligning command:

```
> Bismark -genome ~/hg19/ sample.fq
```

Thereafter, Bismark is used for calling the methylation level for each CpG site. For this, we used “methylation extractor”. This function is used for calling every single analyzed C base annotated with its location in the genome and its context e.g., CpG, CHH, or CHG.

```
> bismark -o sample.bam ~/hg19/ -1 sample_1.fq -2 sample_2.fq
```

This command produces a “coverage” format file that will be used in the downstream analysis to calculate the methylation level. Coverage file is a text tab-delimited file contains all the CpG sites that are called from RRBS reads, the file composed of the columns: chromosome, position, strand, count methylated and count unmethylated. The methylation levels were calculated by β equation (3.3). The regional methylation level for promoter, gene body, and intergenic regions were determined by calculating the mean of methylation level of all CpG sites in that region.

$$\beta = \frac{\text{Methylated } C}{\text{Methylated } C + \text{Unmethylated } C} \quad (3.3.)$$

In order to reduce the bias and avoid false positives of CpG calls, only CpG bases with depth ≥ 3 sites in promoter or gene body regions were used for downstream analysis (32). Read depth of each base was called using the “depth” function from Samtools. Samtools is a software package used for parsing and manipulating SAM and BAM files such as sorting, merging and PCR duplicates removing (152).

```
> samtools sort sample.bam > sorted.sample.bam
```

```
> samtools depth sorted.sample.bam > sample.depth
```

To annotate CpG sites to the corresponding genes, refseq gene list from UCSC and bedtools were used. Bedtools is a flexible tool that have a wide diverse usage for genome feature analysis such as merging, counting, intersect and complement genomic intervals from bed, sam, and fasta files (161).

3.4 CNV Estimation from RRBS Data

To estimate CNV from RRBS reads, we used HMMcopy at 500kb resolution (162). HMMcopy is a tool using a read-depth (RD) approach for CNV deduction from WGS data. HMMcopy makes a bias-free CNV estimation by correcting mappability bias and GC content in sequencing reads. It implements a hidden markov model (HMM) for a copy number profile segmentation into non overlapping windows that are predicted to have the same copy number state and relate it to the biological CNV events. HMMcopy

takes an input of three WIG format files that are generated by a linux based HMMcopy_util tool, namely GC content, readcount and mappability values for fixed width non overlapping windows across the reference genome (162). WIG (wiggle) file format is a text-based file that displays continuous data such as probability scores, GC content, and transcriptome data. On the other hand, BIGWIG file is an extended WIG file format which facilitates working with big data such as reference genome. The three wigs file were generated by: mapCounter, gcCounter and read counter commands using HMM_util. Briefly, mapCounter tool was used for calculating the average mappability for fixed width non overlapping windows across whole sequences of reference genome in a bigwig file. We have downloaded the bigwig file “wgEncodeCrgMapabilityAlign100mer.bigWig” for hg19 from UCSC genome browser (<https://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/>).

```
> mapCounter -w 500000 wgEncodeCrgMapabilityAlign100mer.bigWig > map.wig
```

gcCounter tool was used for calculating the GC content for fixed width non overlapping windows across whole sequences of reference genome in a fasta file (hg 19).

```
> gcCounter -w 500000 genome.fa > gcCounter.wig
```

readCounter tool was used for calculating the read counts for fixed width non overlapping windows across whole sequences in aligned reads bam file.

```
> readCounter -w 500000 sample.sorted.bam > sample.wig
```

The GC content and mappability values were corrected before performing HMM prediction on the reads. The correction procedure includes:

- Filter out bins with 0 read and 0 GC content, filter out outlier bin, smoothing windows by loess with a small span on GC curve.
- Corrected GC (cor.gc) is calculated by correcting GC content in each bin.

Corrected mappability (cor.map) is calculated by correcting the mappability of each bin. “Copy” is the log₂ of cor.map that will be used in CNV prediction in HMM.

```
> mydata = wigsToRangedData(readfile = "sample.wig",gcfile =
"gcCounter.wig",mapfile = "map.wig")
> corrected_data=correctReadcount(mydata)
```

HMMcopy also plots a GC bias graph and corrected CNV graphs over the genome (Appendix 1).

The prediction procedure includes HMM segmentation function takes in “copy” values and predicts the regions of equal copy number into segments, then assigns a biological copy number state of each region using HMM. HMMsegment consists of two parts; the first part performs iteratively Expectation-Maximization algorithm to find the optimal parameters, the second part is to perform Viterbi algorithm that conducts the actual segmentation of the data and output segmented state (162).

HMMcopy reports predictions as a table containing the chromosomal segment location annotated with corrected mean number, and state number between (1-6):

- 1: homozygous deletion
- 2: heterozygous deletion
- 3: neutral
- 4: increased copy number
- 5: heterozygous duplication
- 6: homozygous duplication

Moreover, HMMcopy provides a graph that visualizes the CNV states and their mean distributions over the whole genome of all chromosomes (Figure 3.5).

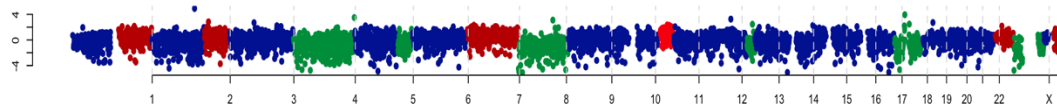


Figure 3.5. An example graph of CNV estimation by HMMcopy throughout all chromosomes.

3.5 Correlation Calculations

To infer the association between the different omics, we used Pearson correlation coefficient (r), which is a commonly used correlation measure with continuous variables (163). We calculated the Pearson correlation coefficient (r) between; promoter methylation and gene expression, gene body methylation and gene expression, gene expression and CNV. The promoter region was defined as 1 kb downstream and 1kb upstream of TSS, while gene body region was defined to be from 1kb downstream of TSS to TES (Figure 3.6). The CNV values that are used for correlation calculations are the corrected mean of segments provided by HMMcopy. While, for gene expression data, $\log_2(\text{fpkm}+1)$ values were used for correlation calculations.



Figure 3.6. The Promoter and the gene body regions of each gene

3.6 Bayesian Networks

Bayesian Networks (BN) are probabilistic graphical models that represent the joint probability distribution in a factorized way (25). A BN composed of a graphical structure with a set of parameters. BNs are defined as directed acyclic graphs (DAG) consisting of nodes and directed edges. It is constructed as a set of conditional independence assumptions between the variables and its non-descendants given its parents. The

parameters represent the conditional probability distributions between variables that connected directly by edges giving their causal relationship (20). The pointing side of the edges shows the direction of the causing. For example, if an arc goes from A node heading to B node, meaning that A is the causing of B. In other words, B is conditionally dependent on A (figure 3.7). The conditional probability of A given B is represented by $P(A | B)$. Formally, the sets of variables A and B are said to be conditionally independent given the set C if $P(A | B,C) = P(A | C)$.

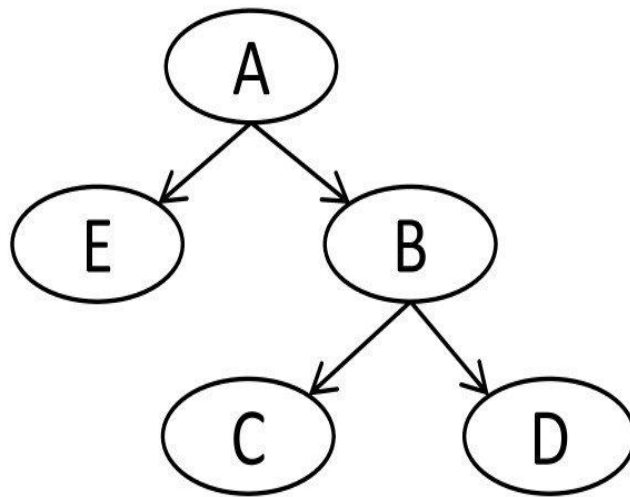


Figure 3.7. Simple Bayesian network structure (DAG).

There are several ways to build a BN model: The structure is defined manually by expert information or the parameters are estimated by using maximum likelihood or Bayesian method (164). The performance of BN models can be tested and compared using scoring methods such as relative likelihood, Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) (165). AIC score is used to examine the compatibility of their structures, and then a relative likelihood approach was used to compare the goodness of fit of one network to another. N1 and N2 are two networks, if $AIC(N1) \leq AIC(N2)$ then the relative likelihood of N2 respect to N1 is:

$$\exp\left(\frac{AIC(N1) - AIC(N2)}{2}\right) \quad (3.4.)$$

Akaike information criterion (AIC) is an estimator score used for estimating the quality of the model on the number of parameters (k) and the maximum likelihood estimation for the model(L) (Equation 3.5).

$$AIC = 2k - 2\ln(L) \quad (3.5.)$$

Bayesian information criterion (BIC) is a criterion for model selection that is very related to AIC score but differs in depending on the number of observations as an extra criterion. It is calculated by the formula (3.6).

$$BIC = k\ln(n) - 2\ln(L) \quad (3.6.)$$

To dissect the causal relationship association among the three omics, we built three BN structures to model the causal relationship between CNV, gene expression and DNA methylation. The first model “CEM”, assumes that there is a causal relation from CNV to gene expression, and from gene expression to DNA methylation in a serial connection (Figure 3.8-A). The second model “CME” assumes that there is a causal relation from CNV to DNA methylation, and from DNA methylation to gene expression in a serial connection (Figure 3.8-B). The third model “Indep” assumes that there is a causal relation from CNV to both DNA methylation and gene expression independently in a diverging connection (Figure 3.8-C).

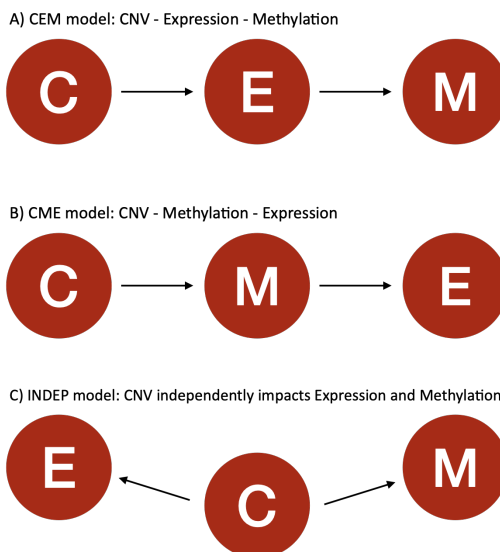


Figure 3.8. The three constructed BN: A) CEM, B) CME, C) Indep

Because BN performs better with discrete values, the data of gene expression and DNA methylation levels were discretized (166). FPKM values were discretized into three categories (low, moderate, and high) expressions. The cutoffs were set based on the same categories used in EBI atlas database (199); (0.2-10): low, (10-1000): moderate, (>1000): high expression. DNA methylation levels were also discretized into three categories (hypo, neutral, and hyper) methylated. The cutoffs were set to be (0-0.2): hypomethylated, (0.2-0.8): neutral, (>0.8): hypermethylation. For DNA copy number, we used CNV states that are provided by HMMcopy. The states ranged between 1-6: (1): homozygous deletion, (2): heterozygous deletion, (3): neutral, (4): gain, (5): amplification, (6): high level amplification.

The parameters of the three models were fitted using maximum likelihood estimation (MLE). BNlearn R package was used for BN model construction, fitting and scoring (164). As well, BIC score was also calculated, and only models that have followed the same model by both AIC and BIC were kept. Moreover, model comparisons were done by using BIC score only. The model with the lowest score was taken to be the best fitted model. The comparison for a model to another were considered by taking the

difference between the two models with the lowest and the second lowest BIC (equation 3.7) and the strength of the evidence were adjusted according to ΔBIC (167) as :

0-2: indicates weak evidence,

2-6: positive evidence,

6-10: strong evidence,

>10: very strong evidence.

$$\Delta\text{BIC} = \text{BIC}_{\text{second lowest}} - \text{BIC}_{\text{lowest}} \quad (3.7.)$$

The models were applied on protein-coding genes. Only genes that have been detected in all cells were chosen for model testing. In order to have as much as possible of genes, we eliminated sample 17 because of the low number of genes detected (8226 genes) from RRBS data, which in turn lowers the number of mutual genes to be tested. Figure (3.9) shows the number of detected genes from RRBS data in all cells, comparing the number of detected genes in sample 17 to the other 24 samples. As a result, 2661 genes were used to test BN models.

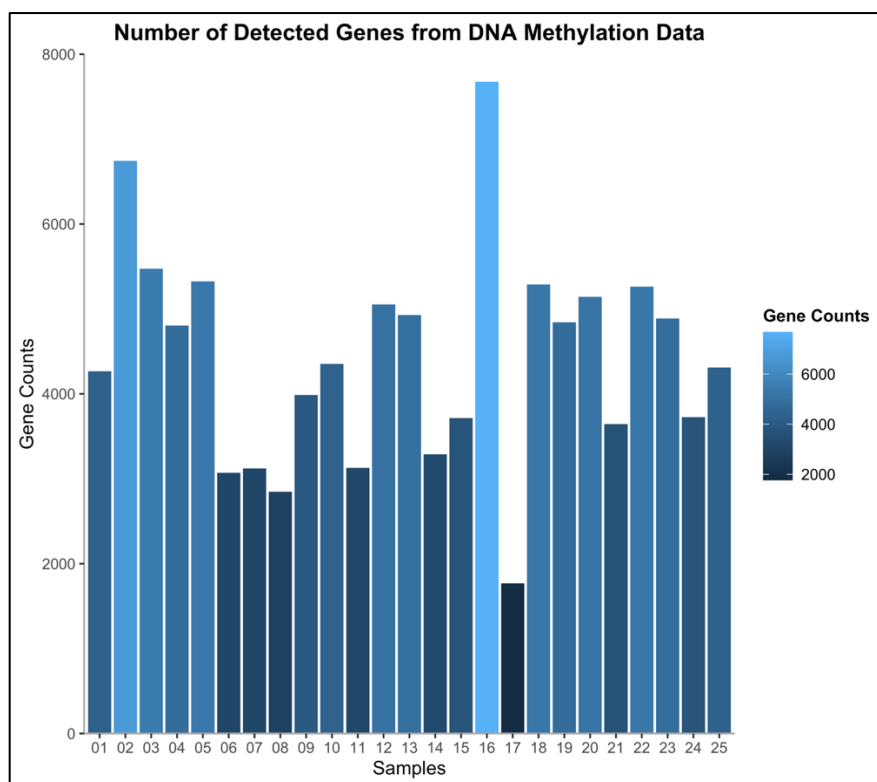


Figure 3.9. Number of the detected genes from RRBS data of each sample.

3.7 Gene Set Enrichment Analysis

3.7.1 CBioportal Database

In order to find whether the detected genes have been reported in HCC before, we searched the genes in cbioportal database (168). Cbioportal is a platform that contains a comprehensive large-scale of cancer genomics data including data from TCGA, ICGC, and published sequence studies from academic and commercial institutions (167). The Cancer Genomic Atlas (TCGA) is the largest and richest cancer data of 200 different types of cancer collected from about 12,000 patients around the world (169). TCGA data are composed of many omics data including DNA sequencing, RNA sequencing, CNV, DNA methylation, and proteomic data. The whole TCGA data are implemented in cbioportal database (168, 170).

4. RESULTS

4.1 Gene Expression Levels

As mentioned in section 3.2, we have used Tophat-Cufflinks pipeline to analyze RNA sequencing data. We annotated the sequences to hg19 reference genome with protein coding Refseq genes. Table (4.1) shows the mapped reads, mapped ratio, and the number of detected genes of RNA-seq data for 25 samples. At the end of the pipeline, 18584 genes were yielded. In order to select the expressed genes that have been detected in all cells, we eliminate both the genes with (NA) value and the genes that have zero value in all cells, which resulted in 15326 genes.

Table 4.1. Sequence Information of RNA Sequencing Data Samples.

Sample	Total number of analyzed reads	Mapped read pairs	Mapping efficiency	Mapped genes (fpkm>0)
1	3956277	1973751	68.0%	5618
2	4005566	2216595	72.2%	8606
3	3751180	1948646	69.4%	6934
4	3612890	1958354	71.6%	9348
5	4074177	2080378	68.3%	7892
6	7133651	3418379	65.5%	4617
7	5750626	2507828	60.0%	3088
8	3445214	1663878	66.5%	5314
9	3567104	1683551	65.1%	6097
10	3706688	1783222	67.0%	6930
11	3864532	1634657	59.9%	4818
12	6636352	3813552	72.8%	7414
13	6015670	2408480	57.0%	4875
14	6365473	3330118	68.2%	3645
15	7018374	3737093	69.8%	6368
16	11163622	7020724	77.1%	7859
17	9983774	4585241	59.7%	5016
18	6186861	3423019	71.8%	7398
19	6986571	4000465	72.3%	6540
20	6607400	3049466	63.4%	5581
21	5470092	2207147	58.0%	2630
22	7102096	3384499	65.9%	7349
23	9066635	4786858	70.2%	7784
24	1908044	905168	64.9%	4007
25	3088062	1213749	55.8%	4033

As a part of the pipeline, the sequencing reads were aligned to hg19 reference genome (UCSC assembly) with Tophat. Next, Cuffquant tool was used to estimate the gene expression of each sample. Then, Cuffnorm tool was used to normalize the values from Cuffquant so that all gene expression values become on the same scale. Cuffnorm tool reported FPKM values of each gene for all samples. For the downstream analysis, we have used $\log_2(FPKM+1)$ values for gene expression levels. Figure (4.1) shows the gene expression ($\log_2(\text{fpkm}+1)$) distributions for each sample.

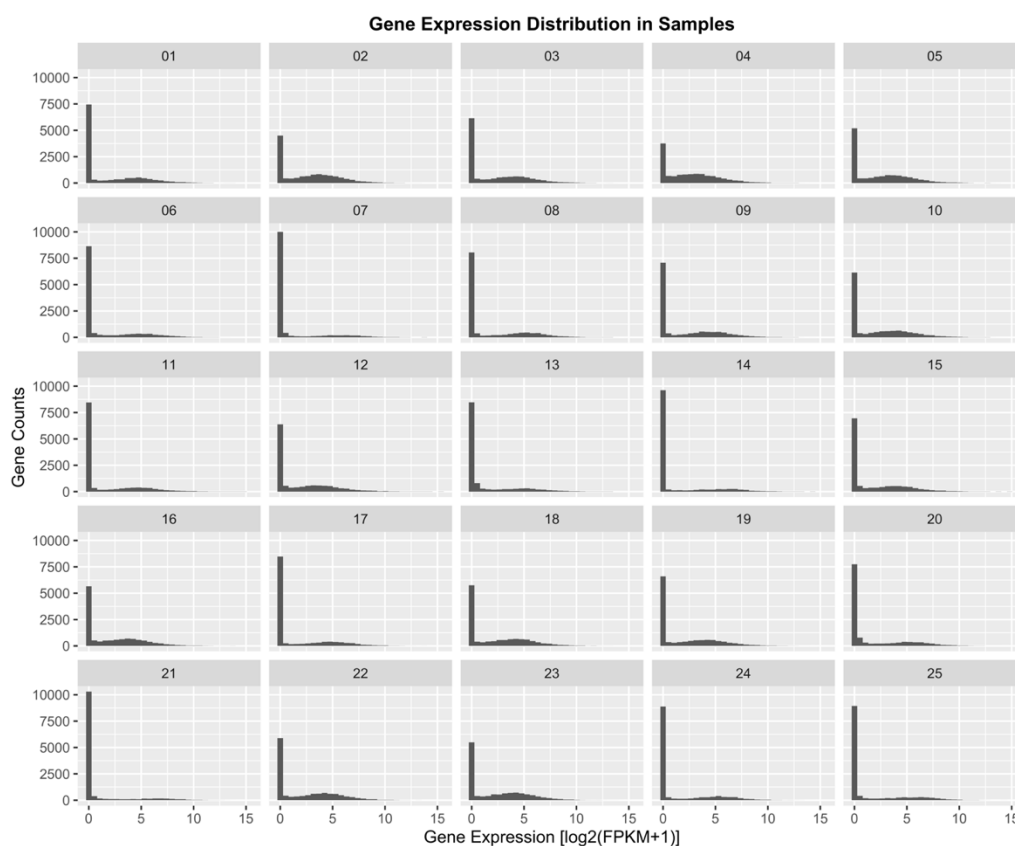


Figure 4.1. Gene expression distribution in all samples.

4.2 DNA Methylation Levels

For DNA methylation analysis Bismark software was used for the alignment and CpG sites calling (previously described in section 3.3) of RRBS data. Table (4.2) shows the sequencing information including the total number of analyzed CpG sites

and the total number of detected genes in each sample. On average, about 13100 genes were detected on all cells. Sample 16 had the highest detected genes with 15818 genes, while Sample 17 had the least detected genes with 8226 genes.

Table 4.2. Sequence Information of DNA Methylation Data.

Sample	Total number of analyzed reads	Mapping efficiency	Total number of C's analysed	Total C's in CpG context (methylated and unmethylated)	Total C's in CpG context (methylated and unmethylated) (depth ≥ 3)	Total number of detected genes
1	5682216	6.0%	16111942	3608134	690810	13851
2	6821806	15.5%	51571705	10532456	1159830	15156
3	7462253	7.4%	26660719	5510918	884191	14515
4	3195983	13.3%	18732458	4211753	1012511	14755
5	8115618	7.6%	30127778	6353989	734467	14006
6	5939977	4.6%	13266640	2492453	465859	12388
7	5951204	4.7%	13471964	2700915	393998	11362
8	6193519	3.5%	10531074	2181766	348531	11509
9	5472350	5.9%	15719599	3328962	749362	14136
10	5664762	6.3%	17649347	3648788	737629	13921
11	7887331	4.5%	17188932	3534484	328956	10998
12	8109880	7.3%	27725731	6020939	637289	13485
13	6050471	10.0%	28002614	5560722	720851	13829
14	10215873	3.5%	16413258	3541333	345575	11282
15	8258172	4.9%	18112819	3463948	435799	12225
16	9905917	13.5%	61163128	13431814	1677730	15818
17	6193280	4.2%	12755078	2510035	185173	8226
18	7432772	7.5%	26231957	5290348	732962	14072
19	6714080	7.7%	24631107	4873392	616350	13493
20	6855087	8.2%	25252120	5218707	780469	14233
21	7519129	5.2%	17807460	3151685	500155	12926
22	7388531	7.9%	26280431	5726413	753525	13991
23	8689899	6.7%	27273514	5966027	613082	13337
24	9126636	4.0%	18124466	3465984	422334	11956
25	9126636	4.0%	18124466	3465984	676406	13652

The results of CpG sites showed that CpG island (CGI) were highly covered, that is, about 50% of the called CpG sites in each sample were related to CGI regions (Figure 4.2). Comparably, promoter regions were highly covered than other genomic regions, that is,

about 45%-50% of the called CpG sites in each sample were related to promoter regions (Figure 4.3).

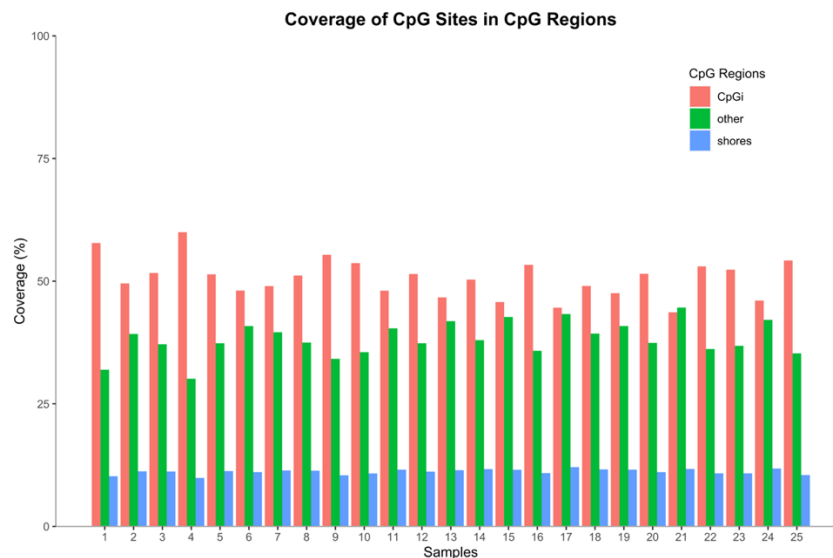


Figure 4.2. Coverage of CpG sites in CpG Regions of Each Sample.

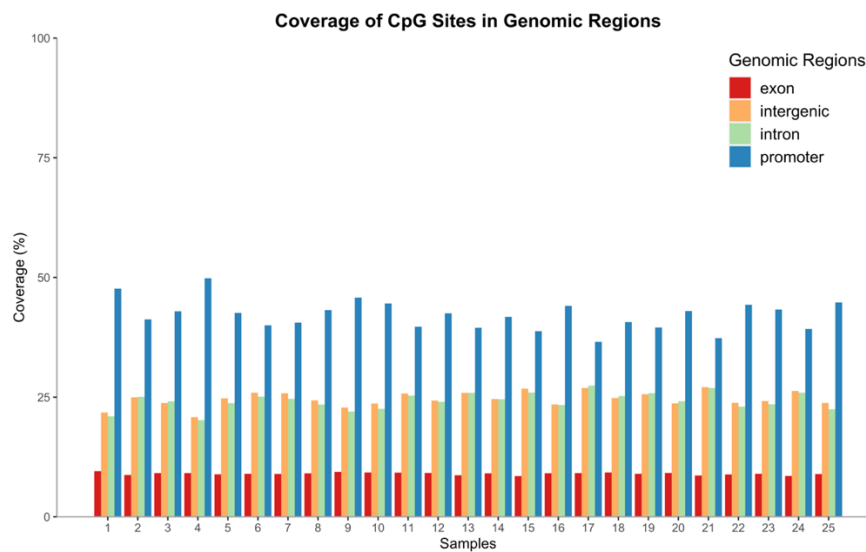


Figure 4.3. Coverage of CpG sites in Genomic Regions of Each Sample.

The results showed a hypomethylation in all 25 samples, especially in promoter regions and a higher methylation level in gene body regions. Figure (4.4) shows the methylation level of all samples in different genomic regions.

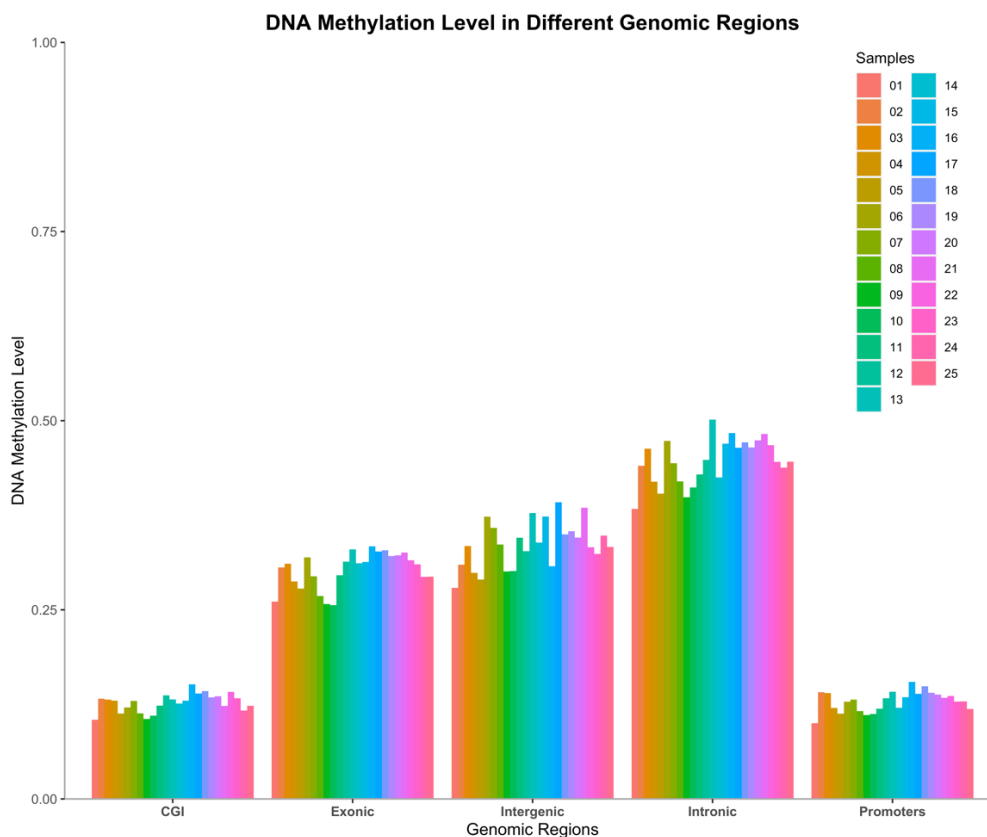


Figure 4.4. Methylation Level of Different Genomic Regions of Each Sample.

4.3 Copy Number Variation

As mentioned in section 3.4, HMMcopy a R package was used to estimate CNVs from RRBS data. From CNV results, all 25 samples showed an amplification in chromosome 7 and q arm of chromosome 1. I also observed a deletion in chromosome 8 and chromosome 4. Moreover, the samples (3,4,6,9,12,13,15,16,18,19,20,21,22,23,25) showed a loss in chromosome 13, while samples (2,6,13,14,15,20,21,22,23) showed an amplification in chromosome 6, Figure (4.5), shows the CNV pattern of genomes of some

samples (all samples are in Appendix 2), the red color represents the amplification, blue represents the neutral, and the green represents the deletion in copy number.

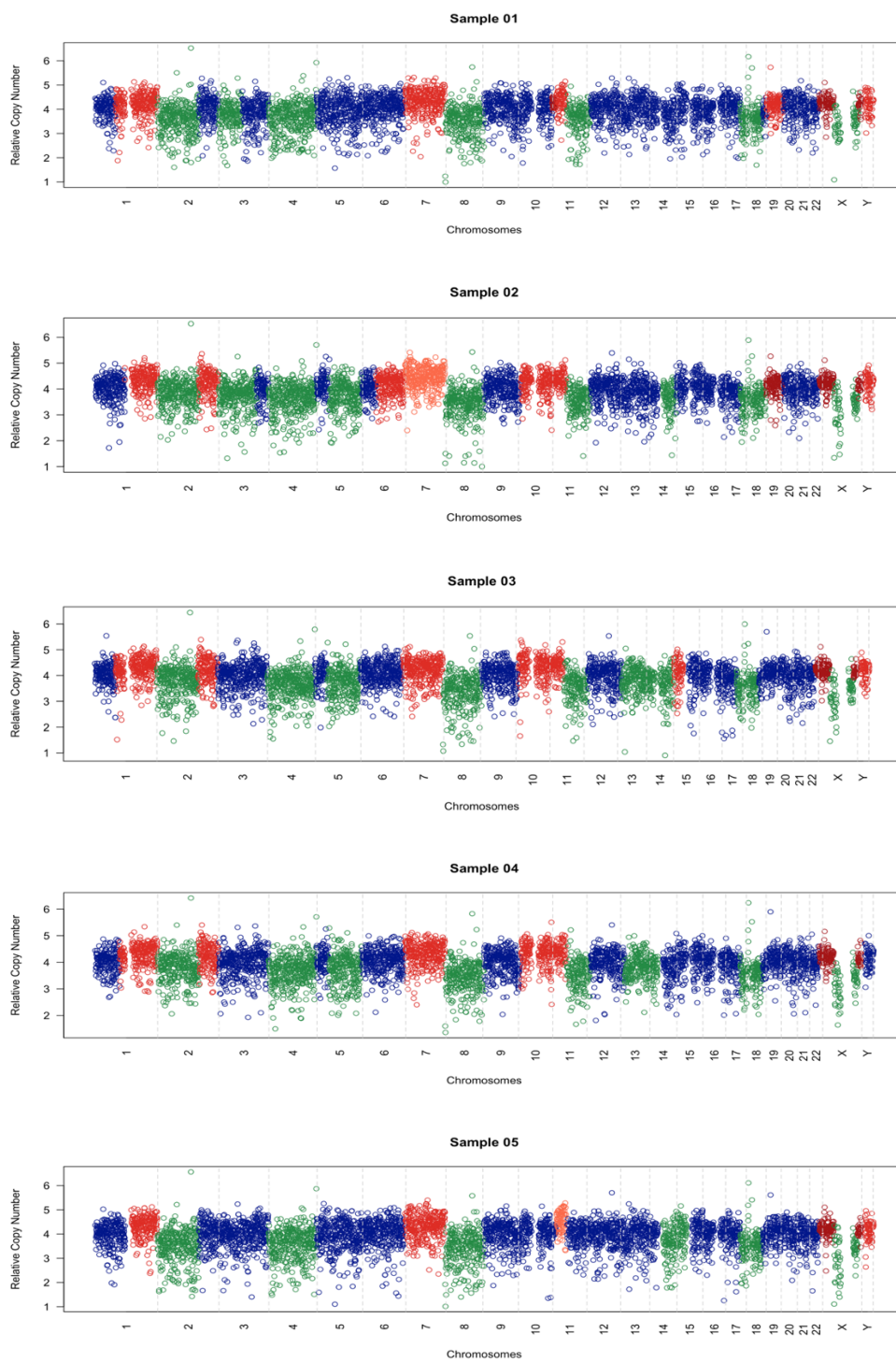


Figure 4.5. CNV pattern of some samples.

4.4 Correlation Between Omics

For the correlation between DNA methylation and gene expression, we observed a negative relationship in the promoter region with Pearson correlation coefficient (r) of -0.1387 ± 0.08 (mean \pm SD). We also observed a positive correlation with Pearson correlation coefficient (r) of 0.3136 ± 0.07 (mean \pm SD) gene body region. Figure (4.6) shows the Pearson correlation coefficient (r) between DNA methylation and gene expression in the two regions (promoter and gene body) of each sample. In this correlation, only genes that have been detected from both RRBS and RNA-seq data were used. For each sample, we found that all genes that have been detected from RRBS were detected from RNA-seq data. So, the number of genes that have been used in this calculation are the same in table 4.2.

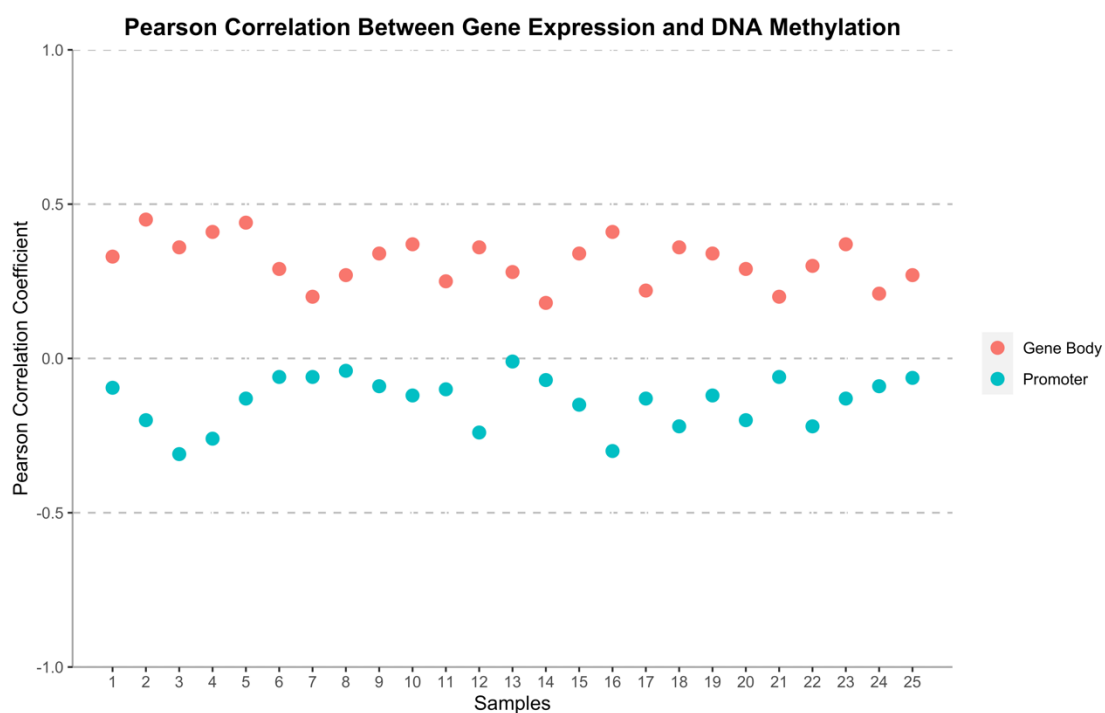


Figure 4.6. Pearson correlation coefficient between DNA methylation and gene expression in promoter and gene body regions.

For the correlation between DNA copy number and gene expression, we found a highly positive correlation with Pearson correlation coefficient (r) of 0.821 ± 0.07 (mean \pm SD).

The DNA copy number that is used in this correlation, is the median of the copy number in each segment that is provided by HMMcopy. The Figure (4.7) shows the Pearson correlation coefficient (r) between gene expression levels and DNA copy number in each sample. This correlation was calculated on 15326 genes.

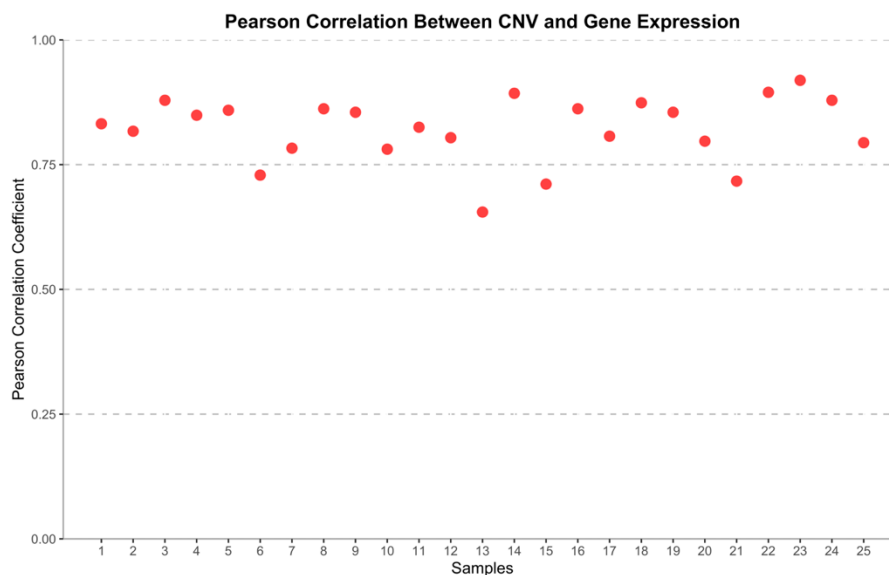


Figure 4.7. Pearson correlation coefficient between CNV and Gene Expression

4.5 Causality Analysis Using BN

To explore the causal relationship between omics, we used Bayesian Network to analyze the relationship between DNA methylation and gene expression by considering that CNV is the trigger of the causal relationship since its state is not modifiable (44). FPKM values were discretized into three categories (low, moderate, and high) expression. The cutoffs were set based on the same categories used in EBI Atlas Database: (0.5-10 FPKM): low, (11-1000 FPKM): moderate, (>1000): high expression. DNA methylation levels were also discretized into three categories (hypo, neutral, and hyper) methylated. The cutoffs were set to be (0-0.2): hypomethylated, (0.21-0.8): neutral, (>0.8): hypermethylation. For DNA copy number, we used CNV states that are provided by HMMcopy. The states ranged between 1 to 6: (1): homozygous deletion, (2): heterozygous deletion, (3): neutral, (4): gain, (5): amplification, (6): high level amplification.

Three BN models were constructed using maximum likelihood estimation. Also, for model scoring, both AIC and BIC scores were overlapped to decide which of the three BN models (described in 3.6 section) is the most likely to represent the data for each set of variables. The three models are INDEP, CME, and CEM (figure 3.9). In “INDEP” model, CNV affects independently DNA methylation and gene expression. In “CME” model, CNV affects DNA methylation, which in turn affects gene expression. In “CEM” model, CNV affects gene expression, which then affects DNA methylation.

The models were applied to protein-coding genes that have been detected to have a variation in gene expression, methylation and copy number throughout all samples. As previously described in section 3.6, sample 17 were eliminated in this analysis because the number of detected genes from DNA methylation data were very low which in turn could affect the number of mutual genes to be analyzed.

The relative likelihood was used to compare the goodness of fit of one BN to another (section 3.6). We only kept the models when the best model is at least ten times more likely to be than the second-best model for both AIC and BIC. The results showed that 21 genes were best fitted to one of the three models according to relative likelihood. Table (4.3) shows the AIC, BIC and the relative likelihood of the 21 genes. Out of 21 genes, CME model was best fitted to 16 genes, CEM to 4 genes and 1 gene was best fitted to INDEP model.

Table 4.3. Genes with verified BN models and AIC and BIC score of the three models and relative likelihood values.

GENE	Best fitted Model	AIC				BIC			
		INDEP	CME	CEM	Relative Likelihood	INDEP	CME	CEM	Relative Likelihood
ULK1	CEM	-38,545833	-39,344477	-45,416305	20,8199952	-42,66902102	-44,056693	-49,539494	15,50869
HLA.B	CME	-53,750063	-61,104094	-52,870194	39,5282517	-57,87325105	-65,816309	-56,993382	53,06561
SLC26A11	CME	-33,199123	-37,928193	-33,266463	10,2868364	-37,32231098	-42,640409	-37,389652	13,8098
PPP2R5A	CME	-57,358823	-62,132136	-55,660909	10,8770671	-63,838185	-68,611432	-60,962151	10,87707
COPZ1	CME	-45,843344	-51,342025	-46,312274	12,3650728	-49,96653258	-56,054241	-50,435462	16,59978
ANKS1B	CEM	-27,478388	-27,789182	-35,332987	43,4626654	-30,42352303	-30,734317	-38,278122	43,46267
TIM44	CME	-45,711793	-55,407433	-46,172043	101,260331	-49,83498126	-60,119648	-50,295232	135,9393
A1BG	CME	-48,646964	-57,110518	-49,579232	43,1914803	-52,77015269	-61,822734	-53,70242	57,9834
ALDH2	CME	-42,748862	-48,761806	-42,236788	20,2159522	-46,87205085	-53,474022	-46,359976	27,13937
GON4L	CME	-51,229639	-55,902784	-50,209636	10,3457206	-57,70893461	-62,382081	-55,510878	10,34572
FBLN1	INDEP	-46,487603	-37,538125	-36,982996	87,7717069	-49,43273775	-40,483259	-39,92813	87,77171
C3	CME	-43,248846	-50,508169	-43,232024	37,7000397	-46,19398098	-53,453303	-46,177158	37,70004
MAP3K6	CME	-32,941672	-37,924072	-32,487296	12,075757	-37,0648607	-42,636287	-36,610485	16,21138
ZNF695	CEM	-35,189638	-34,458313	-43,021397	50,1931979	-39,90185346	-38,581501	-47,144586	37,38861
DSTN	CME	-39,918266	-46,743217	-39,345106	30,3402594	-44,04145406	-51,455432	-43,468295	40,73098
TF	CME	-37,08112	-48,345884	-37,060452	279,326647	-41,20430819	-53,058099	-41,183641	374,9885
ANP32B	CME	-31,597169	-38,09578	-31,962265	21,4721645	-34,5423031	-41,040915	-34,9074	21,47216
PRCC	CME	-54,56772	-60,138928	-53,755422	16,2096074	-61,04701625	-66,618224	-59,056664	16,20961
GNS	CME	-43,765062	-48,624238	-43,765062	11,3542046	-47,88824994	-53,336453	-47,88825	15,24271
MGAT4C	CEM	-22,358108	-22,358108	-35,559292	735,530238	-25,30324284	-25,303243	-38,504426	735,5302
UBR4	CME	-45,888437	-53,753649	-45,888437	51,0398303	-50,01162523	-58,465865	-50,011625	68,5196

On the other hand, we focused on BIC scores only, the model with the lowest score was taken to be the best fitted model. The comparison for a model to another were considered by taking the difference between the two models with the lowest and the second lowest BIC (equation 3.6) and the strength of the evidence were adjusted according to ΔBIC as:

0-2: indicates weak evidence,

2-6: positive evidence,

6-10: strong evidence,

>10: very strong evidence.

$$\Delta\text{BIC} = \text{BIC}_{\text{second lowest}} - \text{BIC}_{\text{lowest}} \quad (3.6.)$$

The results showed that out of 1830 genes have a valid BIC score after eliminating genes with $\Delta\text{BIC} > 6$ (low evidence), 207 genes were left. CEM model was best fitted to 169 genes, CME to 34 genes, and INDEP model to 4 genes. Appendix 3 contains the BIC scores and the names of all genes. Table (4.4) shows the number of genes were best fitted to each model according to BIC scores only.

Table 4.4. Number of genes best fitted to each model and the evidence strength according to BIC scores only

Evidence strength (ΔBIC)	CEM	CME	INDEP	TOTAL
Strong (6-10)	91	26	1	118
Very strong (>10)	78	8	3	89
TOTAL	169	34	4	207

4.6 Gene Set Enrichment Analysis

4.1.1. CBioportal Database

We split the genes into 2 groups: Group 1 contains the genes that have best fitted to BN models according to AIC and BIC scores with relative likelihood ≥ 10 (21 genes), and Group 2 has the genes with a valid model according to BIC score only (207 genes). For Group 1 genes, cbioportal showed that the genes have been reported in 83% of 396

cases in HCC studies. Figure 4.8 shows HCC related studies from cbioportal that have reported the detected genes. At the left side of each row shows the genes and the percentage of how much this gene is reported in studies while the bars at the right side represents each study with different omic type. The grey bar means the gene is not mentioned in that study and the colored bar means that the gene is reported in that study. Cbioportal also reports alteration frequency. For Group 1, it showed that the genes are reported in about 48% of the studies with high gene expression, 30% with multiple alterations, and 2% with CNV amplification.

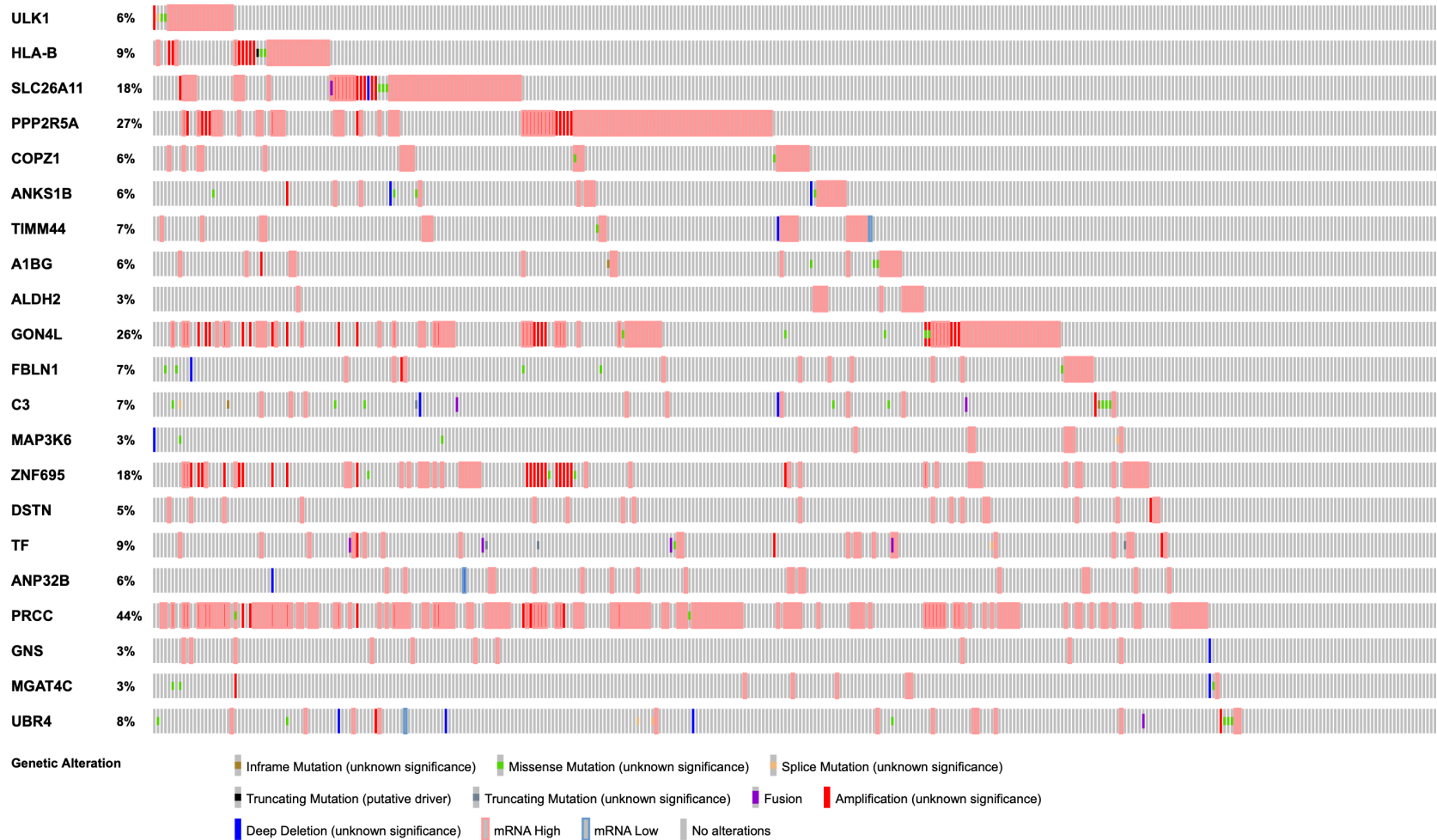


Figure 4.8. The gene list of Group 1 and the related HCC studies reported by cbiportal

For genes in group 2, we tested each model separately in the database. Genes of CME model (169 genes) have been reported in 99% of the cases of HCC. Figure 4.9 shows some of the genes and the study that they have been reported them are shown (all genes are in Appendix 4). Also, alteration frequency summary table of cbioportal have shown that the genes have been reported in 75% of studies with multiple alterations, %1 with mutations, 1% with copy number amplification, 2% with low mRNA, and 20% with high mRNA.

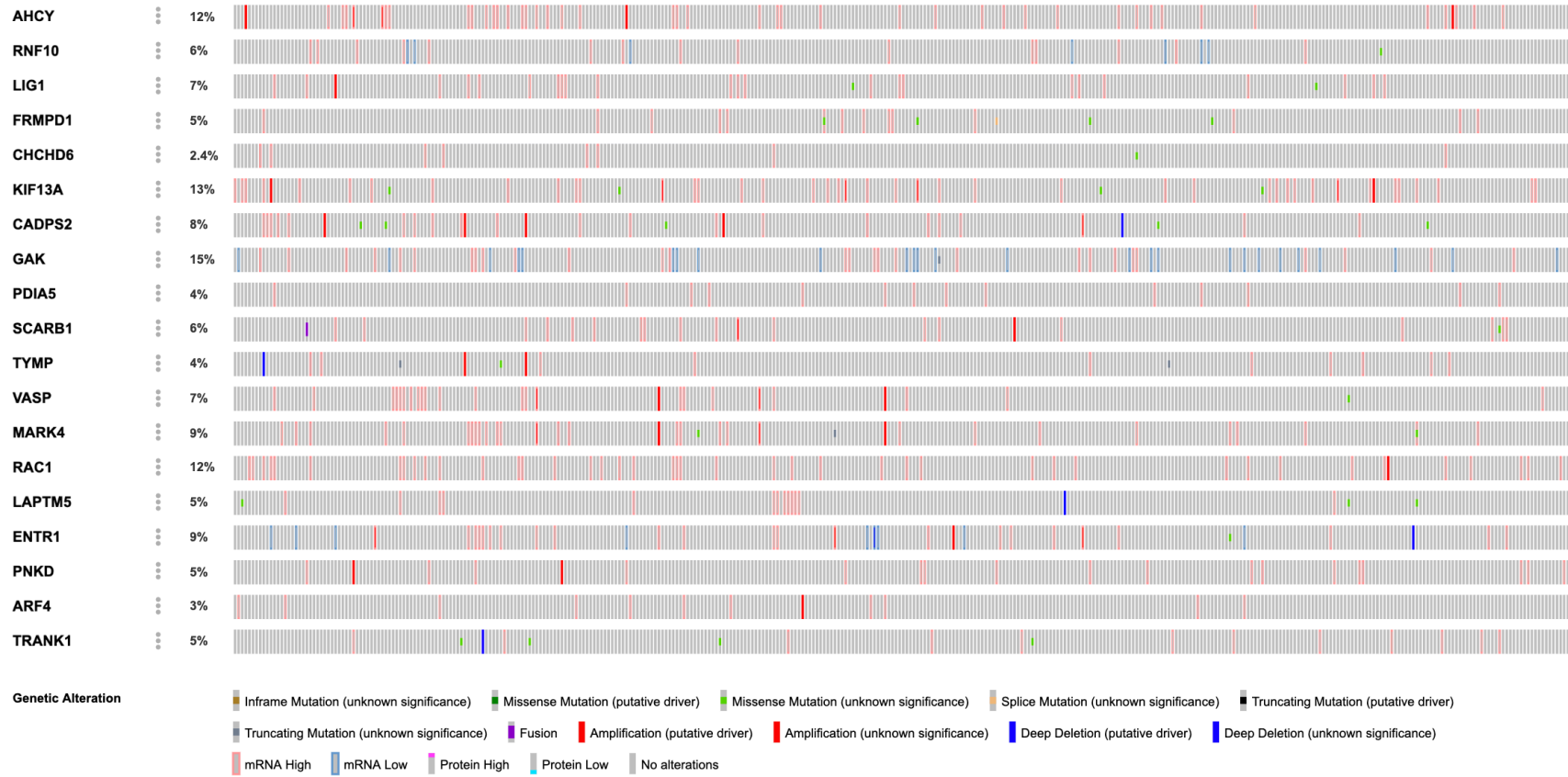


Figure 4.9. The gene list of Group 2 (CME) and the related HCC studies reported by cbiportal

Moreover, cbioportal have shown that some genes have matched to pathways in HCC. Table 4.5 shows the affected pathways and the genes that have been matched to each pathway.

Table 4.5. Genes of Group 2 (CME) that matched to pathways in HCC (reported by cbioportal).

Pathway	Genes Matched
BLCA-2014-RTK-RAS-PI(3)K-pathway	STK11,FGFR3
HIPPO	STK11,FGFR3
RTK-RAS	FGFR3,RAC1
LUSC-2012-RTK-RAS-PI(3)K-pathway	STK11,FGFR3
HNSC-2015-RTK-RAS-PI(3)K-pathway	FGFR3
SKCM-2015-RTK-RAS-PI(3)K-pathway	RAC1
STAD-2014-RTK-RAS-PI(3)K-pathway	JAK2
PI3K	STK11
LUAD-2014-RTK-RAS-PI(3)K-pathway	STK11

Genes reported for following the CEM model (34 genes) have been shown in about 90% of the cases of HCC. Figure 4.10 shows the genes and the studies that they have been reported. Also, alteration frequency summary statistics tables of cbioportal have shown that the genes have been reported in 40% of the studies with multiple alterations, 40% with high mRNA 5% with low mRNA, 3% with copy number amplification, 1% with mutations, and 1% with copy number deletion.

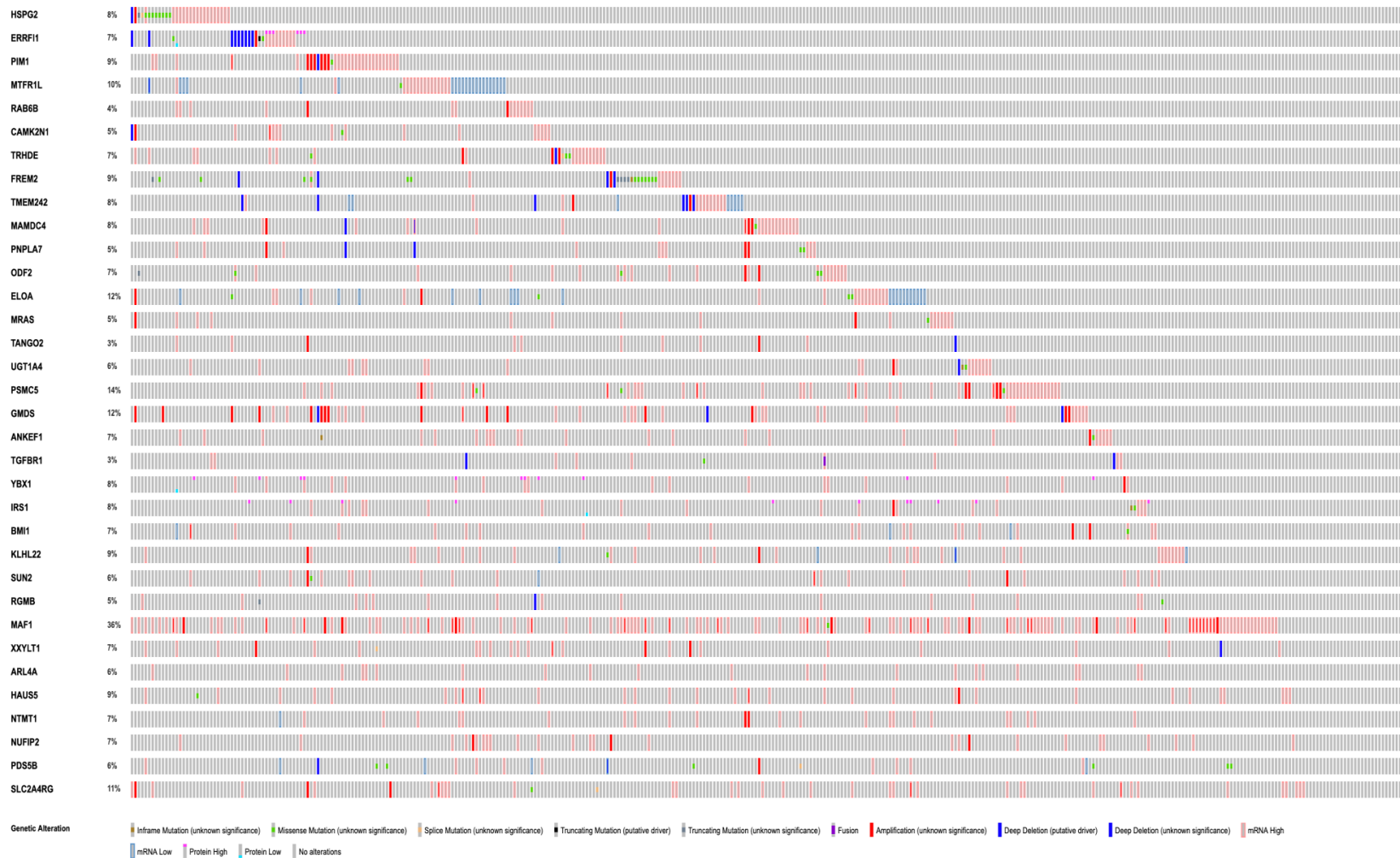


Figure 4.10. The gene list of Group 2 (CEM) and the related HCC studies reported by cbioportal

Moreover, cbioportal have shown that some genes have matched to pathways in HCC. Table 4.6 shows the affected pathways and the genes that have been matched to each pathway.

Table 4.6. Genes of Group 2 (CEM) that matched to the affected pathways in HCC (reported by cbioportal)

Pathway	Genes Matched
COADREAD-2012-TGF-B-signaling-pathway	TGFBR1
TGF-Beta	TGFBR1
RTK-RAS	ERRFI1

For 4 genes following the INDEP model, have been reported in 20% of the cases of HCC. Figure 4.11 shows the genes and the studies that they have been reported. Also, alteration frequency summary of cbioportal have shown that the genes have been reported in 6% of studies with mutations, 1% with multiple alterations, 1% with copy number deletion, 1% with copy number amplification and 11% with high mRNA.

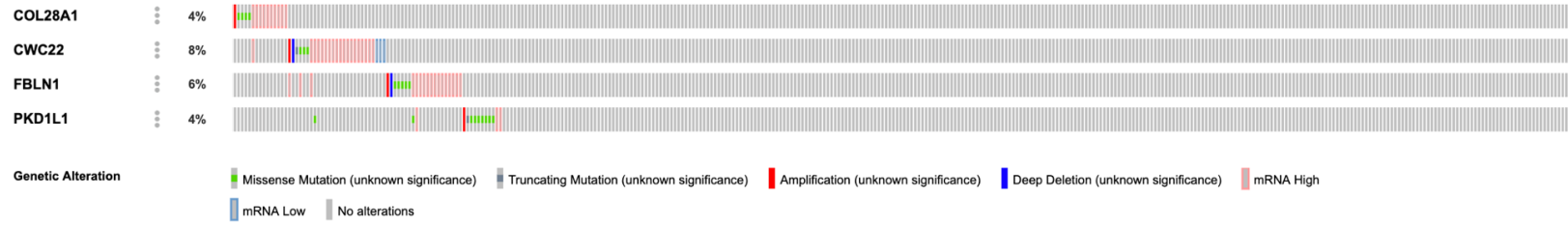


Figure 4.11. The gene list of Group 2 (INDEP) and the related HCC studies reported by cbiportal

5. DISCUSSION AND CONCLUSION

In this study, by integrating the genome, epigenome, and transcriptome of HCC single cell data, we attempt to infer causality in HCC. By fitting BN to gain more insight into HCC susceptibility and progression.

Firstly, we analyzed the RNA-seq and RRBS data on their own. We calculated methylation levels and FPKM values. From RRBS data, the results showed that CpG islands and promoters have a higher coverage than other genomic regions. This explains the nature of the data of RRBS method which uses MspI enzyme that have a restriction site mostly found in CpG islands near promoters, assisting RRBS method to cover highly CpG regions (171). Besides, we observed a low methylation level over the genome of each sample especially in the promoter region. This observation consists with studies that detected a global hypomethylation in HCC (172, 173).

Thereafter, CNV were estimated from RRBS data by using HMM method. All 25 samples showed an amplification in chromosome 7 and q arm of chromosome 1. These amplifications were previously detected in HCC study (174). We also observed a deletion in chromosome 8 and chromosome 4. These two deletions have been reported in literature in HCC (175, 176). As well, some cells showed a loss in chromosome 13, while some others showed an amplification in chromosome 6. This consists with some HCC studies in literature (177, 178).

Then, we investigated the correlation between omics, by calculating Pearson correlation coefficient (r). Between gene expression and DNA methylation we observed a negative correlation in the promoter regions and a positive correlation in gene body regions. These correlations indicate that the DNA methylation in the promoter region might regulate the expression of the corresponding gene, and the methylation in the gene body region is also involved in this regulation (44). In addition, we observed a high correlation between gene expression and CNV, which means that CNV might have an impact on the gene expression levels by changing the gene dosages (179).

Next, we integrated the genome, epigenome, and transcriptome data in order to analyze the causality between omics. By using BN, we explored three different model

alternatives. INDEP models where CNV affects gene expression and DNA methylation independently from each other; CEM model where CNV affects gene expression then DNA methylation; CME model where CNV affects DNA methylation then gene expression.

Selecting the genes according to both AIC and BIC scores with relative likelihood ≥ 10 , yielded 21 genes to follow a specific BN model (CME: 16 genes, CEM: 4 genes, INDEP: 1 gene). Cbioportal showed that all the genes have been previously reported in HCC studies (figure 4.2). All genes were reported in at least one study with high gene expression. In addition, when search these genes in genome wide studies, we found that HLA gene have been related to HCC in 4 different GWAS studies (180-183). Variations in HLA gene that have strongly related to HBV infection and development of liver cirrhosis and HBV-related HCC (183).

On the other hand, when we focused on BIC score only, we got 207 genes followed a BN model (CEM: 34 genes, CME: 179 genes, INDEP: 4 genes). Similarly, all the detected genes were reported in HCC studies. Moreover, Cbioportal have matched some genes to pathways that have been reported to be affected in HCC. Genes that follow CME model: STK11 ($\Delta\text{BIC}=47.35$), FGFR3 ($\Delta\text{BIC}=12.04$), JAK2 ($\Delta\text{BIC}=13.48$), and RAC1 ($\Delta\text{BIC}=6.11$) have been matched to PI3K signaling pathway that play a role in the survival and the rapid growth of HCC tumor (184-186). Also, TGFBR1 gene which follows CEM model ($\Delta\text{BIC}=38.25$) have been matched to TGF-Beta signaling pathway. TGF-Beta signaling pathway has a major contribution in HCC pathogenesis and tumor development, is considered as a master regulator for cell proliferation and differentiation (187, 188). Furthermore, we found that the genes FGFR3 ($\Delta\text{BIC}=12.04$) and RAC1 ($\Delta\text{BIC}=6.11$) that follow CME model and ERFF1 gene ($\Delta\text{BIC}=4328.38$) which follow CEM model have been matched to the same pathway, RTK-RAS signaling pathway (Figure 5.1). Many studies have shown that RTK-RAS signaling pathway plays a major role in HCC proliferation, survival and apoptosis (189-191).

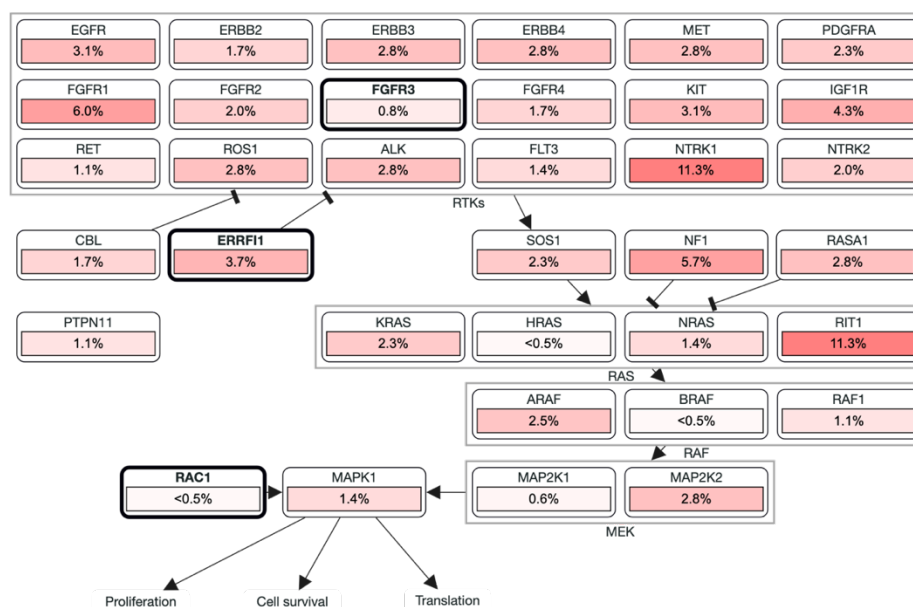


Figure 5.1. Genes in RTK-RAS signaling pathway, highlighting FGFR3 (CME), ERRF1 (CEM), and RAC1 (CME) genes.

In this thesis, we showed different genes which follow different BN models take place in different (or same) pathways that play major roles in HCC tumorigenesis. By this way, the heterogeneity of the omics and their regulations with each other have been shown. This method might help explore the genes that are related to HCC by defining their models according to the relationship between the different omics. Many multi-omics integration methods such as supervised, unsupervised, multi-dimensional scaling, cluster of clusters have been used in HCC studies. These studies (mentioned in literature review) focused on finding the relationship between omics by correlating or clustering them. Here, we suggest the BN to be used to explore the causal relationship between three omics in sequencing data at a single cell level. We found 207 genes with significant model, many of these genes have been reported previously to be related to HCC in either GWAS or sequencing data. Lastly, we introduce this method as a method that can provide a deeper insight and understanding about HCC cells. It can be developed to implement more omic data (e.g., proteomic) and to be used with other cancer types or complex diseases.

5.1 Limitations and Possible Improvements

- In this study, the used data were generated by using one of the first single cell sequencing methods, which means that a quality issue was presented. For example, the mapping efficiency of RRBS data was very low. Which in turn, might affect our calculations of DNA methylation level and CNV estimations.
- The genomic region was set to be from TSS to TES. However, it could also cover the regions that are away from TSS in order to involve the enhancer and regulator regions.
- We only analyzed protein-coding genes; non-coding genes could also be involved.
- For BN construction, DNA methylation and gene expression values were discretized according to some classifications; for gene expression the classification in EBI expression Atlas were used. This might lead to losing some information or affecting the final results.
- Other omics data such as proteomics can be included which might give us more comprehensive results.
- This method could be applied on a larger dataset, that might help in increasing the validation of the data.
- Moreover, it can be used on different data of different types of cancer or complex diseases.

6. REFERENCES

1. Jehan Z. Chapter 1 - Single-Cell Omics: An Overview. In: Barh D, Azevedo V, editors. *Single-Cell Omics*: Academic Press; 2019. p. 3-19.
2. Ravasio A, Myaing MZ, Chia S, Arora A, Sathe A, Cao EY, et al. Single-cell analysis of EphA clustering phenotypes to probe cancer cell heterogeneity. *Communications Biology*. 2020;3(1):429.
3. Cheung P, Vallania F, Warsinske HC, Donato M, Schaffert S, Chang SE, et al. Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. *Cell*. 2018;173(6):1385-97.e14.
4. Enge M, Arda HE, Mignardi M, Beausang J, Bottino R, Kim SK, et al. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell*. 2017;171(2):321-30.e14.
5. Hurria A, Jones L, Muss HB. Cancer Treatment as an Accelerated Aging Process: Assessment, Biomarkers, and Interventions. *American Society of Clinical Oncology Educational Book*. 2016(36):e516-e22.
6. Kamies R, Martinez-Jimenez CP. Advances of single-cell genomics and epigenomics in human disease: where are we now? *Mammalian Genome*. 2020;31(5):170-80.
7. Ecker S, Pancaldi V, Valencia A, Beck S, Paul DS. Epigenetic and Transcriptional Variability Shape Phenotypic Plasticity. *BioEssays*. 2018;40(2):1700148.
8. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*. 2016;17(2):S15.
9. Berger B, Peng J, Singh M. Computational solutions for omics data. *Nature Reviews Genetics*. 2013;14(5):333-46.
10. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009;10(1):57-63.
11. Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. *Nature*. 2019;571(7766):489-99.
12. Altelaar AFM, Munoz J, Heck AJR. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*. 2013;14(1):35-48.
13. Eddy S, Mariani LH, Kretzler M. Integrated multi-omics approaches to improve classification of chronic kidney disease. *Nature Reviews Nephrology*. 2020;16(11):657-68.
14. Knox SS. From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell International*. 2010;10(1):11.
15. Hasin Y, Seldin M, Lusic A. Multi-omics approaches to disease. *Genome Biology*. 2017;18(1):83.
16. Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Frontiers in Oncology*. 2020;10:1030.

17. Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ, et al. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Frontiers in Genetics*. 2019;10:995.
18. Kim D, Li R, Dudek SM, Ritchie MD. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData Mining*. 2013;6(1):23.
19. Tini G, Marchetti L, Priami C, Scott-Boyer M-P. Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in Bioinformatics*. 2019;20(4):1269-79.
20. Zeng ISL, Lumley T. Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science). *Bioinformatics and Biology Insights*. 2018;12:1177932218759292.
21. de Tayrac M, Lê S, Aubry M, Mosser J, Husson F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*. 2009;10:32-.
22. Yuan Y, Savage RS, Markowitz F. Patient-Specific Data Fusion Defines Prognostic Cancer Subtypes. *PLOS Computational Biology*. 2011;7(10):e1002227.
23. Koller D, Friedman N. Probabilistic graphical models: principles and techniques: MIT press; 2009.
24. Pearl J. Causality: Models, reasoning and inference cambridge university press. Cambridge, MA, USA. 2000;9:10-1.
25. Bielza C, Larrañaga P. Bayesian networks in neuroscience: a survey. *Frontiers in Computational Neuroscience*. 2014;8:131.
26. Liang S-B, Fu L-W. Application of single-cell technology in cancer research. *Biotechnology Advances*. 2017;35(4):443-9.
27. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018 [press release]. 2018.
28. Izci Y. Chapter 4 - Single-Cell Technology for Human Gliomas. In: Barh D, Azevedo V, editors. *Single-Cell Omics*: Academic Press; 2019. p. 55-68.
29. Clark T, Maximin S, Meier J, Pokharel S, Bhargava P. Hepatocellular Carcinoma: Review of Epidemiology, Screening, Imaging Diagnosis, Response Assessment, and Treatment. *Current Problems in Diagnostic Radiology*. 2015;44(6):479-86.
30. Lurje I, Czigany Z, Bednarsch J, Roderburg C, Isfort P, Neumann UP, et al. Treatment Strategies for Hepatocellular Carcinoma – a Multidisciplinary Approach. LID - 10.3390/ijms20061465 [doi] LID - 1465. (1422-0067 (Electronic)).
31. Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. *BioMed Research International*. 2018;2018:9836256.
32. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Research*. 2016;26(3):304-19.

33. Biswapriya BM, Carl L, Michael O, Laura AC. Integrated omics: tools, advances and future approaches. *Journal of Molecular Endocrinology*. 2019;62(1):R21-R45.
34. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. *Nature*. 2017;550(7676):345-53.
35. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*. 2017;12(12):e0190152.
36. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature*. 2016;537(7620):347-55.
37. Fondi M, Liò P. Multi -omics and metabolic modelling pipelines: Challenges and tools for systems microbiology. *Microbiological Research*. 2015;171:52-64.
38. Mochida K, Shinozaki K. Advances in Omics and Bioinformatics Tools for Systems Analyses of Plant Functions. *Plant and Cell Physiology*. 2011;52(12):2017-38.
39. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*. 2015;16(2):85-97.
40. Pathak RR, Davé V. Integrating Omics Technologies to Study Pulmonary Physiology and Pathology at the Systems Level. *Cellular Physiology and Biochemistry*. 2014;33(5):1239-60.
41. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics*. 2017;8:84.
42. Haas R, Zelezniak A, Iacovacci J, Kamrad S, Townsend S, Ralser M. Designing and interpreting ‘multi-omic’ experiments that may change our understanding of biology. *Current Opinion in Systems Biology*. 2017;6:37-45.
43. Howey R, Shin S-Y, Relton C, Davey Smith G, Cordell HJ. Bayesian network analysis incorporating genetic anchors complements conventional Mendelian randomization approaches for exploratory analysis of causal relationships in complex data. *PLOS Genetics*. 2020;16(3):e1008198.
44. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife*. 2013;2:e00523.
45. Liu G, Dong C, Liu L. Integrated Multiple “-omics” Data Reveal Subtypes of Hepatocellular Carcinoma. *PLOS ONE*. 2016;11(11):e0165457.
46. Miao R, Luo H, Zhou H, Li G, Bu D, Yang X, et al. Identification of prognostic biomarkers in hepatitis B virus-related hepatocellular carcinoma and stratification by integrative multi-omics analysis. *Journal of Hepatology*. 2014;61(4):840-9.
47. Yildiz G. Integrated multi-omics data analysis identifying novel drug sensitivity-associated molecular targets of hepatocellular carcinoma cells. *Oncol Lett*. 2018;16(1):113-22.
48. Li Y, Zhuang H, Zhang X, Li Y, Liu Y, Yi X, et al. Multiomics Integration Reveals the Landscape of Prometastasis Metabolism in Hepatocellular Carcinoma. *Mol Cell Proteomics*. 2018;17(4):607-18.

49. Cavalli M, Diamanti K, Pan G, Spalinskas R, Kumar C, Deshmukh AS, et al. A Multi-Omics Approach to Liver Diseases: Integration of Single Nuclei Transcriptomics with Proteomics and HiCap Bulk Data in Human Liver. *OMICS: A Journal of Integrative Biology*. 2020;24(4):180-94.
50. Liu F, Qin L, Liao Z, Song J, Yuan C, Liu Y, et al. Microenvironment characterization and multi-omics signatures related to prognosis and immunotherapy response of hepatocellular carcinoma. *Experimental Hematology & Oncology*. 2020;9(1):10.
51. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nature genetics*. 2012;44(6):694-8.
52. Duan M, Hao J, Cui S, Worthley DL, Zhang S, Wang Z, et al. Diverse modes of clonal evolution in HBV-related hepatocellular carcinoma revealed by single-cell genome sequencing. *Cell Research*. 2018;28(3):359-73.
53. Roy PS, Saikia BJ. Cancer and cure: A critical analysis. *Indian J Cancer*. 2016;53(3):441-2.
54. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 2018;68(6):394-424.
55. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*. 2015;136(5):E359-E86.
56. Soriano V, Tefferi A. Prevention of liver cancer with new curative hepatitis C antivirals: Real-world challenges. *Cancer*. 2018;124(8):1647-9.
57. Waly Raphael S, Yangde Z, YuXiang C. Hepatocellular Carcinoma: Focus on Different Aspects of Management. *ISRN Oncology*. 2012;2012:421673.
58. Pei Y, Zhang T, Renault V, Zhang X. An overview of hepatocellular carcinoma study by omics-based methods. *Acta Biochimica et Biophysica Sinica*. 2009;41(1):1-15.
59. Society AC. *Cancer Facts & Figures 2019*. Atlanta: American Cancer Society 2019.
60. Yang JD, Hainaut P, Gores GJ, Amadou A, Plymoth A, Roberts LR. A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nature Reviews Gastroenterology & Hepatology*. 2019;16(10):589-604.
61. Rawla P, Sunkara T, Muralidharan P, Raj JP. Update in global trends and aetiology of hepatocellular carcinoma. *Contemp Oncol (Pozn)*. 2018;22(3):141-50.
62. El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology*. 2012;142(6):1264-73.e1.
63. Kara F, Kesinkilic B, Baran Deniz E, Turkyilmaz M, Dundar S, Kavak Ergun A, et al. Türkiye Kanser İstatistikleri. Ankara: Türkiye Cumhuriyeti Sağlık Bakanlığı Halk Sağlığı Genel Müdürlüğü; 2018.
64. Alacacioglu A, Somali I, Simsek I, Astarcioglu I, Ozkan M, Camci C, et al. Epidemiology and Survival of Hepatocellular Carcinoma in Turkey: Outcome of Multicenter Study. *Japanese Journal of Clinical Oncology*. 2008;38(10):683-8.

65. Can A, Dogan E, Bayoglu IV, Tatli AM, Besiroglu M, Kocer M, et al. Multicenter Epidemiologic Study on Hepatocellular Carcinoma in Turkey. *Asian Pacific Journal of Cancer Prevention*. 2014;15(6):2923-7.
66. Yi S-W, Choi J-S, Yi J-J, Lee Y-h, Han KJ. Risk factors for hepatocellular carcinoma by age, sex, and liver disorder status: A prospective cohort study in Korea. *Cancer*. 2018;124(13):2748-57.
67. Forner A, Reig M, Bruix J. Hepatocellular carcinoma. *The Lancet*. 2018;391(10127):1301-14.
68. Daher S, Massarwa M, Benson AA, Khoury T. Current and Future Treatment of Hepatocellular Carcinoma: An Updated Comprehensive Review. *J Clin Transl Hepatol*. 2018;6(1):69-78.
69. Tang A, Hallouch O, Chernyak V, Kamaya A, Sirlin CB. Epidemiology of hepatocellular carcinoma: target population for surveillance and diagnosis. *Abdominal Radiology*. 2018;43(1):13-25.
70. Golabi P, Fazel S, Otgonsuren M, Sayiner M, Locklear CT, Younossi ZM. Mortality assessment of patients with hepatocellular carcinoma according to underlying disease and treatment modalities. *Medicine (Baltimore)*. 2017;96(9):e5904-e.
71. The Cancer of the Liver Italian Program I. A new prognostic system for hepatocellular carcinoma: A retrospective study of 435 patients. *Hepatology*. 1998;28(3):751-5.
72. Heinrich B, Czauderna C, Marquardt JU. Immunotherapy of Hepatocellular Carcinoma. *Oncology Research and Treatment*. 2018;41(5):292-7.
73. Laurent-Puig P, Zucman-Rossi J. Genetics of hepatocellular tumors. *Oncogene*. 2006;25(27):3778-86.
74. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*. 2015;19(1A):A68-A77.
75. Cancer Genome Atlas Research Network. Electronic address wbe, Cancer Genome Atlas Research N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*. 2017;169(7):1327-41.e23.
76. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L. New Goals for the U.S. Human Genome Project: 1998-2003. *Science*. 1998;282(5389):682.
77. Robertson KD. DNA methylation and human disease. *Nature Reviews Genetics*. 2005;6(8):597-610.
78. Feinberg AP, Tycko B. The history of cancer epigenetics. *Nature Reviews Cancer*. 2004;4(2):143-53.
79. Noble D. Conrad Waddington and the origin of epigenetics. *The Journal of Experimental Biology*. 2015;218(6):816.
80. Waddington CH. CANALIZATION OF DEVELOPMENT AND THE INHERITANCE OF ACQUIRED CHARACTERS. *Nature*. 1942;150(3811):563-5.
81. Counce SJ, Ede DA. The Effect on Embryogenesis of a Sex-linked Female-Sterility Factor in *Drosophila melanogaster*. *Journal of Embryology and Experimental Morphology*. 1957;5(4):404.

82. Wu Ct, Morris JR. Genes, Genetics, and Epigenetics: A Correspondence. *Science*. 2001;293(5532):1103.
83. Baedke J. The epigenetic landscape in the course of time: Conrad Hal Waddington's methodological impact on the life sciences. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*. 2013;44(4, Part B):756-73.
84. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. *Nature Reviews Genetics*. 2007;8(4):253-62.
85. Eshraghi AA, Liu G, Kay S-IS, Eshraghi RS, Mittal J, Moshiree B, et al. Epigenetics and Autism Spectrum Disorder: Is There a Correlation? *Frontiers in Cellular Neuroscience*. 2018;12:78.
86. Bird A. DNA methylation patterns and epigenetic memory. *Genes & development*. 2002;16(1):6-21.
87. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology*. 2013;38(1):23-38.
88. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*. 2007;39(4):457-66.
89. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, et al. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLOS Genetics*. 2010;6(9):e1001134.
90. Goldberg AD, Allis CD, Bernstein E. Epigenetics: A Landscape Takes Shape. *Cell*. 2007;128(4):635-8.
91. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, et al. Global epigenomic reconfiguration during mammalian brain development. *Science (New York, NY)*. 2013;341(6146):1237905-.
92. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. *Nature*. 1993;366(6453):362-5.
93. Doi A, Park I-H, Wen B, Murakami P, Aryee MJ, Irizarry R, et al. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature Genetics*. 2009;41(12):1350-3.
94. Rao X, Evans J, Chae H, Pilrose J, Kim S, Yan P, et al. CpG island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene*. 2013;32(38):4519-28.
95. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*. 2009;41(2):178-86.
96. Venolia L, Gartler SM, Wassman ER, Yen P, Mohandas T, Shapiro LJ. Transformation with DNA from 5-azacytidine-reactivated X chromosomes. *Proc Natl Acad Sci U S A*. 1982;79(7):2352-4.

97. Mohandas T, Sparkes RS, Shapiro LJ. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science*. 1981;211(4480):393.
98. Wolf SF, Dintzis S, Toniolo D, Persico G, Lunnen KD, Axelman J, et al. Complete concordance between glucose-6-phosphate dehydrogenase activity and hypomethylation of 3' CpG clusters: implications for X chromosome dosage compensation. *Nucleic Acids Res*. 1984;12(24):9333-48.
99. Jung M, Pfeifer GP. Aging and DNA methylation. *BMC Biology*. 2015;13(1):7.
100. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*. 2012;13(7):484-92.
101. Kulis M, Esteller M. 2 - DNA Methylation and Cancer. In: Herceg Z, Ushijima T, editors. *Advances in Genetics*. 70: Academic Press; 2010. p. 27-56.
102. Ehrlich M. DNA methylation in cancer: too much, but also too little. *Oncogene*. 2002;21(35):5400-13.
103. Shen J, Wang S, Zhang Y-J, Kappil M, Wu H-C, Kibriya MG, et al. Genome-wide DNA methylation profiles in hepatocellular carcinoma. *Hepatology (Baltimore, Md)*. 2012;55(6):1799-808.
104. Morris TJ, Beck S. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods*. 2015;72:3-8.
105. Warnecke PM, Stirzaker C, Song J, Grunau C, Melki JR, Clark SJ. Identification and resolution of artifacts in bisulfite sequencing. *Methods*. 2002;27(2):101-7.
106. Sadri R, Hornsby PJ. Rapid analysis of DNA methylation using new restriction enzyme sites created by bisulfite modification. *Nucleic Acids Res*. 1996;24(24):5058-9.
107. Merrick WC, Pavitt GD. Protein Synthesis Initiation in Eukaryotic Cells. *Cold Spring Harb Perspect Biol*. 2018;10(12):a033092.
108. Mattick JS, Makunin IV. Non-coding RNA. *Human Molecular Genetics*. 2006;15(suppl_1):R17-R29.
109. San Segundo-Val I, Sanz-Lozano CS. Introduction to the Gene Expression Analysis. In: Isidoro García M, editor. *Molecular Genetics of Asthma*. New York, NY: Springer New York; 2016. p. 29-43.
110. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63.
111. Gowda M, Jantasuriyarat C, Dean RA, Wang G-L. Robust-LongSAGE (RL-SAGE): A Substantially Improved LongSAGE Method for Gene Discovery and Transcriptome Analysis. *Plant Physiology*. 2004;134(3):890.
112. Guigó R. Chapter 2 - The Coding and the Non-coding Transcriptome. In: Walhout AJM, Vidal M, Dekker J, editors. *Handbook of Systems Biology*. San Diego: Academic Press; 2013. p. 27-41.
113. Bumgarner R. Overview of DNA Microarrays: Types, Applications, and Their Future. *Current Protocols in Molecular Biology*. 2013;101(1):22.1.1-1.11.

114. van Hal NLW, Vorst O, van Houwelingen AMML, Kok EJ, Peijnenburg A, Aharoni A, et al. The application of DNA microarrays in gene expression analysis. *Journal of Biotechnology*. 2000;78(3):271-80.
115. Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther*. 2012;22(4):271-4.
116. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nature Reviews Genetics*. 2019;20(11):631-56.
117. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010;464(7289):704-12.
118. Shlien A, Malkin D. Copy number variations and cancer. *Genome Medicine*. 2009;1(6):62.
119. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444-54.
120. Liu M, Fang L, Liu S, Pan MG, Seroussi E, Cole JB, et al. Array CGH-based detection of CNV regions and their potential association with reproduction and other economic traits in Holsteins. *BMC Genomics*. 2019;20(1):181.
121. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*. 2011;12(5):363-76.
122. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 2013;14(11):S1.
123. Korbelt JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, NY)*. 2007;318(5849):420-6.
124. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*. 2009;6(11):S13-S20.
125. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*. 2009;6(9):677-81.
126. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865-71.
127. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*. 2012;28(21):2711-8.
128. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*. 2009;19(9):1586-92.
129. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*. 2012;28(21):2747-54.

130. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*. 2015;112(17):5473.
131. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43(1110):11.0.1-.0.33.
132. Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M. Read count approach for DNA copy number variants detection. *Bioinformatics*. 2012;28(4):470-8.
133. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5(4):557-72.
134. Nijkamp JF, van den Broek MA, Geertman J-MA, Reinders MJT, Daran J-MG, de Ridder D. De novo detection of copy number variation by co-assembly. *Bioinformatics*. 2012;28(24):3195-202.
135. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature*. 2011;470(7332):59-65.
136. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. Detecting copy number variation with mated short reads. *Genome Res*. 2010;20(11):1613-22.
137. Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, et al. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*. 2010;26(10):1277-83.
138. Bosetti C, Turati F, La Vecchia C. Hepatocellular carcinoma epidemiology. *Best Practice & Research Clinical Gastroenterology*. 2014;28(5):753-70.
139. Lurje I, Czigany Z, Bednarsch J, Roderburg C, Isfort P, Neumann UP, et al. Treatment Strategies for Hepatocellular Carcinoma – a Multidisciplinary Approach. *Int J Mol Sci*. 2019;20(6):1465.
140. Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine*. 2020;52(9):1452-65.
141. Team STD. SRA Toolkit. <http://ncbi.github.io/sra-tools/>.
142. Andrews S. FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>; 2010.
143. Liao P, Satten GA, Hu Y-J. PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. *Genetic Epidemiology*. 2017;41(5):375-87.
144. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-20.
145. MacManes M. On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*. 2014;5:13.
146. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*. 2012;7(3):562-78.

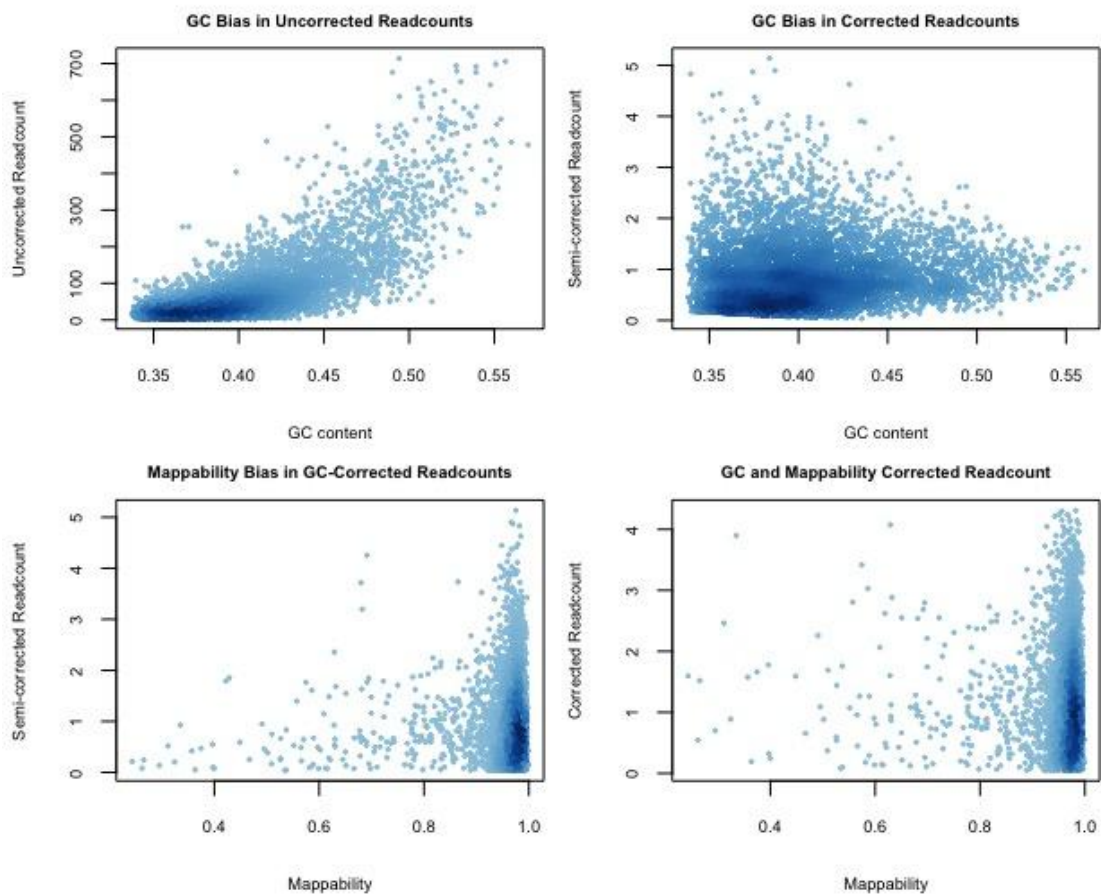
147. Chandramohan R, Wu P-Y, Phan JH, Wang MD. Benchmarking RNA-Seq quantification tools. *Annu Int Conf IEEE Eng Med Biol Soc.* 2013;2013:647-50.
148. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015;31(2):166-9.
149. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics.* 2011;12(1):323.
150. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21.
151. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009;25(9):1105-11.
152. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-9.
153. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology.* 2010;28(5):511-5.
154. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal; Vol 17, No 1: Next Generation Sequencing Data Analysis.* 2011.
155. Hannon GJ. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit; 2010.
156. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics.* 2011;27(11):1571-2.
157. Chen P-Y, Cokus SJ, Pellegrini M. BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics.* 2010;11(1):203.
158. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics.* 2010;26(7):873-81.
159. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPPING program. *BMC Bioinformatics.* 2009;10(1):232.
160. Sun X, Han Y, Zhou L, Chen E, Lu B, Liu Y, et al. A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. *Bioinformatics.* 2018;34(16):2715-23.
161. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841-2.
162. Lai D, Ha G, Shah S. HMMcopy: Copy number prediction with correction for GC and mappability bias for HTS data. R package version 1.32.0. 2020.
163. Liu J, Tang W, Chen G, Lu Y, Feng C, Tu XM. Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai Arch Psychiatry.* 2016;28(2):115-20.
164. Scutari M. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software; Vol 1, Issue 3 (2010).* 2010.
165. Aho K, Derryberry D, Peterson T. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology.* 2014;95(3):631-6.

166. Cobb BR, Rumí R, Salmerón A. Bayesian Network Models with Discrete and Continuous Variables. In: Lucas P, Gámez JA, Salmerón A, editors. *Advances in Probabilistic Graphical Models Studies in Fuzziness and Soft Computing*. Berlin, Heidelberg: Springer, Berlin, Heidelberg; 2007.
167. Lorah J, Womack A. Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling. *Behavior Research Methods*. 2019;51(1):440-50.
168. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401-4.
169. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*. 2013;45(10):1113-20.
170. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling*. 2013;6(269):pl1.
171. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res*. 2005;33(18):5868-77.
172. Hernandez-Vargas H, Lambert M-P, Le Calvez-Kelm F, Gouysse G, McKay-Chopin S, Tavtigian SV, et al. Hepatocellular Carcinoma Displays Distinct DNA Methylation Signatures with Potential as Clinical Predictors. *PLOS ONE*. 2010;5(3):e9749.
173. Toraño EG, Petrus S, Fernandez AF, Fraga MF. Global DNA hypomethylation in cancer: review of validated methods and clinical significance. *Clinical Chemistry and Laboratory Medicine (CCLM)*. 2012;50(10):1733-42.
174. Xu H, Zhu X, Xu Z, Hu Y, Bo S, Xing T, et al. Non-invasive Analysis of Genomic Copy Number Variation in Patients with Hepatocellular Carcinoma by Next Generation DNA Sequencing. *J Cancer*. 2015;6(3):247-53.
175. Guan X-Y, Fang Y, Sham JST, Kwong DLW, Zhang Y, Liang Q, et al. Recurrent chromosome alterations in hepatocellular carcinoma detected by comparative genomic hybridization. *Genes, Chromosomes and Cancer*. 2000;29(2):110-6.
176. Zhou C, Zhang W, Chen W, Yin Y, Atyah M, Liu S, et al. Integrated Analysis of Copy Number Variations and Gene Expression Profiling in Hepatocellular carcinoma. *Sci Rep*. 2017;7(1):10570-.
177. Patil MA, Gütgemann I, Zhang J, Ho C, Cheung S-T, Ginzinger D, et al. Array-based comparative genomic hybridization reveals recurrent chromosomal aberrations and *Jab1* as a potential target for 8q gain in hepatocellular carcinoma. *Carcinogenesis*. 2005;26(12):2050-7.
178. Dore MP, Realdi G, Mura D, Onida A, Massarelli G, Dettori G, et al. Genomic instability in chronic viral hepatitis and hepatocellular carcinoma. *Human Pathology*. 2001;32(7):698-703.

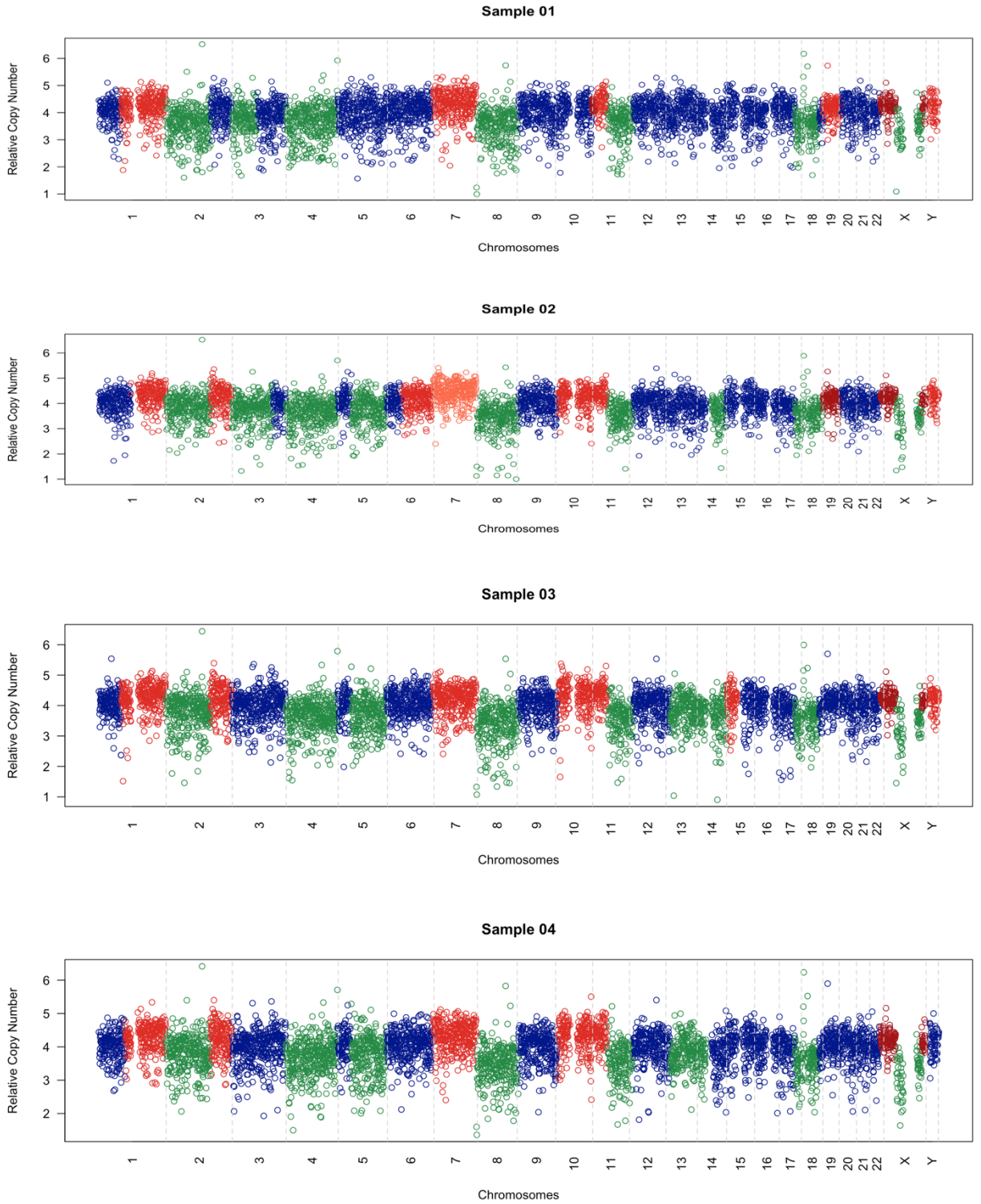
179. Shao X, Lv N, Liao J, Long J, Xue R, Ai N, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Medical Genetics*. 2019;20(1):175.
180. Sawai H, Nishida N, Khor S-S, Honda M, Sugiyama M, Baba N, et al. Genome-wide association study identified new susceptible genetic variants in HLA class I region for hepatitis B virus-related hepatocellular carcinoma. *Sci Rep*. 2018;8(1):7958-.
181. Lee M-H, Huang Y-H, Chen H-Y, Khor S-S, Chang Y-H, Lin Y-J, et al. Human leukocyte antigen variants and risk of hepatocellular carcinoma modified by hepatitis C virus genotypes: A genome-wide association study. *Hepatology*. 2018;67(2):651-61.
182. Jiang D-K, Sun J, Cao G, Liu Y, Lin D, Gao Y-Z, et al. Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma. *Nature genetics*. 2013;45(1):72-5.
183. Li S, Qian J, Yang Y, Zhao W, Dai J, Bei J-X, et al. GWAS identifies novel susceptibility loci on 6p21.32 and 21q21.3 for hepatocellular carcinoma in chronic hepatitis B virus carriers. *PLoS genetics*. 2012;8(7):e1002791-e.
184. Dimri M, Satyanarayana A. Molecular Signaling Pathways and Therapeutic Targets in Hepatocellular Carcinoma. *Cancers (Basel)*. 2020;12(2):491.
185. Zhou Q, Lui VWY, Yeo W. Targeting the PI3K/Akt/mTOR pathway in hepatocellular carcinoma. *Future Oncology*. 2011;7(10):1149-67.
186. Kudo M. mTOR Inhibitor for the Treatment of Hepatocellular Carcinoma. *Digestive Diseases*. 2011;29(3):310-5.
187. Chen J, Gingold JA, Su X. Immunomodulatory TGF- β 2; Signaling in Hepatocellular Carcinoma. *Trends in Molecular Medicine*. 2019;25(11):1010-23.
188. Huang J, Qiu M, Wan L, Wang G, Huang T, Chen Z, et al. TGF- β 1 Promotes Hepatocellular Carcinoma Invasion and Metastasis via ERK Pathway-Mediated FGFR4 Expression. *Cellular Physiology and Biochemistry*. 2018;45(4):1690-9.
189. Yang S, Liu G. Targeting the Ras/Raf/MEK/ERK pathway in hepatocellular carcinoma. *Oncology letters*. 2017;13(3):1041-7.
190. Regad T. Targeting RTK Signaling Pathways in Cancer. *Cancers (Basel)*. 2015;7(3):1758-84.
191. Liu L, Cao Y, Chen C, Zhang X, McNabola A, Wilkie D, et al. Sorafenib Blocks the RAF/MEK/ERK Pathway, Inhibits Tumor Angiogenesis, and Induces Tumor Cell Apoptosis in Hepatocellular Carcinoma Model PLC/PRF/5. *Cancer Research*. 2006;66(24):11851.

7. APPENDIX

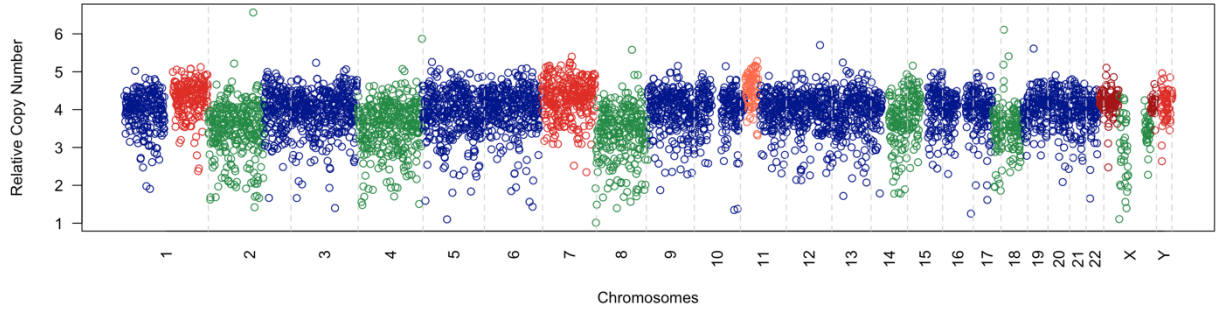
Appendix 1: Before and after correction step of GC-content and mappability by HMMcopy.



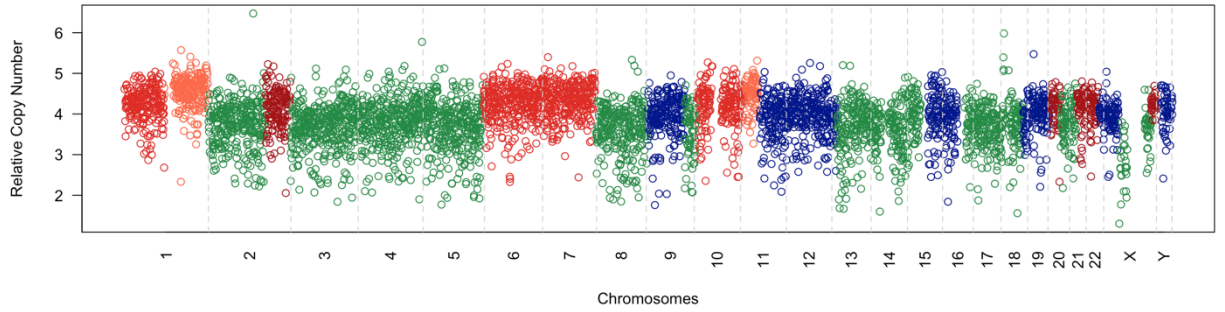
Appendix 2: CNV Pattern of All Samples.



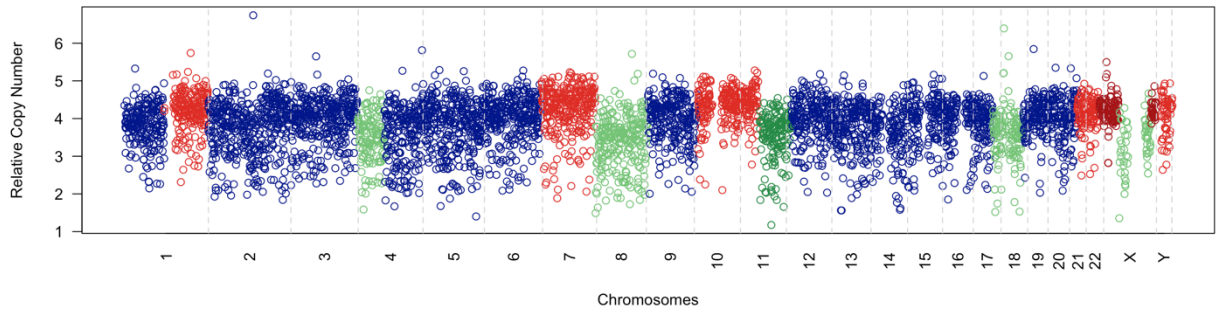
Sample 05



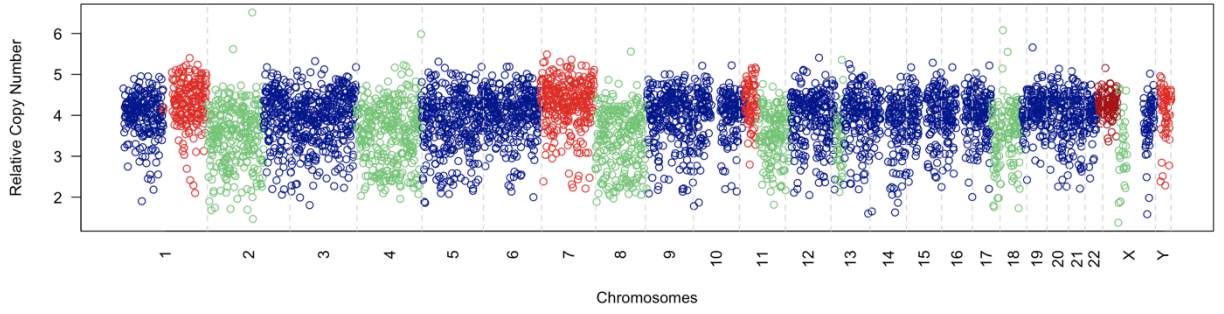
Sample 06



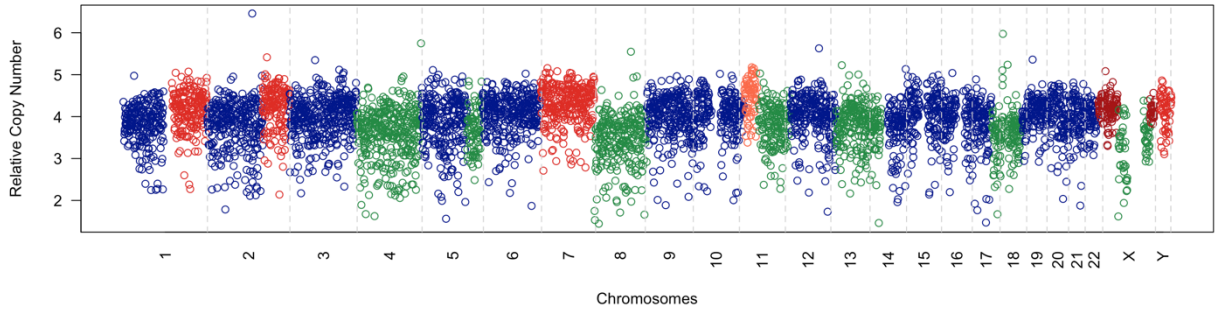
Sample 07



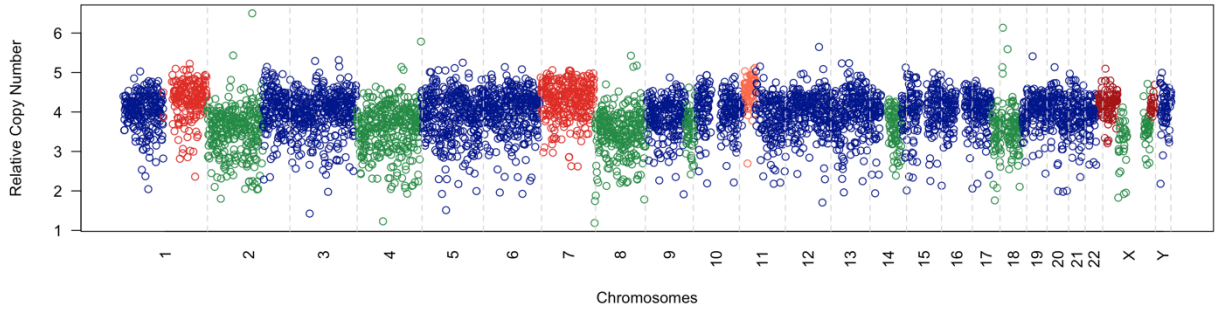
Sample 08



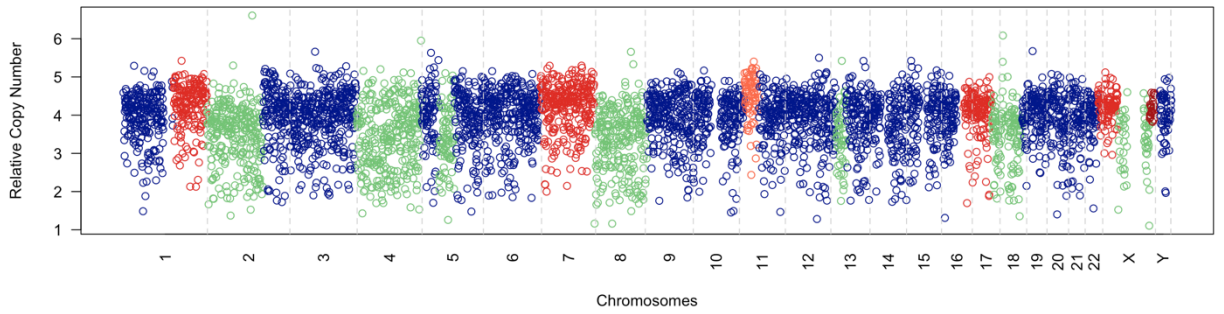
Sample 09



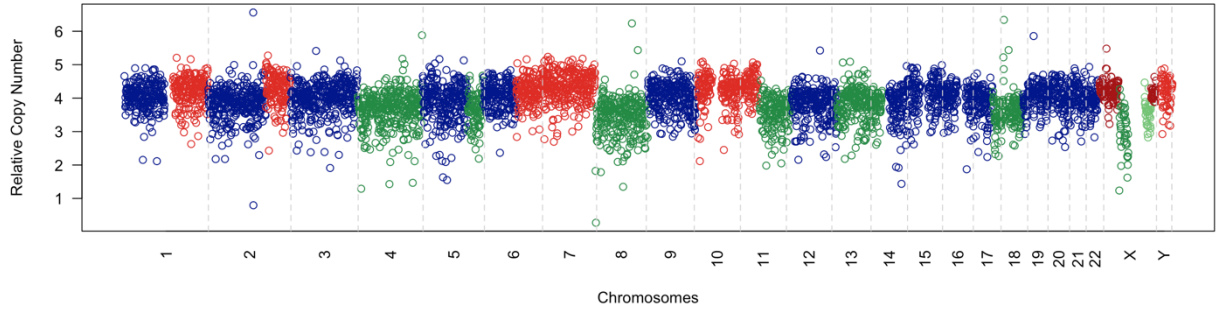
Sample 10



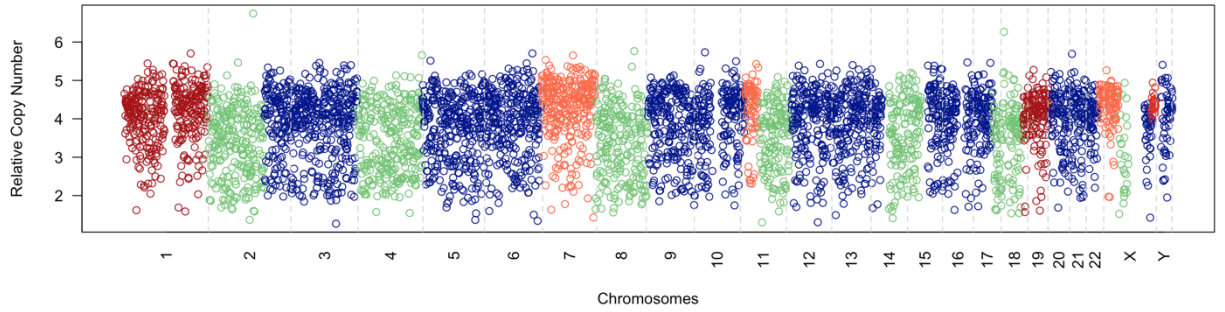
Sample 11



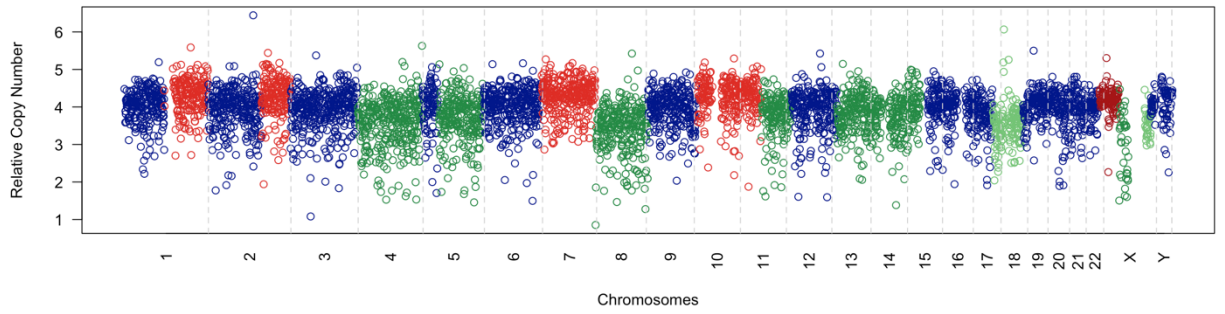
Sample 16



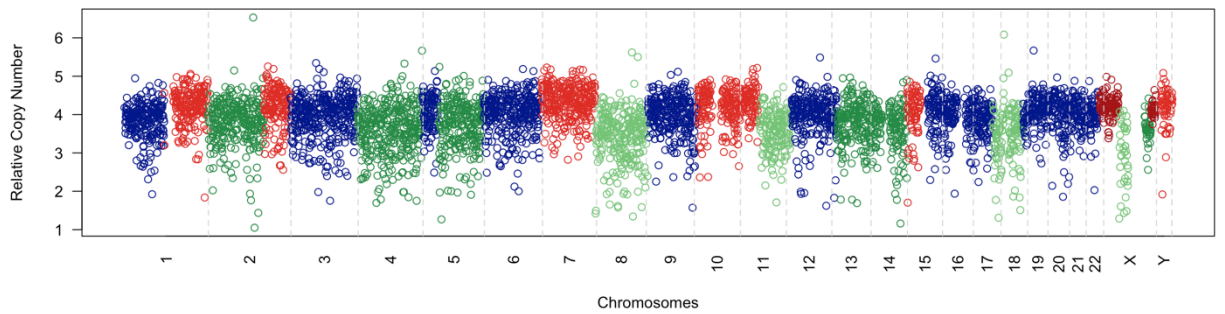
Sample 17



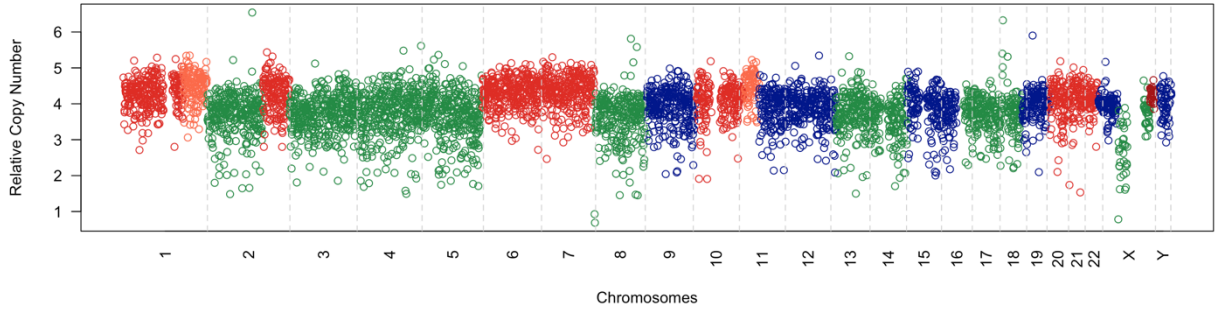
Sample 18



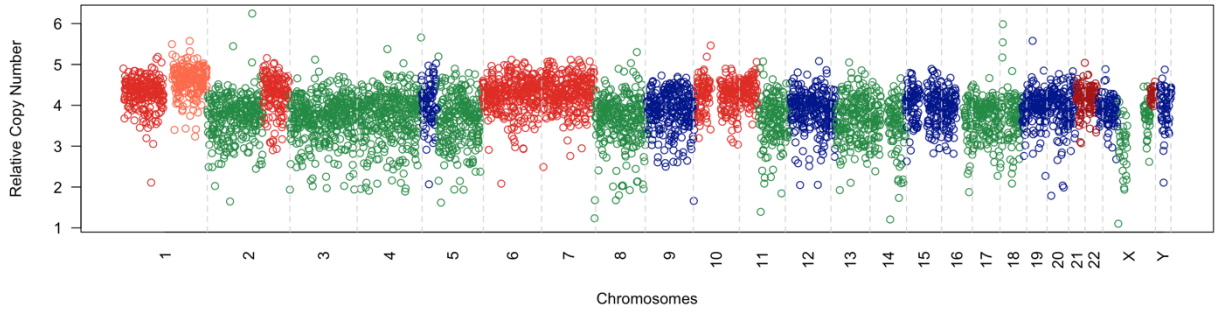
Sample 19



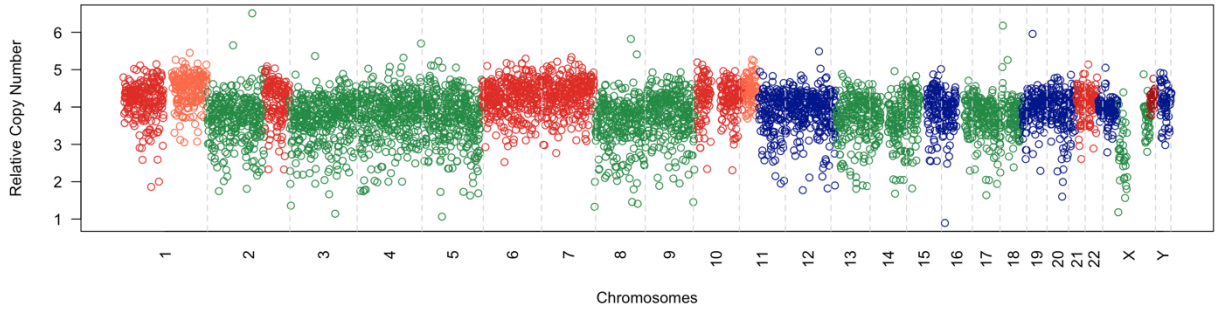
Sample 20



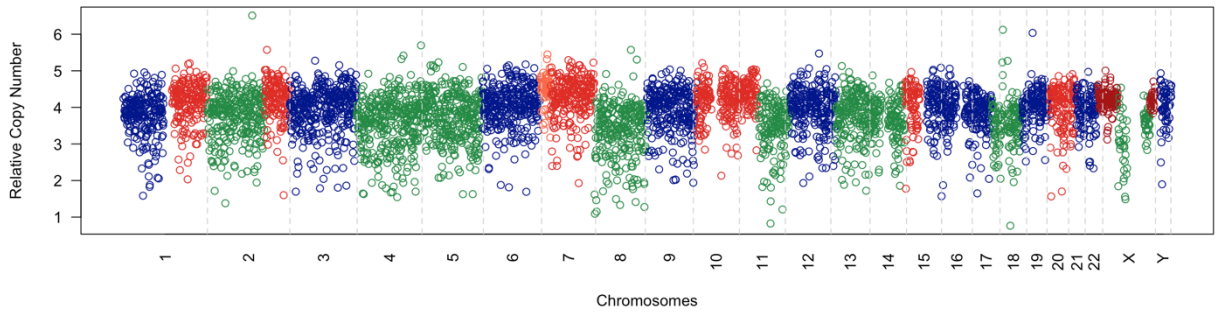
Sample 21



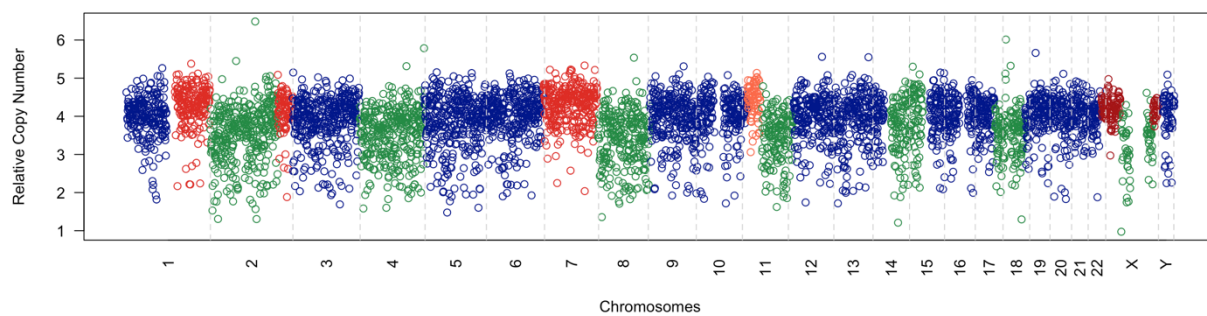
Sample 22



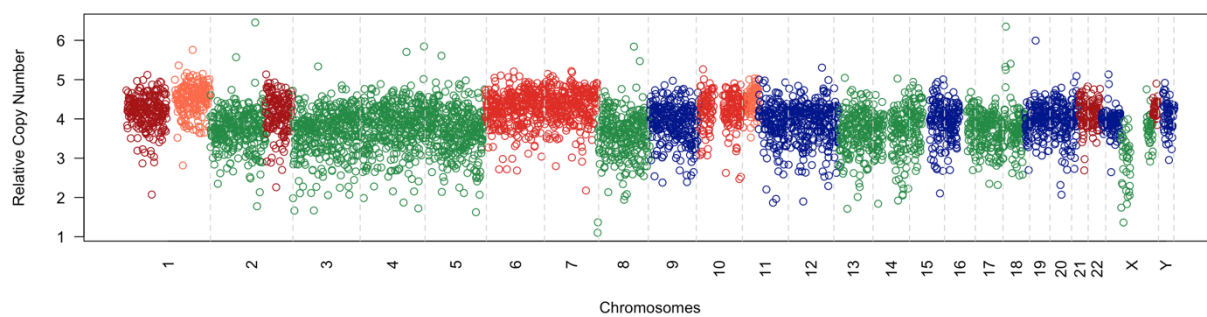
Sample 23



Sample 24



Sample 25



Appendix 3: BIC scores for the three models with gene names, Δ BIC, and best fitted model for each gene.

GENE	BEST MODEL	Indep	CME	CEM	Min.Model	Diff
HDAC4	CME	-2,2289733	-18,479551	-5,7020637	-18,479551	12,7774877
TGFBR1	CEM	-59,09022	-56,842739	-66,378335	-66,378335	7,28811516
LATS2	CME	-39,868038	-48,925034	-38,85295	-48,925034	9,0569962
P3H4	CME	-35,582237	-42,904957	-34,374466	-42,904957	7,32272003
NCKAP5L	CME	-50,69638	-57,526745	-49,964953	-57,526745	6,83036493
UROC1	CME	-47,470606	-62,226466	-50,898225	-62,226466	11,3282413
CATSPERG	CME	-40,13805	-51,706312	-44,541814	-51,706312	7,16449748
MAF1	CEM	-70,127162	-68,444081	-76,968214	-76,968214	6,8410515
EHD2	CME	7,93598202	-11,521332	8,96149495	-11,521332	19,4573145
HELZ2	CME	-6,8587646	-15,723993	-6,0152629	-15,723993	8,86522877
NUFIP2	CEM	-58,370376	-59,022175	-65,508333	-65,508333	6,48615817
HLA-A	CME	-58,056125	-65,12896	-56,785351	-65,12896	7,07283471
HLA-B	CME	-72,27374	-81,637386	-71,091871	-81,637386	9,36364647
HLA-C	CME	-67,592986	-77,134011	-65,980301	-77,134011	9,54102536
HLA-G	CME	-23,080286	-38,663098	-29,231813	-38,663098	9,43128521
SLC4A7	CME	-29,833926	-40,488045	-31,097326	-40,488045	9,39071884
THEMIS2	CME	-33,99145	-60,537583	-34,113619	-60,537583	26,4239642
ITGB4	CME	29,0678127	-29,146277	28,974366	-29,146277	58,1206427
SNRNP70	CME	-6,0085445	-14,698472	-6,4150288	-14,698472	8,2834435
KIF13A	CME	-46,502361	-53,038304	-46,093555	-53,038304	6,53594354
TTL11	CME	-14,006259	-48,017534	-18,578979	-48,017534	29,4385544
ARL4A	CEM	-62,684045	-59,091975	-69,277898	-69,277898	6,59385296
ATXN10	CME	-82,636159	-90,782049	-82,251582	-90,782049	8,14589083
ANKEF1	CEM	-25,726159	-40,127815	-47,722586	-47,722586	7,59477108
LAPTM5	CME	-56,000843	-63,220911	-57,111232	-63,220911	6,10967946
KIF16B	CME	-47,193052	-60,656727	-45,318361	-60,656727	13,4636755
COPG2	CME	-11,842149	-18,544694	-10,371415	-18,544694	6,70254568
CHCHD6	CME	-68,502955	-75,340588	-68,801841	-75,340588	6,53874713
SP5	CME	-27,810869	-43,78184	-28,468223	-43,78184	15,3136174
FZD1	CME	4,26501558	-21,936174	3,54813329	-21,936174	25,4843076
AK4	CME	-53,752006	-61,38631	-52,894717	-61,38631	7,63430357
ANO8	CME	-20,708698	-39,585338	-18,663597	-39,585338	18,8766392
PSTPIP1	CME	41,600559	6,84325264	43,9090974	6,84325264	34,7573064

ZNF700	CME	-46,804593	-61,114382	-44,843713	-61,114382	14,3097884
TSHZ2	CME	-47,007186	-58,972682	-45,425708	-58,972682	11,9654961
ODF2	CEM	-65,405178	-64,609607	-74,438585	-74,438585	9,03340712
BTBD11	CME	8,32123274	-2,4978474	9,76701825	-2,4978474	10,8190802
SUN2	CEM	-55,498792	-54,991637	-62,41162	-62,41162	6,91282775
COL28A1	INDEP	-86,307244	-78,533805	-78,511991	-86,307244	7,77343959
HIVEP2	CME	4,12833587	-10,745175	4,59399462	-10,745175	14,8735109
PSMC5	CEM	-45,457182	-46,417871	-54,094255	-54,094255	7,67638347
SLC2A4RG	CEM	-48,013562	-46,612315	-54,07152	-54,07152	6,05795734
PSMF1	CME	-66,835961	-74,502376	-66,130247	-74,502376	7,66641551
ZNF786	CME	-23,30706	-48,041465	-23,968608	-48,041465	24,0728574
FZR1	CME	-35,555658	-56,106893	-34,657015	-56,106893	20,5512352
CBX2	CME	-34,503671	-42,509904	-33,850562	-42,509904	8,00623258
RSRP1	CME	-55,201769	-63,952997	-52,459862	-63,952997	8,75122796
NFIX	CME	-11,655808	-19,964134	-10,442732	-19,964134	8,30832592
CWC22	INDEP	-27,506789	-16,534352	-15,914477	-27,506789	10,9724373
COPZ1	CME	-70,490997	-82,77623	-69,577861	-82,77623	12,2852322
NPTX1	CME	68,0746423	51,7231343	67,9681455	51,7231343	16,2450112
UGT1A4	CEM	-66,26618	-63,118433	-74,379629	-74,379629	8,11344894
ARHGEF10L	CME	-53,195984	-61,586743	-53,548454	-61,586743	8,03828945
GLIS3	CME	-10,506115	-23,987154	-11,415943	-23,987154	12,5712103
CLPTM1	CME	-66,026292	-74,549658	-66,222747	-74,549658	8,32691134
ACAP3	CME	-39,21396	-61,515604	-38,584288	-61,515604	22,3016447
ATG10	CME	-44,998985	-64,12722	-43,543441	-64,12722	19,1282351
UHRF1	CME	-23,074111	-41,539849	-21,529016	-41,539849	18,4657374
PTPRG	CME	-35,660284	-47,194171	-34,427225	-47,194171	11,5338862
PHKG1	CME	-0,6955023	-23,490236	-1,1380948	-23,490236	22,3521415
MRAS	CEM	-56,802938	-59,372961	-68,02502	-68,02502	8,65205865
RABGAP1	CME	-47,057699	-54,111555	-47,005697	-54,111555	7,05385588
DLG5	CME	-52,634652	-68,579269	-53,802621	-68,579269	14,7766476
CAMK2N1	CEM	-59,424113	-57,683585	-71,408041	-71,408041	11,9839285
LRWD1	CME	-62,244179	-69,951799	-60,71578	-69,951799	7,70761959
FRMD4B	CME	-51,175267	-59,970823	-52,937674	-59,970823	7,03314878
ZNF429	CME	-71,251297	-80,851	-69,724751	-80,851	9,59970248
UST	CME	-14,996172	-34,380256	-13,948147	-34,380256	19,3840844
ENTR1	CME	-67,936161	-74,041247	-66,871744	-74,041247	6,10508596
GPX1	CME	-55,405177	-66,344928	-53,634319	-66,344928	10,9397515

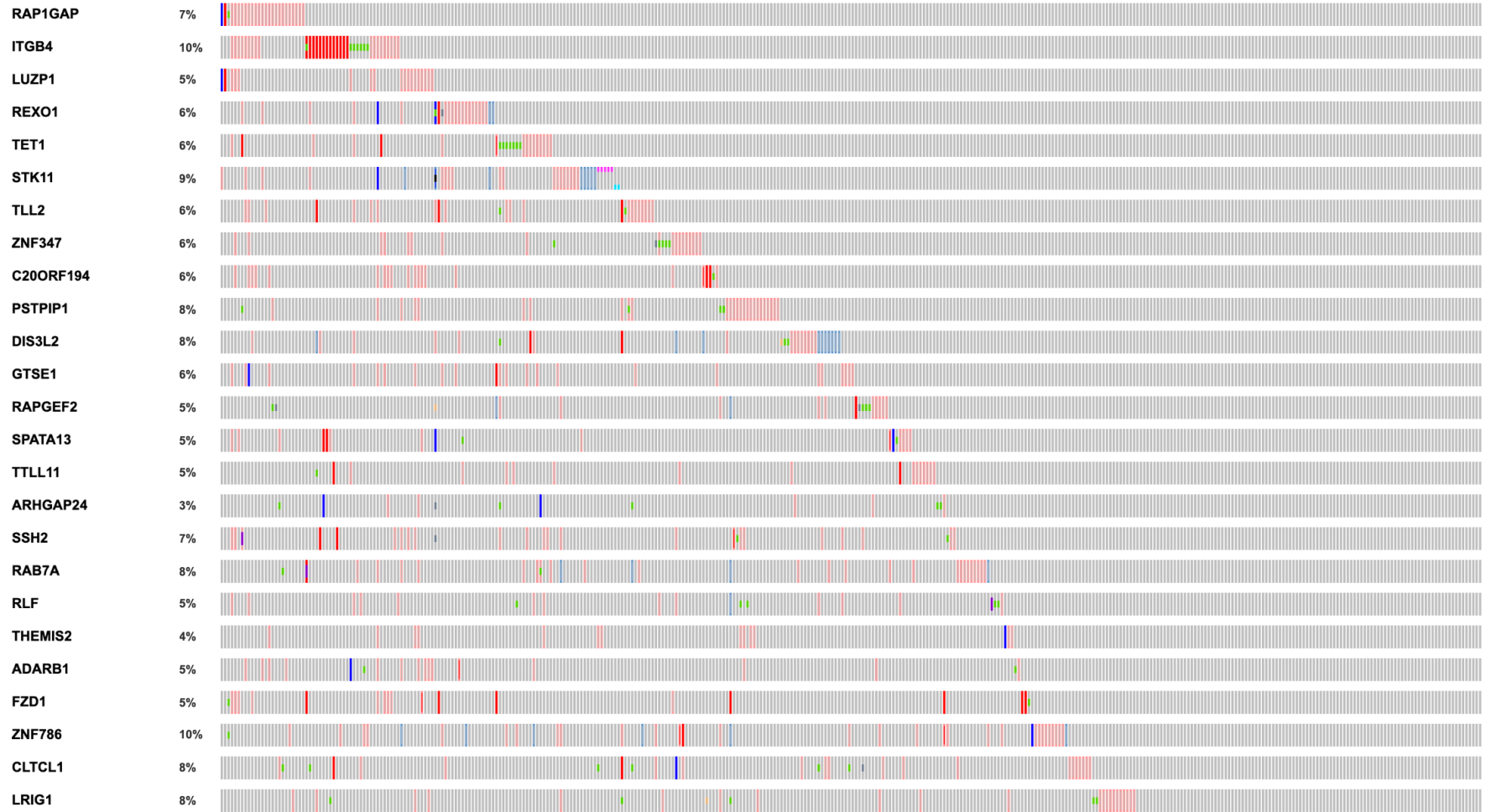
PIGT	CME	-71,033168	-80,197408	-70,246833	-80,197408	9,16423966
TMEM39B	CME	-62,253917	-72,866092	-60,319625	-72,866092	10,6121749
GMDS	CEM	-37,66387	-36,398546	-45,272452	-45,272452	7,60858254
L3MBTL1	CME	-33,970442	-50,661269	-32,050436	-50,661269	16,6908273
COL12A1	CME	80,3598151	71,757414	81,9586348	71,757414	8,60240116
IFFO2	CME	-39,68893	-46,39258	-39,141003	-46,39258	6,7036501
BTBD9	CME	-24,36432	-31,812188	-24,175993	-31,812188	7,44786854
MAMDC4	CEM	-41,167171	-40,143775	-50,515528	-50,515528	9,34835677
FRS2	CME	-44,630675	-56,50779	-43,985698	-56,50779	11,8771155
ANKS1B	CME	2,92743233	-37,252111	-16,069274	-37,252111	21,1828367
PDS5B	CEM	-58,200664	-56,787385	-64,431873	-64,431873	6,23120895
AHCY	CME	-63,320166	-70,007514	-62,124803	-70,007514	6,68734796
CENPP	CME	-52,846249	-60,208354	-51,460716	-60,208354	7,36210421
PIM1	CEM	-24,234289	-22,657799	-38,657848	-38,657848	14,4235589
A1BG	CME	-69,72231	-84,505093	-68,997867	-84,505093	14,7827831
SCARB1	CME	-65,198434	-71,637998	-64,165012	-71,637998	6,43956401
DIS3L2	CME	-30,12329	-63,123493	-28,643228	-63,123493	33,0002032
GNA11	CME	-53,549131	-61,646037	-53,075963	-61,646037	8,09690632
GNA14	CME	-35,475903	-43,212206	-36,001483	-43,212206	7,21072258
PDIA5	CME	-74,537469	-81,646969	-75,141641	-81,646969	6,50532815
RFTN2	CME	-56,840878	-64,228634	-55,385061	-64,228634	7,38775653
NACC1	CME	-58,825715	-66,232324	-58,249417	-66,232324	7,40660867
PLEKHN1	CME	-12,01164	-35,686039	-16,822539	-35,686039	18,8634998
ARFGEF3	CME	3,82533198	-20,671949	2,29568092	-20,671949	22,9676297
ZEB1	CME	-23,083044	-45,295135	-21,281171	-45,295135	22,2120913
PNKD	CME	-57,437771	-63,508796	-55,590406	-63,508796	6,07102529
ALDH2	CME	-65,059629	-78,804653	-64,189316	-78,804653	13,7450234
WDR62	CME	-37,925723	-46,434215	-36,642554	-46,434215	8,50849138
HSPG2	CEM	-5,5158039	-7,7273104	-25,577689	-25,577689	17,8503788
ZDHHC8	CME	-45,718486	-59,026968	-43,565569	-59,026968	13,3084818
GTSE1	CME	-28,43571	-60,784711	-27,023415	-60,784711	32,3490004
RLF	CME	-31,320055	-58,613355	-31,542228	-58,613355	27,0711272
FAM117A	CME	-52,754704	-62,028897	-51,727383	-62,028897	9,2741926
FAM117B	CME	-34,201284	-47,074986	-34,192211	-47,074986	12,8737025
ARF4	CME	-82,241377	-88,286585	-80,80578	-88,286585	6,04520802
XXYL1	CEM	-45,482335	-45,592783	-52,374628	-52,374628	6,78184507
RGMB	CEM	-10,661526	-23,830627	-30,70311	-30,70311	6,87248331

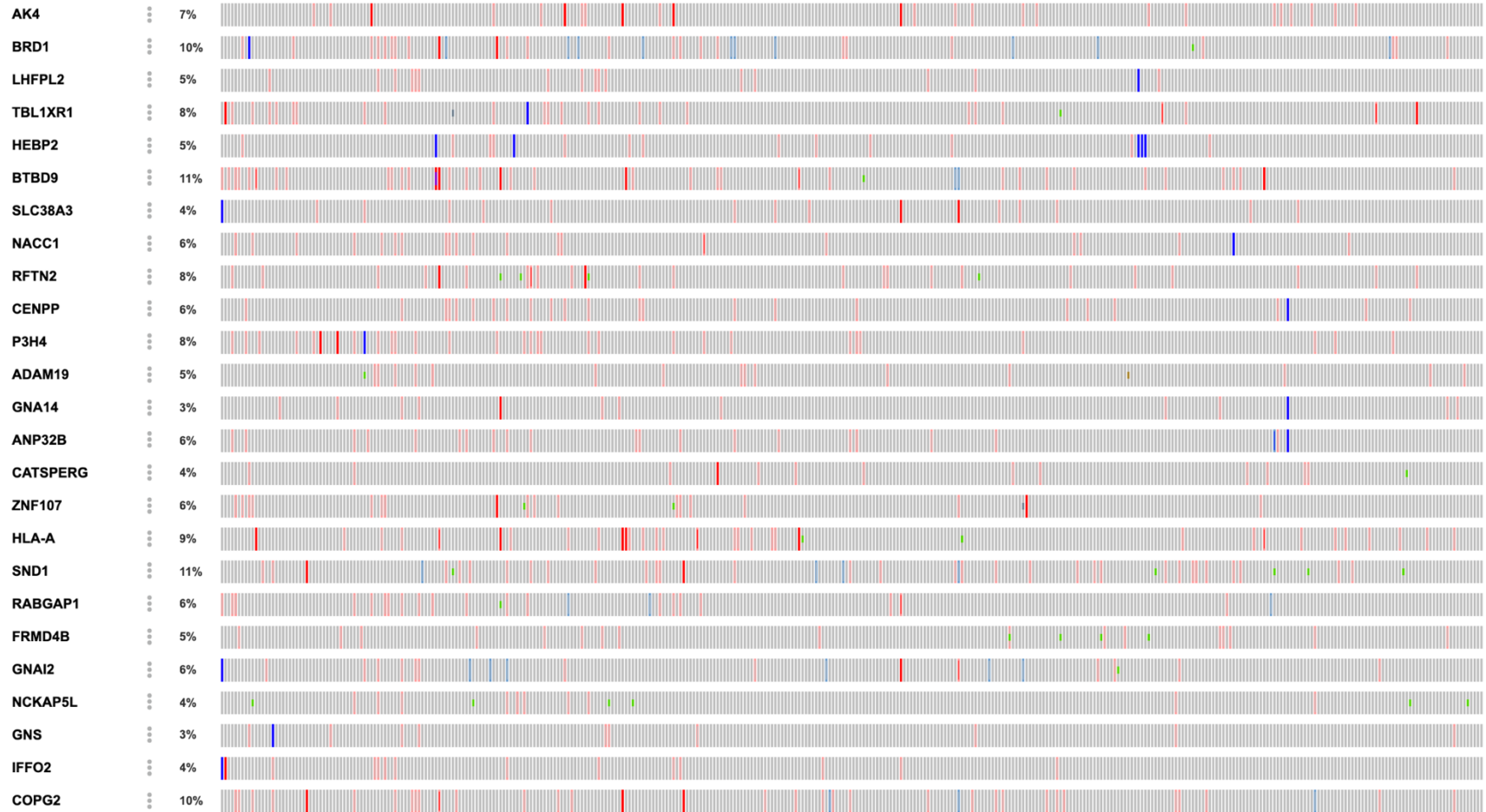
ZNF107	CME	-42,503284	-49,614011	-40,796006	-49,614011	7,11072752
RAB6B	CEM	-31,991675	-32,591917	-45,145562	-45,145562	12,5536449
RAB7A	CME	-52,742203	-79,915618	-51,455665	-79,915618	27,173415
FBXO46	CME	-42,308166	-52,553112	-42,301059	-52,553112	10,2449459
EPS15L1	CME	-48,540818	-63,440134	-48,493977	-63,440134	14,8993162
CHD1L	CME	-61,717998	-72,55296	-59,070026	-72,55296	10,8349616
ARHGGEF26	CME	-57,552618	-68,705615	-59,913875	-68,705615	8,79173977
ZNF43	CME	-19,236161	-31,696003	-23,12337	-31,696003	8,57263304
EIF4E3	CME	-33,108193	-51,184209	-33,799795	-51,184209	17,3844137
VASP	CME	-66,89586	-73,065628	-65,572146	-73,065628	6,16976798
ZNF608	CME	-57,018104	-69,479125	-55,571152	-69,479125	12,4610209
TANGO2	CEM	-55,976774	-55,55838	-64,523827	-64,523827	8,54705271
ELOA	CEM	-68,459181	-68,479901	-77,284425	-77,284425	8,80452404
FBLN1	INDEP	-77,628118	-63,326392	-62,736788	-77,628118	14,3017262
CLTCL1	CME	-6,9003265	-30,119774	-4,8268754	-30,119774	23,2194478
C3	CME	-41,174471	-50,282256	-41,930392	-50,282256	8,35186382
FAM210B	CME	-60,486713	-68,254596	-60,097645	-68,254596	7,76788363
ZNF91	CME	-63,150081	-78,583043	-61,677107	-78,583043	15,4329616
SIPA1L3	CME	-4,7935985	-13,95454	-3,7219214	-13,95454	9,16094176
C20orf194	CME	-33,803365	-69,445382	-31,216279	-69,445382	35,6420166
NTMT1	CEM	-59,766259	-59,240064	-66,277886	-66,277886	6,51162718
STOX2	CME	-33,302117	-55,151312	-40,914258	-55,151312	14,2370534
CADPS2	CME	-50,007022	-56,534209	-49,856638	-56,534209	6,5271871
JAK2	CME	-33,72356	-47,204693	-32,463885	-47,204693	13,4811328
ZNF675	CME	-60,561751	-71,271151	-58,335623	-71,271151	10,7094006
KLHL22	CEM	-66,500597	-64,506666	-73,448537	-73,448537	6,9479401
AP1S3	CME	-17,006715	-38,919654	-15,433197	-38,919654	21,9129397
PHLDB2	CME	-54,263327	-75,223849	-52,561066	-75,223849	20,9605215
BRD1	CME	4,50320836	-3,0659156	5,23430589	-3,0659156	7,56912394
GNAI2	CME	-65,534079	-72,719364	-65,747314	-72,719364	6,97205006
BMI1	CEM	-59,768611	-58,603519	-66,727551	-66,727551	6,9589396
DHX34	CME	-46,251003	-60,390347	-45,902786	-60,390347	14,1393432
YBX1	CEM	-53,157482	-53,255511	-60,344493	-60,344493	7,0889822
KMT5C	CME	-12,987651	-29,146431	-11,806874	-29,146431	16,1587798
ITPR1	CME	-12,523131	-25,028828	-14,520238	-25,028828	10,50859
TET1	CME	-8,380825	-57,94433	-9,7698197	-57,94433	48,17451
MARK4	CME	-58,369235	-64,497784	-57,428473	-64,497784	6,1285489

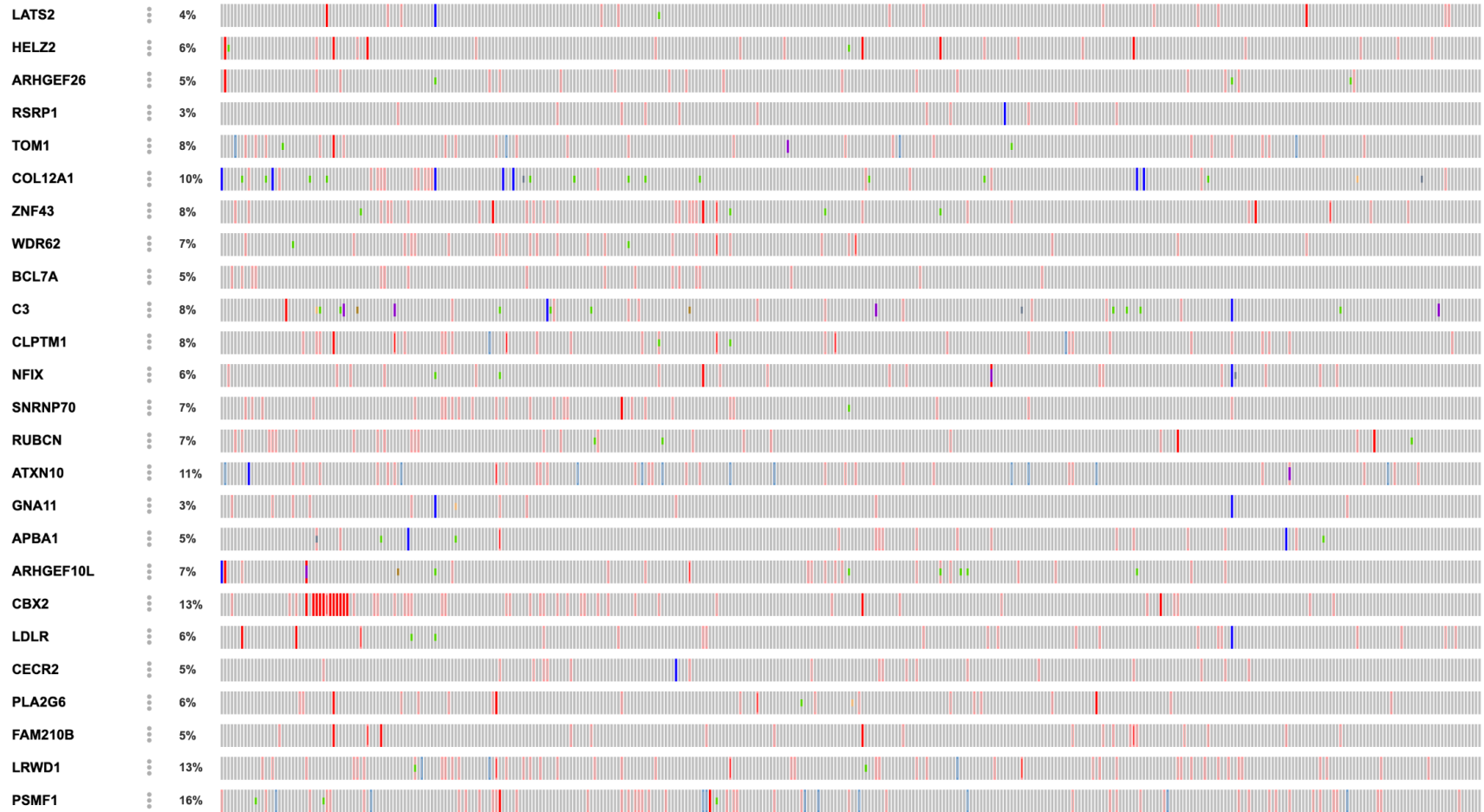
TOM1	CME	-78,011028	-86,634044	-73,619212	-86,634044	8,62301611
DSTN	CME	-61,097433	-73,851504	-58,952216	-73,851504	12,7540704
GAK	CME	-46,649746	-53,161432	-45,472577	-53,161432	6,51168576
ST6GALNAC 4	CME	-52,43062	-75,007581	-51,320688	-75,007581	22,5769602
RABL6	CME	-53,784171	-65,367156	-52,852869	-65,367156	11,5829841
MTFR1L	CEM	-61,247151	-60,980077	-73,939109	-73,939109	12,6919576
ADAM19	CME	-37,097275	-44,368317	-35,938057	-44,368317	7,27104174
TMEM242	CEM	-37,162515	-36,923777	-46,623105	-46,623105	9,46058923
REXO1	CME	10,9929448	-41,46522	11,6227675	-41,46522	52,4581648
BCL7A	CME	-21,806597	-30,791758	-22,306923	-30,791758	8,48483478
FGFR3	CME	-40,410875	-52,455586	-36,515396	-52,455586	12,0447113
TF	CME	-64,65454	-75,487241	-63,423622	-75,487241	10,8327006
RAP1GAP	CME	10,3989626	-52,40638	7,22874549	-52,40638	59,6351251
LUZP1	CME	-14,396593	-70,232264	-14,073252	-70,232264	55,8356708
RUBCN	CME	-38,044632	-46,193446	-34,705239	-46,193446	8,14881401
BIK	CME	-50,30767	-66,092135	-49,169156	-66,092135	15,7844655
ZSWIM5	CME	47,5456159	33,7123011	43,7407852	33,7123011	10,0284841
CECR2	CME	-22,563199	-30,343147	-20,103582	-30,343147	7,77994801
LHFPL2	CME	-16,589221	-26,403711	-18,84433	-26,403711	7,55938116
ANP32B	CME	-39,466503	-46,675387	-39,254521	-46,675387	7,20888463
LIG1	CME	-60,958603	-67,61533	-59,481212	-67,61533	6,65672748
LDLR	CME	-61,605096	-69,401363	-61,151765	-69,401363	7,79626689
TYMP	CME	-9,0532903	-15,391125	-6,5918043	-15,391125	6,33783498
ZNF347	CME	-28,016818	-64,120411	-26,623693	-64,120411	36,1035934
FRMPD1	CME	-20,567942	-27,787365	-21,190238	-27,787365	6,59712687
FOXD2	CME	-29,08813	-44,054082	-29,43678	-44,054082	14,6173015
PKD1L1	INDEP	-64,743319	-50,734456	-49,122527	-64,743319	14,0088628
RAC1	CME	-67,40579	-73,520123	-65,146622	-73,520123	6,11433276
FREM2	CEM	-39,203644	-37,188776	-49,476526	-49,476526	10,2728815
APBA1	CME	-47,005814	-55,089307	-45,91292	-55,089307	8,08349259
PNPLA7	CEM	-19,250545	-18,932948	-28,375081	-28,375081	9,12453577
HAUS5	CEM	-34,87385	-42,620966	-49,185277	-49,185277	6,56431054
GNS	CME	-73,874646	-80,630569	-72,769256	-80,630569	6,7559226
TRHDE	CEM	-42,894131	-41,984873	-54,110587	-54,110587	11,2164561
PLA2G6	CME	-51,552487	-59,323276	-50,539879	-59,323276	7,77078934
ESR1	CME	1,4961795	-13,431316	2,35437046	-13,431316	14,9274958
STK11	CME	6,09601194	-43,755401	3,59221162	-43,755401	47,3476124

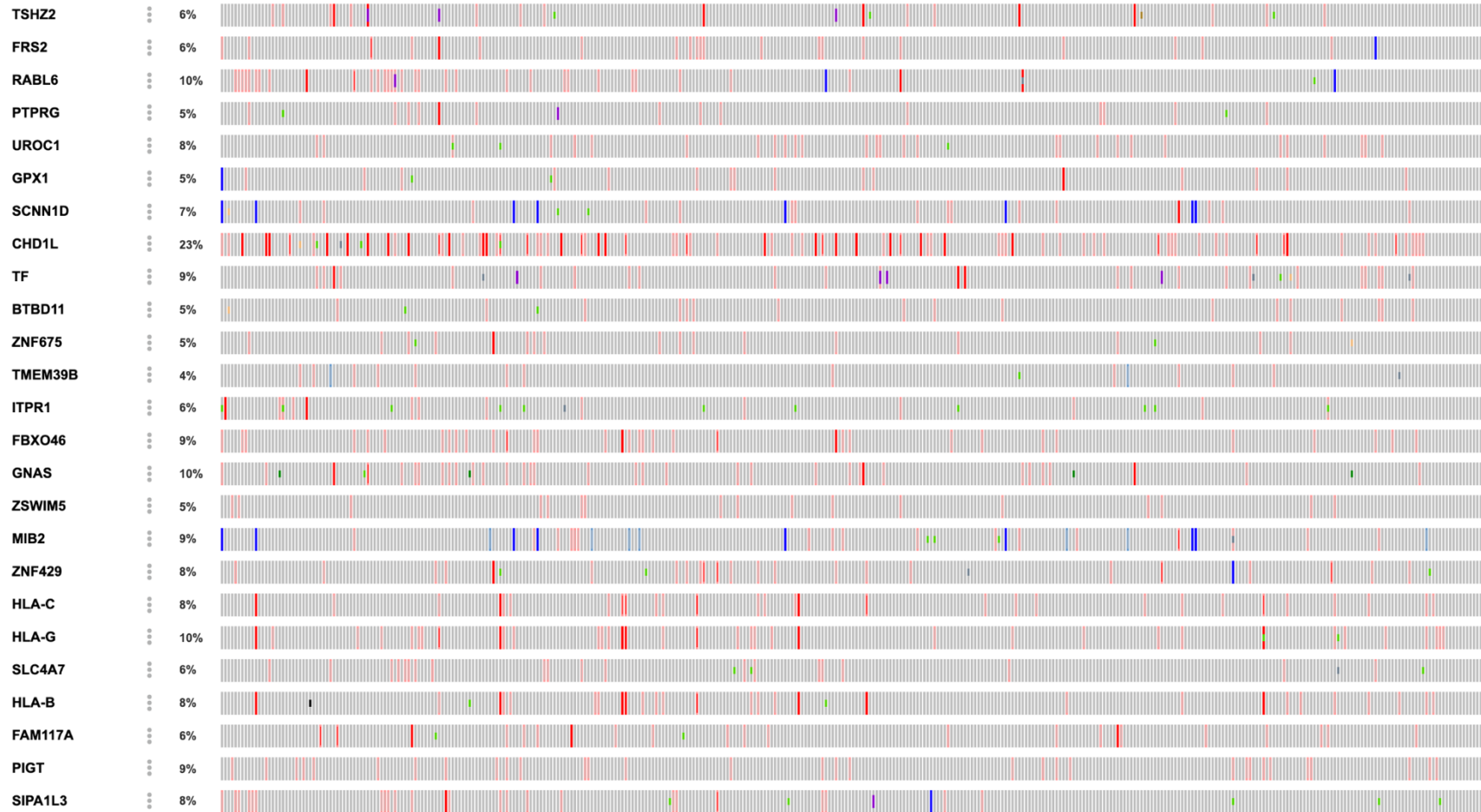
PRKCZ	CME	-16,755149	-34,271145	-20,132416	-34,271145	14,1387288
SND1	CME	-46,065388	-53,123913	-44,844091	-53,123913	7,05852517
SPATA13	CME	-35,608789	-65,388853	-35,677769	-65,388853	29,7110841
LRIG1	CME	-37,833024	-61,033158	-36,156772	-61,033158	23,2001341
TLL2	CME	11,8601275	-29,840538	13,6405709	-29,840538	41,7006655
IRS1	CEM	-30,015979	-28,963193	-37,104606	-37,104606	7,0886266
MYDGF	CME	-58,3155	-70,336916	-58,235188	-70,336916	12,0214155
MIB2	CME	-30,047178	-39,82836	-27,610242	-39,82836	9,78118212
PDE12	CME	-57,428404	-71,827493	-59,854041	-71,827493	11,9734527
SCNN1D	CME	-43,369011	-54,268854	-41,585817	-54,268854	10,8998438
HEBP2	CME	-34,181102	-41,711505	-32,448559	-41,711505	7,53040248
FANCD2	CME	-0,5609667	-15,195588	0,27097665	-15,195588	14,6346215
SIM2	CME	31,592153	14,6796178	29,9142257	14,6796178	15,2346079
ATXN2	CME	-46,295638	-59,556942	-45,216419	-59,556942	13,261304
ADARB1	CME	-26,375595	-52,125751	-26,481395	-52,125751	25,644356
TRANK1	CME	-10,399262	-16,415184	-9,6113325	-16,415184	6,01592134
SSH2	CME	18,2865896	-9,3089578	20,364836	-9,3089578	27,5955474
RNF10	CME	-78,825819	-85,486912	-77,10017	-85,486912	6,66109211
ERRFI1	CEM	-61,723329	-60,317221	-78,469228	-78,469228	16,7458998
GNAS	CME	-34,922358	-44,969172	-33,820296	-44,969172	10,0468134
RAPGEF2	CME	-27,028451	-57,824884	-25,756708	-57,824884	30,7964332
SLC38A3	CME	-85,161886	-92,598675	-82,551823	-92,598675	7,43678913
TBL1XR1	CME	-53,694264	-61,800256	-54,261038	-61,800256	7,53921769
STK17B	CME	-41,028115	-55,624315	-40,59359	-55,624315	14,5962001
UBR4	CME	-59,332041	-72,315623	-59,648909	-72,315623	12,6667141
ARHGAP24	CME	-24,572983	-54,974007	-25,69302	-54,974007	29,2809871

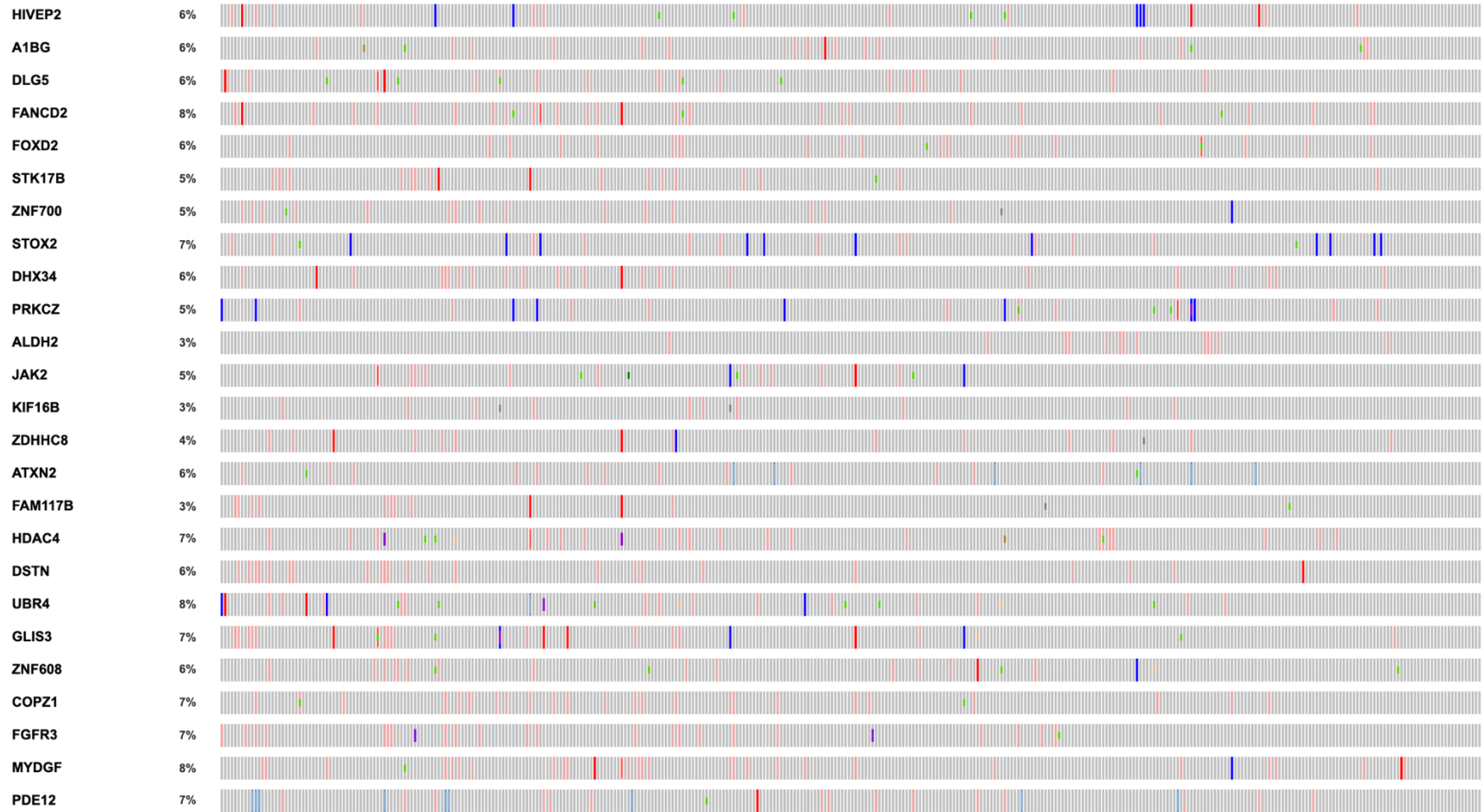
APPENDIX 4: The gene list of Group 2 (CME) and the related HCC studies reported by cbioportal

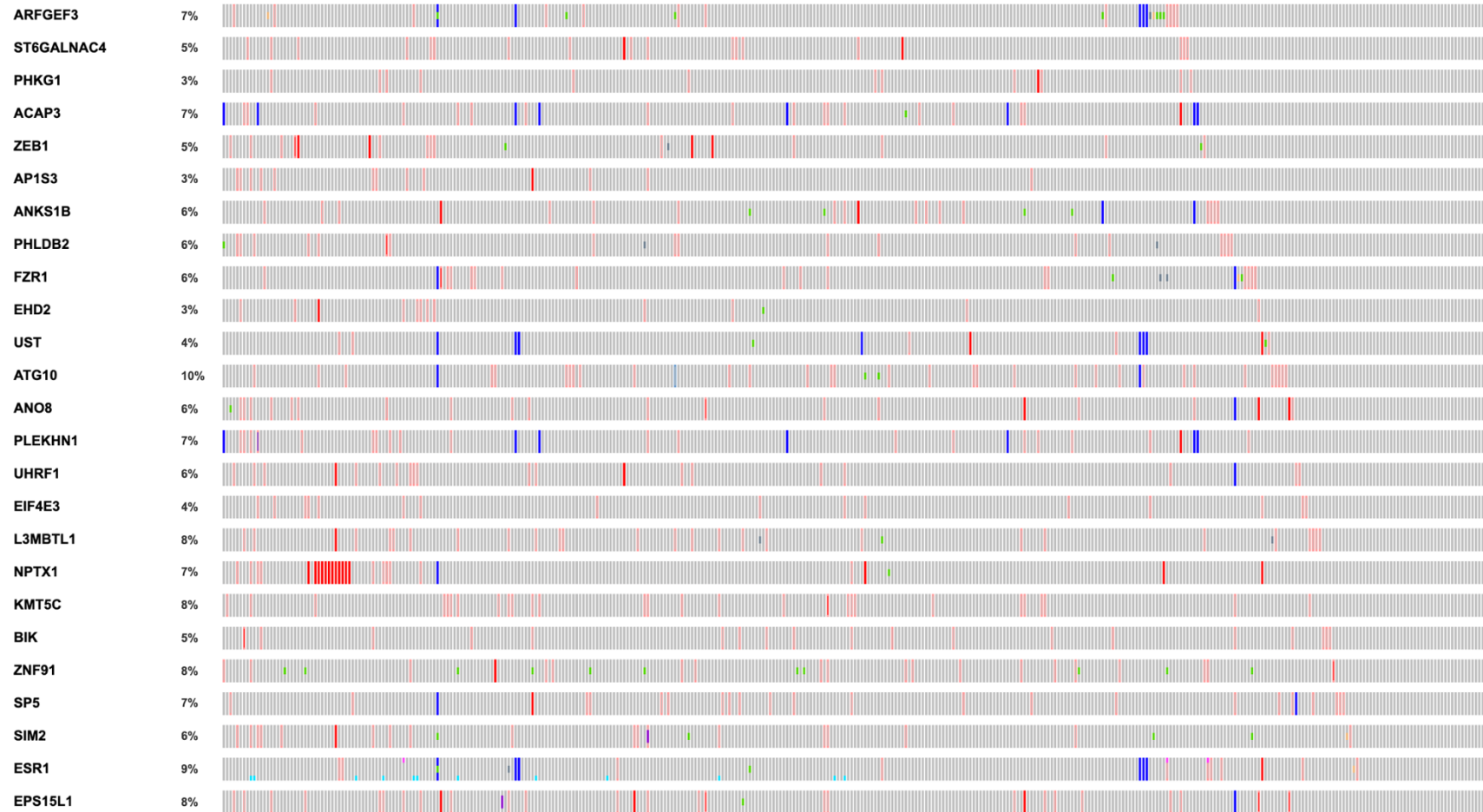


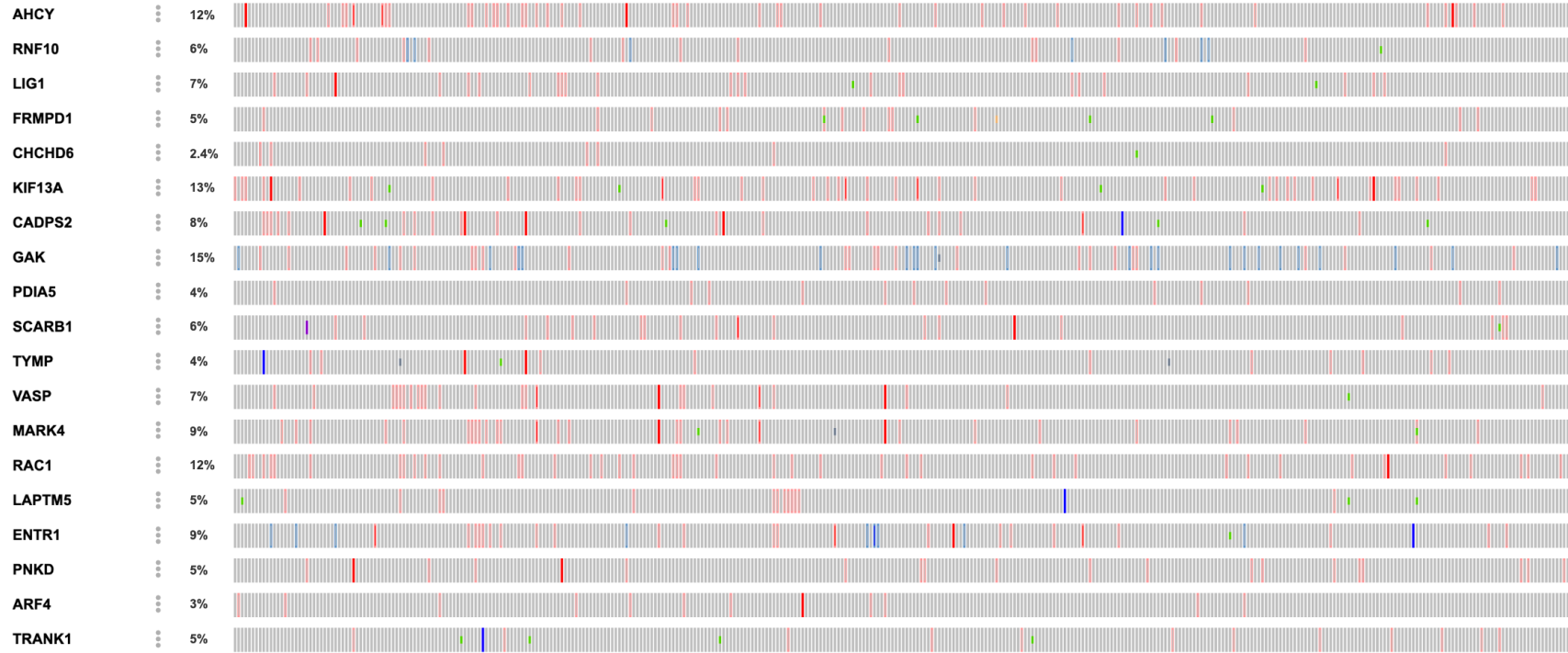













APPENDIX 5: Digital Receipt




Dijital Makbuz

Bu makbuz ödevinizin Turnitin'e ulaştığını bildirmektedir. Gönderiminize dair bilgiler şöyledir:

Gönderinizin ilk sayfası aşağıda gönderilmektedir.

Gönderen:	Muntadher Zahid Jihad
Ödev başlığı:	Muntadher Jihad Tez 2021
Gönderi Başlığı:	BAYESIAN NETWORKS FOR OMICS DATA ANALYSIS IN HEPAT...
Dosya adı:	thesis_after_defense_turnitin.docx
Dosya boyutu:	4.75M
Sayfa sayısı:	65
Kelime sayısı:	13,865
Karakter sayısı:	74,843
Gönderim Tarihi:	04-Mar-2021 10:18AM (UTC+0300)
Gönderim Numarası:	1523916231



Copyright 2021 Turnitin. Tüm hakları saklıdır.

APPENDIX 6: Thesis Originality Report

BAYESIAN NETWORKS FOR OMICS DATA ANALYSIS IN HEPATOCELLULAR CARCINOMA SINGLE-CELL SEQUENCING			
ORJİNALLİK RAPORU			
% 11	% 7	% 9	% 2
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ
BİRİNCİL KAYNAKLAR			
1	bmcbioinformatics.biomedcentral.com İnternet Kaynağı		% 1
2	www.nature.com İnternet Kaynağı		% 1
3	elifesciences.org İnternet Kaynağı		% 1
4	Hou, Yu, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang, and Jirun Peng. "Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas", Cell Research, 2016. Yayın		<% 1
5	"Handbook of Nutrition, Diet, and Epigenetics", Springer Science and Business Media LLC, 2019 Yayın		<% 1
6	www.ncbi.nlm.nih.gov İnternet Kaynağı		<% 1

APPENDIX 7: Girişimsel Olmayan Klinik Araştırmaları Etik Kurulu



T.C.
HACETTEPE ÜNİVERSİTESİ
Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu

Sayı : 16969557-503

Konu :

17.03.2020

Dr. Öğr. Üyesi İdil YET
Sağlık Bilimleri Enstitüsü
Biyoinformatik Anabilim Dalı
Öğretim Üyesi

Sayın Dr. Öğr. Üyesi YET,

Kurulumuza değerlendirilmek üzere sunduğunuz GO 20/232 kayıt numaralı ve "*Bayes Ağlarını Kullanarak Karaciğer Kanserinin Tekil Hücrelerinin Omikler Arasındaki İlişkisinin Belirlenmesi*" başlıklı proje Kurulumuzun 17.03.2020 tarihli toplantısında değerlendirilmiş olup, açık erişimli veri tabanı kullanılarak 25 karaciğer kanser hücresinde üç farklı omik verisi arasındaki ilişkilerin bilgisayar temelli yöntemlerle değerlendirileceği anlaşılmıştır. Gönüllü insanlar üzerinde gerçekleştirilecek nitelikte olmayan bu tip çalışmalar Etik Kurulların kapsamı dışında kalmaktadır.

Bu yazı ilgili protokolün etik açıdan incelendiğini belirtilmek için Etik Kurul kararı yerine geçmek üzere hazırlanmıştır.

Prof. Dr. Ayşe Lale DOĞAN
Başkan

EK :
Toplantı Katılım Tutanağı.

Hacettepe Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu
06100 Sıhhiye-Ankara
Telefon: 0 (312) 305 1082 • Faks: 0 (312) 310 0580 • E-posta: goetik@hacettepe.edu.tr

Ayrıntılı Bilgi için: