GENOME WIDE VARIATION ANALYSIS OF FORMALIN FIXED PARAFFIN EMBEDDED PULMONARY METASTATIC TUMOR SAMPLES OF OSTEOSARCOMA PATIENTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZELHA NİL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOLOGY

JANUARY 2012

Approval of the thesis:

GENOME WIDE VARIATION ANALYSIS OF FORMALIN FIXED PARAFFIN EMBEDDED PULMONARY METASTATIC TUMOR SAMPLES OF OSTEOSARCOMA PATIENTS

submitted by **ZELHA NİL** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Musa Doğan _____
Head of Department, **Biology**

Prof. Dr. Ufuk Gündüz
Supervisor, **Biology Dept., METU** _____

Examining Committee Members:

Prof. Dr. Fikret Arpacı _____
Medical Oncology Dept., GATA

Prof. Dr. Ufuk Gündüz _____
Biology Dept., METU

Prof. Dr. Semra Kocabıyık _____
Biology Dept., METU

Assist. Prof. Dr.Yeşim Aydın Son _____
Informatics Dept., METU

Assist. Prof. Dr. Sreeparna Banerjee _____
Biology Dept., METU

Date: *10/01/2012*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Zelha Nil

Signature :

ABSTRACT

GENOME-WIDE VARIATION ANALYSIS OF FORMALIN FIXED PARAFFIN EMBEDDED PULMONARY METASTATIC TUMOR SAMPLES OF OSTEOSARCOMA PATIENTS

Nil, Zelha

M.Sc., Department of Biology

Supervisor: Prof. Dr. Ufuk Gündüz

January 2012, 201 pages

Osteosarcoma (OS) is a type of cancer that starts in the bone. It generally occurs in the cells called osteoblasts which form matrix of the bone. It is the most common malignant tumor of bone with an incidence rate of 19% among all cancer types. The vast majority of OS patients have pulmonary metastases at the time they are diagnosed, and about half develop lung disease later. Moreover, pulmonary metastatic tumors lead to poor prognosis and increased death rate. Although mutations in the genes coding for p53, Rb, fos and myc were detected in pulmonary metastatic tumors of OS, there is no unique genetic pathway identified for progression of pulmonary metastasis.

In this research, a genome wide association study (GWAS) using FFPE samples from lung tissue of 9 patients with pulmonary metastatic OS was

performed. Among 358 associated SNPs, rs6499861, rs10884554 and rs12154602 were found to be associated with metastatic OS most significantly. Moreover, second wave analysis of GWAS results provided the significant genes and pathways associated with metastatic OS.

A methodology for copy number aberration and LOH analysis of SNP array data of a FFPE sample was generated using R-aroma package. Results were obtained by three different methods, namely, CalMaTe, TumorBoost and Virtual Normal algorithm. Among these, CalMaTe was found to produce less noisy data than VN Algorithm during total copy number segmentation. LOH analysis could only be performed for one sample with the second method due to poor data quality of the other samples.

According to the results of copy number aberration and LOH analysis of one tumor sample T8, copy number gains in 1p31.1, 6p21.32, 7p14.3, 11q22.1, 12p12.1, and 18q12.1 chromosomal regions and copy number losses in 2p16.2, 8q24.13, 17q23.3 and 17q21.31 chromosomal regions have been found. Moreover, LOH events were observed in 2q14.3, 11q13.4, 18p11.21, 19q12, 20p13 and 23q21.1 chromosomal regions.

Identification associated SNPs and significant copy number changes may be helpful in investigation of potential diagnostic and prognostic markers in metastatic osteosarcoma.

Keywords: Osteosarcoma, SNP, Mutation, Pulmonary metastasis

ÖZ

OSTEOSARKOM HASTALARININ PARAFİNLENMİŞ AKCİĞER METASTATİK TÜMÖR ÖRNEKLERİNDE GENOM DÜZEYİNDEKİ ÇEŞİTLİLİKLERİN ANALİZİ

Nil, Zelha
Yüksek Lisans, Biyoloji Bölümü
Tez Danışmanı: Prof. Dr. Ufuk Gündüz

Ocak 2012, 201 sayfa

Osteosarkom (OS) kemik dokusunda başlayan bir kanser türüdür. Genellikle, kemik matrisini oluşturan osteoblastlarda meydana gelir. Kanser türleri içinde yüzde on dokuzu oluşturan OS en yaygın malign kemik tümörüdür. OS hastalarının büyük çoğunluğu teşhis konulduğunda akciğer metastazı göstermektedir ve hastaların yarısı daha sonra akciğer hastalığına yakalanmaktadır. Ayrıca, pulmoner metastatik tümörler, kötü prognoz ve ölüm oranlarında artışa sebep olmaktadır. Pulmoner metastatik tümörlerde p53, Rb, Fos and Myc proteinlerini kodlayan genlerde mutasyonlar gözlenmiş olmasına rağmen, henüz pulmoner metastaza özgün bir genetik yolak belirlenememiştir. Bu sebeple, OS pulmoner metastazına sebep olan moleküler mekanizmaların araştırılması gereklidir.

Bu çalışmada, 9 hastadan elde edilmiş parafinlenmiş akciğer metastatik OS tumor örneklerinde genomboyu ilişkilendirme analizi gerçekleştirilmiştir. İlişkili bulunan 358 SNP arasından, istatistiksel olarak en önemli ilişkileri gösteren SNPler, rs6499861, rs10884554 ve rs12154602'dır. Ek olarak, genom boyutunda ilişkilendirme sonuçlarının ikincil analizi yapılmış, ve SNPlerle ilişkili gen ve biyolojik yolaklar tespit edilmiştir.

Parafinlenmiş örneklerin kopya sayısı değişiklikleri ve heterozigosite kaybı analizi için, aroma R paketi kullanılarak, bir metodoloji geliştirilmiştir. Bu metodolojide kullanılan 3 yöntemin sonuçlarına göre, elde edilen toplam kopya sayısı segmentasyon verileri, CalMaTe metodu kullanıldığında VN algoritmasına göre daha az gürültülü bulunmuştur. İkinci metod kullanılarak yapılan heterozigosite kaybı analizi, sadece tek bir örnek için sonuç vermiştir.

Kopya sayısı anormallikleri ve heterozigosite kaybı analizi tek bir tumör örneği üzerinde yapılmıştır. 1p31.1, 6p21.32, 7p14.3, 11q22.1, 12p12.1, ve 18q12.1 kromozom bölgelerinde kopya sayısı artışı ve 2p16.2, 8q24.13, 17q23.3 ve 17q21.31 kromozom bölgelerinde kopya sayısı azalışı gözlenmiştir. Ayrıca, 2q14.3, 11q13.4, 18p11.21, 19q12, 20p13 ve 23q21.1 kromozom bölgelerinde heterozigosite kaybı gözlenmiştir.

İlişkili SNPlerin ve önemli kopya sayısı değişikliklerinin belirlenmesi, metastatik osteosarkom için potansiyel tanı ve prognostik belirteçlerinin araştırılmasına yardımcı olacaktır.

Anahtar kelimeler: Osteosarkom, SNP, Mutasyon, Pulmoner metastaz

To my family and me,

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my supervisor Prof. Dr. Ufuk Gündüz for her support throughout this study. I would also thank to Prof. Dr. Fikret Arpacı for his valuable contributions, support and encouragement in this research.

I feel great appreciation to Prof. Dr. Mükerrem Safalı for his sincere guidance and confidence throughout this study.

I would like to thank Assist. Prof. Dr. Yeşim Aydın Son for her support, guidance and contributions; to her former students Dr. Gürkan Üstünkar and Onat Kadıoğlu, M.Sc.; and to all the members of Informatics Institute, especially to Kerem and Adnan.

I am grateful to Assist. Prof. Dr. Henrik Bengsston for his help during data analysis, and I greatly appreciate him for his great work with aroma.affymetrix.

Examining committee members Prof. Dr. Semra Kocabıyık and Assist. Prof. Dr. Sreeparna Banerjee are greatly acknowledged for their participation and valuable comments.

I am grateful to Tuğba Keskin, Gülistan Tansık and Burcu Özsoy for their support and friendship throughout the whole university years. I also thank to Yaprak Dönmez, Pelin Sevinç, Gülşah Pekgöz, Esra Güç, Esra Kaplan, Çağrı Urfalı, Sevilay Akköse, and Özlem Darcansoy İşeri for their friendship during my master study. Also special thanks to all other members of Lab206 Team for their support and for providing a stimulating environment: Ahu İzgi, Neşe Çakmak, Aktan Alpsoy, Murat Erdem, Okan Tezcan, Çiğdem Şener,

Rouhollah Khodadust, Gözde Ünsoy. All special project students that I have worked with are greatly acknowledged for their friendship and contributions throughout this study.

I also thank to Hadiye Demir, Emin Eker and Ozan Tuğluk for their support and friendship throughout my study.

My dear friends "187/8", Ayşegül Özgen, Ahmet Aydoğan, Mengü Türk, and Onur Özalan were always with me, supporting me during this thesis and my whole life. I cannot thank enough to them.

Throughout the whole university years, Dinçer Çevik was always with me, supporting me, helping me… Thank you "incelikli hayta".

My dear friend, Efsun Ağırtaş, is greatly acknowledged and appreciated for her help during this thesis, even though she had lost one of her ears after that, she is still with me…

I would like to express my great appreciation to my sisters Ümran Nil and Ferda Eser Nil. I am deeply indebted to my mother Hatice Nil and father Mehmet Cüneyt Nil for their endless love, trust and support in every step of my life.

This study was supported by the Research Funds of METU Grant No: BAP-08-11-2010-R-19 and GATA Grant No: AR2010/10. Also, I gratefully acknowledge the fellowship I had through my master years from The Scientific and Technical Research Council of Turkey (TUBITAK) - BİDEB.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ vi

ACKNOWLEDGEMENTS ix

TABLE OF CONTENTS xi

LIST OF TABLES xvii

LIST OF FIGURES xx

LIST OF ABBREVIATIONS xxiii

CHAPTERS

1 INTRODUCTION 1

1.1 Cancer 1

1.2 Sarcomas of the Bone 2

1.2.1 Malignant Bone Tumors 4

1.2.1.1 Osteosarcoma 4

1.2.1.1.1 Clinical Characteristics 5

1.2.1.1.2 Molecular Characteristics of Osteosarcoma 6

1.2.1.1.3 Metastasis of Osteosarcoma 10

1.2.1.1.4 Treatment 11

1.2.1.1.4.1 Surgery 11

1.2.1.1.4.2 Chemotherapy 12

1.2.1.1.4.3 Radiation Treatment 13
1.3 Formalin Fixed Paraffin Embedded Samples in Molecular Studies..... 13
1.4 Molecular Profiling..... 14
1.4.1 Microarray Technology..... 15
1.4.1.1 Applications of Microarrays 16
1.4.1.1.1 Gene Expression Profiling Analysis..... 17
1.4.1.1.2 Genome-Wide Association Studies..... 17
1.4.1.1.2.1 GWAS Tools..... 20
1.4.1.1.3 Copy Number Variation and Loss of Heterozygosity Analysis……22
1.4.1.1.3.1 Copy Number State Analysis Tools 25
1.4.2 Affymetrix GeneChip Human Mapping 250K SNP Array Platform……..26
1.5 AIM OF THE STUDY..... 28
2 MATERIALS AND METHODS 30
2.1 MATERIALS 30
2.1.1 Patients 30
2.1.2 Tissues..... 31
2.1.3 Chemicals and Reagents..... 31
2.1.4 Primers 32
2.1.5 Microarrays..... 33
2.2 METHODS..... 34
2.2.1 Treatment Protocol of Patients 34
2.2.2 Staging and Classification of Tumor Specimens..... 36

2.2.3 Preparation of Tissues for DNA Isolation 37
2.2.3.1 Manual Microdissection 37
2.2.3.2 Deparaffinization and Rehydration 38
2.2.4 DNA Isolation from Tissues 39
2.2.5 Quality Assessment of Isolated DNA 40
2.2.5.1 Spectrophotometric Analysis of DNA 41
2.2.5.2 Agarose Gel Electrophoresis 41
2.2.5.3 Gene Specific Polymerase Chain Reaction 42
2.2.5.4 Randomly Amplified Polymorphic DNA Polymerase Chain Reaction 43
2.2.6 Microarray Analysis 44
2.2.6.1 Restriction Enzyme Digestion 45
2.2.6.2 Ligation 46
2.2.6.3 PCR 48
2.2.6.4 PCR Product Purification and Elution 49
2.2.6.5 Quantitation and Normalization 51
2.2.6.6 Fragmentation of PCR Products 52
2.2.6.7 Labeling of Fragmented PCR Products 53
2.2.6.8 Target Hybridization 54
2.2.6.9 Washing and Staining 56
2.2.6.10 Scanning and Preliminary Analysis 59
2.2.6.11 Microarray Data Analysis and Visualization of Results 61
2.2.6.11.1 Genotyping: BRLMM Algorithm 61

2.2.6.11.2 Genome-Wide Association Analysis and SNP Prioritization 63

2.2.6.11.2.1 Data Preprocessing and Cleaning 64

2.2.6.11.2.2 GWAS: SNP, Gene and Pathway Association 66

2.2.6.11.2.3 SNP Prioritization 68

2.2.6.11.3 Genome-Wide Copy Number Aberration and Loss of Heterozygosity Analysis 69

2.2.6.11.3.1 Preprocessing 73

2.2.6.11.3.2 Probe Summarization 74

2.2.6.11.3.3 Post-Processing 75

2.2.6.11.3.4 Raw Copy Number Calculation 76

2.2.6.11.3.5 Copy-Number Segmentation 79

3 RESULTS AND DISCUSSION 81

3.1 The Patients 81

3.2 DNA Isolation and Optimization 82

3.3 Quality Assessment of Isolated DNA 83

3.3.1 Spectrophotometric Measurement and Agarose Gel Electrophoresis 83

3.3.2 Gene Specific PCR for *β-Actin* Gene and RAPD-PCR 85

3.4 Optimization of the Microarray Protocol 88

3.5 Microarray Data Analysis 91

3.5.1 Preliminary Data Analysis: Genotyping and Genotype Data Quality 91

3.5.2 Genome-Wide Association Analysis: Disease, Gene and Pathway Association 93

3.5.2.1 Data Preprocessing and Cleaning 93

3.5.2.2 GWAS: SNP, Gene and Pathway Association 98

3.5.2.3 SNP Prioritization 109

3.5.3 Copy Number Variation and Loss of Heterozygosity Analysis .. 114

3.5.3.1 Comparison of the Three Methods for CNV and LOH Analysis……...115

3.5.3.2 Comparison and Biological Interpretation of CNA and LOH Results of Tumor Sample T8 120

4 CONCLUSION 136

4.1 Recommendations..... 140

REFERENCES 142

APPENDICES

A. PHOTOGRAPHS OF HEMATOXYLIN-EOSIN STAINED SECTIONS……..157

B. BUFFERS AND SOLUTIONS..... 161

C. STATISTICAL BACKGROUND FOR SNP - COMPLEX DISEASE ASSOCIATION ANALYSIS 164

D. FILE STRUCTURE OF PLINK STATISTICS FILES..... 167

E. AHP SCORING SCHEME..... 168

F. AROMA PACKAGE R VIGNETTES FOR CNV ANALYSIS METHODS……..170

G. TOP 100 SIGNIFICANT SNPs, GENES AND PATHWAYS..... 179

H. CNA AND LOH RESULTS OBTAINED BY THREE METHODS 187

I. RAW CN AND BAF VS GENOMIC POSITION GRAPHS OF CHROMOSOMES WITH CNA FOR SAMPLE T8 196

LIST OF TABLES

TABLES

Table 1.1 Main Cancer Types 2
Table 1.2 General Classification of Bone Tumors 3
Table 1.3 Minor and major copy number states presented as the conjunction of information regarding total copy number (columns) and heterozygosity status (rows) 24
Table 2.1 Primers used in gene specific PCR and RAPD-PCR, and amplicon sizes. 33
Table 2.2 The contents of 250K Sty Assay Kit. 33
Table 2.3 Amplification conditions for *β-Actin* gene and RAPD-PCR 44
Table 2.4 Reagents used in Sty I digestion master mix. 46
Table 2.5 Digestion conditions for Sty I enzyme. 46
Table 2.6 a) Reagents used in ligation master mix. b) Ligation conditions for Sty I adaptor. 47
Table 2.7 The reagents used in target amplification PCR reaction. 48
Table 2.8 Amplification conditions for target amplification PCR reaction. 49
Table 2.9 Fragmentation conditions. 52
Table 2.10 Reagents used in labeling master mix. 54
Table 2.11 Reaction conditions for labeling. 54
Table 2.12 Reaction mixture preparation for hybridization. 55
Table 2.13 Washing/Staining protocol for mapping arrays. 58
Table 2.14 Manufacturer's Parameters for Genotyping Analysis 61
Table 3.1 Clinicopathologic properties of the patients 81
Table 3.2 Quantity and Quality Values of DNA Samples 83

Table 3.3 QC and SNP call rates of each sample..... 92
Table 3.4 Performance of 250K Sty Mapping Assay..... 93
Table 3.5 Individual SNP *p*-values of association of GWAS..... 100
Table 3.6 Top 20 significantly enriched genes according to combined *p*-value. 102
Table 3.7 Top 20 significant pathways according to combined *p*-value..... 105
Table 3.8 Significantly associated genes and corresponding SNPs in cation transport and skeletal development pathways. 106
Table 3.9 Top 20 SNPs according AHP prioritization of GWAS. 111
Table 3.10 Association of Prioritized SNPs to Genes and Diseases in the Databases 112
Table 3.11 Summary of differences between three methods of CNA and LOH analysis 116
Table 3.12 Number of regions of CNA and LOH obtained by three methods119
Table 3.13 Noise level estimators for CalMaTe and VN Algorithm 120
Table 3.14 CNA regions obtained for sample T8 121
Table 3.15 Genes found in the CNA regions..... 126
Table 3.16 LOH regions obtained by Tumorboost 130
Table 3.17 Genes found in the LOH regions. 132
Table 4.1 CNA regions of sample T8 with corresponding genes in those regions 139
Table 4.2 LOH regions of sample T8 with corresponding genes in those regions 139
Table D.1 File structure of PLINK descriptive statistics files. 167
Table E.1 Scoring scheme of SNPs..... 168
Table G.1 Top 100 SNPs according unadjusted *p*-value..... 179
Table G.2 Top 100 genes according combined *p*-value. 180
Table G.3 Top 100 pathways according combined *p*-value. 183
Table H.1 CNA segments obtained by CalMaTe 187

Table H.2 CNA segments obtained by VN algorithm 188

LIST OF FIGURES

FIGURES

Figure 1.1 Karyotype of conventional osteosarcoma showing multiple structural and numerical aberrations (Hameed & Dorfman, 2011)..... 7
Figure 1.2 Key molecular factors in Osteosarcoma (Clark, Dass, & Choong, 2008)..... 8
Figure 1.3 Summary of validation of results of a GWAS (Patel & Ye, 2011). 20
Figure 1.4 Copy number states of a tumor-normal match sample. 22
Figure 1.5 GeneChip Mapping Assay Overview. 27
Figure 2.1 Treatment protocol. CT: chemotherapy; IPA: ifosfamide, cisplatin and adriamycin; G-CSF: granulocyte–colony-stimulating factor; IEA: ifosfamide, etoposide, and adriamycin. 35
Figure 2.2 Distribution of PCR products into purification plate 50
Figure 2.3 The back view of cartridges and application of tough spots..... 56
Figure 2.4 Equipments in GeneChip System. 57
Figure 2.5 Genotyping console workflow. 60
Figure 2.6 BRLMM algorithm workflow (T. D. Model et al., 2006). 62
Figure 2.7 Logic Workflow of METU-SNP Software System (Gurkan Üstünkar, 2011)..... 63
Figure 2.8 Flowchart of CNA and LOH analysis from SNP arrays..... 72
Figure 2.9 Reference selection for VN Algorithm. 77
Figure 2.10 Locus-level estimates of a) TCNs and b) ASCNs with corresponding CNSs. 79
Figure 3.1 DNAs isolated from patient samples and MCF7 breast cancer cell line on 1% agarose gel..... 84

Figure 3.2 PCR products on 2% agarose gel. 86
Figure 3.3 PCR products on 2% agarose gel. 87
Figure 3.4 Restriction Enzyme digestion products a) Standard protocol b) Modified protocol..... 89
Figure 3.5 PCR amplification gel of adaptor ligated fragments. 89
Figure 3.6 Fragmentation product of sample N1. 90
Figure 3.7 Array image and quality controls. a) A whole array image b) Checkerboard pattern throughout the array c) B2 oligonucleotide control at left bottom corner 90
Figure 3.8 Graphical display of QC metrics 91
Figure 3.9 Number of SNPs per chromosome 94
Figure 3.10 MAF distribution of SNPs as a proportion vs number of SNPs. The values on the bars represent the exact number of SNPs for each proportion range. 95
Figure 3.11 Proportion of missing SNPs for respective samples vs number of samples. 95
Figure 3.12 Proportion of individuals which is missing for respective SNPs vs number of SNPs. 96
Figure 3.13 Manhattan Plot of negative logarithm of unadjusted *p*-value of association for individual SNPs and their distribution on individual chromosomes. 99
Figure 3.14 Chromosomal distribution of prioritized SNPs. 110
Figure 3.15 Difference between the CalMaTe and VN Algorithm in terms of signal to noise ratio in TCN signals. 118
Figure 3.16 The panels for a) Chr1 b) Chr2 c) Chr6 d) Chr7 e) Chr8 f) Chr11 g) Chr12 h) Chr17 i) Chr18 corresponding to segmentation results which are drawn as relative copy number vs physical position. 125
Figure 3.17 Allele B fractions with specified regions of LOH a) Chr2 b) Chr11 c) Chr18 d) Chr19 e) Chr20 and f) Chr23 (X chromosome). 132

Figure A.1 a) Tumor tissue, b) Normal tissue of Patient 1..... 157
Figure A.2 a) Tumor tissue, b) Normal tissue of Patient 2..... 158
Figure A.3 a) Tumor tissue, b) Normal tissue of Patient 3..... 158
Figure A.4 a) Tumor tissue, b) Normal tissue of Patient 4..... 158
Figure A.5 a) Tumor tissue, b) Normal tissue of Patient 5..... 159
Figure A.6 a) Tumor tissue, b) Normal tissue of Patient 6..... 159
Figure A.7 a) Tumor tissue, b) Normal tissue of Patient 7..... 159
Figure A.8 a) Tumor tissue, b) Normal tissue of Patient 8..... 160
Figure A.9 a) Tumor tissue, b) Normal tissue of Patient 9..... 160
Figure A.10 a) Tumor tissue, b) Normal tissue of Patient 10..... 160
Figure I.1 The panels for a) Chr1 b) Chr2 c) Chr6 d) Chr7 e) Chr8 f) Chr11 g) Chr12 h) Chr17 i) Chr18 are corresponding to raw CN and BAF graphs drawn vs physical position.. 200
Figure I.2 Segmentation results of sample T8 over all chromosomes. 201

LIST OF ABBREVIATIONS

ACNE	Allele-Specific Copy Number Estimation
ACS	American Chemical Society
ACS	American Cancer Society
AHP	Analytic Hierarchy Process
Aroma	An R Object Oriented Microarray Analysis Environment
BAF	Allele B Fraction
BSA	Bovine Serum Albumin
CN	Copy Number
CN-LOH	Copy Number Neutral Loss of Heterozygosity
CNA	Copy Number Aberration
CNP	Copy Number Polymorphism
CNS	Copy Number State
CNV	Copy Number Variation
CRMA	Copy number estimation using Robust Multichip Analysis
CRMA v2	Copy number estimation using Robust Multichip Analysis Version 2
dH_2O	Distilled water
DMSO	Dimethyl sulfoxide
dNTP	Deoxyribonucleotide triphosphate
EDTA	Ethylenediaminetetraacetic acid
EGB	Ensembl Genome Browser
EtBr	Ethidium Bromide
FFPE	Formalin Fixed Paraffin Embedded

GADA	Genome Alteration Detection Analysis
GATA	Gülhane Military Medical Academy
GWAS	Genome-wide Association Study
H&E	Hematoxylin-Eosin
LOH	Loss of Heterozygosity
MAF	Minor Allele Frequency
MES	4-Morpholineethanesulfonic acid
OS	Osteosarcoma
QC	Quality control
RAPD-PCR	Randomly Amplified Polymorphic DNA Polymerase Chain Reaction
RIF	Reference Into Function
SNP	Single Nucleotide Polymorphism
TCN	Total Copy Number
TdT	Terminal Deoxynucleotidyl Transferase
TMACL	Tetramethyl Ammonium Chloride
VN	Virtual Normal
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

1.1 Cancer

Cancer is among the most lethal diseases, accounting for approximately 13% of all deaths worldwide according to the 2008 statistics of World Health Organization (Mathers & Loncar, 2006). Cancer is the second most common cause of death in Turkey, exceeded by heart and other chronic diseases (T.C. Saglik Bakanligi, 2006).

Cancer develops as a result of a series of inherited and/or acquired mutations which cause noteworthy changes in the behavior of a single cell and its offspring (Rieger, 2004). Those mutations can be caused by both external agents (tobacco, infectious organisms, chemicals, and radiation) and internal agents (inherited mutations, hormones, immune conditions, and mutations that occur from metabolism). These factors can operate together or sequentially to initiate or promote carcinogenesis (American Cancer Society, 2011). By the

growth of cancerous tissue, cancer cells metastasize to distant organs with the invasion of nearby tissue and vascular system.

Cancer has more than hundred different types named according to initiation site of cancer. In fact, there are five broad categories of cancer, which are stated in Table 1.1 (National Cancer Institute, 2011).

Table 1.1 Main Cancer Types

Cancer Type	Origin
Carcinoma	Skin or tissues that line or cover internal organs
Leukemia	Blood-forming tissue such as the bone marrow
Lymphoma and myeloma	Cells of the immune system
Central nervous system cancers	Tissues of the brain and spinal cord
Sarcoma	Bone, cartilage, fat, muscle, blood vessels, or other connective or supportive tissue

1.2 Sarcomas of the Bone

Cancers of the soft tissue or bone are called as sarcomas which comprise over forty different types of cancers. Sarcomas accounts for less than 1% of adult cancer diagnoses, and there are nearly 2,400 new cases per year in the bone (American Cancer Society, 2011). Sarcomas are considered as primary bone cancers and are categorized according to their initiation site. For instance, chondrosarcomas start in cartilage, osteosarcomas start in bone, and

fibrosarcomas start in fibrogenic tissue. The most common type of bone sarcoma is osteosarcoma (OS) comprising roughly 35% of bone tumors. The second most common type in adults is chondrosarcoma, accounting for 30% of bone sarcomas, and the second most common type in children is Ewing's sarcoma. Remaining types are very rare, each accounting for less than 1% of all bone sarcomas. Moreover, within these major types, there are further subtypes (Malaver, Helman, & Brian, 2008). Different types of sarcomas of the bone are listed in Table 1.2.

Table 1.2 General Classification of Bone Tumors

Histologic Type*	**Benign**	**Malignant**
Hematopoietic (41.4%)	—	Myeloma
	—	Reticulum cell sarcoma
Chondrogenic (20.9%)	Osteochondroma	Primary chondrosarcoma
	Chondroma	Secondary chondrosarcoma
	Chondroblastoma	Dedifferentiated chondrosarcoma
	Chondromyxoid fibroma	Mesenchymal chondrosarcoma
Osteogenic (19.3%)	Osteoid osteoma	Osteosarcoma
	Benign osteoblastoma	Parosteal osteogenic sarcoma
Unknown origin (9.8%)	Giant cell tumor	Ewing’s tumor
	—	Malignant giant cell tumor
	—	Adamantinoma
	(Fibrous) histiocytoma	(Fibrous) histiocytoma
Fibrogenic (3.8%)	Fibroma	Fibrosarcoma
	Desmoplastic fibroma	—
Notochordal (3.1%)	—	Chordoma
Vascular (1.6%)	Hemangioma	Hemangioendothelioma
	—	Hemangiopericytoma
Lipogenic(_0.5%)	Lipoma	—
Neurogenic (_0.5%)	Neurilemoma	—

*Distribution based on Mayo Clinic experience.
(Dahlin DC., 1978)

1.2.1 Malignant Bone Tumors

Malignant bone tumors occur less commonly than benign tumors, still they are much more risky. This is because in these tumors, cancer cells can metastasize to distant organs through the blood stream or through lymphatic system. The most common spread site of malignant bone tumors is the lungs. Chondrosarcoma, Ewing's sarcoma and Osteosarcoma are the most frequent malignant bone tumors. Among those, OS is the most common one, having an incidence rate 35% (Hameed & Dorfman, 2011).

1.2.1.1 Osteosarcoma

Osteosarcoma is among primary malignant bone tumors and it is characterized by the formation of immature bone or osteoid tissue. According to World Health Organization (WHO) histological classification of bone tumors, OSs are divided into central and surface tumors, and there are subtypes within each group (Schajowicz, Sissons, & Sobin, 1995). In this study, the conventional central high grade OS of bone (classic OS), which represents about 90% of all cases of osteosarcoma, was investigated.

Classic OS accounts for approximately 15% of biopsy-analyzed primary bone tumors. The incidence of classic OS is 3 cases/million population/year, which forms 0.2% of all malignant tumors (Campanacci, 1999).

1.2.1.1.1 Clinical Characteristics

OS has two peak incidences, the first occurring at ages between 15 and 29 and the second peak occurring at ages older than 60 (Picci, 2007). Actually, about 75% of patients with OS are between 15–25 years of age. Tumors observed in older ages usually develop after Paget's disease, radiation or dedifferentiated chondrosarcomas. The incidence of OS in males is higher than the incidence in females, with a ratio of 1.5:1.0 (Malaver et al., 2008).

There are a number of families with several members who developed OS and this suggests the presence of a genetic predisposition to OS. (Hillmann, Ozaki, & Winkelmann, 2000). Until today, patients with hereditary retinoblastoma have presented the strongest genetic predisposition. Those patients are 500 times more likely to suffer from OS than the general population. In 3% to 4% of pediatric OS patients, a constitutional germline mutation in p53 was observed. The majority of patients with germline p53 mutations had a family history of Li-Fraumeni syndrome (Picci, 2007).

The long tubular bones are affected in 80% to 90% of OS cases. The axial skeleton is not so frequently affected, but more often observed in adults. About 85% of extremity tumors occur in femur, tibia and humerus, while less than 1% are observed in hands and feet bones. OS usually initiate in the metaphyseal region in the long bones. Tumors originating in the midshaft and in the epiphysis are very rarely seen (Wu et al., 2009).

Most of the patients with classic OS of the extremities suffer from pain prior to soft tissue swelling because of stretching of the periosteum. Weakening of the bone with development of minute stress fractures can also result in pain. Almost 15% of young patients show pathological fractures leading to pain. The second most common symptom is swelling, which is caused by the soft tissue mass. Systemic symptoms such as weight loss, pallor, fever, anorexia are very rarely observed in OS patients (Hameed & Dorfman, 2011).

Etiology of osteosarcoma is still unknown. The evidence that bone sarcomas can be induced in chosen animals by viruses or cell-free extracts of human osteosarcomas suggests a viral origin (Finkel, Reilly, & Biskis, 1976). Moreover, ionizing radiation is the single environmental agent that is known to cause OS in human. In about 2% of OS cases, radiation is involved (ACS, 2011).

1.2.1.1.2 Molecular Characteristics of Osteosarcoma

Osteosarcomas have complex karyotype contrary to translocation-associated sarcomas such as Ewing's sarcoma, where specific gene fusion proteins promote tumorigenesis. Complex structural and numerical chromosomal aberrations with significant heterogeneity within the same tumor are observed in most of the high-grade OSs by conventional karyotyping (Figure 1.1). In high-grade OS, most commonly rearranged chromosomal regions are 1p11-13, 1q10-12, 1q21-22, 11p15, 12p13, 17p12-13, 10q13, and 22q11-13. Gain of

chromosome 1 and loss of 9, 10, 13, 17, and 6p are among most frequently seen numerical abnormalities (Bridge et al., 1996).

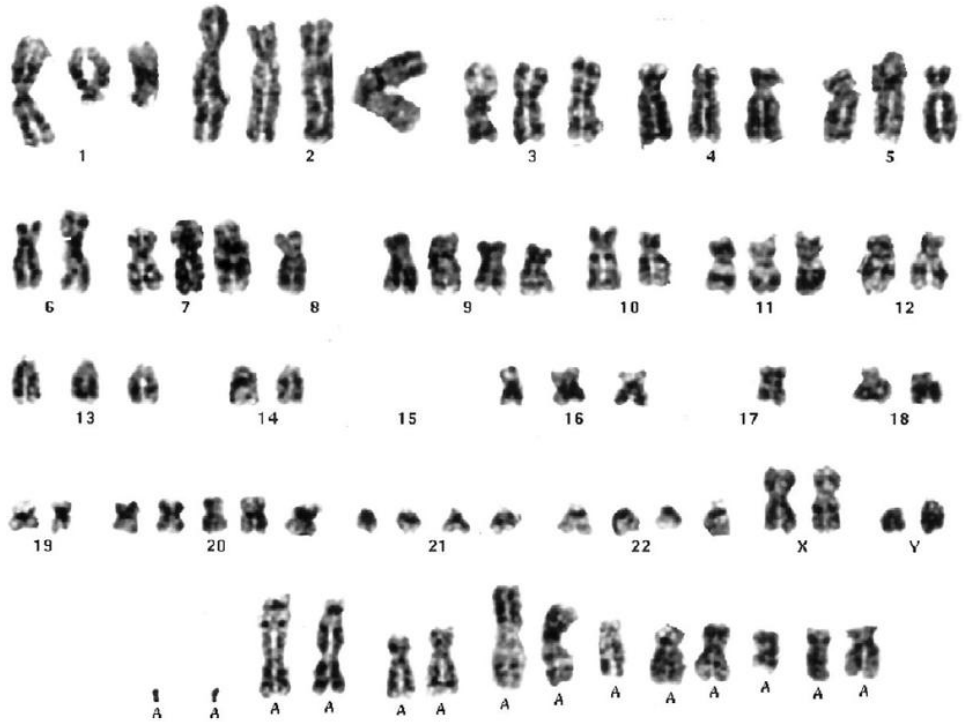


Figure 1.1 Karyotype of conventional osteosarcoma showing multiple structural and numerical aberrations (Hameed & Dorfman, 2011).

After development of array-based comparative genomic hybridization (array-CGH), the complexity of karyotype of OS genome is understood more completely. In a study of 48 OS samples by array-CGH (Man et al., 2004), it was shown that there were more gains than losses and; high-level recurrent amplifications mapped to 1p36.32 (PRDM16), 6p21.1 (CDC5L, HSPCB,

NFKBIE), 8q24, 12q14.3 (IFNG), 16p13 (MGRN1), and 17p11.2 (PMP22, MYCD, SOX1, ELAC27). Moreover, recurrent homozygous deletions mapped to 1q25.1, 3p14.1, 13q12.1, 4p15.1, 6q12, and 6q16.3 were detected in that study. In about 50% of OS cases, maintenance of telomere length through alternate mechanisms was observed. The genomic aberrations and telomere length protection in OS reveal a high degree of chromosomal instability in these tumors (Selvarajah et al., 2008).

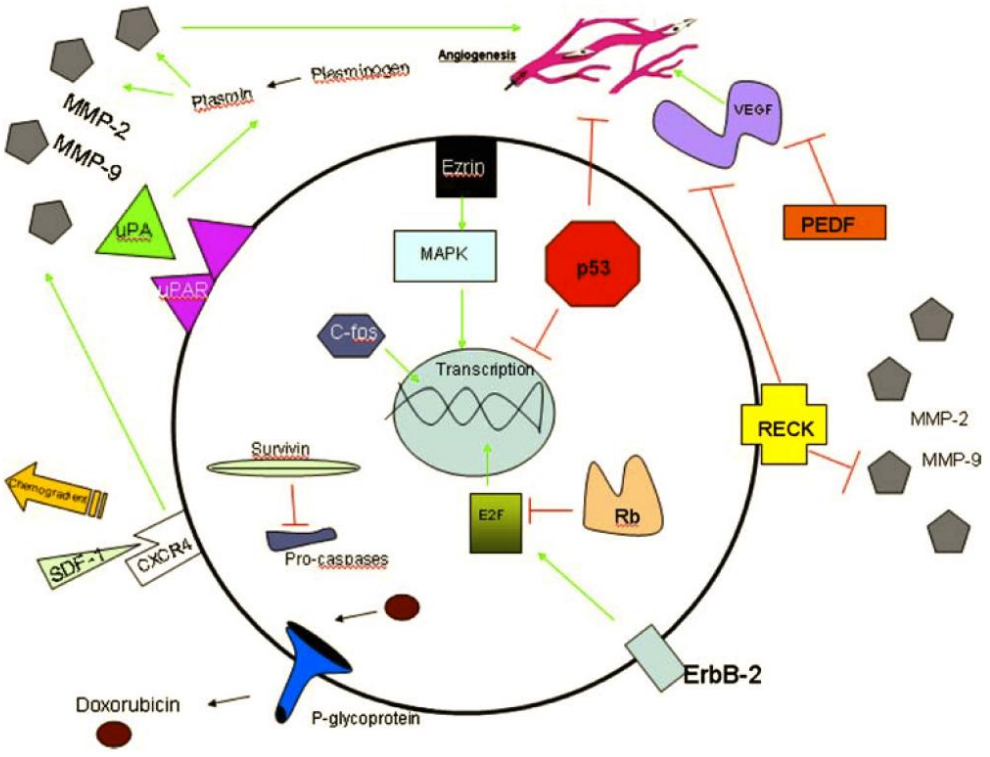


Figure 1.2 Key molecular factors in Osteosarcoma (Clark, Dass, & Choong, 2008).

In recent years with the development of high-throughput molecular assays, prognostic factors of OS have been studied (Figure 1.2). The most investigated molecular pathways in OS include cell adhesion molecules (cysteine-rich protein with Kazal motifs (RECK), ezrin, CD44, matrix metalloproteinases (MMPs), cadherin), cell cycle–associated genes (URG4, SKP2), tumor suppressors (RB, p53, p16 and p14, NF-2, TGF), apoptosis (p53, BAX, cytochrome C, livin, survivin, HSP90), angiogenesis factors (VEGF, CD34, CD31, factor VIII, integrin, microvessel density, pigment epithelium derived factor (PEDF), migration inhibition factor), oncogenes (C-MYC, FOS/JUN, MDM2, CDK4, cyclin D1, Her2/neu, Wnt, MET, FGFR2, telomerase), receptor tyrosine kinases (Her2/Neu, PDGF, C-Kit), and chemotherapy resistance–associated factors such as P-glycoprotein, ERCC gene polymorphism, ECM remodeling, and osteoclast differentiation genes by expression signature and other resistance associated genes (Strauss, Ng, Mendoza-Naranjo, Whelan, & Sorensen, 2010).

In brief, many molecular targets and pathways having role in OS tumorigenesis have been shown in the past three decades. However, these potential molecular markers have to be validated along with clinical trials in order to better interpret the biology of OS and supervise future management of the patients (Hameed & Dorfman, 2011).

1.2.1.1.3 Metastasis of Osteosarcoma

Unlike carcinomas, bone sarcomas spread almost exclusively through the blood; because bones do not have a lymph system. Hematogenous dissemination is marked by pulmonary involvement in its early stage and secondarily by bone involvement. Bone metastasis is rarely the first sign of dissemination (3.9%). (Malaver et al., 2008)

In fact, the most common metastasis in patients with OS is to the lungs (ACS, 2011). 50% of the patients are found to have metastatic disease at the initial consultation or at follow-up. In addition, 11–20% of the patients have primary pulmonary metastases (X. Chen et al., 2009).

At present, the long-term survival rate of metastatic OS is poor despite the improvements in the treatment of OS. Therefore, high-risk metastatic factors such as older age, large tumor size and elevated LDH level have to be considered well because they can be a sign for metastases to the lungs. Besides, in order to improve survival rate of metastatic patients, further studies should be done to identify novel therapeutic approaches, such as new chemotherapy agents, gene therapy, stereotactic radiotherapy and immunotherapy. (Wu et al., 2009)

1.2.1.1.4 Treatment

There has been great development in the treatment of OS during the past several decades. In the 1960s, there was only one treatment named as amputation which includes removal of the limb with the tumor (Bridge et al., 1996). In those years, only a small fraction of patients lived more than 2 years. Till now, it has been found that chemotherapy before and after surgery have a great advantage to increase the survival rates of the patients (Rech et al., 2004). Nowadays, the types of treatment used for OS include surgery, chemotherapy, and radiotherapy. Each is detailed below.

1.2.1.1.4.1 Surgery

The main goal of surgery is to diagnose and remove cancer. Therefore, surgery for OS is of two types. The first one is the biopsy performed to confirm the cancer. The second one is the surgery to take out the tumor. This second surgery can be either the kind that saves the extremities, which is called limb-sparing, or that removes the cancer and all or part of an arm or leg, which is called amputation. Generally, the type of surgery depends on the place of the tumor. Some tumors are much harder to treat and hence, surgery cannot be possible for those. These include tumors at the base of the skull, or in the spine or pelvis (Malaver et al., 2008).

1.2.1.1.4.2 Chemotherapy

Chemotherapy for OS can be applied both before and after surgery. Most of the time multiple drugs are given together (ACS, 2011). Before routine use of systemic chemotherapy for OS, fewer than 20% of patients had survival rates higher than 5 years. Moreover, 50% of patients showed relapsed disease almost exclusively in the lungs, within 6 months after surgery. In 1980s, two randomized clinical studies comparing surgery alone to surgery followed by chemotherapy was performed and indicated that the addition of systemic chemotherapy enhanced survival in patients with localized high-grade OS (Malaver et al., 2008). These studies implied that most of the patients with localized tumors have micrometastatic disease, and that systemic chemotherapy improve the survival by fighting against those micrometastases (Strauss et al., 2010). In the past two decades, standard treatment has become the application of neoadjuvant (presurgical) and adjuvant (postsurgical) chemotherapy. Additionally, the five most important drugs used for OS treatment include high-dose methotrexate (HD-MTX), adriamycin (ADM), etoposide, cisplatin (CDDP), and ifosfamide (IFOS) (Hattinger, Pasello, Ferrari, Picci, & Serra, 2010). In this study, all patients had pulmonary metastatic disease. For such patients, there is no accepted standard treatment approach; however, in the literature, it is suggested that currently available aggressive multiagent chemotherapy with complete surgical resection can be applied to such patients (Harting & Blakely, 2006).

1.2.1.1.4.3 Radiation Treatment

Radiation therapy is not a primary treatment way for OS, but this can change with the development of novel technologies. Radiation therapy is generally performed on patients who have refused surgery, need palliation, or have lesions in axial locations. For the tumors of the axial skeleton and facial bones, surgery can only be applied in limited form. For this reason, radiotherapy has great importance for such tumors. Moreover, when function and cosmesis preservation is vital, a combination of limited surgery and radiotherapy can be used (Malaver et al., 2008).

1.3 Formalin Fixed Paraffin Embedded Samples in Molecular Studies

In the present study, formalin fixed paraffin embedded (FFPE) samples of pulmonary metastatic OS were utilized. In fact, the FFPE samples in histopathology archives have a great importance because these archives contain a historical collection of almost every disease. Those FFPE samples were first used in the development of numerous immunohistochemical assays which are now used in routine diagnostics. Recently, those samples are used as a source of DNA for molecular analysis and the detection of infectious agents. (Lehmann & Kreipe, 2001). With the molecular analysis of this archival tissue, the correlations between molecular findings, the response to treatment and the clinical outcome could be investigated. In addition, these findings may guide future studies analyzing native or freshly frozen biopsies (Paik, Kim, Song, & Kim, 2005).

Nevertheless, DNA obtained from FFPE samples are generally in low quality. This is because chemical modifications caused by formalin fixation and paraffinization at high temperatures can lead to fragmentation of DNA. Moreover, storage time also increases the fragmentation further. For this reason, there is a great effort to develop novel reliable techniques to isolate DNA from FFPE samples (Shi et al., 2004). As a result of this effort, some new molecular technologies were developed and the DNA samples obtained from the archive specimens can be used in many different molecular assays such as PCR. However, the high-throughput molecular techniques, such as microarrays, require a reliable recovery of DNA from the FFPE tissue and there is also new extraction methodologies developed for this purpose (Lewis, Maughan, Smith, Hillan, & Quirke, 2001).

1.4 Molecular Profiling

Biological processes are complex pathways which involves actions of many molecules together. Therefore, researchers are after the "big picture" and molecular profiling can be used as a general approach for such studies. In fact, molecular profiling refers to collective description of molecular approaches that measure genome wide mutations and expression of multiple genes and proteins on biological samples simultaneously (Lockhart & Winzeler, 2000). This multidimensional measurement of biological processes has become possible over the last fifteen years with the development of appropriate technological platforms (Ioannidis, 2007). The high-throughput molecular profiling is generally performed by using microarrays of different types. In

theory, the results obtained from molecular profiling of each patient can provide optimized, individualized management of patients. For example, toxicity may be decreased by avoiding unnecessary treatment in patients, and efficacy may also be optimized by selecting the patients, who get the maximal benefit, for treatment.

1.4.1 Microarray Technology

Microarray is a technology which allows researchers to study different conditions by analyzing the expression of many genes or mutations throughout the genome in a single reaction (X. Li et al., 2008). By this technology, scientists can explore basic aspects of growth and development, and investigate the principal genetic causes of many human diseases, such as cancer.

DNA microarrays are small, solid supports onto which the sequences from thousands of different genes, i.e. probes, are immobilized at fixed locations in an orderly way. Actually, those probes are printed, or synthesized directly onto the support. The probes can be DNA, cDNA, or oligonucleotides (Ting Lee, 2004). There are actually two basic types of DNA microarrays. The first one is transcriptomic which uses either oligonucleotides or cDNAs as probes in order to identify gene expression levels of many genes. The other one is genomic, which uses DNA or oligonucleotides as probes, in order to identify polymorphisms, genomic copy number variations (CNVs) and mutations.

1.4.1.1 Applications of Microarrays

Microarrays, specifically DNA microarrays, can be used in many different areas. They help in the discovery of novel genes, and expression levels of genes under different conditions. Moreover, they can be used to investigate the genes responsible for different diseases such as heart diseases, infectious disease, mental illness, and especially cancer (Min, Choi, Lee, & Yoo, 2009). With the help of DNA microarrays, it is now possible to classify the types of cancer on the basis of the patterns of gene activity in the tumor cells. Furthermore, DNA microarray technology can be applied as a tool in pharmacogenomics in order to find out the correlations between responses to drugs and the genetic profiles of the patients. Likewise in pharmacogenomics, DNA microarray technology can be used in toxicogenomics for the study of the impact of toxins on the cells and their passing on to the progeny by establishing correlation between responses to toxins and genetic profiles of cells (Grant & Hakonarson, 2008).

There are two broad applications of DNA microarrays, which can be systematically used in all of the areas stated above. The first one is gene expression analysis and the second one is genome-wide association and mutation analysis. Both are detailed in subsequent sections.

1.4.1.1.1 Gene Expression Profiling Analysis

The process by which mRNA, and eventually protein, is synthesized from the DNA template of each gene is called as gene expression. Some protein coding genes are expressed under all circumstances at a constant level. However, other genes are only expressed at particular moments and under particular external conditions. Understanding which genes are expressed under which condition is vital in order to get information about the biological processes in the cell (Ting Lee, 2004). Gene expression microarrays contain thousands of probes, either cDNAs or oligonucleotides, which are a part of corresponding transcripts of different genes. With the help of this microarray technology, scientists can measure the expression of thousands of genes simultaneously under specific pre-defined conditions such as study of gene expression profile of drug resistant cancer cells (Darcansoy, 2009).

1.4.1.1.2 Genome-Wide Association Studies

Genome-wide association study (GWAS) can be defined as analysis or scan of numerous common genetic variants in different individuals in order to decide if a particular variant is associated with the trait under study (Johnson & O'Donnell, 2009). GWASs are important tools for finding environmental and genetic risk factors responsible for human complex diseases. In those studies, genetic markers within the genomes of many different people are scanned in order to find out those risk factors. In a such study, if alleles or haplotypes of a particular genetic marker are observed significantly more frequent in cases than

controls, this marker is said to be associated with the phenotype under study (Manolio, 2010). In fact, GWASs only identify the disease associated genetic markers in DNA, rather than specifying causal genes.

The genetic markers frequently used in GWAS include SNPs, microsatellites and copy number variations (CNVs). SNPs are single nucleotide changes in the sequence of DNA differing between members of a species or homologous chromosomes of an individual. SNPs occur every 100 to 300 bases throughout the genome, within both coding and non-coding regions (Komar, 2009). Compared to other markers, SNPs are less mutable and highly diverse. These features of SNPs make their use in GWAS more feasible (Frazer et al., 2007). With the foundation of International HapMap Project and public databases of genetic information, SNPs are started to be commonly used in GWAS (The International HapMap Consortium, 2005). Since December 2010, over 1,200 human GWAS have performed to examine almost 200 diseases and they have found nearly 4,000 SNP associations (Johnson & O'Donnell, 2009).

A summary of a GWAS is represented in Figure 1.3. Accordingly, the first step in GWAS is genotyping the SNPs in the genome of both cases and controls. This is performed with the help of genotyping microarrays which can genotype whole genome in a single experiment. There are many different platforms, i.e. SNP arrays, offered by Illumina and Affymetrix and most of them can genotype almost 1 million SNPs. In present study, an Affymetrix genotyping array covering 250,000 SNPs has been used (Affymetrix, 2011a). After genotyping, allele frequency of each SNP is calculated. If the allele frequency of an SNP is much higher in cases compared to controls, the odds ratio of that SNP will be higher than 1, indicating an association. In other words, the

proportion of individuals in the case group with a specific allele is higher than the proportion of individuals in the control group having the same allele. Moreover, to detect the significance of the odds ratio, *p-values* for each SNP is calculated using a simple chi-squared test. Calculations are generally made by using bioinformatics tools such as PLINK (Purcell et al., 2007). After calculation of odds ratios and *p-values* for all SNPs, Manhattan plot is drawn from the negative logarithm of the *p-value* as a function of genomic location, i.e. chromosomal location. During analysis, the *p-value* significance threshold is defined. In general, to be considered as significant among millions of SNPs, the *p-value* of particular SNP should be very low, ($10^{-7 \text{ or } -8}$) (Clarke et al., 2011). Results of GWAS indicate that most of the associated SNPs are found in non-coding regions, or introns, on the chromosome between genes (Manolio, 2010). After the discovery of associated SNPs, a validation experiment with the same methodology is performed on an independent sample set of the same or larger size than the sample analyzed in the GWAS (Hardy & Singleton, 2009). After that, SNPs are selected for replication in another independent cohort. The number of selected SNPs determines the genotyping method. If the number is as few as 10, alternate genotyping methods such as TaqMan PCR can be used to detect the presence of risk loci in a sample. The SNPs which could be successfully replicated in all sets are further analyzed for risk prediction for the phenotype under study and functional studies are performed on those SNPs.

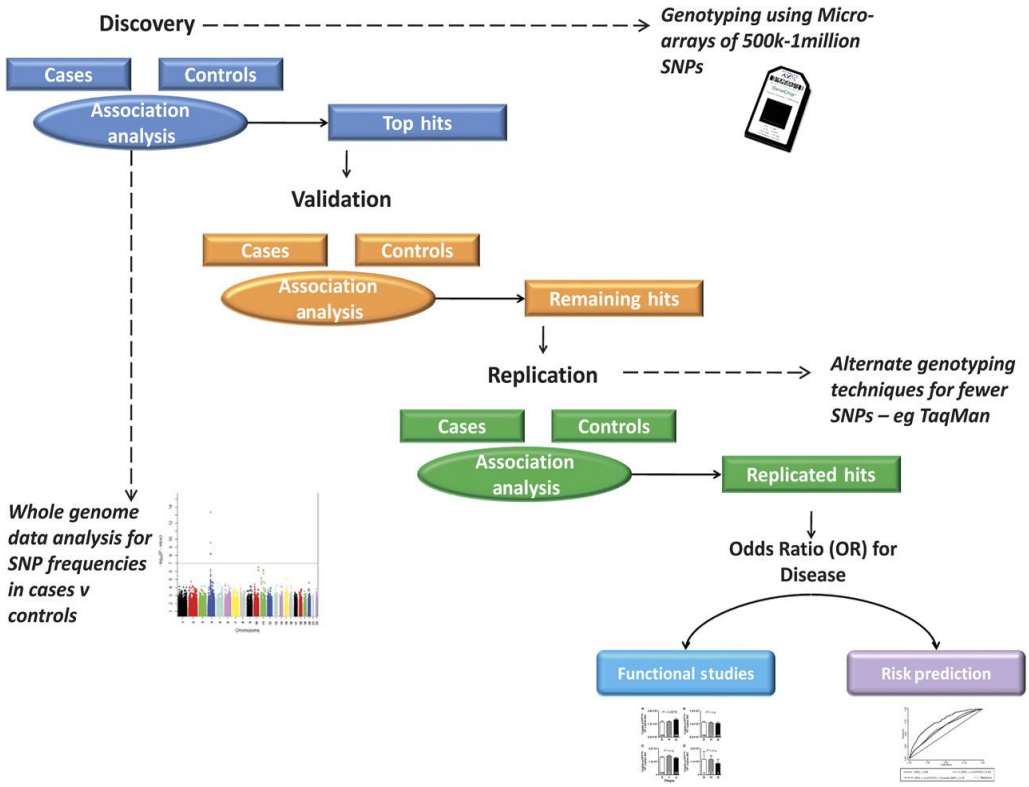


Figure 1.3 Summary of validation of results of a GWAS (Patel & Ye, 2011).

1.4.1.1.2.1 GWAS Tools

GWAS, in which thousands of SNPs throughout the genome are genotyped in cases and controls, is recently the most accepted strategy to identify genomic regions associated with common complex diseases (Distefano & Taverna, 2011). With the improvements in technology, various statistical tools have developed, namely PLINK (Purcell et al., 2007), BEAGLE (B. L. Browning & Browning, 2007) and, METU-SNP (Ustünkar & Aydın Son, 2011).

PLINK is an open source C/C++ GWAS toolset that is developed by Shaun Purcell at the Center for Human Genetic Research. PLINK can handle large data sets including thousands of markers genotyped for thousands of samples. It offers numerous functions for GWAS comprising: Data management, summary statistics for quality control, population stratification, basic association tests, copy number variant analysis, meta-analysis, result annotation and reporting.

BEAGLE is a software program used for genetic association analysis, haplotypes phase inferring, and genotype imputation which is statistical method to substitute a calculated value for a missing genotype data point. In fact, genotype imputation recently has become a part of GWAS in order to raise the power of existing marker sets by replacing ungenotyped data to increase the coverage of SNPs in case control data sets beyond what has been genotyped. Several software programs have been developed so far to account for genotype imputation, like BEAGLE, which is written in Java and works on most of the computing platforms (e.g., Windows, Linux, Unix, Mac and Solaris).

METU-SNP is a java based integrated software system, which can be used in various platforms. It utilizes public databases (dbSNP, Entrez Gene, KEGG, Gene Ontology) to prioritize SNPs according to their genomic location, evolutionary conservation, biological significance and statistical significance by Analytic Hierarchy Process (AHP). It also performs second-wave analysis which intends to find genes and pathways related to disease associated SNPs. In addition to prioritization, representative SNP selection is performed from prioritized SNPs, in which most informative SNPs for the disease under study

are selected (Gürkan Üstünkar, Özöğür-Akyüz, Weber, Friedrich, & Aydın Son, 2011).

1.4.1.1.3 Copy Number Variation and Loss of Heterozygosity Analysis

In addition to GWAS analysis, SNP arrays can also be used for studying copy number variations (CNVs), copy number aberrations (CNAs) and loss of heterozygosity (LOH) (W. Sun et al., 2009). The copy number states (CNS) of a tumor normal match sample that can be obtained from such an SNP array data are outlined in Figure 1.4.

.

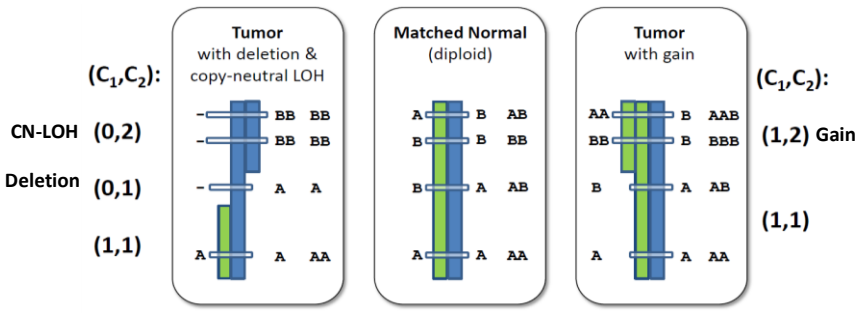


Figure 1.4 Copy number states of a tumor-normal match sample.

Numerous studies have discovered submicroscopic copy number variation of DNA segments ranging in size from kilobases (kb) to megabases (Mb). Those inheritable CNVs or copy number polymorphisms (CNPs) include deletions, insertions, duplications and complex multisite variants and they are found in human and all other mammals (Redon et al., 2006). On the other hand, CNAs can be defined as acquired somatic alterations and they are frequently observed in tumor tissues. When compared to CNVs, CNAs are generally longer and cover a significant proportion of the genome (W. Sun et al., 2009). With the help of SNP arrays which provide high resolution and allele specific information, it is possible to detect CNAs in case-control studies. In order to detect CNAs, the signal intensities of the match and mismatch probes of SNP markers are compared with the values from another individual or the matched normal sample of the individual with the phenotype and copy number states (CNS) of each locus is determined. The CNS of a given locus (j) is defined as a pair of numbers (a_j, b_j), where $a_j \geq 0$ and $b_j \geq 0$ and are respectively the larger and the smaller of the two parental copy numbers at this locus. By definition we have $b_j \leq a_j$, and $c_j = a_j + b_j$ is the total copy number. The quantities a_j and b_j are called major and minor copy numbers, respectively. With this principle, the relative copy number per locus in cases is determined and these locus level estimates are combined to obtain region level estimates by segmentation tools (H. Lu, Schölkopf, & Zhao, 2011). In order to achieve noise reduction and segmentation during CNA detection, various statistical tools have been developed, such as aroma package (Henrik Bengtsson, 2004), which will be explained in the following section (Carter, 2007).

With the help of allele specific information obtained by SNP arrays, LOH detection for case-control studies can also be performed. LOH can be defined as a form of allelic imbalance caused by either complete loss of an allele

[CNS= (1, 0)] or an increase in copy number of one allele relative to the other [CNS= (2, 1)] (Table 1.3). These two conditions both lead to a decrease in heterozygosity index of particular locus, indicating LOH (H. Lu et al., 2011). Moreover, SNP arrays can also detect copy number neutral LOH (CN-LOH) resulted from uniparental disomy (UPD) in which one allele or whole chromosome from one parent is missing leading to reduplication of the other parental allele [CNS=(2,0)] (Jacobs et al., 2007). In a disease situation, UPD may be pathological if the wild type allele is missing and two copies of the mutant allele are present. LOH detection with SNP arrays allows identification of patterns of allelic imbalance with prospective prognostic and diagnostic uses. This function of SNP arrays have a huge potential in cancer diagnostics because LOH is a well-known feature of most human cancers (Sellick et al., 2004). Current studies with SNP arrays have revealed that both solid tumors (OS, gastric cancer, liver cancer etc.) and hematologic malignancies (ALL, MDS, CML etc.) have a high rate of LOH caused by genomic deletions or UPD and genomic gains. With the help of such studies, mechanisms of these diseases can be investigated and new treatment approaches can be developed such as creation of targeted drugs (Wong et al., 2004).

Table 1.3 Minor and major copy number states presented as the conjunction of information regarding total copy number (columns) and heterozygosity status (rows).

	Deletion	**Neutral**	**Gain**
LOH	(0,1)	(0,2)	(a, 0) with $a \geq 3$
Heterozygosity	(0,0)	(1,1)	(a, b) with $1 \leq b \leq a$ and $a + b > 2$

1.4.1.1.3.1 Copy Number State Analysis Tools

There are a number of methods to analyze SNP array in the context of CNA and LOH studies such as VanillaICE, PennCNV, QuantiSNP, and BirdSuite. However, those methods are designated for normal cells and well-designed for CNV studies. Moreover, they do not effectively describe the copy number states in cancer cells because either allele-specific amplifications are not considered, or the difference between normal and CN-LOH cannot be distinguished, or they can only detect rare CNAs. For this reason, tools that can combine total and allele specific signals are needed in order to call CNS in cancer case-control studies. One of the tools that take into account cancer specific considerations is the aroma package (An R Object Oriented Microarray Analysis Environment) developed by Henrik Bengtsson at Lund Institute of Technology, Lund University (Henrik Bengtsson, 2004) and in the present study aroma has been used for CNV and LOH analysis.

Aroma is a package for low-level analysis of microarrays, which includes calibration and normalization, of single and two- or multiple- channel microarray data. With aroma, unlimited number of arrays of all Affymetrix chip types, e.g. expression arrays, exon arrays and SNP chips, can be analyzed simultaneously. It works directly with the raw data files, i.e. .CEL and .CDF files. Preprocessing step includes background correction, allelic cross-talk calibration, quantile normalization, and nucleotide-position normalization. Post-processing includes PCR fragment length normalization and/or GC-content normalization. It can perform both paired and non-paired copy-number analysis for all Affymetrix genotyping arrays. It uses various segmentation

methods such as CBS, GLAD, GADA and HaarSeg. At the end of the analysis, it provides dynamic HTML reports, such as ChromosomeExplorer (HapMap demo, Tumor/Normal Demo) (Henrik Bengtsson, 2004).

1.4.2 Affymetrix GeneChip Human Mapping 250K SNP Array Platform

In this study, StyI array platform of the GeneChip® Human Mapping 500K Array Set has been used. This set is comprised of two arrays, each capable of genotyping an average of 250,000 SNPs. One array uses the Nsp I restriction enzyme (~262,000 SNPs), while the second uses Sty I (~238,000 SNPs). In both arrays, 50 SNPs are used as controls (Affymetrix, 2011a).

In Sty I array system, 250 ng of genomic DNA samples are cut by Sty I restriction enzyme and ligated to adaptors that recognize the cohesive four base-pair (bp) overhangs. A generic primer recognizing the adaptor sequence is used to amplify adaptor-ligated DNA fragments. PCR conditions have been optimized to amplify fragments in the 200 to 1,100 bp size range. The amplified DNA is then fragmented, labeled, and hybridized to a GeneChip 250K Array (Figure 1.5) (Affymetrix, 2011b).

There are more than 6.5 million features on Sty I array, and each feature contain more than one million copies of a 25-bp oligonucleotide probe, synthesized in parallel by photolithography technique. Each SNP is cross-

examined by 6- or 10- probe quartets where each probe quartet is made up of a perfect match and a mismatch probe for each allele. In total, there are 24 or 40 different 25-bp oligonucleotides per SNP (Jacobs, 2006).

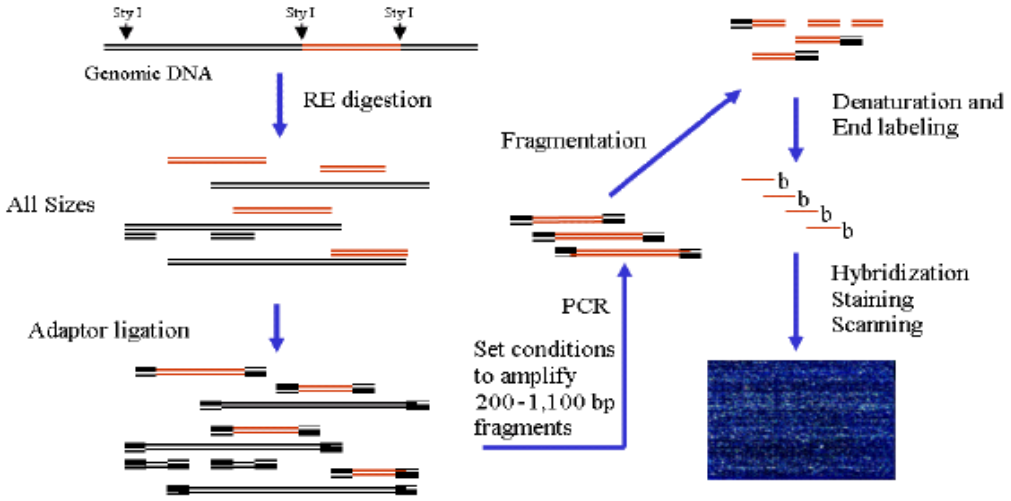


Figure 1.5 GeneChip Mapping Assay Overview.

1.5 AIM OF THE STUDY

Osteosarcoma is the most common malignant tumor of bone with an incidence rate of 19% among all cancer types. The vast majority of OS patients have pulmonary metastases at the time they are diagnosed, and about half develop lung disease later. Moreover, pulmonary metastatic tumors lead to poor prognosis and increased death rate. Although mutations in the genes coding for P53, RB, FOS and MYC were detected in pulmonary metastatic tumors of OS, there is no unique genetic pathway identified for progression of pulmonary metastasis. Therefore, it is essential to investigate cellular and molecular mechanisms underlying pulmonary metastasis of osteosarcoma. The objectives of this study were:

- To identify the SNPs associated with pulmonary metastasis of OS.
- To perform a second-wave analysis to detect gene and pathway association of significantly associated SNPs.
- To detect genome-wide copy number aberrations in pulmonary metastatic OS tumor samples comparing those with their respective normal tissues.
- To identify genome-wide LOH events occurred in tumor samples compared to matched normals.

- To provide a microarray data analysis workflow for application of FFPE-derived DNA samples in order to produce reliable genotype, copy number, and LOH predictions.

In summary, this study aimed to identify potential prognostic markers of metastatic OS by developing a data analysis workflow from available tools for the use of FFPE samples in high-throughput array analysis.

CHAPTER 2

MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 Patients

Ten patients with high grade, pulmonary metastatic osteosarcoma (OS) were included in this study after providing written informed consent and after the approval of the Institutional Ethic Committee of Gülhane Military Medical Academy (GATA). The required information about the patients and the histopathologic properties of the tumors were recorded from the files of the patients. These adult patients were older than age 15 years, with an Eastern Cooperative Oncology Group (ECOG) performance score of 0–1; no previous history of cancer; no history of other previous osteosarcoma-related treatments; and normal cardiac, renal, and liver functions. Patients older than age 60 were excluded from the current study.

2.1.2 Tissues

Ten tissue samples of OS were derived from Turkish patients with pulmonary metastatic cancer during diagnosis of advanced disease with relapse. All biopsies were performed by the same thoracic surgeon using an open incisional biopsy. Pathologic samples were evaluated by the same pathologist. All samples were lung tissues which were stored as formalin fixed paraffin embedded (FFPE) blocks in the Archive of Pathology Department of GATA. FFPE blocks were taken out of the Archive with the consent of the Ethical Committee of GATA.

2.1.3 Chemicals and Reagents

American Chemical Society (ACS) grade xylene was purchased from Sigma-Aldrich, USA. Extra pure xylene was obtained from Merck, Germany.

Molecular grade absolute ethanol, agarose and ethidium bromide (EtBr) were purchased from Applichem, Germany.

DNeasy Blood and Tissue Kit and proteinase K were obtained from Qiagen, USA.

DNA ladder mix (100-10000 bp), 50 bp DNA ladder, 6X loading dye, nuclease free water, dNTP set, *Taq* Buffer with $(NH_4)_2SO_4$, $MgCl_2$ and *Taq* DNA polymerase were purchased from Fermentas, Lithuania.

Human Cot-1 DNA (1 mg/ml), 5M Tetramethyl Ammonium Chloride (TMACL), Denhardt's Solution, 20X SSPE, 10% Tween-20, MES Hydrate, MES sodium salt, DMSO and EDTA were obtained from Sigma-Aldrich, USA.

T4 DNA Ligase, Sty I Restriction Enzyme, and Acetylated Bovine Serum Albumin (BSA) were purchased from New England Biolabs, UK.

Agencourt AMPure PCR Purification Kit was obtained from Beckman Coulter Genomics, Switzerland.

Herring Sperm DNA (10mg/ml) was bought from Promega, USA

R-Phycoerythrin Streptavidin was obtained from Invitrogen, USA.

Biotinylated anti-streptavidin antibody was purchased from Vector Labs, USA.

Titanium Taq PCR Kit was gotten from Clontech, Japan.

2.1.4 Primers

β-actin primers (Baran et al., 2007) and RAPD-PCR primers (Siwoski et al., 2002) were purchased from Alpha DNA, Canada. Generic primer was included in Affymetrix GeneChip Human Mapping 250K Sty Assay Kit. Primer sequences and amplicon sizes are given in Table 2.1.

Table 2.1 Primers used in gene specific PCR and RAPD-PCR, and amplicon sizes.

Primer	Sequence	Amplicon Size
***β-actin* Forward**	5' CAGAGCAAGAGAGGCATCCT 3'	209 bp
***β-actin* Reverse**	5' TTGAAGGTCTCAAACATGAT 3'	
RAPD-PCR 1	5' AATCGGGCTG 3'	Variable
RAPD-PCR 2	5' GAAACGGGTG 3'	Variable
Generic primer	5'ATTATGAGCACGACAGACGCCTGAT CT3'	Variable

2.1.5 Microarrays

GeneChip Human Mapping 250K Sty Arrays and GeneChip Human Mapping 250K Sty Assay Kits were purchased from Affymetrix, USA. Kit contents are listed in Table 2.2.

Table 2.2 The contents of 250K Sty Assay Kit.

Reagent	Volume
Adaptor Sty (50µM)	25 µl
PCR primer (100 µM)	450 µl
Reference genomic DNA (50ng/µl)	30 µl
GeneChip® Fragmentation Reagent	25 µl
10X Fragmentation Buffer	250 µl
GeneChip® DNA Labeling Reagent (30 mM)	60 µl
Terminal Deoxynucleotidyl Transferase (TdT) (30 U/µl)	110 µl
5X Terminal Deoxynucleotidyl Transferase Buffer	420 µl
Oligonucleotide Control Reagent	60 µl

Adaptor Sty sequence was as follows:

5'ATTATGAGCACGACAGACGCCTGATCT3'

3'AATACTCGTGCTGTCTGCGGACTAGAGWWCp5'

2.2 METHODS

2.2.1 Treatment Protocol of Patients

All phases of the treatment protocol were performed in the Department of Medical Oncology, Gülhane Military Medical Academy (GATA) (Arpaci et al., 2005). They are schematized in Figure 2.1 and are as follows.

Pre-operative treatment: After a histopathologic diagnosis was established, cisplatin, doxorubicin, ifosfamide, and mesna were administered for 2 cycles over the course of 3 days, every 3 weeks. For stem cell collection (leukapheresis), Granulocyte–colony-stimulating factor (G-CSF) was given twice a day for 4 days beginning on the 14th day after the second cycle of chemotherapy. Mononuclear cells (MNC) were collected on the day after the last day of G-CSF administration. MNCs were cryopreserved with a final dimethylsulphoxide (DMSO) concentration of 10%. Preoperative high dose chemotherapy was administered 3 weeks after completion of the second cycle of chemotherapy, using a high-dose combination of ifosfamide with mesna administered until 24 hours after the last dose of the drug in addition to carboplatin and etoposide. All drugs were given between days 1–6. After 2

days of rest without chemotherapy, the cryopreserved cells were reinfused intravenously to the patient.

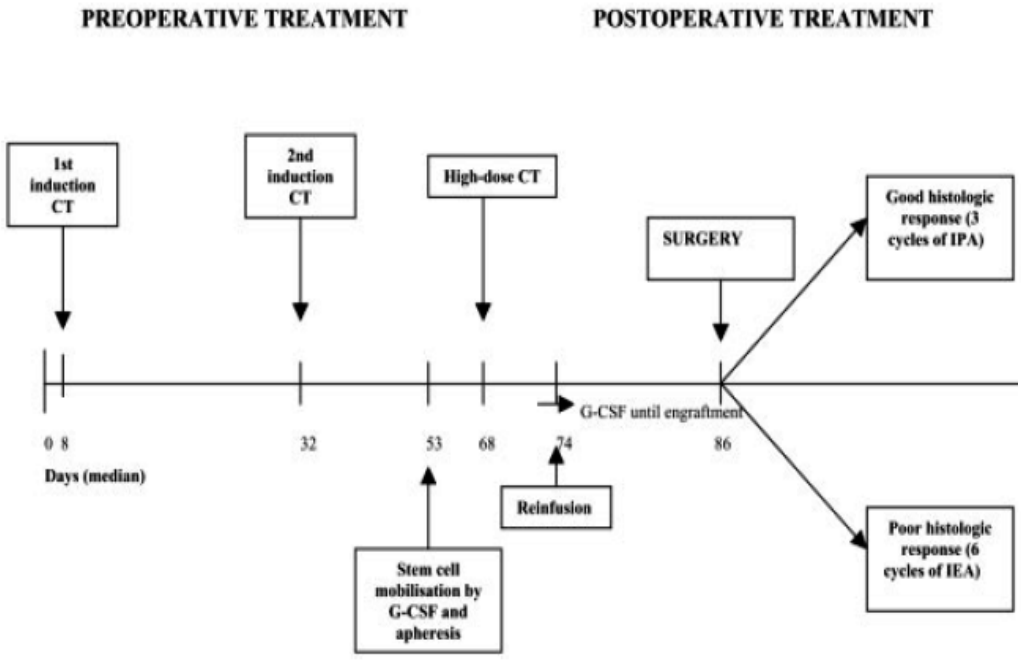


Figure 2.1 Treatment protocol. CT: chemotherapy; IPA: ifosfamide, cisplatin and adriamycin; G-CSF: granulocyte–colony-stimulating factor; IEA: ifosfamide, etoposide, and adriamycin.

Surgery: In all patients, surgery was performed after the leukocyte and platelet counts were normalized. Of the 10 patients, 10 underwent limb-sparing surgery (LSS). A slice of the tumor was extracted from the excised tissue containing the whole tumor. After decalcification, formalin fixation with 10% formic acid

and paraffinization were performed in order to prepare paraffin blocks of the tissue samples.

Postoperative treatment: Adjuvant therapy was initiated 2–3 weeks after the completion of surgery. For patients in whom the tumor necrosis rate exceeded 90%, the chemotherapy regimen was the same as that used for induction chemotherapy (ifosfamide, cisplatin, and doxorubicin [IPA]), and was administered for 3 treatment cycles every 3 weeks. Patients in whom the tumor necrosis rate was less than 90% received 6 cycles of ifosfamide, etoposide, and doxorubicin on Days 1–3, every 3 weeks.

Follow-up and observation: After the completion of all chemotherapy treatments, physical examination, whole blood count, blood chemistry, direct X-ray of the operated area, chest X-ray and/or thoracic CT, and CT or MRI of the lesion (if needed) and whole-body bone scintigraphy were performed every 3 months for 2 years, every 6 months for 3 years, and annually for 5 years thereafter.

2.2.2 Staging and Classification of Tumor Specimens

The histopathology of pulmonary metastatic osteosarcoma tumor specimens was examined by an expert pathologist (Prof. Dr. Mükerrem Safalı, Department of Pathology, Gülhane Military Medical Academy, Ankara)

according to World Health Organization (WHO) specifications. The clinical staging of samples was done according to the American Joint Committee on Cancer TNM system.

2.2.3 Preparation of Tissues for DNA Isolation

Before genomic DNA isolation, normal and tumor cells were selected and pretreated. Selection and pretreatment procedures are stated below.

2.2.3.1 Manual Microdissection

Microdissection involves procurement of pure populations of cells from heterogeneous histological sections (Erickson, Gillespie, & Emmert-Buck, 2008). In manual microdissection, the area of interest (normal and tumor) was identified by examination of hematoxylin-eosin (H&E) stained 5 µm sections (Appendix A) by an expert pathologist (Prof. Dr. Mükerrem Safalı, Department of Pathology, Gülhane Military Medical Academy, Ankara). Five 10 µm sections per sample were prepared. The area of interest (normal and tumor) on each 10 µm section was identified by superimposing the marked H&E-stained slide. The region outside the area of interest was scraped off from the slides with a scalpel. In other words, selected normal and tumor regions were left on the slides by manual microdissection.

2.2.3.2 Deparaffinization and Rehydration

Five 10 µm sections per sample were heated to 56^{o}C for 20 minutes to melt the paraffin. After heating, they were transferred to a reservoir containing 100% xylene and incubated for 5 minutes. This incubation was repeated in fresh xylene for 2 more times. The slides were allowed to air-dry briefly in a fume hood. A minimum of 1 cm^2 of tumor surface area was scraped from each slide. The scraped tissue was placed into a 1.5-ml nuclease-free eppendorf tube (Greiner), and was then washed twice with 1ml of ACS grade xylene, to ensure that all the paraffin was removed. Samples were then centrifuged in a microcentrifuge (ThermoScientific, USA) for 3 minutes at maximum speed, and the xylene was removed.

After deparaffinization, the samples were rehydrated by sequential treatment with 1ml of 100%, 75%, and 50% ethanol. The first ethanol wash was centrifuged for 3 minutes at 14,000g and supernatant was discarded. The samples were then washed with 75% ethanol and were centrifuged for 3 minutes at 14,000g. After removal of supernatant, the 50% ethanol wash was performed and samples were centrifuged for 5 minutes at 14,000g. After discarding supernatant, the tubes containing the tissue samples were allowed to air-dry briefly before DNA extraction was carried out.

2.2.4 DNA Isolation from Tissues

Formalin Fixed Paraffin Embedded tissues are the most available material for routine diagnostics in pathology laboratories worldwide. Those specimens represent a huge resource in order to discover and evaluate prognostic DNA markers. However, some DNA samples are highly fragmented or damaged because of preservation procedures, i.e., formalin fixation and paraffinization (Siwoski et al., 2002). Therefore, studying on these samples requires special handling in molecular assays. In our study, following de-waxing in xylene and rehydration in ethanol, genomic DNA was extracted from paired normal and tumor FFPE samples by using DNeasy Blood and Tissue Kit. Some modifications were made to the manufacturer's protocol (Lyons-Weiler, Hagenkord, Sciulli, Dhir, & Monzon, 2008) in order to isolate utilizable DNA from FFPE samples. All the steps of protocol, excluding proteinase K treatment, were performed at room temperature.

The air-dried tissue samples were homogenized in 300 µl of buffer ATL and 100 µl of Qiagen proteinase K (600 mAU/ml). The samples were incubated overnight at 56^{o}C in an air incubator (Heidolph, Germany), with shaking at 60 rpm. After tissue lysis, 400 µl of buffer AL was added to the sample and then vortexed for 15 seconds. The samples were incubated at 70^{o}C for 10 minutes to inactivate proteinase K. 400 µl of molecular grade 100% ethanol was then added to the samples and mixed by vortexing. Approximately half the volume of the sample was placed into the spin column provided by the manufacturer. The sample was then centrifuged for 1 minute at 8000g, and the flow-through was discarded. The second half of the sample was then added to the spin column and the step was repeated. The AW1 and AW2 wash steps were

performed according to the standard DNeasy protocol, using 500 µl of each wash buffer. AW1 wash was followed by a 1-minute centrifugation at 8000g, and the AW2 wash was followed by a 3-minute 14,000g centrifugation. The spin column was placed into a new 1.5 ml eppendorf tube. 50 µl Buffer AE was directly applied to the column membrane. Following 1-minute incubation, the column was centrifuged for 1 minute at 8000g. Another 50 µl Buffer AE was added and, incubation and centrifugation steps were repeated for the elution of DNA, for a final volume of 100 µl.

2.2.5 Quality Assessment of Isolated DNA

FFPE specimens generally yield DNA with low-quality. Moreover, for these nonrenewable, archival samples, specimen amount is often a limiting factor. Therefore, there is a need to assess the suitability of isolated DNA by using minimal amount of material. There are four different methods to evaluate the quality of DNA for microarray experiments as stated below (Siwoski et al., 2002).

2.2.5.1 Spectrophotometric Analysis of DNA

All DNA samples were quantified with NanoDrop 2000c spectrophotometer (ThermoScientific, USA). All samples processed for downstream analysis in this study had an OD-260/280 ratio between 1.7 and 2.0.

2.2.5.2 Agarose Gel Electrophoresis

The degree of fragmentation of DNA samples was measured by horizontal agarose gel electrophoresis.

One gram of agarose was dissolved in 100 ml 1X TAE buffer (Appendix B) and boiled in a microwave oven. After cooling, 7 µl EtBr solution (Appendix B) was added to the gel solution. The gel solution was poured into electrophoresis plate and the comb was placed. After solidification of the gel, DNA samples were loaded. In order to load, 7.5 µl of each DNA sample was mixed with 1.5 µl 6X loading dye (Appendix B). DNA ladder mix (100-10000 bp) was also loaded. Electrophoresis was run with 1 l of 1X TAE buffer for 1h at 80V. The gel was visualized under UV light and a photograph was taken.

2.2.5.3 Gene Specific Polymerase Chain Reaction

Gene specific PCR for *β-Actin* gene was performed to assess the efficiency of DNA samples to amplify a product of a specific size by PCR. Primers for *β-Actin* and PCR conditions were previously described by Baran *et al*. (Baran et al., 2007).

The mixture was prepared in the sterile 0.5 ml eppendorf tubes. 25 µl reaction mix contained 1X *Taq* Buffer with $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 0.25 mM dNTP mix, 2 µM from each forward and reverse primers, 0.5 µl template and 0.5 unit *Taq* DNA polymerase. Amplification conditions are represented in Table 2.4. A non-template control which contains nuclease free water instead of template was also included to identify any contamination. PCR reaction was carried out in Thermal Cycler (ThermoScientific, USA). PCR products were detected by horizontal agarose gel electrophoresis. Preparation of agarose gels were described above. 2% (w/v) agarose gel was prepared by dissolving 2 g agarose in 100 ml 1X TAE buffer. 10 µL of PCR sample was mixed with 2 µL 6X loading dye and loaded. 50 bp DNA Ladder was also loaded at each run. Electrophoresis was run with l l of 1X TAE buffer for 1 h at 90V. Gel was observed under UV light and photographed.

2.2.5.4 Randomly Amplified Polymorphic DNA Polymerase Chain Reaction

RAPD-PCR was performed to assess the efficiency of DNA samples to amplify products of variable sizes by PCR. Primers for RAPD-PCR and PCR conditions were previously described by Siwoski *et al*. (Siwoski et al., 2002).

The mixture was prepared in the sterile 0.5 ml eppendorf tubes. 25 µl reaction mix contained 1X Taq Buffer with $(NH_4)_2SO_4$, 2 mM $MgCl_2$, 0.2 mM dNTP mix, 20 µM from each RAPD-PCR primers, 0.5 µl DNA template and 1 unit Taq DNA polymerase. Amplification conditions are represented in Table 2.3. A non-template control which contains nuclease free water instead of template was also included to identify any contamination. PCR reaction was carried out in Thermal Cycler (ThermoScientific, USA). PCR products were detected by horizontal agarose gel electrophoresis. Preparation of agarose gels were described above. 2% (w/v) agarose gel was prepared by dissolving 2 g agarose in 100 ml 1X TAE buffer. 10 µl of PCR sample was mixed with 2 µl 6X loading dye and loaded. 50 bp DNA Ladder was also loaded at each run. Electrophoresis was run with l l of 1X TAE buffer for 1 h at 90V. Gel was observed under UV light and photographed.

Table 2.3 Amplification conditions for *β-Actin* gene and RAPD-PCR

	***β-Actin* gene**	**RAPD-PCR**
Initial denaturation	94°C, 5 min	94°C, 5 min
Denaturation	94°C, 30 sec	94°C, 1 min
Annealing	53°C, 45 sec	35°C, 1 min
Extension	72°C, 1 min	72°C, 2 min
Final extension	72°C, 5 min	72°C, 5 min
Cycle number	35	45

2.2.6 Microarray Analysis

As stated above in Section 2.2.3, DNAs isolated from FFPE samples have some restrictions for use in molecular assays. Fragmentation caused by preservation procedures is the main problem. However, microarray experiments require high quality DNA. Therefore, studying on these samples requires some modifications in the protocol and data analysis. In our study, paired normal and tumor DNA samples were analyzed on the Affymetrix GeneChip Human Mapping 250K Sty SNP Array, containing 238,300 SNPs. Some modifications were made in order to optimize the standard protocol for FFPE DNA samples (Lyons-Weiler et al., 2008). Those modifications are explained in each section of microarray analysis in detail.

Microarray analysis protocol contains ten stages as stated below.

1. Restriction enzyme digestion
2. Ligation
3. PCR

4. PCR product purification and elution
5. Quantitation and normalization
6. Fragmentation
7. Labeling
8. Target hybridization
9. Washing and staining
10. Scanning

2.2.6.1 Restriction Enzyme Digestion

DNA samples were digested with restriction enzyme called Sty I which produces four base-pair (bp) overhangs after digestion. The standard 250K Sty assay failed to produce sufficient PCR product from fragmented DNA samples for array hybridization. Therefore, during this stage, initial DNA amount for each sample was increased from 250 ng to 1000 ng by increasing the starting volume of the each DNA sample. In order to get efficient digestion, each component of digestion master mix increased two times. Moreover, all steps were performed on ice.

From each DNA sample, required volume containing 1000 ng of DNA was taken to a 0.5 ml eppendorf tube. Digestion master mix was prepared into a 1.5 ml eppendorf tube according to the required volumes of reagents as stated in Table 2.4. 31.5 µl digestion mix was added to each sample. After vortexing, samples were placed into thermal cycler and 500K digest program was run. Digestion conditions were modified according to the enzyme amount in the

reaction. Standard and modified digestion conditions were given in Table 2.5. After digestion, all digestion products were stored at -20^{o}C to be used in the next stage.

Table 2.4 Reagents used in Sty I digestion master mix.

Reagent	**1 sample standard**	**1 sample modified**
AccuGENE® Water	11.55 µl	23,1 µl
NE Buffer 3 (10X)	2 µl	4 µl
BSA (100X; 10 mg/ml)	0.2 µl	0,4 µl
Sty I (10 U/µl)	1 µl	4 µl
Total	14.75 µl	31,5 µl

Table 2.5 Digestion conditions for Sty I enzyme.

Temperature	**Time standard**	**Time modified**
37^{o}C	2 hours	16 hours
65^{o}C	20 min	20 min
4^{o}C	Hold	Hold

2.2.6.2 Ligation

Following Sty I digestion, DNA samples were ligated to adaptors that recognize the cohesive four bp overhangs. In this step, in order to continue

with standard protocol, all digestion products were concentrated. Moreover, all ligation steps were performed on ice.

In order to proceed with ligation, digestion products were concentrated to 19.55 µl from 36.5 µl by air drying samples on a heater microcentrifuge. Ligation master mix was then prepared into a 1.5 ml eppendorf tube according to the required volumes of reagents as stated in Table 2.6a. 5.25 µl ligation mix was added to each digestion product. After vortexing, samples were placed into thermal cycler and 500K ligate program was run. Ligation conditions were given in Table 2.6b.

Table 2.6 a) Reagents used in ligation master mix. b) Ligation conditions for Sty I adaptor.

a)

Reagent	**1 sample**
Adaptor Sty I (50 µM)	0.75 µl
T4 Ligase Buffer (10X)	2.5 µl
T4 DNA Ligase (400U/µl)	2 µl
Total	5.25 µl

b)

Temperature	**Time**
16°C	180 min
40°C	20 min
4°C	Hold

After ligation, samples were diluted with the addition of 75 µl AccuGENE® Water and were stored at -20°C to be used in the next stage.

2.2.6.3 PCR

A generic primer that recognizes the adaptor sequence is used to amplify adaptor ligated DNA fragments. By this way, amplification of the whole genome was provided. In order to get sufficient amount of PCR product for hybridization stage, three PCR reactions are required according to the manufacturer's protocol. In this study, however, four PCR reactions were performed to get sufficient amount of PCR product. Therefore, four 0.5 ml PCR-eppendorf tubes were prepared for each sample.

90 µl PCR mixture was prepared for each sample and 10 µl ligation product was added to the mixture. The reagents and PCR conditions are given in Table 2.7 and Table 2.8, respectively.

Table 2.7 The reagents used in target amplification PCR reaction.

Reagent	**For 1 reaction**
AccuGENE® Water	39.5 µl
TITANIUM Taq PCR	10 µl
GC-Melt (5M)	20 µl
dNTP (2.5 mM each)	14 µl
PCR Primer 002 (100 µM)	4.5 µl
TITANIUM Taq DNA Polymerase (50X)	2 µl
Ligation product	10 µl
Total	100 µl

Table 2.8 Amplification conditions for target amplification PCR reaction.

	Temperature	Time
Initial denaturation	94°C	3 min
Denaturation	94°C	30 sec
Annealing	60°C	45 sec
Extension	68°C	15 sec
Final extension	68°C	7 min
Cycle number	40	

PCR products were detected by horizontal agarose gel electrophoresis. Preparation of agarose gels were described above. 2% (w/v) agarose gel was prepared by dissolving 2 g agarose in 100 ml 1X TAE buffer. 3 µl of PCR sample was mixed with 1 µl 6X loading dye and loaded. 50 bp DNA Ladder was also loaded at each run. Electrophoresis was run with 1 l of 1X TAE buffer for 40 min at 120V. Gel was observed under UV light and photographed. The protocol was continued with the samples which gave products between 110-200 bp.

2.2.6.4 PCR Product Purification and Elution

During this stage, PCR product purification was carried out by using Agencourt AMPure PCR Purification Kit according to the manufacturer's protocol.

Each PCR product was distributed in equal volumes into three wells of a purification plate, which is demonstrated in Figure 2.2. In our modified protocol specific to FFPE DNA samples, there were 4 PCR tubes for each sample; therefore, 12 wells were necessary for each sample.

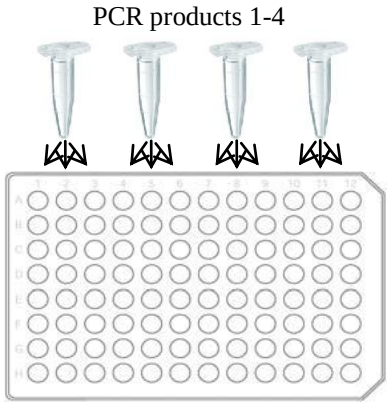


Figure 2.2 Distribution of PCR products into purification plate

To each well, 1.8µl AMPure XP per 1.0 µl of PCR product was added and mixed by pipette until getting a homogenous mixture. After incubation at room temperature for 5 minutes in order to let the PCR products to bind paramagnetic beads, the purification plate was placed onto an Agencourt STRIPlate 96 Super Magnet Plate for 2 minutes to separate beads from the solution. The cleared supernatant was aspirated from the purification plate and discarded. The beads and PCR products were washed in 2 volumes of 70% Ethanol to remove contaminants for 2 times. Ethanol was aspirated and discarded at each time. The purification plate was incubated at Room Temperature for 5 minutes to ensure all traces of Ethanol were removed. The

purification plate was taken off from the magnet plate. 40 µL of elution buffer was added to each well of the purification plate and mixed homogenously. The purification plate was placed onto Magnet Plate for 1 minute to separate beads from the solution. The eluant or supernatant contained the purified PCR products. The twelve wells of each sample were collected in one eppendorf tube. Samples were stored at -20^{o}C to be used in the next stage.

2.2.6.5 Quantitation and Normalization

In this step, samples were prepared for the next stage which is fragmentation. Forty five µl of DNA with 2 µg/µl concentration is necessary for fragmentation.

After PCR purification, samples were concentrated by vaporizing the liquid part on a heater centrifuge. It was essential that there should be no remaining liquid in tubes so that with the addition of 45 µl elution buffer, only this amount of product was present in each tube. After elution, samples were quantified with NanoDrop 2000 Spectrophotometer (ThermoScientific, USA). Once the concentration of each reaction is determined, samples were normalized to 2 µg/µl by adding RB Buffer. Normalization was done according to the formula (Equation 2.1) below:

X µL RB Buffer = 45 µL − (Y µL purified PCR product)

Where:

Y = The volume of purified PCR product that contains 90 µg

The value of Y is calculated as:

Y µL purified PCR product = (90 µg) ÷ (Z µg/µL)

Z = the concentration of purified PCR product in µg/µL (2.1)

2.2.6.6 Fragmentation of PCR Products

During this stage the purified, normalized PCR products were fragmented using DNase I (fragmentation reagent). All steps were performed on ice and as fast as possible in order to get uniform fragmentation.

The program on thermal cycler was adjusted before the experiment and thermal cycler was preheated to 37^{o}C. The 55 µl fragmentation mixture was prepared by addition of 45 µl purified PCR product (90µg), 5 µl fragmentation buffer (10X) and 5 µl diluted fragmentation reagent (0.05 U/µl). Each tube placed onto thermal cycler and fragmentation program was run (Table 2.9).

Table 2.9 Fragmentation conditions.

Temperature	Time
37^{o}C	35 min
95^{o}C	15 min
4^{o}C	Hold

Once the program finished, fragmentation products were run on a 2% agarose gel. 4 µl sample was mixed with 2 µl 6X loading dye and loaded to gel. 50 bp DNA Ladder was also loaded at each run. Electrophoresis was run with l L of 1X TAE buffer for 40 min at 120V. Gel was observed under UV light and photographed. The protocol was proceeded with the samples which gave products approximately 180 bp.

2.2.6.7 Labeling of Fragmented PCR Products

The fragmented samples were labeled using the GeneChip® DNA Labeling Reagent. During this stage, to the ends of fragmented PCR products, biotin reactive nucleotides were added with the help of Terminal Deoxynucleotidyl Transferase (TdT) enzyme.

The Labeling Master Mix was prepared according to the manufacturer's protocol and the required volumes of reagents are given in Table 2.10. 50.5 µl fragmented DNA was added to 19.5 µl labeling mix. The samples were placed onto the thermal cycler and the labeling program was run. The labeling conditions were given in Table 2.11.

Table 2.10 Reagents used in labeling master mix.

Reagent	For 1 sample
Reagent TdT Buffer (5X)	14 µL
GeneChip® DNA Labeling Reagent (30 mM)	2 µl
TdT enzyme (30 U/µL)	3.5 µl
Total	19.5 µl

Table 2.11 Reaction conditions for labeling.

Temperature	Time
37°C	4 h
95°C	15 min
4°C	Hold

2.2.6.8 Target Hybridization

Each sample was loaded onto a GeneChip® Human Mapping 250K Sty Array. Before starting experiment, hybridization oven was preheated to 49°C and adjusted to 60 rpm.

Hybridization Master Mix was prepared according to the manufacturer's protocol and the required volumes of reagents are given in Table 2.12. 190 µl mix was added to 70 µl labeled sample. Samples were denatured on a thermal cycler for 10 minutes at 95°C and they were hold at 49°C.

Table 2.12 Reaction mixture preparation for hybridization.

Reagent	For 1 reaction
MES (12X; 1.25 M)	12 µl
Denhart's Solution (50X)	13 µl
EDTA (0.5 M)	3 µl
Herring Sperm DNA (10 mg/ml)	3 µl
OCR, 0100	2 µl
Human Cot-1 DNA® (1 mg/mL)	3 µl
Tween-20 (3%)	1 µl
DMSO (100%)	13 µl
Tetramethyl Ammonium Chloride (TMACL, 5 M)	140 µl
Total	190 µl

During denaturation, arrays were taken out from +4°C and they were allowed to warm to room temperature by leaving on the bench top 10 to 15 minutes. Once they were ready, a 200 µL pipette tip was inserted into the upper right septum of each array. The samples which were at 49°C were taken out from thermal cycler and each 200 µl sample was immediately loaded lower left septum of each array (Figure 2.3). The septa were then covered with tough-spots.

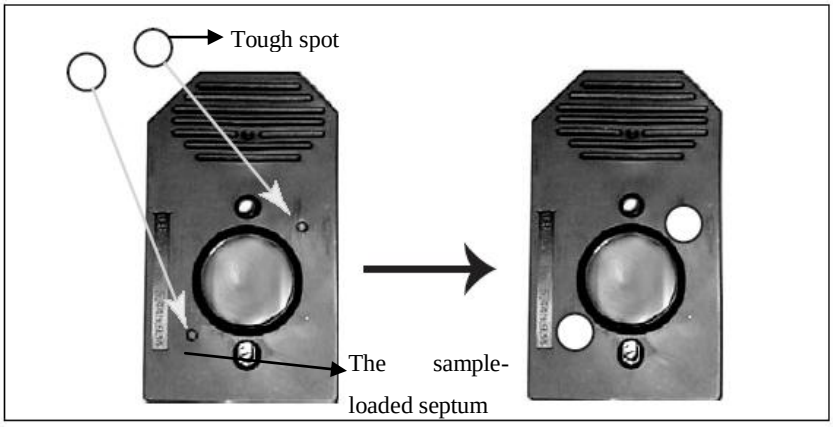


Figure 2.3 The back view of cartridges and application of tough spots.

Arrays were loaded into an oven tray and the tray was then immediately placed into the hybridization oven (Figure 2.4). Samples were left to hybridize for 16 to 18 hours at 49°C, with rotating at 60 rpm.

2.2.6.9 Washing and Staining

In this stage, cartridges were washed and stained in order to be made ready for scanning. First of all, two different wash solutions, Wash A and Wash B, (Appendix B) were prepared freshly. They were filtered through a 0.2 µm filter and their pH was adjusted to 8.0. Those solutions were stored at room temperature.

The Fluidics Station 450 (in Figure 2.4) is used to wash and stain the probe arrays; it is operated using GeneChip Operating Software. There are four modules in this station. After turning on the fluidics station, prior to placing the arrays into the station, priming was done by prime_450 protocol of the machine in order to ensure that the lines of the fluidics station are filled with the appropriate buffers and the fluidics station is ready to run fluidics station protocols. By the priming protocol, the intake buffer reservoir A was changed to Wash A: Non-Stringent Wash Buffer, and intake buffer reservoir B was changed to Wash B: Stringent Wash Buffer.

Figure 2.4 Equipments in GeneChip System.

Once priming was complete, four different solutions for staining, which are array holding buffer, stain buffer, Streptavidin Phycoerythin (SAPE) stain solution and antibody stain solution, were prepared. Preparation instructions for these solutions are given in Appendix B.

Once hybridization was complete, the hybridization cocktail was removed from the probe array and the probe array was completely filled with 270 µl of Array Holding Buffer. 820 µl of Array Holding Buffer was added to each microcentrifuge tube. One tube is needed per module. The three remaining staining solutions were placed into three different vials.

Table 2.13 Washing/Staining protocol for mapping arrays.

Steps	Protocol
Post Hyb Wash #1	6 cycles of 5 mixes/cycle with Wash Buffer A at 25°C
Post Hyb Wash #2	24 cycles of 5 mixes/cycle with Wash Buffer B at 45°C
Stain	Stain the probe array for 10 minutes in SAPE solution at 25°C
Post Stain Wash	6 cycles of 5 mixes/cycle with Wash Buffer A at 25°C
2nd Stain	Stain the probe array for 10minutes in Antibody Stain Solution at 25°C
3rd Stain	Stain the probe array for 10 minutes in SAPE solution at 25°C
Final Wash	10 cycles of 6 mixes/cycle with Wash Buffer A at 30°C. The final holding temperature is 25°C
Filling Array	Fill the array with Array Holding Buffer.

In the Fluidics Station workstation, sample numbers and sample information was entered into system computer. After data entry, "Mapping500Kv1_450" protocol was selected to control the washing and staining of the probe array. The three vials containing stain solutions were loaded into the fluidics station. The washing/staining protocol was run after the probe arrays were placed into the station. The Affymetrix staining protocol for mapping arrays is a three stage process consisting of SAPE staining, followed by antibody amplification step and a final staining with SAPE. Washing/Staining protocol was outlined in Table 2.13. Once staining was complete, the array was filled with Array Holding Buffer prior to scanning.

2.2.6.10 Scanning and Preliminary Analysis

The probe arrays were loaded into GeneChip® Scanner 3000 7G. Scanning program was run according to the manufacturer's instructions. By using GCOS (GeneChip® Operating Software), scanning protocol was selected and started.

After completion of scanning, image analysis was done. Quality of image data was checked according to quality control criteria which are B2 oligonucleotide control presence, and checkerboard pattern at each corner and throughout the array. The raw data files in .CEL format were generated and transferred into Genotyping Console 4.0 Software in order to get quality control reports and to do whole-genome genotyping analysis to obtain genotype information of each sample. Genotyping console general workflow is given in Figure 2.5.

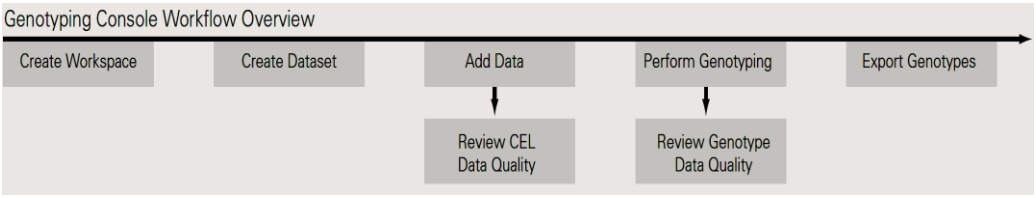


Figure 2.5 Genotyping console workflow.

In order to check the quality of CEL files, QC (Quality Control) Call rate of each sample were generated. In manufacturer's protocol, samples with QC call rate higher than QC Call Rate Threshold 86 are studied in further analysis. However, for FFPE samples, threshold have to be decreased because the data from those samples generally yield QC call rate around 70 as stated in Lyons' research (Lyons-Weiler et al., 2008). We defined our threshold as 60, hence, the samples with at least 60 QC call rate could be analyzed in our study. Graphical display of QC metrics across samples was created to identify outliers, including 50 signature SNPs for tracking sample identity.

Genotyping analysis was performed with the samples having higher QC call rate than QC threshold 60. In this analysis, genotype files in CHP format were generated using BRLMM algorithm to be used in whole-genome association analysis.

2.2.6.11 Microarray Data Analysis and Visualization of Results

2.2.6.11.1 Genotyping: BRLMM Algorithm

Genotyping analysis was performed with the samples having higher QC call rate than QC threshold 60 by using manufacturer's default parameters (Table 2.14). In this analysis, genotype files in CHP format were generated using Bayesian Robust Linear Model with Mahalanobis distance classifier (BRLMM) genotyping algorithm (T. D. Model, Mapping, Set, Model, & Snp, 2006). The BRLMM algorithm workflow is given in Figure 2.6.

Table 2.14 Manufacturer's Parameters for Genotyping Analysis

BRLMM Algorithm	Default Settings
Score Threshold*	0.5
Block Size	0
Prior Size#	10000
DM Threshold¥	0.17

*The maximum value of confidence for which the algorithm will make a genotype call.
How many probe sets to use for determining prior.
¥ DM confidence threshold used for seeding clusters.

After genotyping analysis, genotype results were exported in PLINK (Purcell et al., 2007) ready format to be used in GWAS performed by METU-SNP software. The PLINK files include Pedigree and Map format files. Map format contains either 3 columns representing chromosome number, reference

sequence ID for SNP and base pair location, or 4 columns with an additional column of genetic distance in centimorgans.

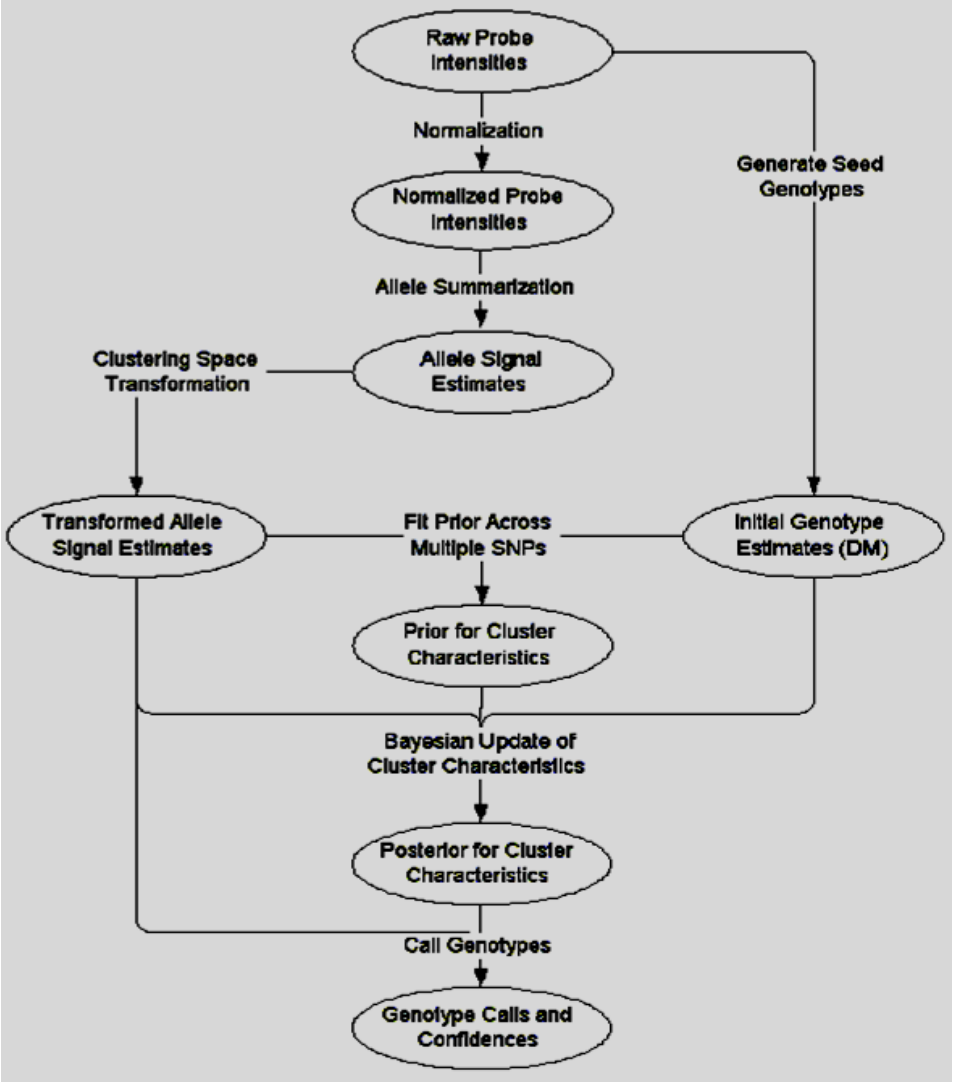


Figure 2.6 BRLMM algorithm workflow (T. D. Model et al., 2006).

2.2.6.11.2 Genome-Wide Association Analysis and SNP Prioritization

Genome-wide association analysis followed by representative SNP selection was performed by METU-SNP software system (Gurkan Üstünkar, 2011). The workflow of the GWAS and SNP prioritization is represented in Figure 2.7. Each step is detailed in subsequent sections.

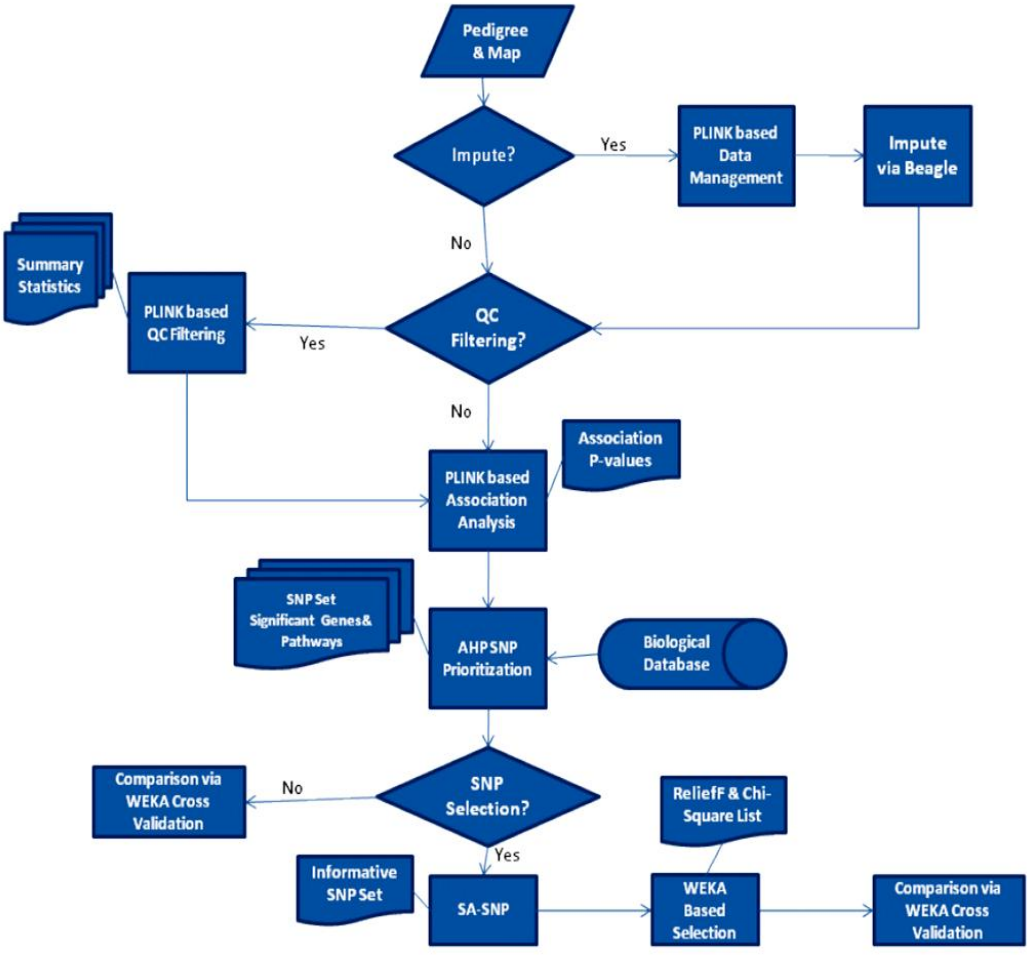


Figure 2.7 Logic Workflow of METU-SNP Software System (Gurkan Üstünkar, 2011).

2.2.6.11.2.1 Data Preprocessing and Cleaning

In the first step of GWAS analysis, data preprocessing was performed in order to remove data points which are under thresholds. In this step, PLINK based quality control filtering and imputation via BEAGLE (B. L. Browning & Browning, 2007) was performed by METU-SNP.

In order to filter out SNPs/individuals, some QC thresholds for PLINK to work on are defined. They are minor allele frequency, missingness and Hardy Weinberg equilibrium. Definitions of these thresholds and their importance in SNP complex disease association analysis are given in Appendix C. Default values of QC thresholds for the analysis with METU-SNP were 0.05 for Minor Allele Frequency (MAF), 0.1 for SNP Missingness Rate, 0.1 for Individual Missingness Rate and 0.001 for Hardy Weinberg equilibrium and, those default thresholds were designed according to commonly used values in various GWAS studies. In our study, the thresholds of MAF, SNP Missingness Rate, Individual Missingness Rate were changed to 0.01, 0.4 and 0.3, respectively, because the data obtained from FFPE samples were noisy and also not in high quality in contrast to the data from fresh samples (Purcell, 2010). With the help of analysis with PLINK, whole genome genotyping data is split into individual chromosome files to be used as input files for BEAGLE in imputation process. Moreover, the descriptive statistics files including freq.frq, missing.imiss, missing.lmiss, hardy.hwe were created as outputs during PLINK based QC analysis. The structures of these files are given in Appendix D.

After QC analysis, data is imputed for missingness by BEAGLE which is integrated into METU-SNP. Imputation is a statistical method to predict a calculated value for a missing data point. In GWAS, imputation is used to substitute missing or ungenotyped data if genotyping fails in certain number of typed SNPs and so SNP coverage of the samples in case-control studies is increased beyond what has been genotyped. After imputation, three output files for each chromosome are formed by BEAGLE. The first one is the phased file (.phased.gz) which provides imputed missing data. The second and third files are the genotype probabilities file (.gprobs.gz) and allelic r^2 file (.r2), respectively. These two files signify how accurate the imputation process has been for a particular marker. There are two columns in allelic r^2 files: The marker identifier and estimated squared correlation ($0 \leq r^2 \leq 1$) (which is between the allele dosage with highest posterior probability in the genotype probabilities file and the true allele dosage for the marker). If allelic r^2 value is high enough, more accurate genotype imputation has been obtained. In METU-SNP, default threshold for allelic r^2 is 0.95 to include only "well" imputed markers in subsequent analysis. However, as stated previously, our data was not in high quality. For this reason, we had to specify the threshold for imputation as 0.75.

The output of the data-preprocessing and cleaning step was PLINK based binary data file which contained the imputed genotype data satisfying QC thresholds. This file was used in association analysis in the next step.

2.2.6.11.2.2 GWAS: SNP, Gene and Pathway Association

In GWA step of analysis, the main objective is to find the significantly associated SNP biomarkers with the disease phenotype. In the second-wave GWAS analysis, the calculated *p*-values of the disease associated SNPs are used to associate gene and pathways based on the combined *p*-value approach (Peng et al., 2010).

In initial GWAS, a PLINK based association analysis was run in order to find significantly associated SNPs. There are three different statistical methods offered by METU-SNP to calculate the *p*-values, namely Uncorrected, Bonferroni and False Discovery Rate (Appendix C). The latter two methods make adjustments for multiple testing. For our study, we have chosen to utilize uncorrected *p*-values. In this step, there is again a threshold to be determined for *p*-value. Depending on this threshold, SNPs are labeled as significant or not. In our study, the threshold *p*-value is specified as default 0.05 to measure statistical significance of individual SNPs.

Following initial GWAS, the calculated *p*-values for individual SNPs are used in a second-wave GWAS analysis in order to identify significant genes and pathways which will be used in SNP prioritization step. In order to label a gene as significant, METU-SNP offers three thresholds: (1) combined *p*-value, (2) min SNP *p*-value and (3) max SNP *p*-value. Fisher's combination test is used to obtain combined *p*-value for a gene (Appendix C). For a gene to be called as significant, the combined value of it should be less than the threshold. In METU-SNP, we can also specify thresholds regarding the *p*-values of

individual SNPs associated with the gene. By this way, we can limit how big the minimum *p*-value of the SNPs or maximum *p*-value of the SNPs associated with the gene. In our study, we chose the combined *p*-value threshold as the default 0.05, min *p*-value threshold as 0.005 and max *p*-value threshold as 0.5.

In the second step of second-wave GWAS analysis, we determine the significant pathways. METU-SNP also offers three thresholds for this step: (1) combined *p*-value, (2) number of significant genes and (3) proportion of significant genes. Fisher's exact test is used to obtain combined *p*-value for a pathway (Appendix C). If combined value for the pathway is less than the threshold, pathway is labeled as significant. In METU-SNP, pathways are regarded as a combination of genes. Hence, the latter two thresholds are used to determine the significance of the pathway. In our study, we chose the combined *p*-value threshold as the default 0.05, the threshold for number of significant genes as 3. We did not define the third threshold, proportion of significant genes, in order to obtain all pathways that have association with the disease.

At the end of GWAS analysis, three output files were created by METU-SNP: (1) snp.txt, (2) gene.txt and (3) pathway.txt.

2.2.6.11.2.3 SNP Prioritization

In this step of the analysis, the information of statistically significant SNPs, genes and pathways are used to determine the prioritized set of SNPs. METU-SNP uses the AHP based SNP prioritization approach. This approach employs an SNP scoring scheme (Appendix E) that uses genetic information and biological functionality information in the METU-SNP database, and statistical information obtained by GWAS. Using these, AHP scores are calculated for each SNP. In fact, AHP scores are generated only for the SNPs that meets the individual SNP *p*-value threshold (less than or equal to the user specified *p*-value, 0.05 in our study). Then, those SNPs are ranked according to their AHP scores. The relevant SNPs are then filtered either depending on the calculated *p*-values or AHP scores or both.

At the end of this step, we obtain the most significant SNPs, genes and pathways in results panel. In the SNPs panel, SNP rsID, *p*-value and significance are written for the first user defined *n* SNPs. In our study we defined *n* as 2050. In Gene's panel, Entrez gene ID, full name and cytolocation are listed for the most significant genes. In Pathways panel, pathway ID, pathway system (database), full pathway title, pathway URL (web site for the particular pathway) and gene count for the pathway are listed.

2.2.6.11.3 Genome-Wide Copy Number Aberration and Loss of Heterozygosity Analysis

Genome-wide CNA and LOH analysis was performed with R-based (Venables & Smith, 2011) package: An R Object Oriented Microarray Analysis Environment for Affymetrix Microarrays (aroma.affymetrix) (Henrik Bengtsson, 2004). Aroma is a package for low-level analysis, which includes calibration and normalization, of single and two- or multiple- channel microarray data. Unlimited number of arrays of all Affymetrix chip types, e.g. expression arrays, SNP chips, exon arrays and so on, can be analyzed simultaneously. It works directly with the raw data files, i.e. .CEL and .CDF files. Preprocessing step includes background correction, allelic cross-talk calibration, quantile normalization, and nucleotide-position normalization. Postprocessing includes PCR fragment length normalization and/or GC-content normalization. It can perform both paired and non-paired copy-number analysis for all Affymetrix genotyping arrays. It uses various segmentation methods such as CBS, GLAD, GADA, PSCBS and HaarSeg. At the end of the analysis, it provides dynamic HTML reports, such as ChromosomeExplorer (HapMap demo, Tumor/Normal Demo).

In this study, different aroma vignettes and extensions were tried, and three methods which fit our data best were chosen. The R-vignettes of all methods utilized in this study are given in Appendix F and also they can be found on the aroma-project website.

The main difference of the three methods used in this study was in the quality of data they can analyze. The threshold for data quality is the respective SNP call rate. For a data to be called as high quality, it should have at least 95% SNP call rate. The first method, A Calibration Method to Improve Allele-Specific Copy Number Estimates from SNP Microarrays (CalMaTe) (Henrik Bengtsson, Neuvial, Cnrs, & Olshen, 2011), uses high quality data which is structured as tumor and reference sets in the same folder, i.e. nonpaired set. There is no allele specific segmentation method that can be applied after it and for this reason we have applied a total copy number segmentation method called Circular Binary Segmentation (CBS). The second method, Allele Specific Copy number estimation using Robust Multichip Analysis Version 2 (ASCRMA v2) followed by Tumorboost (Henrik Bengtsson, Neuvial, & Speed, 2010), utilizes high quality paired data which is a set of tumor and normal samples taken from the same patient. By this way, each tumor sample is compared to its respective normal match in downstream analysis. After CN analysis, we have applied Paired Parent Specific CBS (PSCBS) for segmentation. The last method, CRMA v2 (Henrik Bengtsson, Wirapati, & Speed, 2009) followed by Virtual Normal (VN) algorithm for reference selection, can analyze low quality data, such as the one obtained from FFPE samples, as nonpaired data set in which tumors and normals are found together. During this method, all normal samples are used as reference for each tumor and segmentation is performed with Genome Alteration Detection Analysis (GADA). In downstream analysis, each tumor is analyzed separately comparing to total normal reference.

The overall workflow of the CNA and LOH analysis from SNP arrays with the chosen three methods are given in the Figure 2.8. It can be divided into five steps, which are explained in detail in following sections, namely:

1. Preprocessing
2. Probe summarization
3. Post processing
4. Raw copy number calculation
5. Copy number segmentation

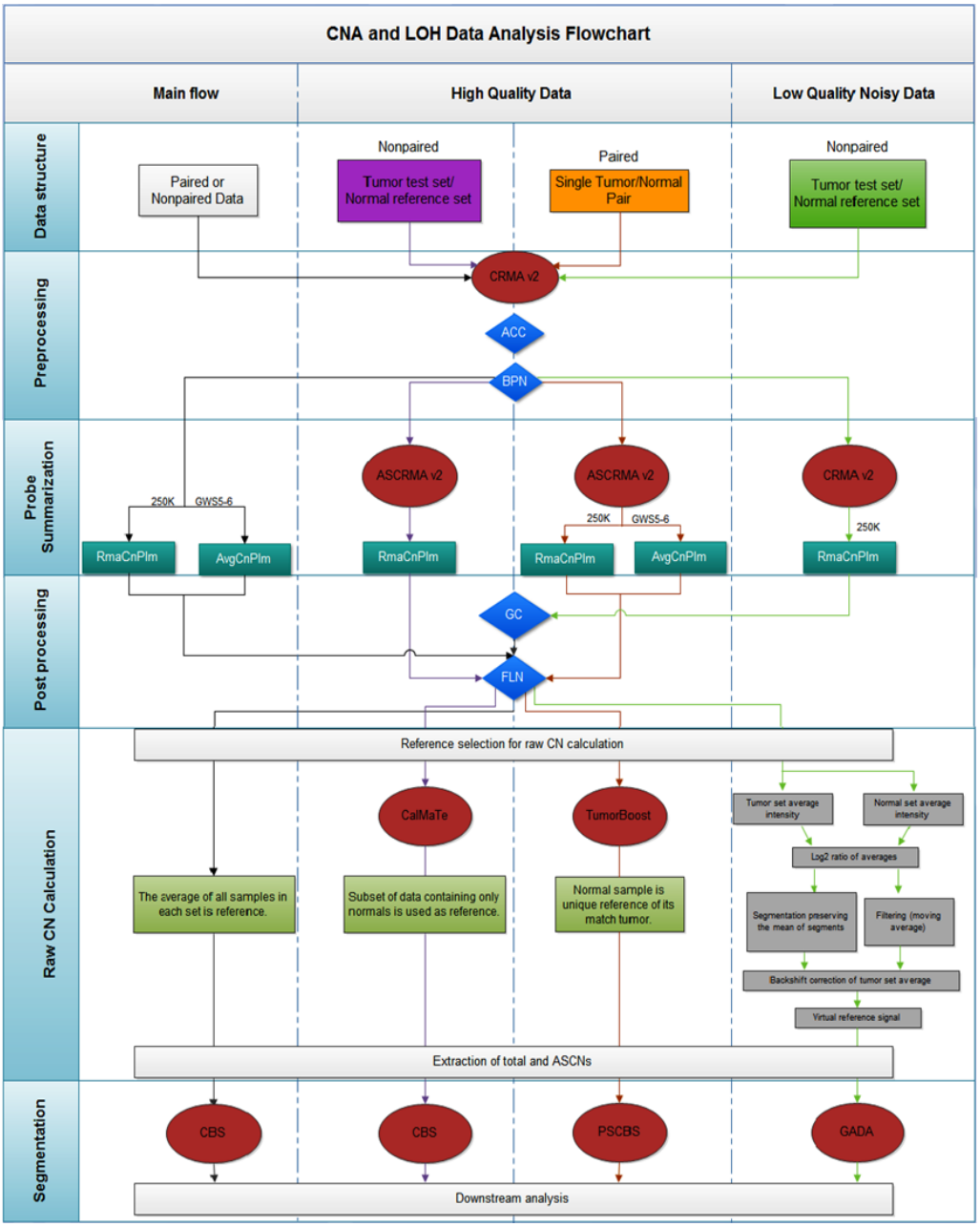


Figure 2.8 Flowchart of CNA and LOH analysis from SNP arrays. (CRMA v2: Copy number estimation using Robust Multichip Analysis Version 2, ACC: Allelic crosstalk, BPN: Base position normalization, CalMaTe: A Calibration Method to Improve Allele-Specific Copy Number Estimates, ASCRMA v2: Allele specific CRMA v2, GC: GC content correction, FLN: Fragment length normalization, CBS: Circular binary segmentation, PSCBS: Paired parent specific CBS, GADA: Genome alteration detection analysis; Zoom in Figure 2.9).

2.2.6.11.3.1 Preprocessing

In all the three methods, we have utilized CRMA v2 method (Henrik Bengtsson et al., 2009) for data preprocessing which includes 2 major steps. The first one is allelic crosstalk calibration (ACC) and the second one is base position normalization (BPN).

Allelic crosstalk is cross-hybridization between probes targeting the two alleles, Perfect match A (PM_A) and Perfect match B (PM_B), of a SNP. It occurs because of the close similarity between PM_A and PM_B probes which differ by one nucleotide in the position that corresponds to the SNP locus in the genome (H Bengtsson, Irizarry, Carvalho, & Speed, 2008). ACC also includes background correction which is the removal of the effects of non-specific binding or spatial heterogeneity across the array. By ACC, the probe level signals are made comparable across samples (H. Lu et al., 2011). This calibration can be applied to each array separately.

BPN is the second step in preprocessing. In fact, it is normalization for nucleotide-position probe sequence effects. This correction is based on probe sequence affinity which can be attributed to its sequence composition. For this reason, probe sequence affinity was modeled as a function of nucleotide position to control (i) little fluctuations in probe affinities across arrays, and (ii) differences in PM_A and PM_B affinities (Benilton Carvalho, Bengtsson, Speed, & Irizarry, 2007). By this normalization, probe affinities are adjusted per array while the signal levels are equalized for several arrays.

2.2.6.11.3.2 Probe Summarization

The next step in the CNA-LOH analysis is the extraction of a signal proportional to the CN of each allele, called as probe summarization. In this step, normalized probe level signals are combined into locus level estimates by fitting a log additive or median model of intensities (H. Lu et al., 2011). The three methods used in this study differed in this step by using different algorithms, namely CRMA v2 and ASCRMA v2.

CRMA v2 applies two different summarization methods which is dependent on the chip type used in the experiment. If the array is a GenomeWide SNP (GWS 5 or 6) array, the summarization method is performed by a single-array algorithm (AvgCnPlm) in which a robust average of the signal of the probes for each SNP is calculated. Conversely, if the chip type is 10K, 100K or 250K, CRMA v2 apply a multi-array log additive model (RmaCnPlm) to increase signal-to-noise ratio. These two algorithms make probe summarization for each allele jointly by adding the signals in order to obtain total copy numbers in the next steps. However, in most of the studies, allele specific signals are required to be able to call the regions of LOH, which is essential to identify aberrant regions of the genome inactivating tumor suppressors. For those studies, a subversion of CRMA v2, ASCRMA v2, is employed by only changing the specific argument from TRUE to FALSE (combineAlleles=TRUE > combineAlleles=FALSE). By this way, summarization is performed for each allele separately (Henrik Bengtsson et al., 2009). In our study, since we have used 250K StyI array, we performed RmaCnPlm for both CRMA v2 and ASCRMA v2.

2.2.6.11.3.3 Post-Processing

After probe summarization, post processing is applied to handle signal bias related to size and other properties of sequences where probes are located. There are two steps in post processing: GC Content Normalization (GC) and Fragment Length Normalization (FLN).

Observed intensity of each probe is correlated to its GC content, i.e. increasing with increasing GC content. Some complex relationships with the nucleotide sequences of the probes have been also observed. Moreover, as stated in Section 2.2.6.1 and 2.2.6.3, during array protocol, whole genome is digested with a RE and amplified with PCR after adaptor ligation. The lengths of the fragments vary depending on the enzyme. This variation leads to changes in the efficiency of PCR for each fragment length. Therefore, locus-specific copy number estimates may be correlated with the fragment length. Since the fragments can be identified by the genome annotation, it is easy to estimate and correct for such effects. These two parameters, GC content and FL, vary across assays and between hybridizations; therefore, they should be corrected so as to make comparisons across samples meaningful and more accurate (H. Lu et al., 2011).

For the first and second method, ASCRMA v2 followed by CalMaTe and Tumorboost, GC content correction was not applicable because of the algorithms themselves. Hence, we only performed FLN for these methods, (Henrik Bengtsson et al., 2010). The third method, CRMA v2 followed by VN algorithm, included both corrections (Lisovich et al., 2011).

2.2.6.11.3.4 Raw Copy Number Calculation

In this step of the analysis, raw (full-resolution) copy numbers are obtained, either total or allele specific depending on the algorithm used. In order to calculate copy numbers, a reference should be selected and the three methods used in this study were differed in this reference selection.

The first method, CalMaTe, has an option to use only a subset of samples such as normal samples as reference. After subset selection among all samples, it performs a calibration step to decrease the signal to noise ratio by normalization of tumor ASCNs with normal ASCNs. With this normalization, it increases the power to detect change points of CNs. At the end of analysis, total and ASCNs were calculated for each sample (Henrik Bengtsson et al., 2009).

The second method Tumorboost utilizes paired data set and it chooses the normal match of each tumor sample as reference for the particular tumor. By this way, it normalizes ASCNs of tumors with their matched normals. The most important advantages of Tumorboost analysis of each tumor-normal pair for our study are: (i) each pair can be analyzed immediately without needing reference samples, (ii) each patient can be analyzed at once, this may otherwise be a limiting factor in projects with incomplete patient samples, (iii) the negative effect of normal contamination, i.e. the presence of normal cells in tumor sample, is minimized by normalization of each tumor with its respective normal (Henrik Bengtsson et al., 2010). At the end of analysis, total and ASCNs were calculated for each sample.

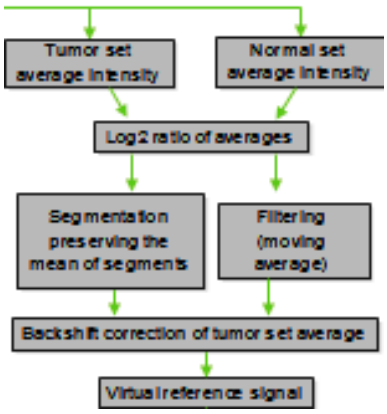


Figure 2.9 Reference selection for VN Algorithm.

The last method utilizes VN algorithm for reference selection (Figure 2.9). The actual purpose of this algorithm is to derive a reliable copy number reference signal specifically from tumor samples in so eliminating the need for a normal reference set. By this way, the noise in the data is also minimized making this method applicable to low quality paired tumor-normal data sets, such as our data, paired FFPE samples. During reference selection by VN algorithm, the first step is taking the average of tumor intensity signals and the average of normal (template) intensity signals separately. The average of tumor signal intensities is used as reference signal at first place. By this way, the high frequency component common in all tumor samples is preserved. When this reference signal is used as denominator in computing the (log2) ratio of CNs instead of the standard reference, it would reduce the noise that high frequency component is leading to. However, if this reference signal is used unmodified, this would suppress the common aberrations of tumor set for each particular tumor sample where they are present, resulting in the false CN aberrations for the remaining tumors. To prevent this, the (log2) ratio between this synthetic

reference signal and normal reference set based signal is computed by treating these two references as a test/normal pair. After that, mean preserving filtering technique to detect the common CN variations in normals is applied. Then, the information about the position and amplitude of the common CN variations is used to modify the tumor based reference signal by backshift correction of tumor set average, resulting in a virtual reference signal. This virtual reference has two key properties: (i) Its high frequency component is well correlated with the one common for all tumors so that the noise in (log2) ratio between tumor raw CN and the reference signal is decreased; (ii) the CN aberrations common for the part or the whole tumor set is not included in the reference signal so that when each tumor sample is compared to the reference signal, the CNA will be detected for this particular tumor (Lisovich et al., 2011). After virtual reference generation, raw total copy numbers are calculated according to this reference at locus level.

The locus level estimates of both TCNs and ASCNs are converted into log2 ratios and graphed after CN calculation. Below, in Figure 2.10, is an example of such graphs. In the first graph, TCN vs genomic position is represented. For an SNP, there are three possible genotypes, namely AA, AB and BB. For a heterozygous SNP, TCN would be 2 by (A, B) = (1, 1). This corresponds 3 segments in Allele B Fraction (BAF) vs genomic position graph. If there is a gain, there would be 4 possible genotypes, AAA, AAB, ABB and BBB. Then, BAF would have 4 possible fractions, as seen in the figure. If there is a CN-LOH, possible genotypes would be AA or BB. In this condition, BAF would be either 0 or 1. From these graphs, we can also observe the CN change points by corresponding genomic position of changes.

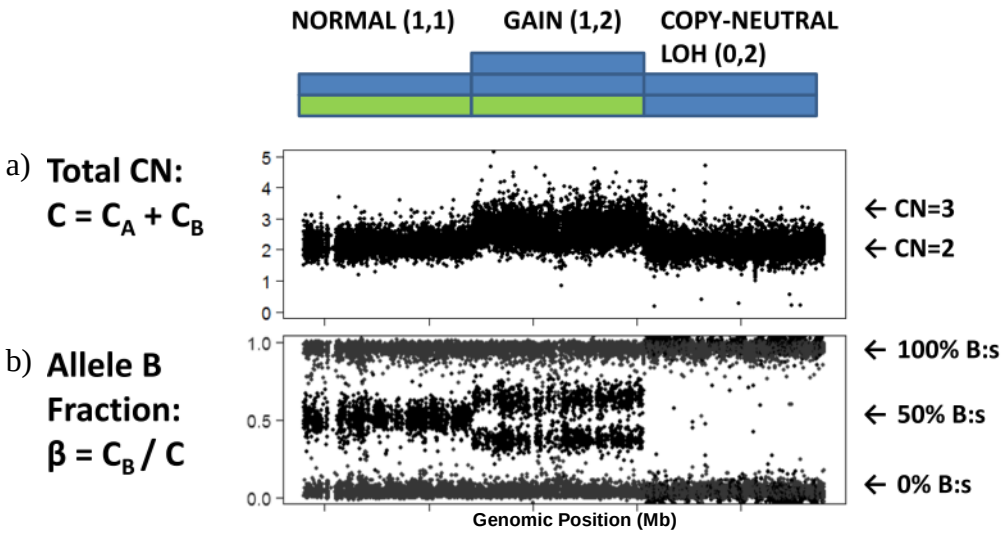


Figure 2.10 Locus-level estimates of a) TCNs and b) ASCNs with corresponding CNSs.

2.2.6.11.3.5 Copy-Number Segmentation

In the previous step, the locus level estimates of copy numbers were obtained by three ways. In this step, the aim is to combine these locus level estimates in order to obtain contiguous regions of genome with the same aberration such as small or large deletions or gains. Region level estimates of CNAs are made by segmentation of total, minor and major copy numbers, or allelic imbalances for LOH analysis (H. Lu et al., 2011).

For the first method, there is a TCN segmentation method called CBS was used. Moreover, a segmentation tool for ASCNs with this method is underdevelopment (Henrik Bengtsson et al., 2011).

The second method we used, ASCRMA v2 followed by Tumorboost, applies Paired parent-specific copy-number segmentation (Paired PSCBS) which performs segmentation using ASCN estimates. It utilizes the original CBS algorithm to detect regions of equal TCN and then uses ASCN information to further segment these regions to identify regions with changes in ASCN. By this way, it can obtain calls for both equal parental copy number and LOH regions (A. B. Olshen et al., 2011).

The third method used in this study utilizes Genome Alteration Detection Analysis (GADA) segmentation to perform TCN segmentation (Pique-Regi, Cáceres, & González, 2010).

CHAPTER 3

RESULTS AND DISCUSSION

3.1 The Patients

In this study, ten patients with metastatic osteosarcoma were included (9 males and 1 female). Characteristics of patients are listed in Table 3.1.

Table 3.1 Clinicopathologic properties of the patients

Patient No.	Age (Yrs)	Gender	Histology	Stage*	Status at follow-up
1	36	Male	Conventional	IIIB	Alive with metastasis
2	26	Male	Conventional	IIIB	Dead
3	30	Male	Conventional	IIIB	Dead
4	30	Male	Conventional	IIIB	Alive with metastasis
5	29	Male	Conventional	IIIB	Alive with metastasis
6	28	Male	Conventional	IIIB	Alive with metastasis
7	24	Male	Conventional	IIIB	Alive with metastasis
8	23	Female	Conventional	IIIB	Alive with metastasis
9	20	Male	Conventional	IIIB	Alive with metastasis
10	34	Male	Conventional	IIIB	Dead

*Staging was performed according to the American Joint Commiteee on Cancer TNM system

All patients had pulmonary metastatic tumors detected by CT scan of chest. The ages of the patients ranged from 23 to 36 years. According to the histopathologic diagnosis, all patients had conventional-type osteosarcoma. FFPE tissue samples were obtained from the patients during lung biopsy.

3.2 DNA Isolation and Optimization

Before genomic DNA isolation, normal and tumor cells were selected by manual microdissection. At least five 10 µm sections per sample were prepared. Following pretreatment with xylene, the interested regions were scraped off from the slides into an eppendorf tube. After DNA isolation, the quantity and quality of isolated DNA was evaluated by spectrophotometric analysis. If the quantity of isolated DNA was less than 50 ng/µl and OD260/280 ratio was not between 1.7 and 2.0, the DNA isolation experiment for the specific sample was repeated by increasing the starting amount of tissue, i.e., with more than five 10 µm sections.

3.3 Quality Assessment of Isolated DNA

3.3.1 Spectrophotometric Measurement and Agarose Gel Electrophoresis

All DNA samples were quantified with NanoDrop 2000c spectrophotometer. The concentrations and OD260/280 ratios of all samples are represented in Table 3.2. All samples processed for downstream analysis had a concentration higher than 50 ng/µl and OD-260/280 ratio between 1.7 and 2.0. From the samples, an average amount of 310 ng/µl DNA was obtained.

Table 3.2 Quantity and Quality Values of DNA Samples

Patient	Tissue Type*	DNA Values			
		OD260	OD280	260/280	ng/µl
1	N	1.08	0.603	1.79	54.0
	T	9.538	4.836	1.97	476.9
2	N	1.9	1.08	1.76	95.0
	T	3.263	1.752	1.86	163.2
3	N	3.096	1.796	1.72	154.8
	T	1.968	1.025	1.92	98.4
4	N	2.348	1.235	1.9	1.4
	T	5.764	2.957	1.95	288.2
5	N	4.205	2.273	1.85	210.2
	T	5.78	2.995	1.93	289.0
6	N	4.773	2.538	1.88	238.6
	T	3.676	1.909	1.93	183.8
7	N	11.452	6	1.91	572.6
	T	13.844	7.236	1.91	692.2

Table 3.2 continued.					
8	N	2.348	1.235	1.9	117.4
	T	30.545	15.824	1.93	1527.3
9	N	2.646	1.486	1.85	124.0
	T	4.923	2.687	1.86	84.5
10	N	4.442	2.428	1.83	222.1
	T	2.601	1.395	1.86	130.0

*** N: Normal; T: Tumor**

The degree of fragmentation of DNA samples was measured by horizontal agarose gel electrophoresis and the photograph is given in Figure 3.1.

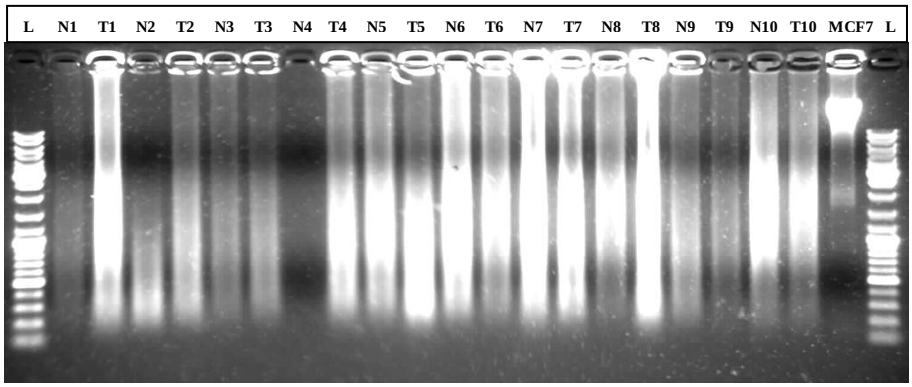


Figure 3.1 DNAs isolated from patient samples and MCF7 breast cancer cell line on 1% agarose gel (N: Normal, T: Tumor, numbers indicate patient number; L:100 bp DNA ladder).

As seen in Figure 3.1, in all lanes, smearing of DNA samples was observed. This was expected because DNA isolated from FFPE samples is generally highly fragmented. However, the samples, namely N1, N2, T2, N3, T3, T9, and N10, gave highly faint smears. This indicates that those samples were much more fragmented than the others or that they had low concentrations of DNA. When DNA concentration values are taken into account, it can be concluded that N1, N2, T3 and T9 gave faint smears because of their low DNA content, 54.0, 95.0, 98.4 and 84.5 ng/µl, respectively. Even though T2, N3 and N10 had high concentrations of DNA (163.2, 154.8 and 222.1 ng/µl, respectively), they also showed pale smears on the gel photo. This is because those samples were so much fragmented that DNA ran out of the gel. Moreover, in lane N4, there was even no smear seen because the concentration of DNA was too low, 1.4 ng/µl. This could also be resulted from too much fragmentation.

3.3.2 Gene Specific PCR for *β-Actin* Gene and RAPD-PCR

Gene specific PCR for *β-Actin* gene was performed to assess the efficiency of DNA samples to amplify a product of a specific size by PCR. The PCR mixture was prepared and amplification was performed in thermal cycler. A non-template control which contains nuclease free water instead of template was also included in all experiments. PCR products were detected by horizontal agarose gel electrophoresis and photograph of the gel is given in Figure 3.2.

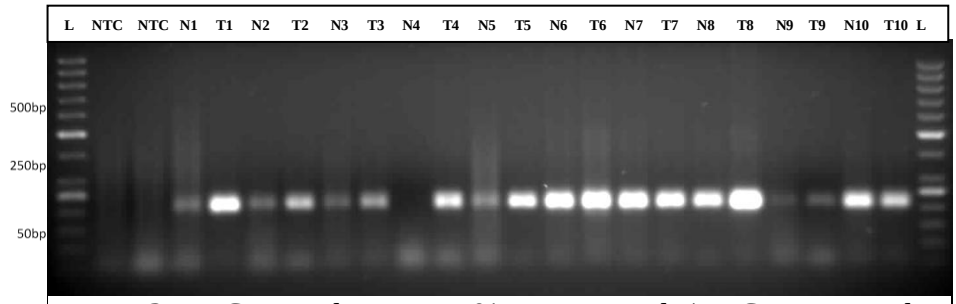


Figure 3.2 PCR products on 2% agarose gel (NTC: No template control, N: Normal, T: Tumor, numbers indicate patient number; L: 50 bp DNA ladder).

Except 4th normal sample (N4), all samples gave product in gene specific PCR for *beta-actin* gene, with changing band intensities in accordance with their concentrations and level of fragmentation as seen in Figure 3.1.

The samples with low DNA content, N1, N2, T3 and T9, showed less intense PCR product bands when compared to the remaining samples. Moreover, the highly fragmented samples, T2, N3 and N10, also gave less intense PCR product bands. Although the samples, N5 and N9, had high concentrations of DNA (210.2 and 124.0 ng/µl, respectively), the PCR product bands were less intense compared to the ones that had similar DNA concentrations. This could also be resulted from more fragmentation of those samples than others.

As a second amplification control, RAPD-PCR was performed to assess the efficiency of DNA samples to amplify products of variable sizes by PCR. The mixture was prepared and reaction was carried out in thermal cycler. PCR

products were detected by horizontal agarose gel electrophoresis and photograph of the gel is given in Figure 3.3.

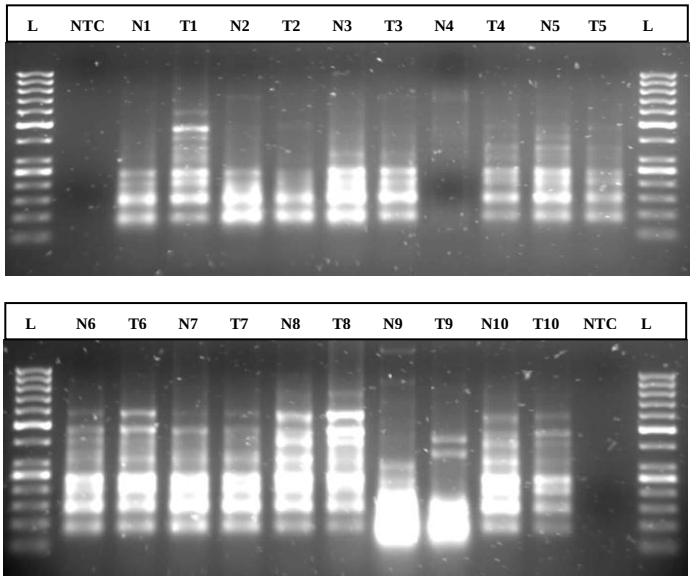


Figure 3.3 PCR products on 2% agarose gel (NTC: No template control, N: Normal, T: Tumor, numbers indicate patient number; L: 50 bp DNA ladder).

Except 4th normal and 9th normal-tumor pair, all samples provided RAPD-PCR expected maximum amplicon sizes ranging between 500 and 800 bp. The sample N4 did not give any product as expected because of very low DNA concentration. The 9th pair gave maximum amplicon size of 250 bp; however, most of the amplicons were between 50 and 150bp. This may be resulted from fragmentation of those DNA samples more than the remaining ones.

Collectively, the results of these quality control experiments suggested that sample N4 should not be used in further experiments. Even though the sample

N9 had sufficient DNA concentration, both PCR controls were not so sufficient, which was also true for the match of it, T9. However, the pair 9 was used in further experiments because on the DNA gel, there were still bands between 300 and 600 bp.

3.4 Optimization of the Microarray Protocol

As stated above in Section 2.2.3, fragmentation caused by preservation procedures is the main problem of DNAs isolated from FFPE samples. Therefore, studying on these samples requires some modifications in the microarray experiment in order to optimize the standard protocol.

In the first stage, Sty I RE digestion, the standard 250K Sty assay failed to produce sufficient PCR product from fragmented DNA samples for array hybridization. Therefore, during this stage, initial DNA amount for each sample was increased by increasing the starting volume of the each DNA sample. In order to get efficient digestion, each component of digestion master mix also increased two times. After RE digestion of the same sample with modified protocol, sufficient digestion product could be obtained. This can be seen in Figure 3.4a and Figure 3.4b.

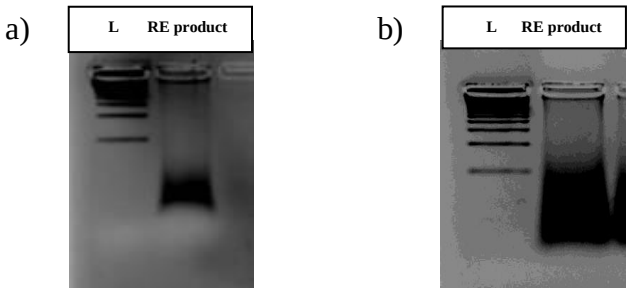


Figure 3.4 Restriction Enzyme digestion products a) Standard protocol b) Modified protocol (L: Ladder)

After RE digestion and ligation of adaptors, PCR was performed by using a generic primer in order to amplify the whole genome simultaneously. In order to get sufficient amount of PCR product for hybridization stage, instead of three, four PCR reactions were performed for each sample. As seen in Figure 3.5, adequate amount of PCR product was obtained for each sample.

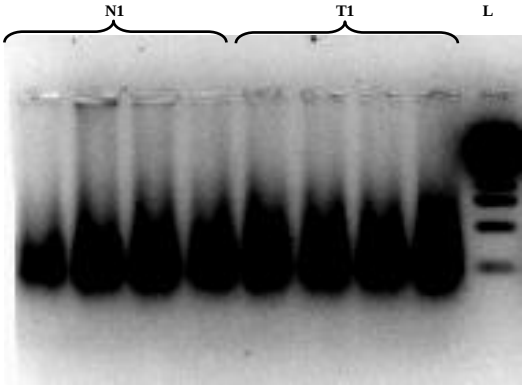


Figure 3.5 PCR amplification gel of adaptor ligated fragments. (N1: Normal 1, T1: Tumor 1, L: Ladder)

After quantitation and normalization of PCR products, fragmentation was performed. DNase I (fragmentation reagent) enzyme was used for random fragmentation of products. A fragmented sample can be seen in Figure 3.6.

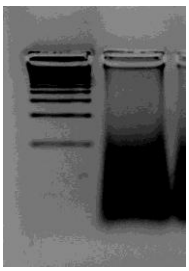


Figure 3.6 Fragmentation product of sample N1. (N1: Normal 1)

After hybridization of samples to chips, scanning was performed. The images were obtained and hybridization quality controls were checked in all images. A sample image can be seen in Figure 3.7.

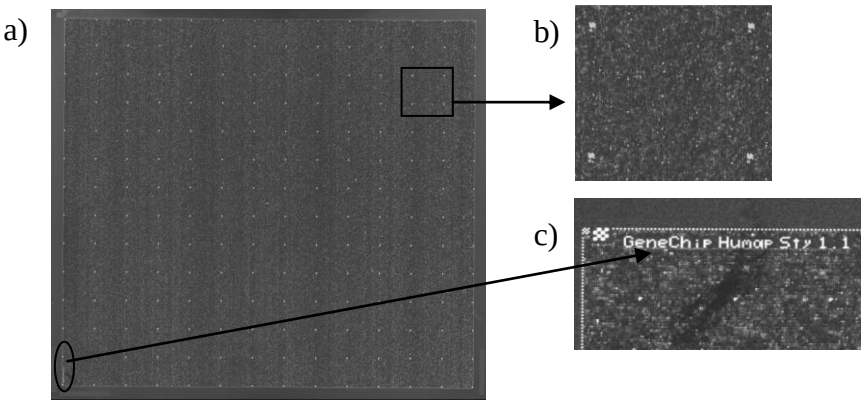


Figure 3.7 Array image and quality controls. a) A whole array image b) Checkerboard pattern throughout the array c) B2 oligonucleotide control at left bottom corner

3.5 Microarray Data Analysis

3.5.1 Preliminary Data Analysis: Genotyping and Genotype Data Quality

The raw data files in .CEL format were generated and transferred into Genotyping Console 4.0 Software. In order to check the quality of CEL files, QC Call rate of each sample were generated (Table 3.4). QC Call Rate Threshold was set to 60 in our study. The samples N9 and T9 had QC call rates less than 60 and; therefore, they were excluded from further analysis. Graphical display of QC metrics across samples was created to identify outliers, including 72 signature SNPs for tracking sample identity (Figure 3.8).

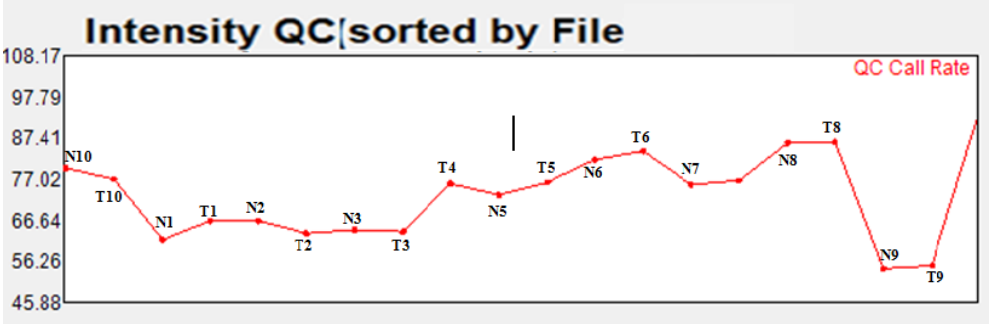


Figure 3.8 Graphical display of QC metrics

Table 3.3 QC and SNP call rates of each sample

P#	Sample name	QC Call Rate	SNP Call Rate
1	N1	62.08	74.32
2	T1	66.86	76.32
3	N2	66.83	74.70
4	T2	63.68	74,26
5	N3	64.63	74.15
6	T3	64.16	74.20
7	T4	76.31	77.67
8	N5	73.44	76.14
9	T5	76.58	77.26
10	N6	82.35	82.03
11	T6	84.51	82.50
12	N7	76.02	77.40
13	T7	77.10	78.35
14	N8	86.56	86.47
15	T8	86.80	84.91
16	N9	54.78	70.81
17	T9	55.58	69.85
18	N10	80.22	83.40
19	T10	77.40	79.18

Genotyping analysis was performed with the samples having higher QC call rate than threshold 60. In genotyping analysis, genotype files in CHP format were generated using BRLMM algorithm to be used in whole-genome association analysis. Genotyping quality control was done by calculation of SNP call rate which is the percentage of SNPs having successful signal. SNP call rates and overall performance of mapping arrays are stated in Table 3.3 and Table 3.4, respectively.

Table 3.4 Performance of 250K Sty Mapping Assay

Average PCR yield (ng/µl ± SD)	2472.60 ± 394.90
SNP average call rate (% ± SD)	78.40 ± 4.02
SNP average call rate in tumor (% ± SD)	78.30 ± 3.55
SNP average call rate in normal (% ± SD)	78.50 ± 4.75
	9 patients, n=17

3.5.2 Genome-Wide Association Analysis: Disease, Gene and Pathway Association

3.5.2.1 Data Preprocessing and Cleaning

During GWAS analysis, the first step included data preprocessing which was performed to remove data points which are under specified thresholds. In this step, PLINK based quality control filtering and imputation via BEAGLE (B. L. Browning & Browning, 2007) was performed by METU-SNP (Ustünkar & Aydın Son, 2011).

Before frequency and genotyping pruning, there were 235,268 SNPs studied on 9 cases and 8 controls of which 15 were males and 2 were females. The distribution of these SNPs across chromosomes is shown in Figure 3.9.

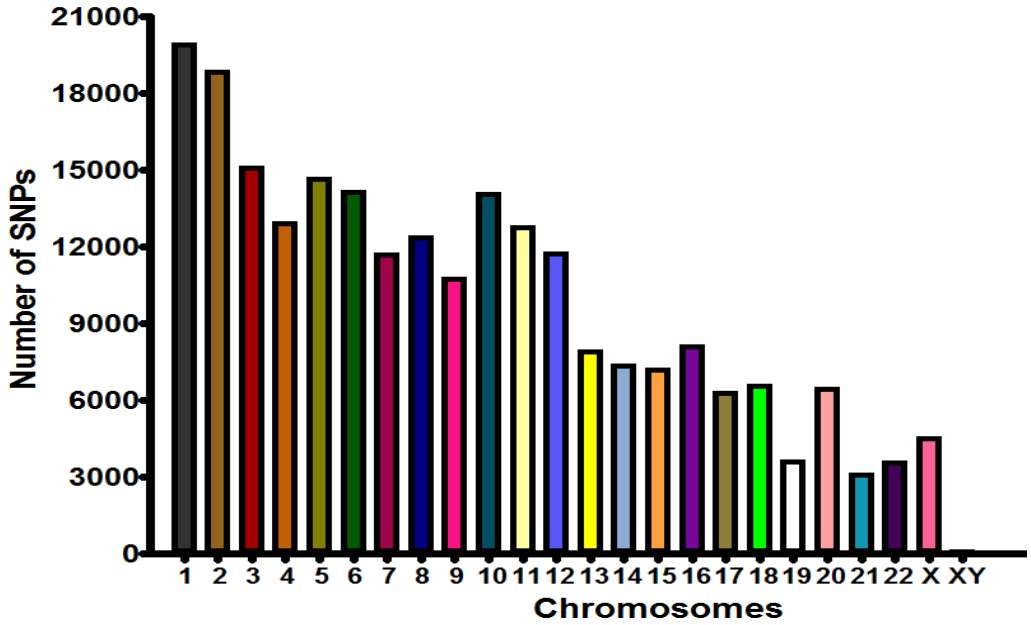


Figure 3.9 Number of SNPs per chromosome

The summary statistics of the raw data (before preprocessing and cleaning) were given in the following graphs. In Figure 3.10, MAF distribution of SNPs was shown. Missingness (genotyping) rate by SNP for all SNPs was drawn as a graph of proportion of individuals which is missing for respective SNPs vs number of missing SNPs (Figure 3.11). Missingness rate by Individual for all samples was shown as proportion of missing SNPs for respective samples vs number of missing individuals.

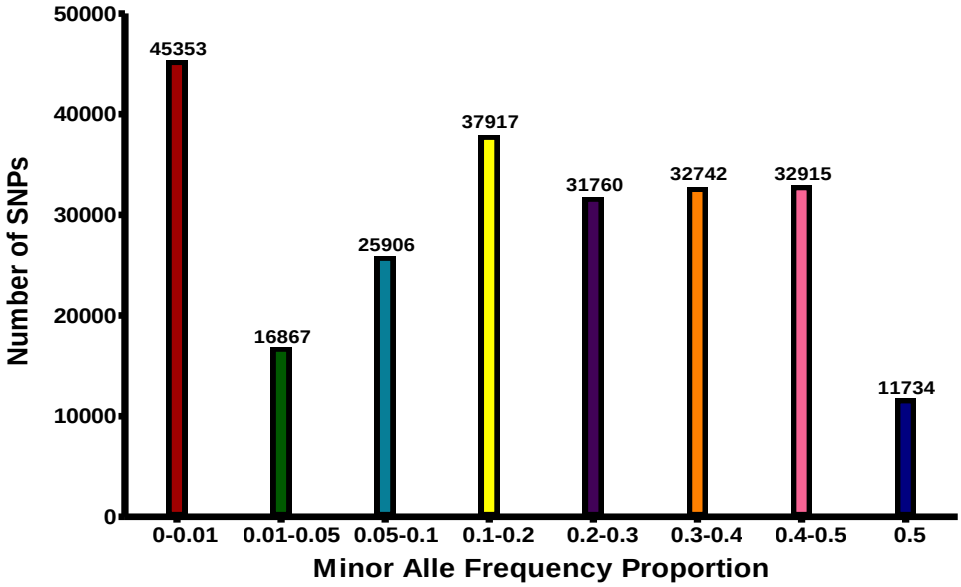


Figure 3.10 MAF distribution of SNPs as a proportion vs number of SNPs. The values on the bars represent the exact number of SNPs for each proportion range.

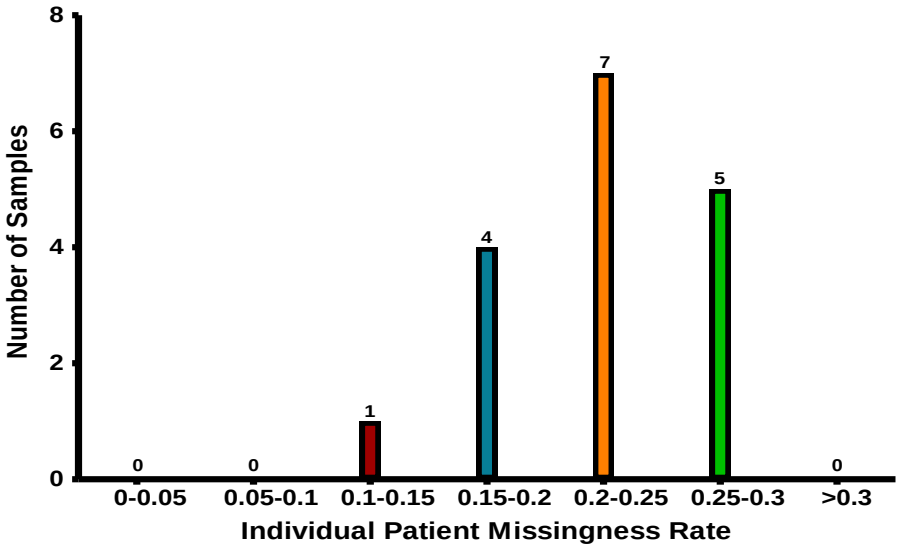


Figure 3.11 Proportion of missing SNPs for respective samples vs number of samples. The values on the bars represent the exact number of missing samples for each proportion range.

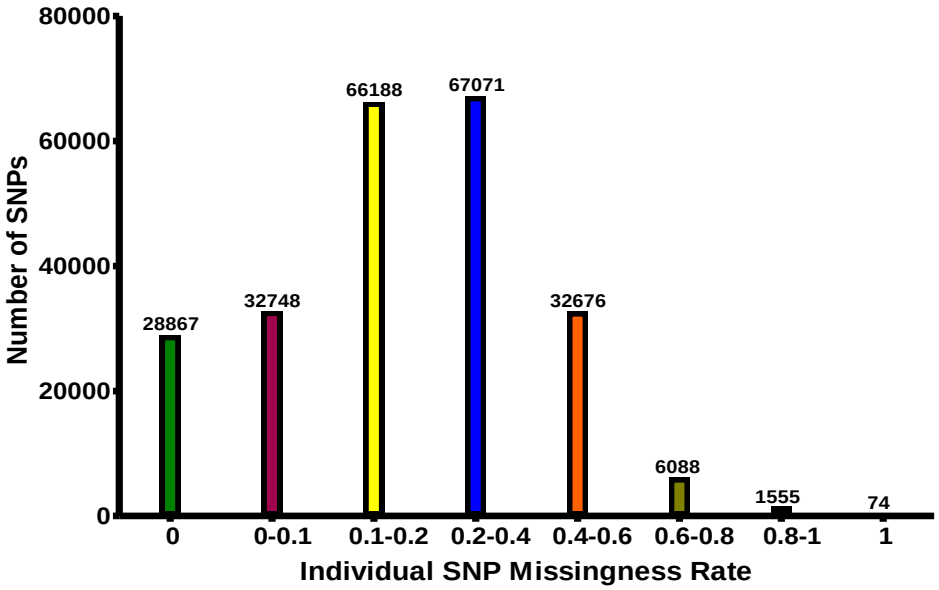


Figure 3.12 Proportion of individuals which is missing for respective SNPs vs number of SNPs. The values on the bars represent the exact number of missing SNPs for each proportion range.

In order to filter out SNPs and/individuals with low quality data, PLINK has defined QC thresholds to work on. When we studied with these default thresholds, all individuals were removed from further analysis. This was mainly because of the data quality of our samples. According to default MAF value, SNPs having MAF ≤ 0.05 should be removed from our data, and this means that 62,220 SNPs should be uninvolved (26% of all) in further analysis (Figure 3.10). With the default value, we would lose 26% of our data points. For this reason, MAF threshold was decreased to 0.01 and this value was imputed during BEAGLE imputation analysis.

According to default individual missingness rate (mind > 0.1), we had to exclude samples which have more than 10% missing genotypes. For our data, with this default mind > 0.1, 17 of 17 samples had to be removed from data set and therefore no individuals left for analysis. As seen in the graph (Figure 3.11), there are no individuals in missingness rate ranges 0-0.05 and 0.05-0.1, and only after the 0.3 mind threshold, all individuals are included. Therefore, we set our mind threshold to 0.3, which was imputed during BEAGLE imputation analysis.

.

According to default SNP missingness rate (geno >0.1), we had to exclude SNPs which are missing in 10% and more than 10% of the samples. For our data, with this default geno > 0.1, 173,653 SNPs had to be removed from data set (Figure 3.12). In order to lower this value, we set our geno threshold to 0.4 which was imputed during BEAGLE imputation analysis.

In order to handle with the missingness we applied imputation via BEAGLE. During the imputation process 74 SNPs with allelic r^2 frequency less than 0.75 have been removed and 235,194 SNPs remained available for the downstream analysis. Among these SNPs, no SNPs are excluded based on Hardy Weinberg Equilibrium threshold ($p < 0.001$). Additionally, 45,353 SNPs were removed due to failure to conform with the frequency test (MAF ≤ 0.01). No individuals are removed considering mind > 0.3. Furthermore, 40,394 SNPs were removed due to failure in genotyping (geno > 0.4). At the end of the preprocessing step, 149,447 SNPs from our genotyping data are selected for subsequent analysis.

The output of the data-preprocessing and cleaning step was PLINK based binary data file which contained the imputed genotype data satisfying QC thresholds. This file was used in association analysis in the next step.

3.5.2.2 GWAS: SNP, Gene and Pathway Association

In the second step of GWAS analysis, the preprocessed data was used for GWAS. The uncorrected *p*-values was selected in statistical analysis. The threshold *p*-value was specified as 0.05 to measure statistical significance of individual SNPs. For combined *p*-value threshold for genes and pathways, the same threshold 0.05 was used. In this study, 358 significantly associated SNPs were obtained. The distribution of these associated SNPs across all chromosomes is shown in Figure 3.13.

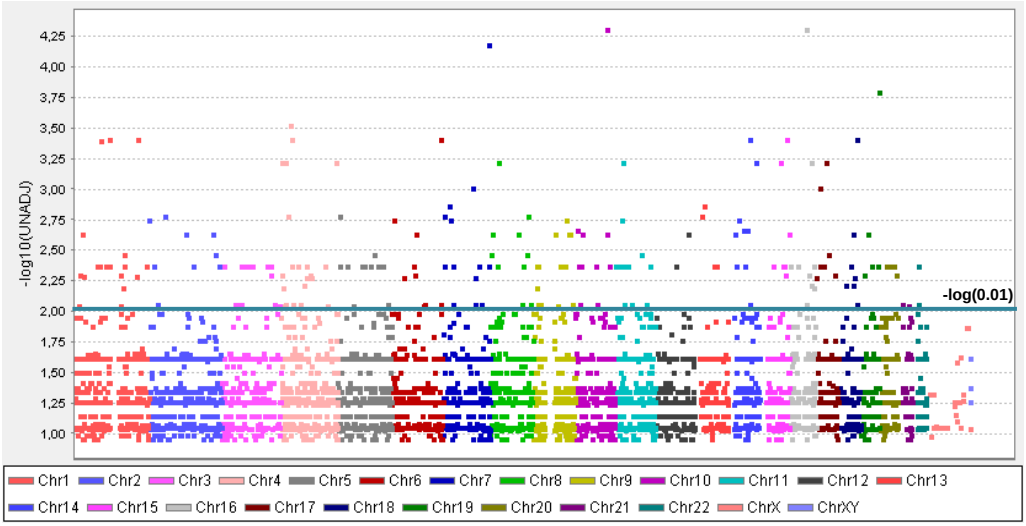


Figure 3.13 Manhattan Plot of negative logarithm of unadjusted *p*-value of association for individual SNPs and their distribution on individual chromosomes.

Top 100 of the significantly associated SNPs are given in Appendix G. Table 3.5 shows the top 20 SNPs which were found to be significantly associated with metastatic OS. The SNPs with smallest *p*-values are rs6499861 (p-value = 4.90E-05), rs10884554 (p-value = 4.90E-05) and rs12154602 (p-value = 6.55E-05).

Table 3.5 Individual SNP *p*-values of association of GWAS.

Rank	Chromosome	SNP ID	Unadjusted *p*-value
1	16	rs6499861	4.90E-05
2	10	rs10884554	4.90E-05
3	7	rs12154602	6.55E-05
4	19	rs7249598	0.000158
5	4	rs28716003	0.0002965
6	15	rs16946687	0.0003873
7	14	rs17120272	0.0003873
8	4	rs6828250	0.0003873
9	1	rs6657368	0.0003873
10	1	rs10923183	0.0003873
11	6	rs16896159	0.0003873
12	18	rs17769881	0.0003873
13	1	rs5744297	0.0003961
14	4	rs3899794	0.0006022
15	14	rs17753747	0.0006022
16	16	rs17313499	0.0006022
17	11	rs2958625	0.0006022
18	15	rs12594008	0.0006022
19	8	rs2614062	0.0006022
20	17	rs9895307	0.0006022

The first SNP in the p-value ranking is rs6499861, which maps to chromosome 16q: 56991496 bp. According to dbSNP database 2011, it is not on a known locus, transcript or coding region. Moreover, it is stated that it has no clinical significance archived in SNP databases and to our knowledge this is the first study that this specific SNP has been found to be associated with a disease, particularly metastatic OS.

The second most significant SNP is rs10884554, which maps to chromosome 10q: 109623417 bp. According to dbSNP database 2011, it is not on a known locus, transcript or coding region. Also, it has no clinical significance indicated. For this reason, this is the first time that this specific SNP has been found to be associated with a disease, metastatic OS.

The last SNP to be mentioned is rs12154602, which maps to chromosome 7q: 154051446. According to dbSNP database, it is found in an intronic region. The gene that covers this intronic region is *DPP6* (dipeptidyl-peptidase 6) which encodes a single-pass type II membrane protein of the S9B family in clan SC of the serine proteases. There are alternate transcriptional splice variants, encoding different isoforms. There is no study involving evidence for an association of this gene to metastatic OS. Therefore, the present study could be the first that reveals such association.

Furthermore, of the top 20 associated SNPs, six were found to map to a gene which are rs12154602, rs17120272, rs6657368, rs10923183, rs5744297, and rs125594008. All of them are found in the intronic regions of the genes so they can be defined as intronic variants.

Following initial GWAS analysis, the calculated *p*-values of individual SNPs were used to perform a second-wave analysis by using combined *p*-values for genes and pathways. In Tables 3.6 and 3.7, top 20 statistically significant genes and pathways ($p < 0.05$) found to be associated with metastatic OS are shown along with their respective *p*-values calculated by using Fisher's combination

test for genes and Hypergeometric test for pathways before AHP prioritization. Top 100 of those genes and pathways are given in Appendix G.

Table 3.6 Top 20 significantly enriched genes according to combined *p*-value.

Entrez Gene ID	Full name	Location	P value
5552	Serglycin	10q22.1	2.04E-04
4487	Msh Homeobox 1	4p16.3-p16.1	6.02E-04
10248	Processing Of Precursor 7 Ribonuclease P/MRP Subunit (S. Cerevisiae)	7q22	9.72E-04
8395	Phosphatidylinositol-4-Phosphate 5-Kinase Type I	9q13	0.001164
6310	Ataxin 1	6p23	0.001506
5836	Phosphorylase Glycogen Liver	14q21-q22	0.002087
401124	Death Domain Containing	4p14	0.002141
55676	Solute Carrier Family 30 (Zinc Transporter) Member 6	2p22.3	0.002280
8110	D4 Zinc And Double PHD Fingers Family 3	14q24.3-q31.1	0.002619
93649	Myocardin	17p11.2	0.003978
729767	Carcinoembryonic Antigen-Related Cell Adhesion Molecule 18	19q13.41	0.004197
25791	Neuronal Guanine Nucleotide Exchange Factor	2q37	0.004270
400954	Echinoderm Microtubule Associated Protein Like 6	2p16.2-p16.1	0.004291
128272	Rho Guanine Nucleotide Exchange Factor (GEF) 19	1p36.13	0.004974
65055	Receptor Accessory Protein 1	2p11.2	0.005249
1179	Chloride Channel Accessory 1	1p31-p22	0.005948
220136	Coiled-Coil Domain Containing 11	18q21.1	0.006051
491	Atpase Ca++ Transporting Plasma Membrane 2	3p25.3	0.006162
5738	Prostaglandin F2 Receptor Negative Regulator	1p13.1	0.006260
57537	Sortilin-Related VPS10 Domain Containing Receptor 2	4p16.1	0.006554

Among the genes which are found to be associated with pulmonary metastasis of OS in this study (Table 3.6), serglycin (*SRGN*) and, D4 Zinc and Double PHD Fingers Family 3 (*DPF3*) genes have been indicated in the literature that

they have association with metastatic nature of cancer cells (Hurnphriessb, Nicodemusoll, Schillerll, & Stevensslll, 1992; X.-J. Li et al., 2011; Hoyal et al., 2005). Li et al. (2011) stated that *SRGN* have high expression in highly metastatic cells of lungs. Hoyal et al. (2005) indicated that SNPs found in the upstream region of the gene region of *DPF3* have association with increased risk of breast cancer development and lymph node metastases.

In addition to *SRGN* and *DPF3*, 4 other genes, which are found to be associated with OS in this study, have characteristics that associate them with cancer as stated in the literature (Xue et al., 2010; VanHouten et al., 2010; White, Varley, & Heighway, 1998). Those genes are Phosphatidylinositol-4-Phosphate 5-Kinase Type I (*PPI4P5K1*), Death Domain Containing 1 (*DDC1*), Echinoderm Microtubule Associated Protein Like 6 (*EML6*), and ATPase Ca++ Transporting Plasma Membrane 2 (*ATP2B2*).

Apart from these specific associations, the expression of those genes in lung tissue can be an indication of an association of those genes to the disease under study. When we look at the top 20 genes, *SRGN*, Msh Homeobox 1 (*MSX1*), and *PPI4P5K1* have been shown to be highly expressed in lung tissue (Frézal, 1998).

From the pathway analysis of our data, it was clearly observed that metabolic pathways such as glycolysis and amino acid degradation were found to be associated with pulmonary metastasis of OS (Table 3.7). This observation is consistent with the metastatic nature of tumors because tumor cells shift their global metabolic programs in favor of tumor invasion, progression, and

metastasis (Y. Hua et al., 2011). Moreover, Sootnik *et al.* (2011) stated that in metastatic tumors of OS, downregulation of key oxidative phosphorylation (OXPHOS) genes occurs in favor of glycolysis (Sottnik, Lori, Rose, & Thamm, 2011). Another study conducted by Y. Hua et al. (2011) indicated that the serum metabolic profile of lung metastasis of OS in mice showed decreased carbohydrate and amino acid metabolism products, but elevated lipid metabolism byproducts. Moreover, from a metabonomic study by Zhang *et al*. (2010), it was observed that the energy metabolism of OS patients was disrupted with the evidence of significantly downregulated TCA cycle and glycolysis, up-regulated lipid metabolism, dysregulated sugar levels, and down-regulated amino acid metabolism (Zhiyu Zhang et al., 2010). In order to prove the functional association of metabolic pathways found in this study to metastatic OS, gene expression, protein expression and metabonomic studies on OS patients should be performed.

Apart from metabolic pathways, there are some other pathways indicating the primary osteo nature of the pulmonary metastatic tumor. Those pathways include cation transport, and skeletal development. Cation transport includes the directed movement of cations, atoms or small molecules with a net positive charge, into, out of or within a cell, or between cells, by means of transporters or pores. In cancer, those transporters are indicated to have role in resistance to chemotherapy in previous studies (Filipski, Mathijssen, Mikkelsen, Schinkel, & Sparreboom, 2009; Yokoo et al., 2008). Moreover, the pathways that are effective in skeletal development are found to be primarily affected in conventional OS (Y. Cai et al., 2010). This could be resulted from the fact that OS generally occurs in the immature skeleton (Akiyama, Dass, & Choong, 2008). In order to investigate these pathways in depth, the genes which were found to be significant in gene association analysis were extracted and the

SNPs in those genes have been found out. These genes and corresponding SNPs are listed in Table 3.8.

Table 3.7 Top 20 significant pathways according to combined *p*-value.

Pathway ID	Pathway System	Pathway Title	Gene Count	P-Value
GO:0004758	GO Function	Serine C-Palmitoyltransferase Activity	2	2.35E-04
GO:0006812	GO Process	Cation Transport	43	4.43E-04
GO:0048103	GO Process	Somatic Stem Cell Division	3	6.95E-04
GO:0016787	GO Function	Hydrolase Activity Acting on Acid Anhydrides	224	0.001368
ASPARTATE-DEG1-PWY	BioCyc	Aspartate Degradation I	4	0.001368
LCTACACAT-PWY	BioCyc	Lactate Oxidation	4	0.001368
P41-PWY	BioCyc	Pyruvate Fermentation to Acetate and Lactate I	4	0.001368
hsa00272	KEGG	Cysteine Metabolism	17	0.001966
ASPARAGINE-DEG1-PWY	BioCyc	Asparagine Degradation I	5	0.002246
GO:0004459	GO Function	L-Lactate Dehydrogenase Activity	5	0.002246
GO:0019642	GO Process	Anaerobic Glycolysis	5	0.002246
GO:0005089	GO Function	Rho Guanyl-Nucleotide Exchange Factor Activity	66	0.002902
GO:0035023	GO Process	Regulation of Rho Protein Signal Transduction	68	0.003284
GO:0005436	GO Function	Sodium:Phosphate Symporter Activity	6	0.003318
GO:0006100	GO Process	Tricarboxylic Acid Cycle Intermediate Metabolic Process	6	0.003318
wiki_37	WikiPathways	Glycolysis and Gluconeogenesis	43	0.003700
GO:0016020	GO Component	Membrane	253	0.003940
GO:0006032	GO Process	Chitin Catabolic Process	7	0.004575
GO:0016740	GO Function	Transferase Activity Transferring Nitrogenous Groups	220	0.004585
GO:0001501	GO Process	Skeletal Development	106	0.004628

Table 3.8 Significantly associated genes and corresponding SNPs in cation transport and skeletal development pathways.

Pathway	Entrez Gene ID	Full Name	SNP ID
Cation Trasnport	6335	Sodium Channel, Voltage-Gated, Type IX, Alpha Subunit	rs6432894
	162514	Transient Receptor Potential Cation Channel, Subfamily V, Member 3	rs12453105
	539	ATP Synthase, H+ Transporting, Mitochondrial F1 Complex, O Subunit	rs2834295
	10050	Solute Carrier Family 17 (Sodium Phosphate), Member 4	rs3799341
	1179	Chloride Channel Accessory 1	rs5744297
	818	Calcium/Calmodulin-Dependent Protein Kinase II Gamma	rs2675671
			rs7080350
	169026	Solute Carrier Family 30 (Zinc Transporter), Member 8	rs10505311
	3749	Potassium Voltage-Gated Channel, Shaw-Related Subfamily, Member 4	rs11578913
	55240	STEAP Family Member 3	rs838072
	9914	ATPase, Ca++ Transporting, Type 2C, Member 2	rs16973771
			rs247811
			rs2326254
	344905	ATPase Type 13A5	rs144194313
	6582	Solute Carrier Family 22 (Organic Cation Transporter), Member 2	rs316000
	10246	Solute Carrier Family 17 (Sodium Phosphate), Member 2	rs2071297
	491	ATPase, Ca++ Transporting, Plasma Membrane 2	rs1318819
			rs12495210
	55676	Solute Carrier Family 30 (Zinc Transporter), Member 6	rs17011863
			rs7580118
Skeletal Development	9061	3'-Phosphoadenosine 5'-Phosphosulfate Synthase 1	rs11945089
			rs3805347
	2121	Ellis Van Creveld Syndrome	rs10516176
			rs10804967
	3487	Insulin-Like Growth Factor Binding Protein 4	rs584828
	5599	Mitogen-Activated Protein Kinase 8	rs9888128
	4487	Msh Homeobox 1	rs3775261
	93649	Myocardin	rs7225754
			rs2052003
	4617	Myogenic Factor 5	rs17007214
	4883	Natriuretic Peptide Receptor C/Guanylate Cyclase C (Atrionatriuretic Peptide Receptor C)	rs976576

The cation transport genes listed in Table 3.8 mainly encodes for cation transporters which are critical for the elimination of many endogenous cations as well as various drugs and toxins. Those transporters are functionally adapted to physiological needs of specific cells. These properties of these transporters make them important in cancer, especially for chemotherapy applications (Filipski et al., 2009; Yokoo et al., 2008). Moreover, some of the transporters in the list have role in cell proliferation pathways, and those are *ATP2C2*, *STEAP3,* and *KCNC4* which have been shown to be associated with tumorigenesis in some type of cancers (Feng et al., 2010; Isobe et al., 2011; Miguel-Velado et al., 2010). However, none of the genes in the list have been associated with metastatic OS. For this reason, this study could be the first that reveals such association.

The genes that were found to be associated with skeletal development pathways in this study are listed in Table 3.8. Among those genes, 3'-Phosphoadenosine 5'-Phosphosulfate Synthase 1 (*PAPSS1*) is found to show high expression in OS cell lines in previous studies. Moreover, Myocardin (*MYOCD*) is also shown to be expressed highly in lung carcinoma and fibroblatic subype of OS. *MSX1* has also been shown to be highly expressed in lung tissue (Frézal, 1998). These expression studies might be an indication of an association of those genes to the disease under study.

Among the genes of skeletal development, Insulin-like growth factor (IGF)-binding protein-4 (*IGFBP4*) is found to be a critical component of bone cell physiology. It was shown that tumorigenic osteoblast cells have decreased amount of IGFBP4 secretion (Durham, Riggs, Harris, & Conover, 1995). Moreover, in previous studies it was indicated that decreased *IGFBP4*

expression could be a step in progression from primary to metastatic cancer in renal cell carcinoma and melanoma because *IGFBP4* may have a role in increasing necrosis and apoptosis, but in decreasing mitosis (Durai et al., 2007; Ueno et al., 2011; J. Z. Yu et al., 2008). However, there is no study involving evidence for an association of this gene to metastatic OS. Therefore, the present study could be the first that reveals such association.

Another gene found to be associated with metastatic OS in this study is Mitogen Activated Protein Kinase-8 (*MAPK8* or *JNK1*). Likewise the other MAP kinases, this kinase has role in a variety of cellular processes such as proliferation, differentiation, transcription regulation and development. Moreover, in literature it was stated that MAPK8 is involved in TNF-α induced and UV radiation induced apoptosis (Matsuguchi, 2009). Furthermore, MAPK signaling pathway has been found to have role in the malignant transformation of osteoblasts and progression of human osteosarcomas (Papachristou, Batistatou, Sykiotis, Varakis, & Papavassiliou, 2003). These findings and the association found in this study indicates that *MAPK8* gene might be involved in metastatic OS; however, expression and functional studies are needed to prove this suggestion.

Myogenic Factor 5 (*MYF5*) gene encodes for a muscle specific regulatory factor. In a study conducted by Percinel *et al*. (2008) it was stated that such regulatory factors are capable of inducing extraskeletal osteosarcoma during local recurrence and lung metastases (Percinel et al., 2008). Therefore, the association of this gene found in this study might be an indication of this capability of Myf5 protein; however, this needs further analysis to be proved functionally.

The last gene belonging to skeletal development pathways was Natriuretic Peptide Receptor C/Guanylate Cyclase C (*NPR3*) encodes a natriuretic peptide receptor. Natriutetic peptides have roles in regulating blood volume and pressure, pulmonary hypertension, cardiac function, and long bone growth. The product of this gene have been found to be important in determining human body height (Estrada et al., 2009). However, there is no study investigating this gene and its mutations in OS. Therefore, this research might be the first that present an association of this gene with metastatic OS.

3.5.2.3 SNP Prioritization

Following GWAS, the AHP based SNP prioritization approach was performed. As mentioned in Section 2, AHP based prioritization algorithm utilize both statistical and genetic information. Using both information, AHP scores were calculated only for the SNPs that meets the individual SNP *p*-value threshold (less than the user specified *p*-value, 0.05 in our study). After *p*-value filtering, 2050 SNPs were left for our data. Then, those SNPs were ranked according to their AHP scores which should be greater than zero. The distribution of top 100 SNPs across chromosomes is given in Figure 3.14. Top 20 SNPs according to their AHP scores are listed in Table 3.9. Top 100 prioritized SNPs are given in Appendix G.

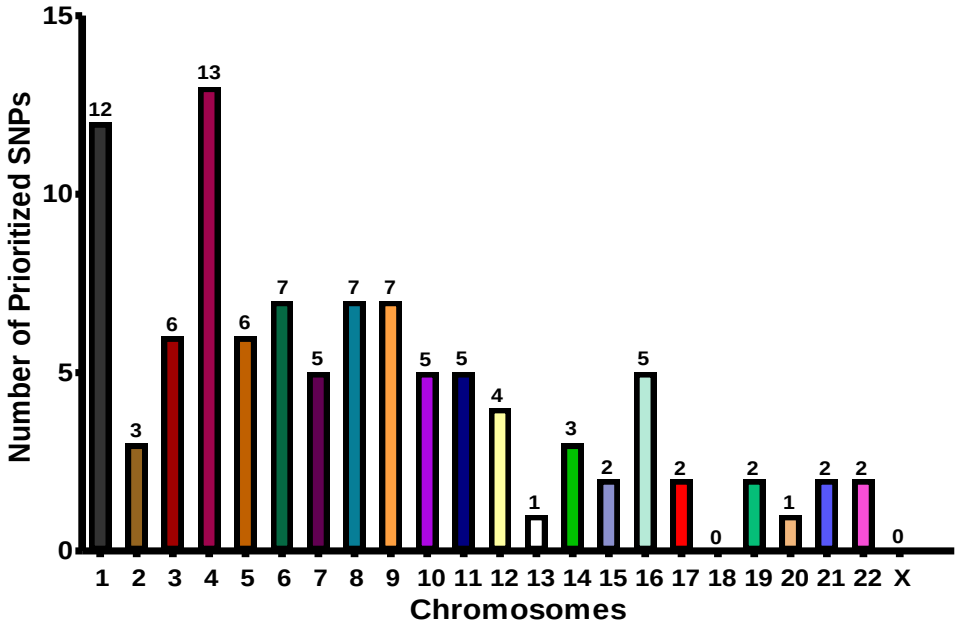


Figure 3.14 Chromosomal distribution of prioritized SNPs.

From the chromosomal distribution of the prioritized SNPs it can be seen that chromosomes 1 and 4 accommodated most of the associated SNPs in contrast to other chromosomes. Numerical and structural chromosomal aberrations of chromosome 1 in OS was previously shown by Murata *et al*. (Murata et al., 1998). Similarly, chromosome 4 was shown to have numerical aberrations in highly malignant OS cases (Mandahl, 1986). These could be the reason why chromosome 1 and 4 showed association for OS in this study. Moreover, a genome-wide copy number aberration analysis of the SNP array data from the same cases was also performed in this study and its results are represented in the next chapter. In that analysis, CN changes in chromosome 1 and 4 were also observed.

Table 3.9 Top 20 SNPs according AHP prioritization of GWAS.

Rank	ID	Chromosome	BasePair	Ahp Score	P-value
1	1159506	6	81098304	0.52954	0.0239
2	2675671	10	75302766	0.52872	0.0873
3	2307349	7	99528943	0.52872	0.0239
4	248381	5	78372981	0.50678	0.0129
5	17675094	16	81506589	0.50678	0.0214
6	10897271	11	61943762	0.50678	0.0239
7	4621357	3	65828746	0.50678	0.0239
8	198430	11	61244266	0.50596	0.0239
9	971074	4	100560884	0.50537	0.0413
10	976576	5	32758256	0.49957	0.00419
11	12495210	3	10449080	0.49170	0.00419
12	1264882	1	111764467	0.49170	0.0239
13	3816052	19	46465637	0.49170	0.0447
14	584828	17	35852756	0.4908	0.0239
15	10115703	9	14712616	0.47011	0.049
16	4919986	21	42663562	0.46900	0.049
17	3742673	14	89805460	0.46894	0.0109
18	3763607	9	70679482	0.46894	0.0196
19	1029665	17	10304916	0.46894	0.0239
20	2148630	1	82188808	0.46894	0.0447

Among the top 20 prioritized SNPs, 19 are found to map genes. However, only rs17675094 and rs2148630 were found to map to genes, cadherin 13 (CDHL13) and latrophilin 2 (LPHN2), which are associated with lung cancer. Both of them are found in the intronic regions of the genes so they can be defined as intronic variants. Table 3.9 shows the association of prioritized SNPs with genes. The METU-SNP prioritization also allows us to obtain disease ontology results from GeneRIF (Gene Reference into Function). A GeneRIF is a summary of annotation to a gene in the NCBI database, containing gene specific information including disease associations. With this function of

METU-SNP, disease annotations of the genes, which were found to be related with the prioritized SNPs found in this study, could also be obtained. These annotations are represented in Table 3.10 by DO id (disease ontology id), disease name and GeneRIF columns.

Table 3.10 Association of Prioritized SNPs to Genes and Diseases in the Databases

Gene ID	Entrez Symbol	DO ID	Disease Name	Gene Reference into Function (RIF)
4176	MCM7	162	Cancer	Increase of the MCM7 is associated with tumor aggressiveness in astrocytoma.
4680	CEACAM6	162	Cancer	For all tumors the amount of CEACAM6 expressed was greater than that of CEACAM5 and reflected tumor histotype.
3559	IL2RA	162	Cancer	The capacity of a CD25-directed immunotoxin to selectively mediate a transient partial reduction in circulating and tumor-infiltrating T regulatory cells in vivo.
4325	MMP16	162	Cancer	Chondroitin-4-sulfate which is expressed on tumour cell surface can function to bind to pro-MMP-2 and facilitate its activation by MT3-MMP-expressing tumour cells to enhance invasion and metastasis.
444	ASPH	162	Cancer	Study demonstrates that high levels of humbug immunoreactivity in colon carcinomas correlate with histologic grade and tumor behavior suggesting that humbug can serve as a prognostic biomarker.
9313	MMP20	162	Cancer	Enamelysin and collagen XVIII were co-localized in the developing enamel matrix and stratum intermedium and in the enamel-like tumor matrix of odontogenic tumors.
79679	VTCN1	162	Cancer	B7-H4 is overexpressed in hyperplastic and malignant endometrial epithelium and is correlated with the T cell number associated with the tumor B7-H4 overexpression may reflect a aggressive biologic potential and play a role in tumor immune surveillance
131	ADH7	462	Malignant Neoplasms	Single Nucleotide Polymorphism in ADH7 gene is associated with upper aerodigestive cancer.

Table 3.10 continued.

3487	IGFBP4	462	Malignant Neoplasms	Overexpression of IGFBP-4 in vivo has been reported to decrease the growth of prostate cancer and altered expression of IGFBP-4 in vivo in colon and other cancers needs to be explored as locally available IGFs appear to stimulate mitogenesis.
5330	PLCB2	462	Malignant Neoplasms	Data indicate that PLC-beta2 expression correlates highly with breast cancer malignancy and suggest that it can be included, as an independent marker among the prognostic indicators in current use.
780	DDR1	462	Malignant Neoplasms	Altered expression of DDR1 may contribute to malignant progression of non-small cell lung carcinoma
818	CAMK2G	1240	Leukemia	CaMKIIgamma is a critical regulator of multiple signaling networks regulating the proliferation of myeloid leukemia cells.
11126	CD160	1240	Leukemia	CD94-expressing cells with cytolytic activity against the recipient's leukemic and tumor cells without enhancement of alloresponse might be able to be expanded from donor G-PBMCs
5016	OVGP1	2871	Endometrial Carcinoma	Gain of oviduct-specific glycoprotein is associated with the development of endometrial hyperplasia and endometrial cancer
491	ATP2B2	4241	Malignant neoplasm of breast	PMCA2 mRNA can be highly overexpressed in some breast cancer cells.
64699	TMPRSS3	6741	Bilateral Breast Cancer	The identification of two novel pathogenic TMPRSS3 mutations (c.646C-->T - R216C)

According to Table 3.10, the prioritized SNPs found in this study have associations with genes which have been found to be associated with different types of cancer in previous studies. This indicates that those genes might have connections with metastatic OS phenotype as well. However, mutation analysis and functional studies are required in order to reveal significant associations of those genes with metastatic OS.

With the help of AHP based prioritization system, prioritized SNP list by integrating significant genes and pathway information from previous GWAS was obtained. Moreover, AHP scoring mechanism integrated disease data and disease-disease interactions to prioritized SNP list, generating associations between SNPs, genes and disease phenotypes. Apart from these results obtained in present study, there is a crucial point to be mentioned. During GWAS, sample size should be sufficient in order to ensure high power to detect genes of risk (W. Y. S. Wang, Barratt, Clayton, & Todd, 2005). Generally, at least 2,000 to 5,000 samples for both case and control groups are required when using general populations (Iles, 2008). In contrast to these huge numbers, in this study there were only 9 cases and 8 controls. Therefore, the presented study should be considered as a pilot study of GWAS of metastatic OS. To further prove the suggestions of this study ideally, large-scale, prospective researches are needed. However, the results obtained in this study could be helpful for researchers to build new hypothesis and conduct studies on metastatic OS.

3.5.3 Copy Number Variation and Loss of Heterozygosity Analysis

The last part of this study included CNA and LOH analysis of SNP array data of ten tumor samples in comparison to their respective normal samples. As stated earlier, all of the samples in this study had low quality (fragmented) DNA because of their FFPE nature. This fragmentation caused poor data quality in microarray experiments. As shown in SNP call rates table (Table 3.3, p.85), all of the samples had an SNP call rate smaller than 95% threshold of

data quality. For this reason, three different methods for CNA and LOH analysis, namely CalMaTe, Tumorboost and VN algorithm, were tried in order to demonstrate which method is most suitable for FFPE data and to obtain most meaningful results of our data. All three methods used in this study were add-ons to R-based package: aroma.affymetrix (Henrik Bengtsson, 2004).

3.5.3.1 Comparison of the Three Methods for CNV and LOH Analysis

A quick summary of differences between three methods are given in Table 3.11. Accordingly, the main differences between first and second method was in the data structure, reference selection and LOH detection. The first method, CalMaTe, could analyze nonpaired high quality data allowing us to get ASCN BAF graphs after CN analysis and to obtain CNA results after CBS segmentation. Since it utilized nonpaired data, during copy number extraction CN normalization was performed using the average of all normal samples as reference. This normalization could solve the normal contamination problem of tumor samples. However, it also causes a restriction in data analysis because the average of all normal samples could include some data points which are actually not present in some of the normals and by normalization, those data points could be lost in the tumors even though they are actually CNAs for those samples. In order to overcome this restriction, a second method, Tumorboost, has been applied and it could analyze each pair separately, performing normalization with match normal sample of each tumor. By this normalization, in addition to normal contamination problem of tumor samples, data point loss could be resolved. Moreover, this method provided LOH results in addition to TCN and ASCN estimates by segmentation with PSCBS. On the other hand,

those two methods are structured according to high quality data and both have restrictions for application to FFPE SNP array low quality data. Considering this situation, a third method, VN algorithm, has been applied, which was able to normalize CNs of tumors separately with a reference obtained from the average of both tumors and normals (Lisovich et al., 2011). Likewise CalMaTe, this method could only perform nonpaired data analysis providing only total CNA results after segmentation with GADA.

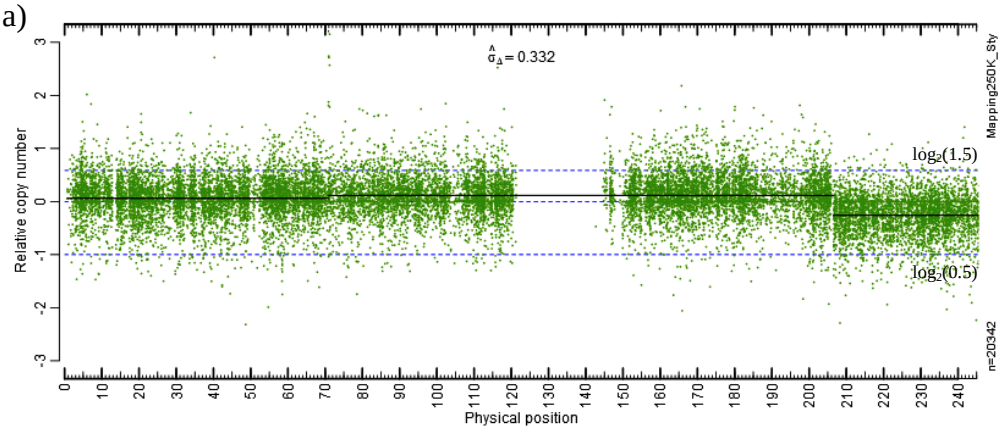
Table 3.11 Summary of differences between three methods of CNA and LOH analysis

Method	Data Structure	Data Quality	Reference selection	TCN	ASCN	CNA	LOH
CalMaTe	Nonpaired	High	Avg of normals	Yes	Yes	Yes	No
Tumorboost	Paired	High	Match normal	Yes	Yes	Yes	Yes
VN algorithm	Nonpaired	Low	VN reference	Yes	No	No	No

After TCN segmentation with the first and third methods, segmentation results were drawn using ChromosomeExplorer within the aroma.affymetrix framework (Henrik Bengtsson, 2004). In order to compare the efficiency of these two methods for TCN segmentation, the noise level along the whole genome for each method was calculated using a robust first-order standard deviation estimator (σ) (H. Lu et al., 2011). This noise level estimator is actually affected by the reference selection in which the two methods differed. In Figure 3.15, there are two panels showing TCN signals of Chromosome1 of

sample T8. The difference between segments involving two successive copy number regions is comparable across panels. Hence, signal to noise ratios can be compared based on the corresponding noise levels. The noise level of the first method (σ_1=0.332) is less than the noise level of the third method (σ_3=0.405). This shows that the first method is more effective during TCN segmentation than the third method for FFPE samples.

a)



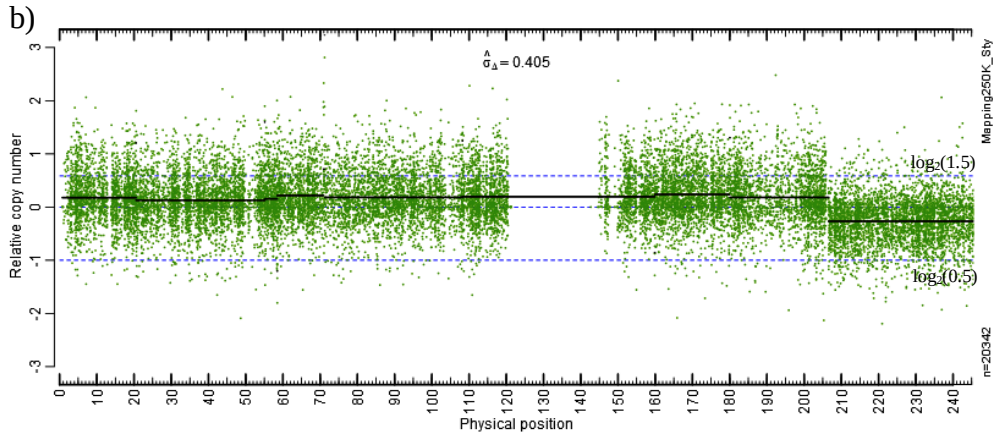


Figure 3.15 Difference between the CalMaTe and VN Algorithm in terms of signal to noise ratio in TCN signals. a) CBS segmentation after CalMaTe; b) GADA segmentation after VN algorithm for Chromosome-1 of sample T8. Green dots: locus-level estimates; black segments: region-level estimates after segmentation, upper blue line: threshold for $\log_2(1.5)$; lower blue line: threshold for $\log_2(0.5)$.

The noise level estimators (σ) for all chromosomes of sample T8 are given in Table 3.12. All estimators obtained from the first method, CalMaTe, were less than the ones obtained from the third method, VN algorithm.

With the second method, Tumorboost, results of only sample T8 was obtained. This was because the data quality threshold (SNP call rate) for the remaining samples were too low to be analyzed with Tumorboost and T8 had the highest data quality with an SNP call rate 84.91%. The detailed results of sample T8 with TCN and BAF graphs are given in subsequent sections.

The overall results obtained by the three methods are represented in Table 3.11. As seen in Table 3.12, the first method, CalMaTe, resulted in 16 chromosomal segments with gain of CN and 8 chromosomal segments with loss of CN over all tumor samples, given in Appendix G. Of which, 11 were belong to sample T8. The second method, Tumorboost, could be run only for sample T8 and showed 6 LOH regions for that sample. The third method, VN algorithm, has shown 294 chromosomal segments containing 152 CN gain and 142 CN loss regions over all tumor samples, given in Appendix G.

Table 3.12 Number of regions of CNA and LOH obtained by three methods

Method	**CNA**		**LOH**
	Loss	**Gain**	
CalMaTe	8	16	Not applicable
Tumorboost (Only for T8)	-	-	6
VN algorithm	142	152	Not applicable

Biological interpretation of all results for all samples is beyond the scope of this thesis. Since sample T8 was the only sample that has given results with all three methods, this interpretation was performed for only T8 in the following section. This also provides comparison of results obtained by these three methods running the same data.

3.5.3.2 Comparison and Biological Interpretation of CNA and LOH Results of Tumor Sample T8

Sample T8 has been analyzed with all three methods. The first and third methods were utilized for TCN segmentation. When the TCN noise levels (σ_{1Avg}=0.334) of all chromosomes of T8 obtained by CalMaTe was compared to the ones (σ_{3Avg}=0.390) obtained by VN algorithm (Table 3.13), CalMaTe has shown less noise levels across all chromosomes than VN algorithm, indicating higher signal to noise ratios. For this reason, only the results of CalMaTe are utilized for biological interpretation of TCN segmentation of sample T8.

Table 3.13 Noise level estimators for CalMaTe and VN Algorithm

Chromosomes	Noise levels (σ_1) for CalMaTe	Noise levels (σ_3) for VN Algorithm
1	0.332	0.405
2	0.334	0.386
3	0.335	0.393
4	0.327	0.410
5	0.337	0.427
6	0.331	0.397
7	0.329	0.399
8	0.343	0.390
9	0.335	0.404
10	0.337	0.380
11	0.325	0.392
12	0.327	0.389
13	0.345	0.406
14	0.332	0.395
15	0.335	0.358
16	0.325	0.378
17	0.339	0.379

Table 3.13 continued.		
18	0.325	0.400
19	0.347	0.325
20	0.347	0.393
21	0.319	0.372
22	0.322	0.348
23	0.361	0.448
Average	0.334	0.390

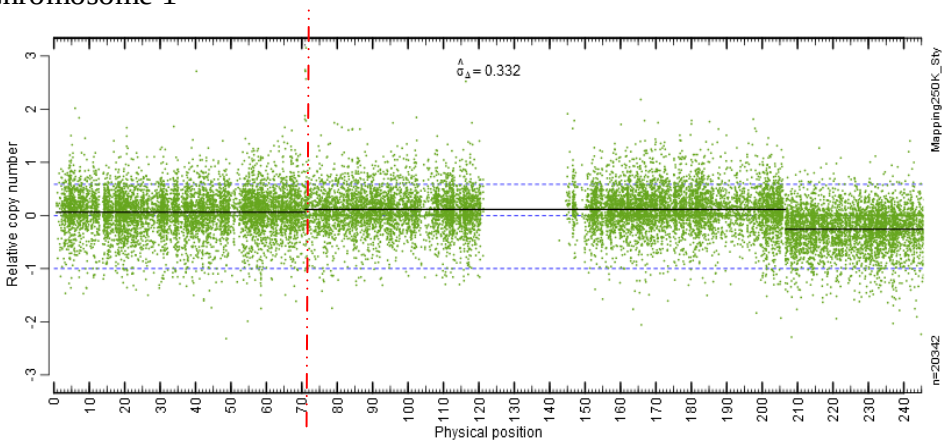
There were 11 regions of CNA found by TCN segmentation with the first method, CalMaTe. In 3.14, all CNA regions are summarized according to call (gain or loss), chromosomal location, mean relative copy number which is log2 ratio of raw copy numbers of tumor and normal (Mean=$\log_2$(raw CN of tumor/raw CN of normal), and number of SNPs found on respective segment.

Table 3.14 CNA regions obtained for sample T8

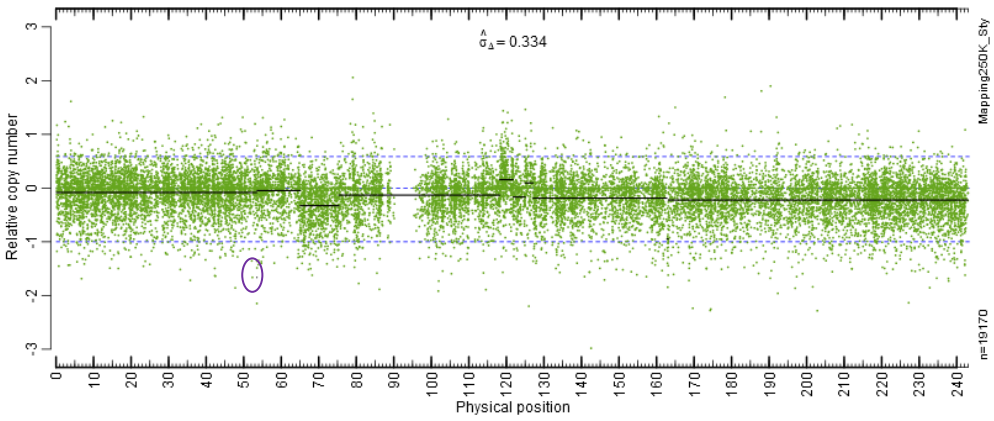
Chromosomes	Start bp	Stop bp	Mean	SNP Count	Call
1	70887099	70911359	3.208	3	Gain
2	53514150	53540600	-1.664	3	Loss
6	32448098	32563997	1.662	3	Gain
7	32629579	32629640	1.885	2	Gain
7	34905983	34905998	5.575	2	Gain
8	126766585	126808810	-1.146	6	Loss
11	101668066	101668330	1.899	2	Gain
12	23185575	23218803	1.198	6	Gain
17	42687933	42687961	-1.753	2	Loss
17	61227637	61227768	-1.946	2	Loss
18	25353854	25354031	1.463	3	Gain

In the following figure, the chromosomes with CNA stated in Table 3.14 are shown. The panels for each chromosome correspond to segmentation results which are drawn as relative copy number vs physical position. The raw total copy number and allele B fraction results of each chromosome with CNA are given in Appendix H.

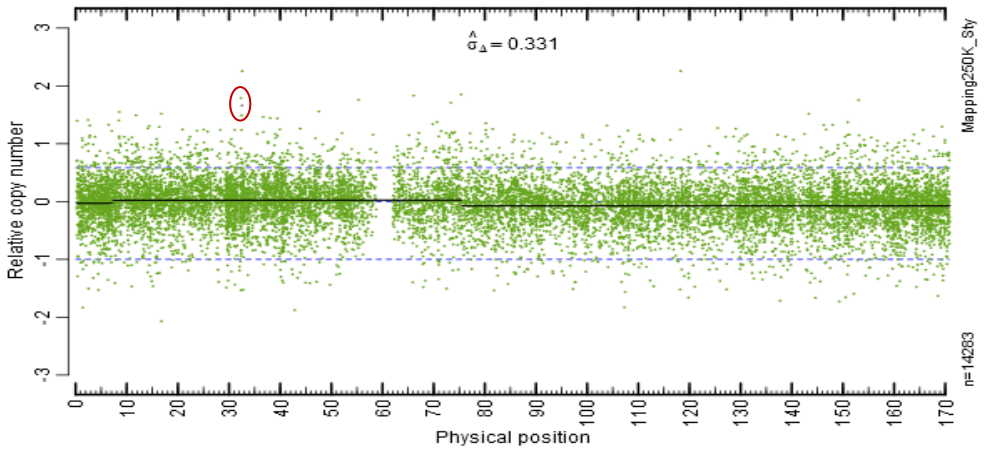
a) Chromosome 1



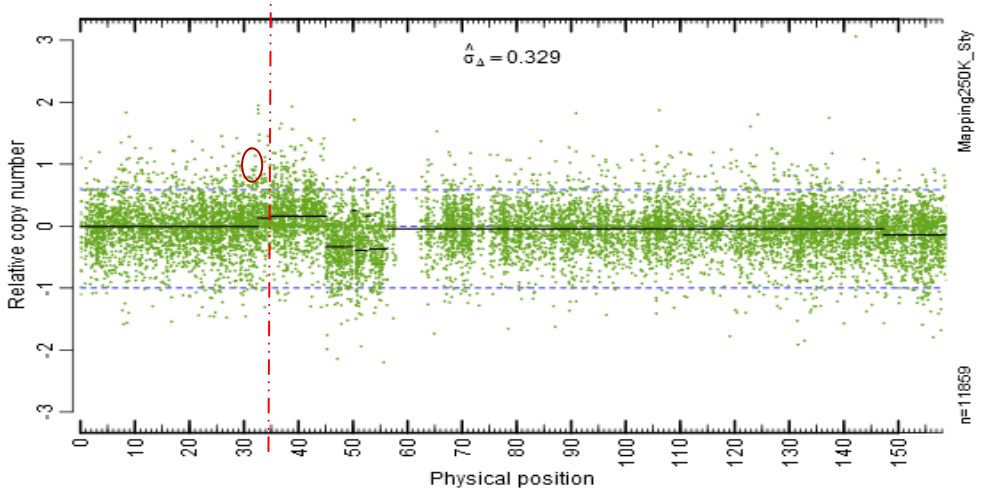
b) Chromosome 2



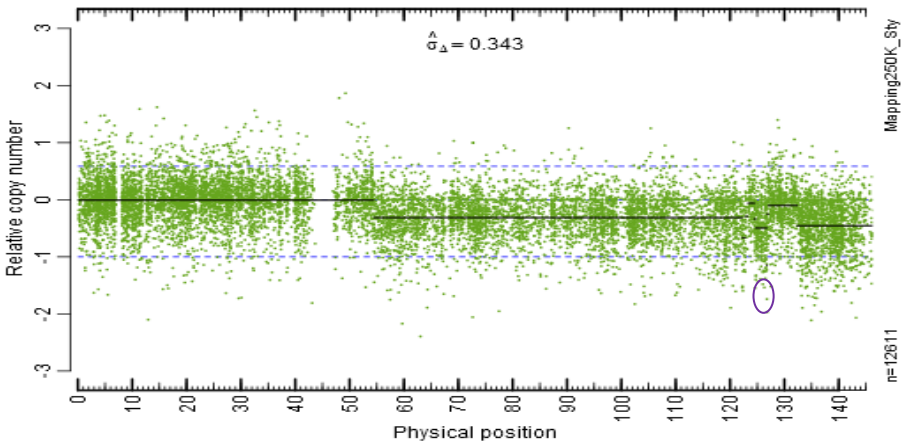
c) Chromosome 6



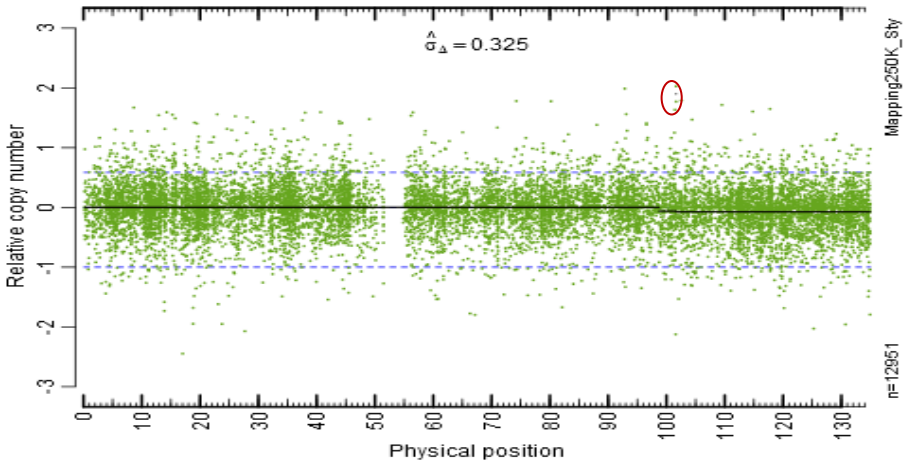
d) Chromosome 7



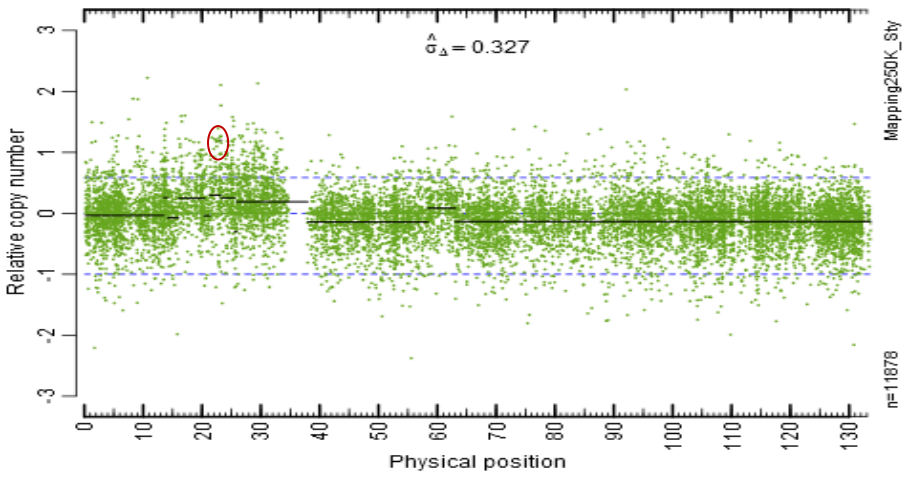
e) Chromosome 8



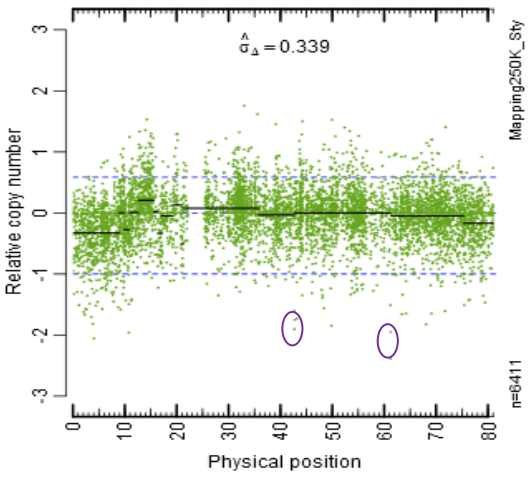
f) Chromosome 11



g) Chromosome 12



h) Chromosome 17



i) Chromosome 18

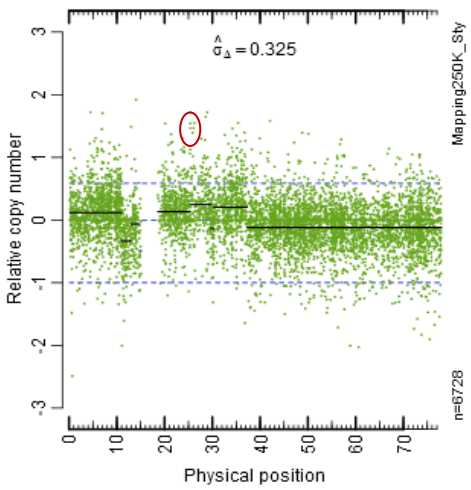


Figure 3.16 The panels for a) Chr1 b) Chr2 c) Chr6 d) Chr7 e) Chr8 f) Chr11 g) Chr12 h) Chr17 i) Chr18 corresponding to segmentation results which are drawn as relative copy number vs physical position. Green dots: locus-level estimates; black segments: region-level estimates after segmentation; upper blue line: threshold for $\log_2(1.5)$; lower blue line: threshold for $\log_2(0.5)$; red dashed line: gain above the thresholds of graph; red circle: gain; purple circle: loss.

In Figure 3.16, the segments obtained by region level estimates are shown. There are two thresholds for a segment to be considered as significant. The first threshold indicates a TCN state of 3 ($\log_2(1.5)=0.585$) and it is represented as upper blue line in the panels. The segments with mean levels above this threshold are considered to be a copy number gain region. The second threshold is for copy number loss regions, indicating a TCN state of 1 ($\log_2(0.5)=-1$) and it is represented as lower blue line in the panels. The segments with mean levels below this threshold are considered to be a copy number loss region. In Figure 3.16, although large segments covering almost half of the chromosomes are observed, those do not exceed the thresholds and so are not considered as significant. According to segmentation results, there were 11 submicroscopic CNAs observed for sample T8, which ranged between 150 bases to 115 kilobases, and they are specified in Figure 3.16 in red (gain) and purple (loss) circles. Those regions were also searched in Ensemble Genome Browser (EGB) (Flicek et al., 2011); and genes corresponding to these segments are summarized in Table 3.15.

Table 3.15 Genes found in the CNA regions

Chr	Loci	Call	Genes
1	p31.1	Gain	RP11-428K13.1-001
2	p16.2	Loss	-
6	p21.32	Gain	Major histocompatibility complex, class II, DR beta 9 (HLA-DRB9) Major histocompatibility complex, class II, DR alpha (HLA-DRA) Butyrophilin-like 2 (MHC class II associated) (BTNL2) Chromosome 6 open reading frame 10 (C6orf10)
7	p14.3	Gain	-
8	q24.13	Loss	-

Table 3.15 continued.			
11	q22.1	Gain	-
12	p12.1	Gain	-
17	q23.3	Loss	Integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61) (ITGB3)
17	q21.32	Loss	Coiled-coil domain-containing protein 46 (CCDC46)
18	q12.1	Gain	-

The first gain segment on Chr1 includes a processed pseudogene, RP11-428K13.1, which is noncoding and is produced by integration of a reverse transcribed mRNA into the genome (Flicek et al., 2011). In the literature, this locus has no association with any disease, especially for metastatic OS. Concordantly, it can be said that this locus can have association with metastatic OS. More studies including fluorescent in situ hybridization (FISH) or real time PCR should be performed to further prove this association.

On the second segment including CN gain on Chr6, there are 4 genes detected. The first one, HLA-DRB9, is an unprocessed pseudogene arising from gene duplication. In previous studies, it has been found that this pseudogene has association with ulcerative colitis (Franke et al., 2010) and diabetes mellitus(D. Song et al., 2002). However, there has been no association found with a cancer, especially metastatic OS. Concordantly, for this locus, this is the first time for an association with metastatic OS found. Further studies should be performed to further prove this association.

The second gene found in the CN gain segment of Chr6 is HLA-DRA gene which is one of the HLA class II alpha chain protein coding genes. The protein

product of this gene is a heterodimer consisting of an alpha and a beta chain, anchored in the membrane. Its central role is presentation of peptides derived from extracellular proteins and it is expressed in antigen presenting cells (APC) of the immune system: B lymphocytes, dendritic cells, macrophages. DRA is the sole alpha chain for DRB1, DRB3, DRB4 and DRB5. In the literature, this gene has been found to be associated with Parkinson's disease (Puschmann et al., 2011), multiple sclerosis (Hafler et al., 2007), hepatocellular carcinoma (W.-ping Lv, Dong, Shi, Huang, & Guo, 2008; Matoba et al., 2005), and leukemia (Promsuwicha & Auewarakul, 2009). Concordantly, this study may be the first that has found association of this locus containing HLA-DRA gene to metastatic OS. However, this association needs to be further proved.

BTNL2 is the third gene found in the CN gain segment of Chr6, which encodes a MHC class II associated protein with unknown function. In previous studies, BTNL2 gene allele variants and SNPs are found to be associated with ulcerative colitis (Franke et al., 2008), sarcoidosis (Wijnen et al., 2011), and tuberculosis(Lian, Yue, Han, Liu, & Liu, 2010). On the other hand, there has been no association for this gene to a cancer type. For this reason, this study may be the first that has found association of this gene to metastatic OS.

The last gene found in the gain segment of Chr6 is C6orf10, uncharacterized protein coding gene. In other words, there has been no specific function of this gene to be found yet. in the literature, C6orf10 has been found to be associated with some diseases, namely systemic lupus, multiple sclerosis, vitiligo, and bone mineral density and fractures (S. A. Chung et al., 2011; Y. Jin et al., 2011; Styrkarsdottir et al., 2008). However, there is no cancer type found to be

associated with this gene. For this reason, this study may demonstrate the first association for this gene with metastatic OS.

The loss segment found on Chr17 contains ITGB3 gene whose protein product is the integrin beta chain beta 3. Generally, integrins are integral cell-surface proteins made up of an alpha chain and a beta chain and have role in cell adhesion as well as cell-surface mediated signaling. A given chain can combine with multiple partners resulting in different integrins. Integrin beta 3 is found along with the alpha IIb chain in platelets. Integrins are all found to be associated with several diseases including cancer. However, there is no clear association of ITGB3 with metastatic OS. In one study, it was found that MAPK signaling pathway was initiated by integrin alphaVbeta3 activation and this increased the cell proliferation in human osteoblast-like cells (Scarlett et al., 2008). Loss of this gene in metastatic OS may have functional importance. For this reason, the association found here should be proved by additional studies.

The second gene found in loss segment of Chr17, CCDC46, is a protein coding gene with unknown function. There has been no association found for this gene with a disease, except present study.

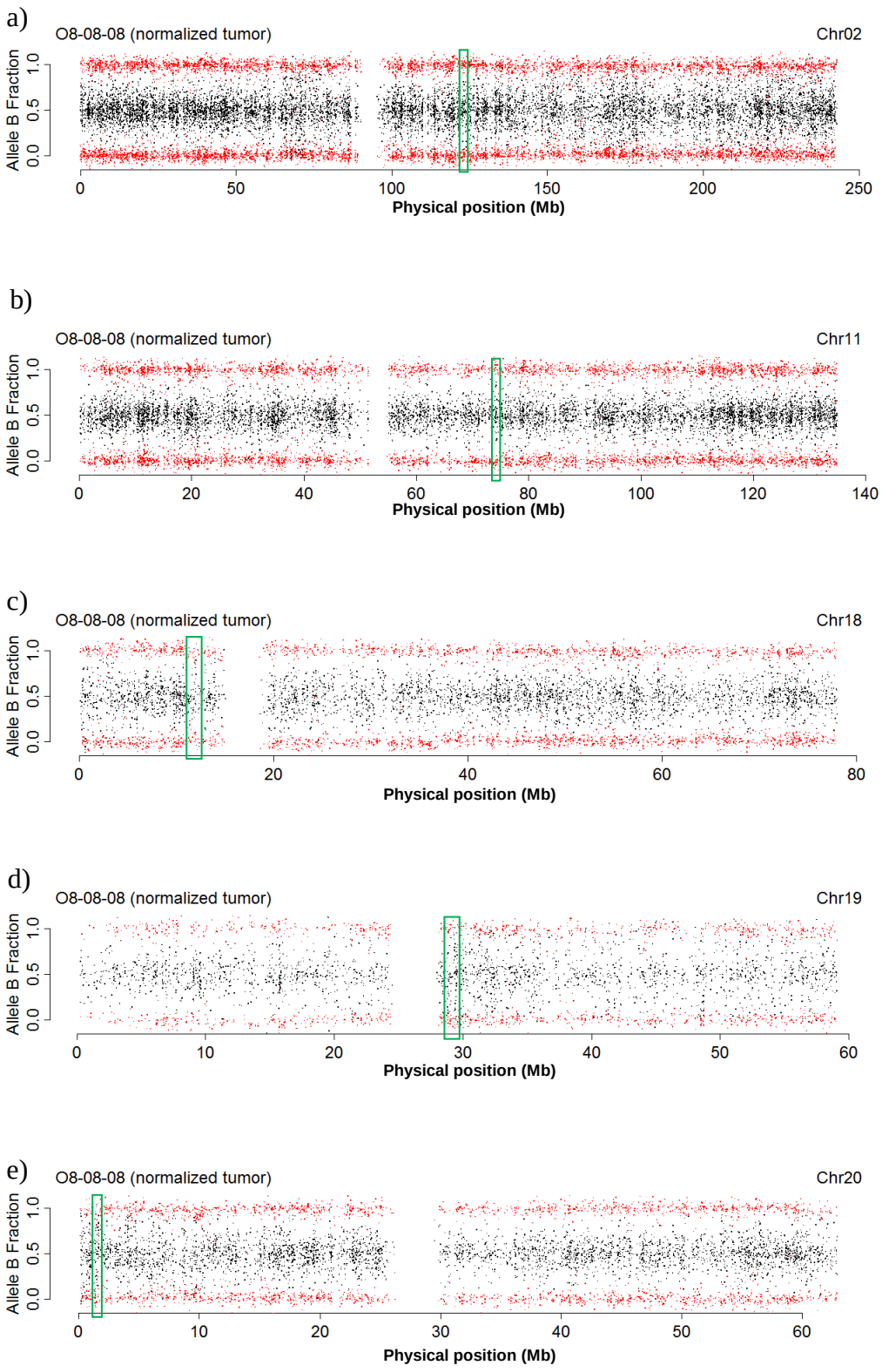
For sample T8, LOH regions could also be obtained by the second method, Tumorboost. All LOH events detected are summarized in Table 3.16. There were 2 LOH segments having deletion of one copy, indicating a CNS of (0, 1) (hemizygous deletion). This one copy deletion in those segments caused an increase in “decrease in heterozygosity index” (DH) to almost 1; 0.71 for the

segment on Chr20, and 0.51 for the segment on Chr23. In other words, heterozygosity of those regions decreased by 71%, and 51%, respectively. Moreover, there were 3 regions with copy number neutral LOH (CN-LOH) with a CNS of (0, 2). Those CN-LOH events could be resulted from uniparental disomy (UPD) because in all of the segments, one allele from one parent was missing and the other parental allele was reduplicated. Those CN-LOH events also led to increase in DH to almost 1; 0.91 for the segment on Chr18, 0.80 for Chr19, and 0.66 for Chr2. The last LOH region included LOH event with gain. In that segment on Chr11, while one allele from one parent is deleted, the other parental allele was amplified itself by 5 times, having a CNS of (0, 5). This amplification of one allele caused DH to increase 0.79.

Table 3.16 LOH regions obtained by Tumorboost

	Chromosomal region (Chr:BP)	**CNS (c_1, c_2)**
LOH with Deletion (0,1)	20:2353420-2359886 23: 79687652-83412609	(0.2, 1.3) (0.3, 1.01)
CN LOH (0,2)	18: 12043153-12256868 19:29591449-29820317 2: 127653629:127679173	(0.01, 2.01) (0.26, 2.26) (0.5, 2.5)
LOH with CN gain; (0,y), y≥3	11:74251268:74251779	(0.65, 5.51)

In Figure 3.17, the LOH regions obtained by Tumorboost are shown, specified on each chromosome by green rectangles. PSCBS segmentation results over all chromosomes are shown in the Appendix H.


a)
O8-08-08 (normalized tumor)
Chr02
Allele B Fraction
0.0
0.5
1.0
0
50
100
150
200
250
Physical position (Mb)
b)
O8-08-08 (normalized tumor)
Chr11
Allele B Fraction
0.0
0.5
1.0
0
20
40
60
80
100
120
140
Physical position (Mb)
c)
O8-08-08 (normalized tumor)
Chr18
Allele B Fraction
0.0
0.5
1.0
0
20
40
60
80
Physical position (Mb)
d)
O8-08-08 (normalized tumor)
Chr19
Allele B Fraction
0.0
0.5
1.0
0
10
20
30
40
50
60
Physical position (Mb)
e)
O8-08-08 (normalized tumor)
Chr20
Allele B Fraction
0.0
0.5
1.0
0
10
20
30
40
50
60
Physical position (Mb)

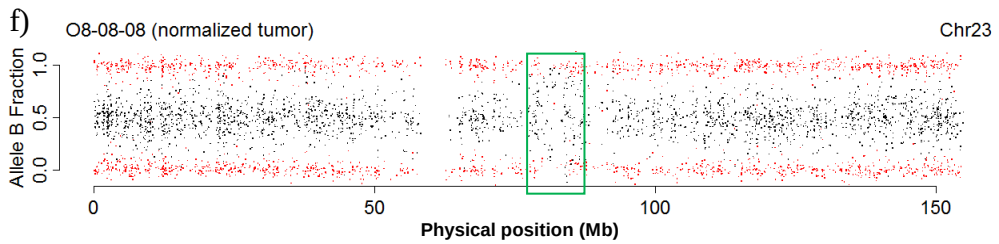


Figure 3.17 Allele B fractions with specified regions of LOH a) Chr2 b) Chr11 c) Chr18 d) Chr19 e) Chr20 and f) Chr23 (X chromosome). Red dots: Homozygous SNPs; black dots: Heterozygous SNPs; green rectangles: LOH regions.

Those LOH regions were searched in Ensemble Genome Browser (EGB) (Flicek et al., 2011); and genes corresponding to these segments are summarized in Table 3.17.

Table 3.17 Genes found in the LOH regions.

Chr	Loci	Genes
2	q14.3	AC114783.1
11	q13.4	Polymerase (DNA-directed), delta 3, accessory subunit (POLD3)
18	p11.21	U6 Spliceosomal RNA (U6) Ankyrin repeat domain 62 (ANKRD62) AP002414.1-201, AP002414.1-001
19	q12	Ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1 (UQCRFS1) AC007786.1
20	p13	-
23	q21.1	Family with sequence similarity 46, member D (FAM46D) Bromodomain and WD repeat domain containing 3 (BRWD3) High mobility group nucleosome binding domain 5 (HMGN5) Ribosomal protein S6 kinase, 90kDa, polypeptide 6 (RPS6KA6)

The first, third and fourth LOH segments found on Chr2, Chr18, and Chr19, respectively, include two large intergenic non-coding RNAs (LicRNAs), AC114783.1 and AC007786.1; one noncoding pseudogene, AP002414.1-201; one novel pseudogene, AP002414.1-001, encoding an uncharacterized protein; and ncRNA gene encoding for U6 Spliceosomal RNA. These ncRNAs and pseudogenes have no association with any disease found in the literature, except for this study.

POLD3 gene is found in LOH region on Chr11. It encodes for POLD3 subunit of DNA polymerase delta complex and takes role in DNA replication. In one study, it was stated that POLD3 have a crucial role in the recycling of PCNA during dissociation-association cycles of polymerase delta during elongation of DNA replication (Y. Masuda et al., 2007). Any mutation in this gene probably causes errors in DNA replication, leading to further mutations. For this reason, functional studies on metastatic OS samples should be performed to prove this mutagenic condition.

In the LOH region on Chr18, ANKRD62 gene has been found. This gene encodes for ankyrin repeats. Those repeats mediate protein-protein interactions in very diverse families of proteins. Any mutation in this gene may cause errors in protein interactions that include ankyrin repeats. This mutation in this gene should be monitored by protein interaction assays in further studies in order to understand the role of this gene in metastatic OS.

In the LOH region detected in Chr19, UQCRFS1 has been found. This gene encodes for a protein that is a key subunit of the cytochrome bc1 complex

(complex III) of the mitochondrial respiratory chain. The study conducted by Ohashi et.al. (2004) suggested that this gene might be involved in development of more aggressive breast cancer phenotype (Ohashi, Kaneko, Cupples, & Young, 2004). This may imply the change in the energy metabolism of cancer cells due to their high energy consumption. These changes in the energy metabolism of cancer cells are also true for metastatic OS (Zhiyu Zhang et al., 2010). Functional studies are needed to prove this energy metabolism change that may be caused by UQCRFS1 gene mutation.

FAM46D gene found in X Chromosome LOH region encodes for a protein with unknown function. However, antibodies against the protein encoded by this gene were found in plasma of only cancer patients, suggesting an association of this gene with cancer (Bettoni et al., 2009). This gene may also be target for detection of metastatic OS; however, further studies would be needed to prove its presence in metastatic OS patients.

The second gene found in X Chromosome LOH region is BRWD3. The protein encoded by this gene includes a bromodomain and a number of WD repeats. It is thought to have a chromatin-modifying function, and may thus play a role in transcription. This gene is found to be associated with translocations in B-cell chronic lymphocytic leukemia patients (Müller, Kuttenkeuler, Gesellchen, Zeidler, & Boutros, 2005). This study suggests that there may be also some association of mutations in this gene with metastatic OS.

The third gene found in the LOH region of ChrX is HMGN5 which encodes a nuclear protein similar to the high mobility group proteins, HMG14 and

HMG1, suggesting a functional role for this protein as a nucleosomal binding and transcriptional activating protein. Moreover, in a study conducted by Jiang *et al*. (2010), it was stated that downregulation of this protein can inhibit the in vitro and in vivo proliferation of prostate cancer cells (Jiang, Zhou, & Zhang, 2010). The functional role of this gene in metastatic OS should be investigated to prove the association with cancer phenotype.

The last gene found in the LOH region of ChrX, RPS6KA6 (also known asRSK4), is a putative tumor suppressor gene, encoding a member of ribosomal S6 kinase family, serine-threonine protein kinases regulated by growth factors. In a study conducted by Thakur *et al*. (2008), it was stated that RSK4 expression may limit the oncogenic, invasive, and metastatic potential of breast cancer cells (Thakur et al., 2008). Moreover, this gene was found to be downregulated in colon carcinogenesis and renal cell carcinomas (López-Vicente et al., 2009). In addition, it is known that the activation status of the X-chromosome may influence the expression levels of X-linked tumor suppressor genes. In our study, LOH occurred in this gene included a one copy deletion, which may affect the expression level of this gene and if the remaining copy of the gene has been inactivated by X-inactivation, this suggests an inactivation of a tumor-suppressor gene in metastatic OS. In order to prove downregulation of this gene in metastatic OS, expression analysis following mutation analysis should be performed in further studies.

CHAPTER 4

CONCLUSION

Osteosarcoma is the most common malignant tumor of bone having an incidence rate of 19% among all cancer types. About half of the OS patients develop pulmonary metastasis which causes poor prognosis and increased death rate among those patients. Although mutations in the some cancer causing genes such as *P53, RB, FOS* and *MYC* were identified in pulmonary metastatic tumors of OS, there is no unique genetic pathway identified for progression of pulmonary metastasis (Wan, Mendoza, Khanna, & Helman, 2005). In this study, it was aimed to identify potential prognostic markers of metastatic OS by performing a GWAS of a small case-control group of patients and by developing a data analysis workflow from available tools for the use of FFPE samples in high-throughput array analysis.

- According to performed GWAS, 358 significantly associated SNPs have been found. The SNPs with smallest *p*-values were rs6499861 (*p*-value = 4.90E-05), rs10884554 (*p*-value = 4.90E-05) and rs12154602 (*p*-value = 6.55E-05). According to dbSNP database, those SNPs have no association with a known disease. Hence, this research might be the first study

indicating the associations of these SNPs with a disease, particularly metastatic OS. Furthermore, of the top 20 associated SNPs, six were found to map to a gene, which are rs12154602, rs17120272, rs6657368, rs10923183, rs5744297, and rs125594008. All of them were intronic variants.

- Second wave analysis of GWAS results provided statistically significant genes associated with metastatic OS. Among those genes, serglycin (SRGN) and, D4 Zinc and Double PHD Fingers Family 3 (DPF3) genes have been indicated to have association with metastatic nature of cancer cells in previous studies. In addition to SRGN and DPF3, 4 other genes which have found to be associated with OS in this study, namely Phosphatidylinositol-4-Phosphate 5-Kinase Type I (PPI4P5K1), Death Domain Containing 1 (DDC1), Echinoderm Microtubule Associated Protein Like 6 (EML6), and ATPase Ca++ Transporting Plasma Membrane 2 (ATP2B2), have been associated with different types of cancer in previous studies. Additionally, SRGN, Msh Homeobox 1 (MSX1), and PPI4P5K1 genes have been shown to be highly expressed in lung tissue. These specific connections of those genes with cancer might suggest possible significant associations with metastatic OS.

- Second wave analysis of GWAS results also provided statistically significant pathways associated with metastatic OS. According to pathway associations, metabolic pathways such as glycolysis and amino acid degradation were found to be associated with pulmonary metastasis of OS, which was consistent with global metabolic shift in favor of tumor progression occurring in the metastatic tumors.

- SNP prioritization with AHP approach was performed in order to combine statistical and genetic information of SNPs. Chromosomes 1 and 4 are the chromosomes which accommodated most of the top 100 prioritized SNPs. Of which 12 SNPs were on Chr1 while 13 SNPs on Chr4. Numerical and structural chromosomal aberrations of chromosomes 1 and 4 in OS were observed in previous studies and also in this research. In addition, 19 of the top 20 prioritized SNPs were found to map genes, and among them, rs17675094 and rs2148630 were found to map to intronic regions of genes, cadherin 13 (CDHL13) and latrophilin 2 (LPHN2), which have been found to be associated with lung cancer in previous studies.

- CNA and LOH analysis workflow were generated in order to produce reliable genotype, copy number, and LOH predictions from FFPE derived DNA samples. The first method, CalMaTe, was found to produce high signal to noise ratios compared to third method during estimation of TCN segments. LOH analysis with the second method, Tumorboost, could only be performed for one of the samples due to restrictions in data quality of the remaining samples.

- There were 11 regions of CNA found by TCN segmentation of sample T8 with the method, CalMaTe. These are summarized in Table 4.1

Table 4.1 CNA regions of sample T8 with corresponding genes in those regions

Chromosome	Loci	Call	Genes
1	p31.1	Gain	RP11-428K13.1-001
2	p16.2	Loss	-
6	p21.32	Gain	HLA-DRB9 HLA-DRA BTNL2 C6orf10
7	p14.3	Gain	-
8	q24.13	Loss	-
11	q22.1	Gain	-
12	p12.1	Gain	-
17	q23.3	Loss	ITGB3
17	q21.32	Loss	CCDC46
18	q12.1	Gain	-

- There were 6 LOH regions found for sample T8 with the method, Tumorboost. Those regions and the genes in those regions are summarized in Table 4.2.

Table 4.2 LOH regions of sample T8 with corresponding genes in those regions

Chromosome	Loci	Genes
2	q14.3	AC114783.1
11	q13.4	POLD3
18	p11.21	U6 ANKRD62 AP002414.1
19	q12	UQCRFS1 AC007786.1

Table 4.2 continued.		
20	p13	-
23	q21.1	FAM46D BRWD3 HMGN5 RPS6KA6 (RSK4)

4.1 Recommendations

The results obtained in this study could be used to build new hypotheses and conduct studies on metastatic OS. Few recommendations are given for future studies:

- To further prove the suggestions of this study ideally, sample size for GWAS could be increased with the addition of publicly available SNP array data with similar hypothesis. This would ensure high power to detect SNPs and genes of risk.

- In order to validate the results of this study, as summarized in Section 1.4.1.1.2 and in Figure 1.3, the most significantly associated SNPs could be screened on an independent group of pulmonary metastatic OS patients with the same methodology presented in this study. The results obtained by this validation step should be replicated on a different group of patients by using small scale SNP screening technologies such as sequencing and TaqMan PCR in order to proceed with functional studies.

- Most significantly associated SNPs, rs6499861, rs10884554 and rs12154602 could be scanned in OS cell lines by mutation analysis and further investigated in a large sample case control study to prove associations more significantly.

- In order to prove the functional association of those genes to metastatic OS, mutation analysis, gene expression, and protein expression studies on OS samples might be performed.

- Considering the restrictions of FFPE data, improvements in the CNV-LOH analysis methods used in this study could be done by changing the thresholds of the algorithms.

- Validation of CNA aberration results could be performed on an independent sample set with both large scale assays such as microarrays and small scale assays such as fluorescent in situ hybridization. Likewise, the validation of LOH results might be done by using alternative genetic markers such as microsatellites in the region of interest.

- In order to obtain high quality data, instead of FFPE samples, fresh frozen samples can be utilized in high-throughput assays. For this, long-term collaboration studies with hospitals could be designed.

REFERENCES

ACS. (2011). Osteosarcoma Overview. Retrieved from http://www.cancer.org/acs/groups/cid/documents/webcontent/003069-pdf.pdf. Last accessed: 24/Aug/2011

Affymetrix. (2011a). GeneChip Human Mapping 500K Array Set | Affymetrix. Retrieved from http://www.affymetrix.com/estore/browse/products.jsp?productId=131459&categoryId=35906&productName=GeneChip-Human-Mapping-500K-Array-Set#1_1. Last accessed: 18/Dec/2011.

Affymetrix. (2011b). *GeneChip ® Mapping 500K Assay Manual*. *Flying*. Last accessed: 18/Dec/2011.

Akiyama, T., Dass, C. R., & Choong, P. F. M. (2008). Novel therapeutic strategy for osteosarcoma targeting osteoclast differentiation, bone-resorbing activity, and apoptosis pathway. *Molecular cancer therapeutics*, *7*(11), 3461-9. doi:10.1158/1535-7163.MCT-08-0530

Arpaci, F., Ataergin, S., Ozet, A., Erler, K., Basbozkurt, M., Ozcan, A., Komurcu, S., et al. (2005). The feasibility of neoadjuvant high-dose chemotherapy and autologous peripheral blood stem cell transplantation in patients with nonmetastatic high grade localized osteosarcoma: results of a phase II study. *Cancer*, *104*(5), 1058-65. doi:10.1002/cncr.21279

Baran, Y., Gür, B., Kaya, P., Ural, A. U., Avcu, F., & Gündüz, U. (2007). Upregulation of multi drug resistance genes in doxorubicin resistant human acute myelogeneous leukemia cells and reversal of the resistance. *Hematology (Amsterdam, Netherlands)*, *12*(6), 511-7. doi:10.1080/10245330701562535

Bengtsson, H, Irizarry, R., Carvalho, B., & Speed, T. P. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics (Oxford, England)*, *24*(6), 759-67. doi:10.1093/bioinformatics/btn016

Bengtsson, Henrik. (2004). *aroma - An R Object-oriented Microarray Analysis environment*. *Preprints in Mathematical Sciences*. Retrieved from http://www1.maths.lth.se/bioinformatics/publications/BengtssonH_2004-aroma.pdf

Bengtsson, Henrik, Neuvial, P., & Speed, T. P. (2010). TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC bioinformatics*, *11*, 245. doi:10.1186/1471-2105-11-245

Bengtsson, Henrik, Neuvial, P., Cnrs, B., & Olshen, A. (2011). Single Tumor-Normal Pair Parent-Specific Copy Number Analysis. *BMC Bioinformatics*.

Bengtsson, Henrik, Wirapati, P., & Speed, T. P. (2009). A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics (Oxford, England)*, *25*(17), 2149-56. doi:10.1093/bioinformatics/btp371

Bettoni, F., Filho, F. C., Grosso, D. M., Galante, P. A. F., Parmigiani, R. B., Geraldo, M. V., Henrique-Silva, F., et al. (2009). Identification of FAM46D as a novel cancer/testis antigen using EST data and serological analysis. *Genomics*, *94*(3), 153-60. doi:10.1016/j.ygeno.2009.06.001

Bridge, J. A., Nelson, M., Mccomb, E., Mcguire, M. H., Rosenthal, H., Vergara, G., Maale, G. E., et al. (1996). Cytogenetic Findings in 73 Osteosarcoma Specimens and a Review of the Literature. *Review Literature And Arts Of The Americas*, *4608*(96).

Browning, B. L., & Browning, S. R. (2007). Efficient Multilocus Association Testing for Whole Genome Association Studies Using Localized Haplotype Clustering. *Power*, *375*, 365-375. doi:10.1002/gepi

Cai, Y., Mohseny, A. B., Karperien, M., Hogendoorn, P. C. W., Zhou, G., & Cleton-Jansen, A.-M. (2010). Inactive Wnt/beta-catenin pathway in conventional high-grade osteosarcoma. *The Journal of pathology*, *220*(1), 24-33. doi:10.1002/path.2628

Campanacci, M. (1999). *Bone and soft tissue tumors: clinical features, imaging, pathology and treatment* (p. 1319). Springer. Retrieved from http://books.google.com/books?id=W-EOMACe0c0C&pgis=1. Last accessed: 31/Nov/2011

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature genetics*, *39*(7 Suppl), S16-21. doi:10.1038/ng2028

Carvalho, Benilton, Bengtsson, H., Speed, T. P., & Irizarry, R. a. (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics (Oxford, England)*, *8*(2), 485-99. doi:10.1093/biostatistics/kxl042

Chen, X., Yang, T.-T., Wang, W., Sun, H.-H., Ma, B.-A., Li, C.-X., Ma, Q., et al. (2009). Establishment and characterization of human osteosarcoma cell lines with different pulmonary metastatic potentials. *Cytotechnology*, *61*(1-2), 37-44. doi:10.1007/s10616-009-9239-3

Chung, S. A., Taylor, K. E., Graham, R. R., Nititham, J., Lee, A. T., Ortmann, W. A., Jacob, C. O., et al. (2011). Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS genetics*, *7*(3), e1001323. doi:10.1371/journal.pgen.1001323

Clark, J. C. M., Dass, C. R., & Choong, P. F. M. (2008). A review of clinical and molecular prognostic factors in osteosarcoma. *Journal of cancer research and clinical oncology*, *134*(3), 281-97. doi:10.1007/s00432-007-0330-x

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature protocols*, *6*(2), 121-33. doi:10.1038/nprot.2010.182

Darcansoy, Ö. (2009). *Investigation of docetaxel and doxorubicin resistance in MCF-7 breast carcinoma cell line*. *Breast*. Middle East Technical University.

Distefano, J. K., & Taverna, D. M. (2011). Technological issues and experimental design of gene association studies. *Methods in molecular biology (Clifton, N.J.)*, *700*, 3-16. doi:10.1007/978-1-61737-954-3_1

Durai, R., Yang, S. Y., Sales, K. M., Seifalian, A. M., Goldspink, G., & Winslet, M. C. (2007). Increased apoptosis and decreased proliferation of colorectal cancer cells using insulin-like growth factor binding protein-4 gene delivered locally by gene transfer. *Colorectal disease : the official*

journal of the Association of Coloproctology of Great Britain and Ireland, *9*(7), 625-31. doi:10.1111/j.1463-1318.2006.01190.x

Durham, S. K., Riggs, B. L., Harris, S. A., & Conover, C. A. (1995). Alterations in insulin-like growth factor (IGF)-dependent IGF-binding protein-4 proteolysis in transformed osteoblastic cells. *Endocrinology*, *136*(4), 1374-80.

Erickson, H. S., Gillespie, J. W., & Emmert-Buck, M. R. (2008). Tissue microdissection. *Methods in molecular biology (Clifton, N.J.)*, *424*, 433-48. Humana Press. doi:10.1007/978-1-60327-064-9_34

Estrada, K., Krawczak, M., Schreiber, S., van Duijn, K., Stolk, L., van Meurs, J. B. J., Liu, F., et al. (2009). A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation. *Human molecular genetics*, *18*(18), 3516-24. doi:10.1093/hmg/ddp296

Feng, M., Grice, D. M., Faddy, H. M., Nguyen, N., Leitch, S., Wang, Y., Muend, S., et al. (2010). Store-independent activation of Orai1 by SPCA2 in mammary tumors. *Cell*, *143*(1), 84-98. doi:10.1016/j.cell.2010.08.040

Filipski, K. K., Mathijssen, R. H., Mikkelsen, T. S., Schinkel, A. H., & Sparreboom, A. (2009). Contribution of organic cation transporter 2 (OCT2) to cisplatin-induced nephrotoxicity. *Clinical pharmacology and therapeutics*, *86*(4), 396-402. American Society of Clinical Pharmacology and Therapeutics. doi:10.1038/clpt.2009.139

Finkel, M. P., Reilly, C. A., & Biskis, B. O. (1976). Pathogenesis of radiation and virus-induced bone tumors. *Recent results in cancer research. Fortschritte der Krebsforschung. Progrès dans les recherches sur le cancer*, (54), 92-103.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., et al. (2011). Ensembl 2011. *Nucleic acids research*, *39*(Database issue), D800-6. doi:10.1093/nar/gkq1064

Franke, A., Balschun, T., Karlsen, T. H., Sventoraityte, J., Nikolaus, S., Mayr, G., Domingues, F. S., et al. (2008). Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nature genetics*, *40*(11), 1319-23. doi:10.1038/ng.221

Franke, A., Balschun, T., Sina, C., Ellinghaus, D., Häsler, R., Mayr, G., Albrecht, M., et al. (2010). Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). *Nature genetics*, *42*(4), 292-4. doi:10.1038/ng.553

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851-61. Nature Publishing Group. doi:10.1038/nature06258

Frézal, J. (1998). GENATLAS online. *Trends in Genetics*, *14*(2), 83. doi:10.1016/S0168-9525(97)01294-8

Grant, S. F. a, & Hakonarson, H. (2008). Microarray technology and applications in the arena of genome-wide association. *Clinical chemistry*, *54*(7), 1116-24. doi:10.1373/clinchem.2008.105395

Hafler, D. A., Compston, A., Sawcer, S., Lander, E. S., Daly, M. J., De Jager, P. L., de Bakker, P. I. W., et al. (2007). Risk alleles for multiple sclerosis identified by a genomewide study. *The New England journal of medicine*, *357*(9), 851-62. doi:10.1056/NEJMoa073493

Hameed, M., & Dorfman, H. (2011). Primary malignant bone tumors--recent developments. *Seminars in diagnostic pathology*, *28*(1), 86-101. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21675380. Last accessed: 05/Dec/2011

Hardy, J., & Singleton, A. (2009). Genomewide association studies and human disease. *The New England journal of medicine*, *360*(17), 1759-68. doi:10.1056/NEJMra0808700

Harting, M. T., & Blakely, M. L. (2006). Management of osteosarcoma pulmonary metastases. *Seminars in pediatric surgery*, *15*(1), 25-9. doi:10.1053/j.sempedsurg.2005.11.005

Hattinger, C. M., Pasello, M., Ferrari, S., Picci, P., & Serra, M. (2010). Emerging drugs for high-grade osteosarcoma. *Expert opinion on emerging drugs*, *15*(4), 615-34. doi:10.1517/14728214.2010.505603

Hillmann, a, Ozaki, T., & Winkelmann, W. (2000). Familial occurrence of osteosarcoma. A case report and review of the literature. *Journal of cancer research and clinical oncology*, *126*(9), 497-502.

Hoyal, C. R., Kammerer, S., Roth, R. B., Reneland, R., Marnellos, G., Kiechle, M., Schwarz-Boeger, U., et al. (2005). Genetic polymorphisms in DPF3 associated with risk of breast cancer and lymph node metastases. *Journal of carcinogenesis*, *4*, 13. doi:10.1186/1477-3163-4-13

Hua, Y., Qiu, Y., Zhao, A., Wang, X., Chen, T., Zhang, Z., Chi, Y., et al. (2011). Dynamic metabolic transformation in tumor invasion and metastasis in mice with LM-8 osteosarcoma cell transplantation. *Journal of proteome research*, *10*(8), 3513-21. doi:10.1021/pr200147g

Hurnphriessb, D. E., Nicodemusoll, C. F., Schillerll, V., & Stevensslll, R. L. (1992). The Human Serglycin Gene. *Journal of Biological Chemistry*, *267*(19), 13558-13563.

Iles, M. M. (2008). What can genome-wide association studies tell us about the genetics of common disease? (E. M. C. Fisher, Ed.)*PLoS genetics*, *4*(2), e33. Public Library of Science. doi:10.1371/journal.pgen.0040033

Institute, N. C. (2011). What is cancer? Retrieved from http://www.cancer.gov/cancertopics/cancerlibrary/what-is-cancer. Last accessed: 24/Aug/2011.

Ioannidis, J. P. a. (2007). Is molecular profiling ready for use in clinical decision making? *The oncologist*, *12*(3), 301-11. doi:10.1634/theoncologist.12-3-301

Isobe, T., Baba, E., Arita, S., Komoda, M., Tamura, S., Shirakawa, T., Ariyama, H., et al. (2011). Human STEAP3 maintains tumor growth under hypoferric condition. *Experimental cell research*, *317*(18), 2582-91. doi:10.1016/j.yexcr.2011.07.022

Jacobs, S. (2006). *Application of FFPE DNA to the Affymetrix GeneChip ® Mapping 500K Arrays*. Retrieved from: http://www.affymetrix.com. Last accessed: 03/Dec/2011

Jacobs, S., Thompson, E. R., Nannya, Y., Yamamoto, G., Pillai, R., Ogawa, S., Bailey, D. K., et al. (2007). Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays. *Cancer research*, *67*(6), 2544-51. doi:10.1158/0008-5472.CAN-06-3597

Jiang, N., Zhou, L.-Q., & Zhang, X.-Y. (2010). Downregulation of the nucleosome-binding protein 1 (NSBP1) gene can inhibit the in vitro and

in vivo proliferation of prostate cancer cells. *Asian journal of andrology*, *12*(5), 709-17. doi:10.1038/aja.2010.39

Jin, Y., Birlea, S. A., Fain, P. R., Gowan, K., Riccardi, S. L., Holland, P. J., Bennett, D. C., et al. (2011). Genome-wide analysis identifies a quantitative trait locus in the MHC class II region associated with generalized vitiligo age of onset. *The Journal of investigative dermatology*, *131*(6), 1308-12. doi:10.1038/jid.2011.12

Johnson, A. D., & O'Donnell, C. J. (2009). An open access database of genome-wide association results. *BMC medical genetics*, *10*, 6. doi:10.1186/1471-2350-10-6

Komar, A. A. (Ed.). (2009). *Single Nucleotide Polymorphisms* (Vol. 578). Totowa, NJ: Humana Press. doi:10.1007/978-1-60327-411-1

Lehmann, U., & Kreipe, H. (2001). Real-time PCR analysis of DNA and RNA extracted from formalin-fixed and paraffin-embedded biopsies. *Methods (San Diego, Calif.)*, *25*(4), 409-18. doi:10.1006/meth.2001.1263

Lewis, F., Maughan, N. J., Smith, V., Hillan, K., & Quirke, P. (2001). Unlocking the archive--gene expression in paraffin-embedded tissue. *The Journal of pathology*, *195*(1), 66-71. doi:10.1002/1096-9896(200109)195:1<66::AID-PATH921>3.0.CO;2-F

Li, X., Quigg, R. J., Zhou, J., Gu, W., Nagesh Rao, P., & Reed, E. F. (2008). Clinical utility of microarrays: current status, existing challenges and future outlook. *Current genomics*, *9*(7), 466-74. doi:10.2174/138920208786241199

Li, X.-J., Ong, C. K., Cao, Y., Xiang, Y.-Q., Shao, J.-Y., Ooi, A., Peng, L.-X., et al. (2011). Serglycin is a theranostic target in nasopharyngeal carcinoma that promotes metastasis. *Cancer research*, *71*(8), 3162-72. doi:10.1158/0008-5472.CAN-10-3557

Lian, Y., Yue, J., Han, M., Liu, J., & Liu, L. (2010). Analysis of the association between BTNL2 polymorphism and tuberculosis in Chinese Han population. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, *10*(4), 517-21. doi:10.1016/j.meegid.2010.02.006

Lisovich, A., Chandran, U. R., Lyons-Weiler, M. a, LaFramboise, W. a, Brown, A. R., Jakacki, R. I., Pollack, I. F., et al. (2011). A novel SNP

analysis method to detect copy number alterations with an unbiased reference signal directly from tumor samples. *BMC medical genomics*, *4*(1), 14. BioMed Central Ltd. doi:10.1186/1755-8794-4-14

Lockhart, D. J., & Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature*, *405*(6788), 827-36. doi:10.1038/35015701

Lu, H., Schölkopf, B., & Zhao, H. (Eds.). (2011). *Handbook of Statistical Bioinformatics*. *Media* (1st ed.). Berlin Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-16345-6

Lv, W.-ping, Dong, J.-hong, Shi, S., Huang, Z.-qiang, & Guo, D.-yu. (2008). [Gene expressions of DNA methyltransferase 1 and human leukocyte antigen-DRalpha in hepatocelluar carcinoma and their clinical significance]. *Beijing da xue xue bao. Yi xue ban = Journal of Peking University. Health sciences*, *40*(5), 543-7.

Lyons-Weiler, M., Hagenkord, J., Sciulli, C., Dhir, R., & Monzon, F. a. (2008). Optimization of the Affymetrix GeneChip Mapping 10K 2.0 Assay for routine clinical use on formalin-fixed paraffin-embedded tissues. *Diagnostic molecular pathology : the American journal of surgical pathology, part B*, *17*(1), 3-13. doi:10.1097/PDM.0b013e31815aca30

López-Vicente, L., Armengol, G., Pons, B., Coch, L., Argelaguet, E., Lleonart, M., Hernández-Losa, J., et al. (2009). Regulation of replicative and stress-induced senescence by RSK4, which is down-regulated in human tumors. *Clinical cancer research : an official journal of the American Association for Cancer Research*, *15*(14), 4546-53. doi:10.1158/1078-0432.CCR-08-3159

Malaver, M., Helman, L. J., & Brian, O. (2008). Sarcomas of Bone. *Cancer Principles & Practice of Oncology* (8th ed., pp. 1794-1834). Wolters Kluwer Health.

Man, T.-K., Lu, X.-Y., Jaeweon, K., Perlaky, L., Harris, C. P., Shah, S., Ladanyi, M., et al. (2004). Genome-wide array comparative genomic hybridization analysis reveals distinct amplifications in osteosarcoma. *BMC cancer*, *4*, 45. doi:10.1186/1471-2407-4-45

Mandahl, N. (1986). Multiple cytogenetic abnormalities in a case of osteosarcoma. *Cancer Genetics and Cytogenetics*, *23*(3), 257-260. doi:10.1016/0165-4608(86)90186-X

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine*, *363*(2), 166-76. doi:10.1056/NEJMra0905980

Masuda, Y., Suzuki, M., Piao, J., Gu, Y., Tsurimoto, T., & Kamiya, K. (2007). Dynamics of human replication factors in the elongation phase of DNA replication. *Nucleic acids research*, *35*(20), 6904-16. doi:10.1093/nar/gkm822

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, *3*(11), e442. doi:10.1371/journal.pmed.0030442

Matoba, K., Iizuka, N., Gondo, T., Ishihara, T., Yamada-Okabe, H., Tamesa, T., Takemoto, N., et al. (2005). Tumor HLA-DR expression linked to early intrahepatic recurrence of hepatocellular carcinoma. *International journal of cancer. Journal international du cancer*, *115*(2), 231-40. doi:10.1002/ijc.20860

Matsuguchi, T. (2009). [JNK signaling in osteoblast differentiation]. *Seikagaku. The Journal of Japanese Biochemical Society*, *81*(8), 703-7.

Miguel-Velado, E., Pérez-Carretero, F. D., Colinas, O., Cidad, P., Heras, M., López-López, J. R., & Pérez-García, M. T. (2010). Cell cycle-dependent expression of Kv3.4 channels modulates proliferation of human uterine artery smooth muscle cells. *Cardiovascular research*, *86*(3), 383-91. doi:10.1093/cvr/cvq011

Min, S., Choi, J. H., Lee, S. Y., & Yoo, N. C. (2009). Applications of DNA Microarray in Disease Diagnostics. *Review Literature And Arts Of The Americas*, *19*(October 2008), 635-646. doi:10.4014/jmb.0803.226

Model, T. D., Mapping, H., Set, A., Model, R. L., & Snp, E. (2006). BRLMM : an Improved Genotype Calling Method for the GeneChip ® Human Mapping 500K Array Set. *ReVision*, 1-18.

Murata, H., Kusuzaki, K., Takeshita, H., Hirasawa, Y., Ashihara, T., Abe, T., & Inazawa, J. (1998). Aberrations of chromosomes 1 and 17 in six human osteosarcoma cell lines using double-target fluorescence in situ hybridization. *Cancer genetics and cytogenetics*, *107*(1), 7-10.

Müller, P., Kuttenkeuler, D., Gesellchen, V., Zeidler, M. P., & Boutros, M. (2005). Identification of JAK/STAT signalling components by genome-

wide RNA interference. *Nature*, *436*(7052), 871-5. doi:10.1038/nature03869

Ohashi, Y., Kaneko, S. J., Cupples, T. E., & Young, S. R. (2004). Ubiquinol cytochrome c reductase (UQCRFS1) gene amplification in primary breast cancer core biopsy samples. *Gynecologic oncology*, *93*(1), 54-8. doi:10.1016/j.ygyno.2004.01.019

Olshen, A. B., Bengtsson, H., Neuvial, P., Spellman, P. T., Olshen, R. a, & Seshan, V. E. (2011). Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics (Oxford, England)*, *27*(15), 2038-46. doi:10.1093/bioinformatics/btr329

Paik, S., Kim, C.-yeul, Song, Y.-kuk, & Kim, W.-seop. (2005). Technology insight: Application of molecular techniques to formalin-fixed paraffin-embedded tissues from breast cancer. *Nature clinical practice. Oncology*, *2*(5), 246-54. doi:10.1038/ncponc0171

Papachristou, D. J., Batistatou, A., Sykiotis, G. P., Varakis, I., & Papavassiliou, A. G. (2003). Activation of the JNK-AP-1 signal transduction pathway is associated with pathogenesis and progression of human osteosarcomas. *Bone*, *32*(4), 364-71.

Patel, R. S., & Ye, S. (2011). Genetic determinants of coronary heart disease: new discoveries and insights from genome-wide association studies. *Heart (British Cardiac Society)*, hrt.2010.219675-. doi:10.1136/hrt.2010.219675

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., et al. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. *European journal of human genetics : EJHG*, *18*(1), 111-7. Nature Publishing Group. doi:10.1038/ejhg.2009.115

Percinel, S., Sak, S.D., Erinanc, H., Savas, B., Sertcelik, A., & Okten, I. (2008). Extraskeletal osteosarcoma with rhabdomyosarcomatous differentiation in local recurrence and lung metastases or so-called malignant mesenchymoma of soft tissue: A phenomenon related with chemotherapy?. *The Internet Journal of Pathology*, *7* (1). Retrieved from http://www.ispub.com/journal/the-internet-journal-of-pathology/volume-7-number-1.html. Last accessed: 05/Feb/2012.

Picci, P. (2007). Osteosarcoma (osteogenic sarcoma). *Orphanet journal of rare diseases*, *2*, 6. doi:10.1186/1750-1172-2-6

Pique-Regi, R., Cáceres, A., & González, J. R. (2010). R-Gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC bioinformatics*, *11*, 380. doi:10.1186/1471-2105-11-380

Promsuwicha, O., & Auewarakul, C. U. (2009). Positive and negative predictive values of HLA-DR and CD34 in the diagnosis of acute promyelocytic leukemia and other types of acute myeloid leukemia with recurrent chromosomal translocations. *Asian Pacific journal of allergy and immunology / launched by the Allergy and Immunology Society of Thailand*, *27*(4), 209-16.

Purcell, S. (2010). PLINK (1.07) Documentation. Retrieved from http://pngu.mgh.harvard.edu/~purcell/plink/dist/plink-doc-1.07.pdf. Last accessed: 07/Jan/2012

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a R., Bender, D., Maller, J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, *81*(3), 559-75. doi:10.1086/519795

Puschmann, A., Verbeeck, C., Heckman, M. G., Soto-Ortolaza, A. I., Lynch, T., Jasinska-Myga, B., Opala, G., et al. (2011). Human leukocyte antigen variation and Parkinson's disease. *Parkinsonism & related disorders*, *17*(5), 376-8. doi:10.1016/j.parkreldis.2011.03.008

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing. Vienna, Austria*. ISBN 3-900051-07-0. Retrieved from: url=http://www.R-project.org. Last accessed: 05/Feb/2012.

Rech, A., Castro, C. G., Mattei, J., Gregianin, L., Di Leone, L., David, A., Rivero, L. F., et al. (2004). [Clinical features in osteosarcoma and prognostic implications]. *Jornal de pediatria*, *80*(1), 65-70.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., et al. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444-54. doi:10.1038/nature05329

Rieger, P. T. (2004). The biology of cancer genetics. *Seminars in oncology nursing*, *20*(3), 145-54. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/15491024. Last accessed: 07/Jan/2012

Savaş, K., & Başkanliği, D. (2006). Sağlik bakanliği. *Cancer*, 2004-2006.

Scarlett, A., Parsons, M. P., Hanson, P. L., Sidhu, K. K., Milligan, T. P., & Burrin, J. M. (2008). Thyroid hormone stimulation of extracellular signal-regulated kinase and cell proliferation in human osteoblast-like cells is initiated at integrin alphaVbeta3. *The Journal of endocrinology*, *196*(3), 509-17. doi:10.1677/JOE-07-0344

Schajowicz, F., Sissons, H. a., & Sobin, L. H. (1995). The World Health Organization's histologic classification of bone tumors. A commentary on the second edition. *Cancer*, *75*(5), 1208-1214. doi:10.1002/1097-0142(19950301)75:5<1208::AID-CNCR2820750522>3.0.CO;2-F

Sellick, G. S., Longman, C., Tolmie, J., Newbury-Ecob, R., Geenhalgh, L., Hughes, S., Whiteford, M., et al. (2004). Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays. *Nucleic acids research*, *32*(20), e164. doi:10.1093/nar/gnh163

Selvarajah, S., Yoshimoto, M., Ludkovski, O., Park, P. C., Bayani, J., Thorner, P., Maire, G., et al. (2008). Genomic signatures of chromosomal instability and osteosarcoma progression detected by high resolution array CGH and interphase FISH. *Cytogenetic and genome research*, *122*(1), 5-15. doi:10.1159/000151310

Shi, S.-R., Datar, R., Liu, C., Wu, L., Zhang, Z., Cote, R. J., & Taylor, C. R. (2004). DNA extraction from archival formalin-fixed, paraffin-embedded tissues: heat-induced retrieval in alkaline solution. *Histochemistry and cell biology*, *122*(3), 211-8. doi:10.1007/s00418-004-0693-x

Siwoski, A., Ishkanian, A., Garnis, C., Zhang, L., Rosin, M., & Lam, W. L. (2002). An efficient method for the assessment of DNA quality of archival microdissected specimens. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, *15*(8), 889-92. doi:10.1097/01.MP.0000024288.63070.4F

Song, D., Liu, Y., Han, Y., Shang, G., Hua, S., Zhang, H., Guo, S., et al. (2002). [Study on the gestational diabetes mellitus and histocompatibility human leukocyte antigen DRB allele polymorphism]. *Zhonghua fu chan ke za zhi*, *37*(5), 284-6.

Sottnik, J. L., Lori, J. C., Rose, B. J., & Thamm, D. H. (2011). Glycolysis inhibition by 2-deoxy-D: -glucose reverts the metastatic phenotype in vitro and in vivo. *Clinical & experimental metastasis*, 865-875. doi:10.1007/s10585-011-9417-5

Strauss, S. J., Ng, T., Mendoza-Naranjo, A., Whelan, J., & Sorensen, P. H. B. (2010). Understanding micrometastatic disease and Anoikis resistance in ewing family of tumors and osteosarcoma. *The oncologist*, *15*(6), 627-35. doi:10.1634/theoncologist.2010-0093

Styrkarsdottir, U., Halldorsson, B. V., Gretarsdottir, S., Gudbjartsson, D. F., Walters, G. B., Ingvarsson, T., Jonsdottir, T., et al. (2008). Multiple genetic loci for bone mineral density and fractures. *The New England journal of medicine*, *358*(22), 2355-65. doi:10.1056/NEJMoa0801197

Sun, W., Wright, F. A., Tang, Z., Nordgard, S. H., Van Loo, P., Yu, T., Kristensen, V. N., et al. (2009). Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic acids research*, *37*(16), 5365-77. doi:10.1093/nar/gkp493

Thakur, A., Sun, Y., Bollig, A., Wu, J., Biliran, H., Banerjee, S., Sarkar, F. H., et al. (2008). Anti-invasive and antimetastatic activities of ribosomal protein S6 kinase 4 in breast cancer cells. *Clinical cancer research : an official journal of the American Association for Cancer Research*, *14*(14), 4427-36. doi:10.1158/1078-0432.CCR-08-0458

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299-320. doi:10.1038/nature04226

Ting Lee, M.-L. (2004). *Analysis of Microarray Gene Expression Data*. *Design* (1st ed.). Boston: Kluwer Academic Publishers. Retrieved from http://kluweronline.com. Last accessed: 01/Jan/2012.

Ueno, K., Hirata, H., Majid, S., Tabatabai, Z., Hinoda, Y., & Dahiya, R. (2011). IGFBP-4 activates the Wnt/beta-catenin signaling pathway and induces M-CAM expression in human renal cell carcinoma. *International journal of cancer. Journal international du cancer*, *129*(10), 2360-9. doi:10.1002/ijc.25899

Ustünkar, G., & Aydın Son, Y. (2011). METU-SNP: An Integrated Software System for SNP-Complex Disease Association Analysis. *Journal of integrative bioinformatics*, *8*(1), 187. doi:10.2390/biecoll-jib-2011-187

VanHouten, J., Sullivan, C., Bazinet, C., Ryoo, T., Camp, R., Rimm, D. L., Chung, G., et al. (2010). PMCA2 regulates apoptosis during mammary gland involution and predicts outcome in breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(25), 11405-10. doi:10.1073/pnas.0911186107

Venables, W. N., & Smith, D. M. (2011). *An Introduction to R. Development* (Vol. 0).

Wan, X., Mendoza, A., Khanna, C., & Helman, L. J. (2005). Rapamycin Inhibits Ezrin-Mediated Metastatic Behavior in a Murine Model of Osteosarcoma. *Cancer Research*, (6), 2406-2411.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., & Todd, J. a. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature reviews. Genetics*, *6*(2), 109-18. doi:10.1038/nrg1522

White, G. R., Varley, J. M., & Heighway, J. (1998). Isolation and characterization of a human homologue of the latrophilin gene from a region of 1p31.1 implicated in breast cancer. *Oncogene*, *17*(26), 3513-9. doi:10.1038/sj.onc.1202487

Wijnen, P. A., Voorter, C. E., Nelemans, P. J., Verschakelen, J. A., Bekers, O., & Drent, M. (2011). Butyrophilin-like 2 in pulmonary sarcoidosis: a factor for susceptibility and progression? *Human immunology*, *72*(4), 342-7. doi:10.1016/j.humimm.2011.01.011

Wong, K.-K., Tsang, Y. T. M., Shen, J., Cheng, R. S., Chang, Y.-M., Man, T.-K., & Lau, C. C. (2004). Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. *Nucleic acids research*, *32*(9), e69. doi:10.1093/nar/gnh072

Wu, P. K., Chen, W. M., Chen, C. F., Lee, O. K., Haung, C. K., & Chen, T. H. (2009). Primary osteogenic sarcoma with pulmonary metastasis: clinical results and prognostic factors in 91 patients. *Japanese journal of clinical oncology*, *39*(8), 514-22. doi:10.1093/jjco/hyp057

Xue, J.-F., Hua, F., Lv, Q., Lin, H., Wang, Z.-Y., Yan, J., Liu, J.-W., et al. (2010). DEDD negatively regulates transforming growth factor-beta1 signaling by interacting with Smad3. *FEBS letters*, *584*(14), 3028-34. doi:10.1016/j.febslet.2010.05.043

Yokoo, S., Masuda, S., Yonezawa, A., Terada, T., Katsura, T., & Inui, K.-ichi. (2008). Significance of organic cation transporter 3 (SLC22A3) expression for the cytotoxic effect of oxaliplatin in colorectal cancer. *Drug metabolism and disposition: the biological fate of chemicals*, *36*(11), 2299-306. doi:10.1124/dmd.108.023168

Yu, J. Z., Warycha, M. A., Christos, P. J., Darvishian, F., Yee, H., Kaminio, H., Berman, R. S., et al. (2008). Assessing the clinical utility of measuring Insulin-like Growth Factor Binding Proteins in tissues and sera of melanoma patients. *Journal of translational medicine*, *6*, 70. doi:10.1186/1479-5876-6-70

Zhang, Zhiyu, Qiu, Y., Hua, Y., Wang, Y., Chen, T., Zhao, A., Chi, Y., et al. (2010). Serum and urinary metabonomic study of human osteosarcoma. *Journal of proteome research*, *9*(9), 4861-8. doi:10.1021/pr100480r

Üstünkar, Gürkan, Özöğür-Akyüz, S., Weber, G. W., Friedrich, C. M., & Aydın Son, Y. (2011). Selection of representative SNP sets for genome-wide association studies: a metaheuristic approach. *Optimization Letters*. doi:10.1007/s11590-011-0419-7

Üstünkar, Gurkan. (2011). *An integrative approach to structured SNP prioritization and representative SNP selection for genome-wide association studies*. *Discovery*. Middle East Technical University.

APPENDIX A

PHOTOGRAPHS OF HEMATOXYLIN-EOSIN STAINED SECTIONS

During manual microdissection, the area of interest (normal and tumor) was identified by examination of hematoxylin-eosin (H&E) stained 5 µm sections under light microscope by an expert pathologist (Prof. Dr. Mükerrem Safalı, Department of Pathology, Gülhane Military Medical Academy, Ankara). The tumor and normal areas selected for each patient are represented below as 10X magnified pictures of H&E stained slides.

Patient 1:

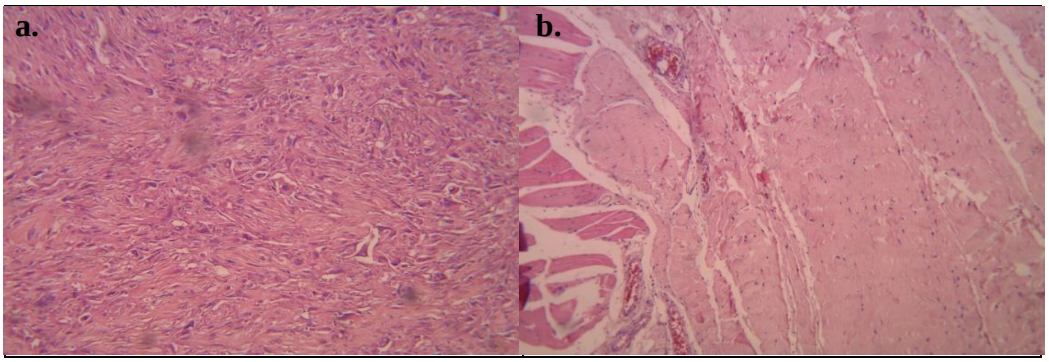


Figure A.1 a) Tumor tissue, b) Normal tissue of Patient 1.

Patient 2:

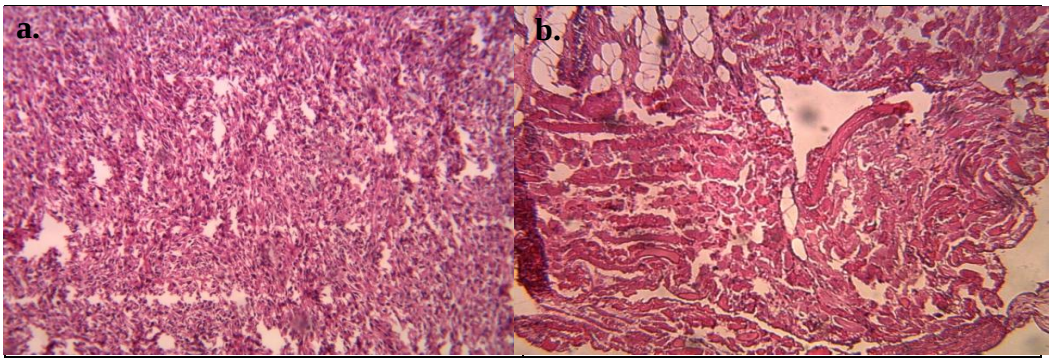


Figure A.2 a) Tumor tissue, b) Normal tissue of Patient 2.

Patient 3:

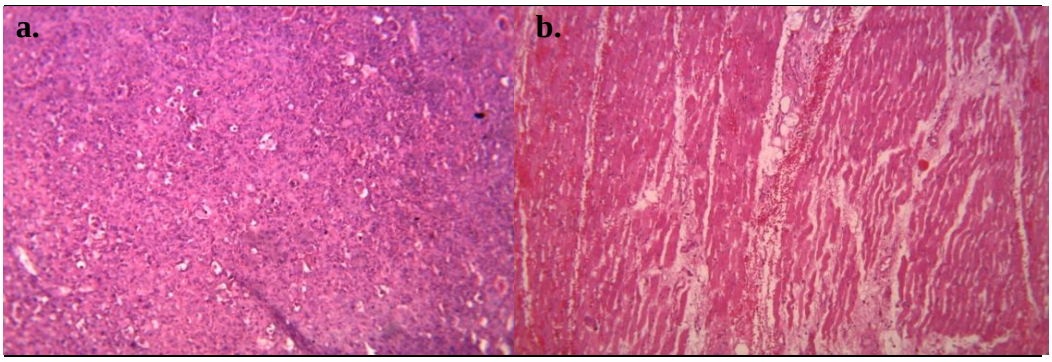


Figure A.3 a) Tumor tissue, b) Normal tissue of Patient 3.

Patient 4:

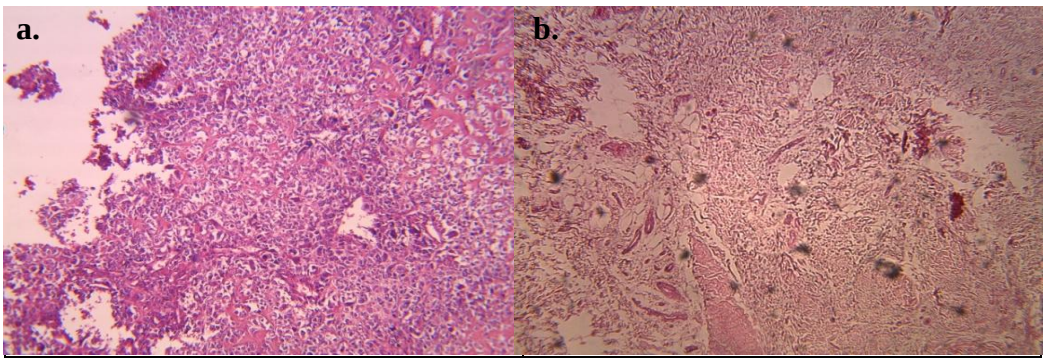


Figure A.4 a) Tumor tissue, b) Normal tissue of Patient 4.

Patient 5:

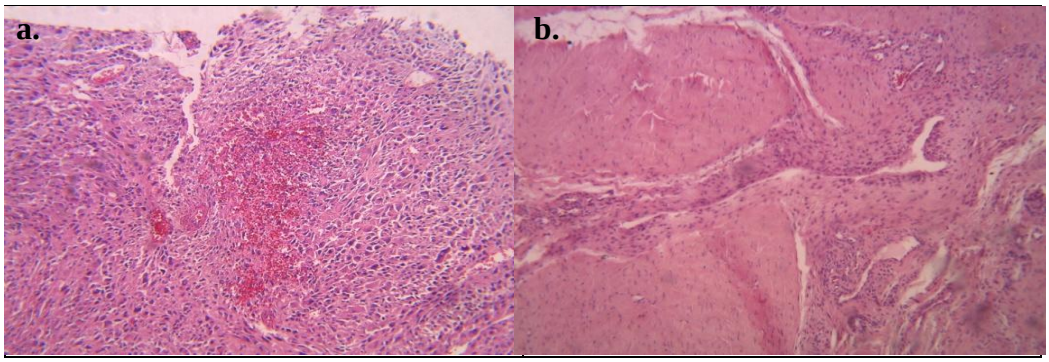


Figure A.5 a) Tumor tissue, b) Normal tissue of Patient 5.

Patient 6:

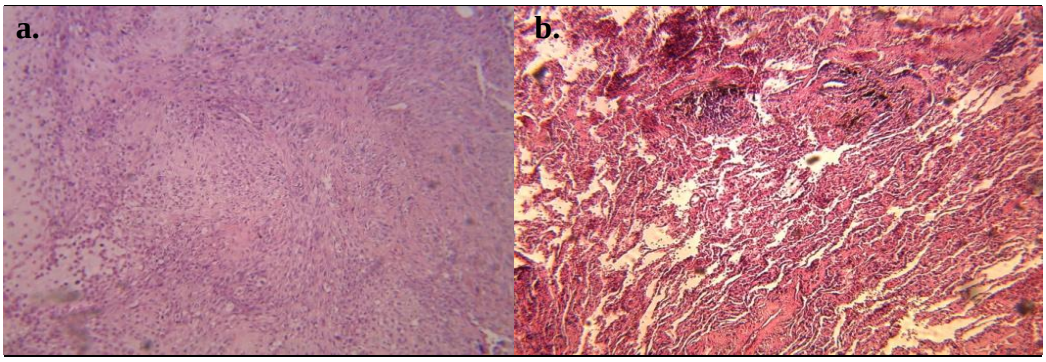


Figure A.6 a) Tumor tissue, b) Normal tissue of Patient 6

Patient 7:

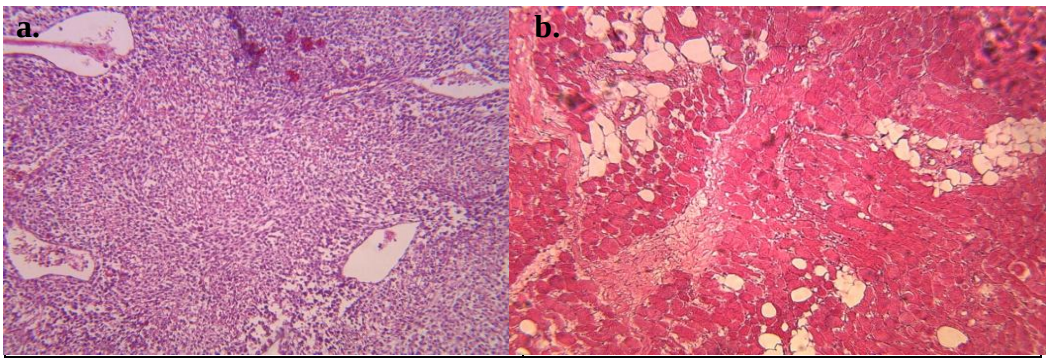


Figure A.7 a) Tumor tissue, b) Normal tissue of Patient 7.

Patient 8:

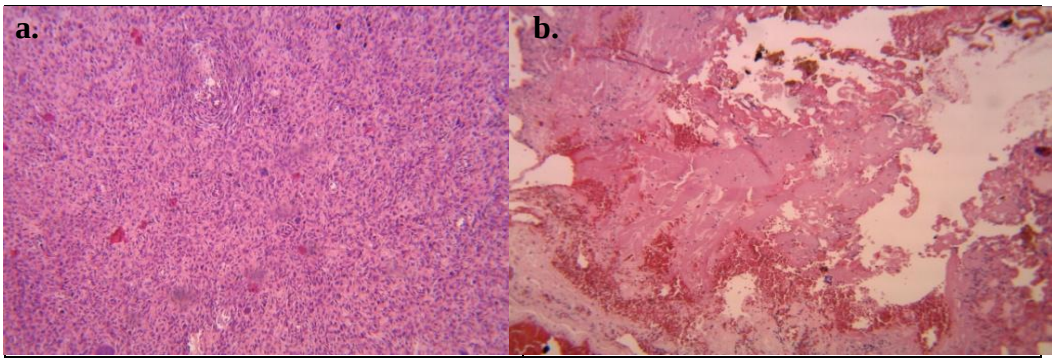


Figure A.8 a) Tumor tissue, b) Normal tissue of Patient 8.

Patient 9:

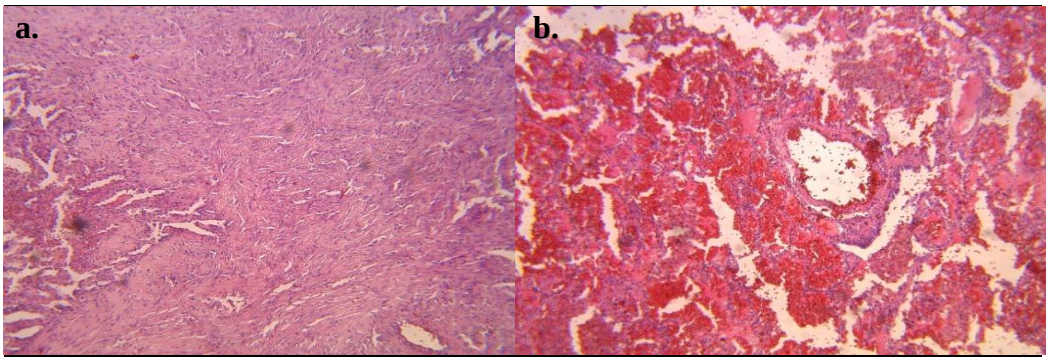


Figure A.9 a) Tumor tissue, b) Normal tissue of Patient 9.

Patient 10:

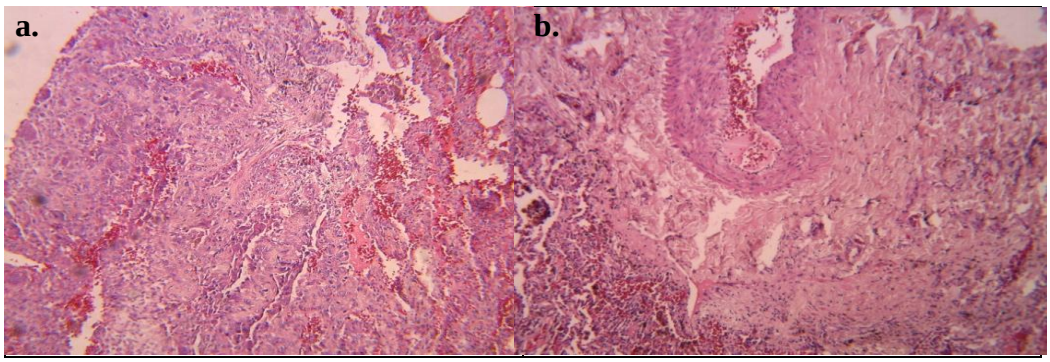


Figure A.10 a) Tumor tissue, b) Normal tissue of Patient 10.

APPENDIX B

BUFFERS AND SOLUTIONS

Agarose Gel Electrophoresis

TAE (Tris-Acetate-EDTA) Buffer (50X, 1 L)

Tris-base (MW: 121.14)	242 g
Glacial acetic acid	57.1 ml
EDTA disodium dehydrate-0.5 M (MW: 372.24)	100 ml

Volume was completed to 1 L with distilled water and pH is adjusted to 8.5. After autoclaved, solution was stored at 4°C.

Ethidium Bromide (EtBr) Solution

EtBr (MW: 394.31)	10 mg
dH_2O	1 ml

Solution was stored in dark, at 4°C.

6X DNA Loading Dye (Fermentas)

Tris-HCl-10mM (pH 7.6)	Bromophenol blue-0.03%
Xylene cyanol FF-0.03%	Glycerol-60%
EDTA-60mM	

Microarray Analysis Solutions

Wash Solutions

Wash A: Non-Stringent Wash Buffer

20X SSPE	300 ml
10% Tween-20	1.0 ml
Water	699 ml

Wash B: Stringent Wash Buffer

20X SSPE	30 ml
10% Tween-20	1.0 ml
Water	969 ml

Staining solutions

Array Holding Buffer

MES Stock Buffer (12X)	8.3 ml
5 M NaCl	18.5 ml
Tween-20 (10%)	0.1 ml
Water	73.1 ml
Total	100 ml

Stain Buffer

H_2O	800.04 µl
SSPE (20X)	360 µl
Tween-20 (3%)	3.96 µl
Denhardt's (50X)	24 µl
Total	1188 µl

SAPE Stain Solution

Stain Buffer	594 µl
1 mg/mL Streptavidin Phycoerythrin (SAPE)	6 µl
Total	600 µl

Antibody Stain Solution

Stain Buffer	594 µl
0.5 mg/mL biotinylated antibody	6 µl
Total	600 µl

APPENDIX C

STATISTICAL BACKGROUND FOR SNP - COMPLEX DISEASE ASSOCIATION ANALYSIS

Hardy-Weinberg Equilibrium

In association studies, each SNP is tested for HWE before association testing with phenotypes. According to HW law, the genotype and allele frequencies of a large, randomly mating population stay constant from generation to generation when migration, mutation, and natural selection do not take place. Hence, HWE can be defined as the stable distribution of frequencies of the genotypes AA, Aa, and aa in the proportions p^2, 2pq, and q^2, respectively, where p and q are frequencies of the alleles A and a. This stable distribution indicates that the maternal and paternal alleles of an individual at a particular locus are statistically independent. A significant departure from HWE for a SNP may point to nonrandom mating and possibly population stratification, nonrandom genotyping error, or missing genotype data.

An SNP is tested for HWE by comparison of the observed genotype counts in a sample with those expected under HWE. During comparison, goodness-of-fit χ^2 test is applied. Allele frequencies (p and q=1−p) of the SNP is calculated by the proportions of alleles in the sample. The expected genotype counts are determined using the HWE expected frequencies Np^2, 2Npq, and Nq^2, where N is the number of individuals genotyped. After that, observed and expected counts are compared by the goodness-of-fit χ^2 test.

For ascertained samples, such as case-control samples, the population prevalence of the trait is expected to be low. For this reason, Hardy-Weinberg testing is only performed in the controls because there should be departure from HWE among cases for any polymorphism associated with phenotype. SNPs with genotypes significantly deviated from HW–expected proportions are excluded from association analyses. Omission of SNPs from association analyses depends on multiple criteria which include the amount of SNPs tested and the call rate (proportion of successfully genotyped SNPs). Generally, SNPs with HWE *p* values less than 0.01 or 0.001 are excluded from association analyses.

Minor Allele Frequency (MAF)

The minor allele frequency is the frequency of the less common allele of a marker in a given population. It has a value between 0 and 0.5. SNPs with a minor allele frequency of 0.05 or greater were studied by the HapMap project.

Fisher's Exact Test

Fisher's exact test is a statistical test used in the analysis of contingency tables where sample sizes are small. In this test, the significance of the deviation from a null hypothesis can be calculated exactly rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity, as with many statistical tests.

Fisher's Combination Test

This test is used for data fusion or meta-analysis (analysis of analyses). By this test, the results from several independent tests having the same hypothesis (H_0) are combined. In GWAS analysis, Fisher's combination test is used as a summary statistic to give an overall value of association for a SNP between studies. Moreover, it is utilized to combine the p values of independent SNPs in order to get combined p value for regions, genes or pathways.

False Discovery Rate

False discovery rate (FDR) control is a statistical method used in multiple hypotheses testing to correct for multiple comparisons. FDR controls the expected proportion of incorrectly rejected null hypotheses (type I errors).

Bonferroni Adjustment

Bonferroni adjustment, likewise (FDR), is a statistical method to adjust the p-values or to correct for multiple testing issues. During multiple hypotheses testing, the possibility of false negative results to be observed is high. In GWAS, Bonferroni adjustment can use either the actual number of tests performed (i.e. SNPs genotyped) or a theoretical value which is the total number of tests possible (i.e. all SNPs). Bonferroni correction method assumes that all the tests (SNPs) are independent.

APPENDIX D

FILE STRUCTURE OF PLINK STATISTICS FILES

During analysis with METU-SNP, basic statistics was applied to our data with the integrated PLINK software. In this step, PLINK generates the descriptive statistics files (freq.frq, missing.imiss, missing.lmiss, hardy.hwe). Table D.1 gives details of these statistics files.

Table D.1 File structure of PLINK descriptive statistics files.

File	Field name	Description
freq.frq	CHR	Chromosome
	SNP	SNP identifier
	A1	Allele 1 code (minor allele)
	A2	Allele 2 code (major allele)
	MAF	Minor allele frequency
	NCHROBS	Non-missing allele count
missing.imiss	FID	Family ID
	IID	Individual ID
	MISS_PHENO	Missing phenotype? (Y/N)
	N_MISS	Number of missing SNPs
	N_GENO	Number of non-obligatory missing genotypes
	F_MISS	Proportion of missing SNPs
missing.lmiss	SNP	SNP identifier
	CHR	Chromosome
	N_MISS	Number of individuals missing this SNP
	N_GENO	Number of non-obligatory missing genotypes
	F_MISS	Proportion of sample missing for this SNP
hardy.hwe	SNP	SNP identifier
	TEST	Code indicating sample
	A1	Minor allele code
	A2	Major allele code
	GENO	Genotype counts: 11/12/22
	O(HET)	Observed heterozygosity
	E(HET)	Expected heterozygosity
	P	H-W p-value

APPENDIX E

AHP SCORING SCHEME

In METU-SNP analysis, SNPs are prioritized according to both genetic information in databases and statistical information obtained by GWAS. The scores each of each function (represents leaf nodes in the hierarchical tree) are given in Table E.1.

Table E.1 Scoring scheme of SNPs.

	Leaf	Description	Score
GWAS Related	0.1	Individual SNP	0.033616
	0.2.1	Significant Gene- Via LD	0.01598
	0.2.2	Significant Gene- Via Direct	0.12099
	0.2.3	Significant Gene - Via Pathway	0.053266
	0.3.1	Significant Pathway Gene - Via LD	0.01465
	0.3.2	Significant Pathway Gene - Via Direct	0.093825
	0.3.3	Significant Pathway Gene - Via Pathway	0.04738
	1.2.1.1	Disease Gene - Via LD	0.036593
	1.2.1.2	Disease Gene - Via Direct	0.186016
	1.2.1.3	Disease Gene - Via Pathway	0.081725
	1.2.2.1.1	Other Gene - Other Disease - Via LD	0.005756
	1.2.2.1.2	Other Gene - Other Disease - Via Direct	0.01818
	1.2.2.1.3	Other Gene - Other Disease - Via Pathway	0.011161
	1.2.2.2.1	Other Gene - Neutral - Via LD	0.00145
	1.2.2.2.2	Other Gene - Neutral - Via Direct	0.004579
	1.2.2.2.3	Other Gene - Neutral - Via Pathway	0.002811
GENETIC	1.1.1	Vertebrate	0.037841
	1.1.2.1	Mammalian - Significant Mouse ECR	0.04532
	1.1.2.2	Mammalian - Other Mammalian	0.023347
	1.3.1.1.1	Non-Coding- UTR-3 - MiRNA Prediction	0.001142
	1.3.1.1.2	Non-Coding- UTR-3 - No MiRNA Prediction	0.000604
	1.3.1.2.1	Non-Coding- UTR-5 - CpG Island	0.002017
	1.3.1.2.2	Non-Coding- UTR-5 - No CpG Island	0.00063
	1.3.1.3	Non-Coding - Intronic	0.000825
	1.3.1.4	Non-Coding - Near Gene 3	0.001476
	1.3.1.5.1	Non-Coding - Near Gene 5 - CpG Island	0.002467
	1.3.1.5.2	Non-Coding - Near Gene 5 - No CpG Island	0.000571

Table E.1 continued.		
1.3.1.6	Non-Coding - Splice3	0.005295
1.3.1.7	Non-Coding - Splice 5	0.006597
1.3.2.1	Coding - Frameshift	0.103733
1.3.2.3.1	Coding - CDS Non Syn - Polyphen Benign	0.001997
1.3.2.3.2	Coding - CDS Non Syn - Possibly Damaging	0.004713
1.3.2.3.3	Coding - CDS Non Syn - Probably Damaging	0.009187
1.3.2.3.4	Coding - CDS Non Syn - Completely Determine	0.024045

APPENDIX F

AROMA PACKAGE R VIGNETTES FOR CNV ANALYSIS METHODS

Vignette is an R jargon for documentation of the codes that are used in running the R packages. They are actually optional supplemental documentation in addition to required boilerplate documentation for R functions and datasets. R vignettes of aroma package that has been utilized in this study are provided in this section.

CalMaTe

```
library("aroma.affymetrix");
library("calmate");
verbose <- Arguments$getVerbose(-8, timestamp=TRUE);

csR <- AffymetrixCelSet$byName("SET1,All",
chipType="Mapping250K_Sty");
print(csR);

dsList <- doASCRMAv2(csR, plm="RmaCnPlm", verbose=verbose);
print(dsList);

cmt <- CalMaTeCalibration(dsList);
print(cmt);

dsCList <- process(cmt, verbose=verbose);
print(dsCList);

names <- getNames(dsList$total);

patientIDs <- c(1N, 2N, 3N, 5N, 6N, 7N, 8N, 9N, 10N);
sampleTypes <- c("tumor", "normal");
```

```
pids <- rep(patientIDs, each=length(sampleTypes));
types <- rep(sampleTypes, times=length(patientIDs));

mat <- cbind(names, pids, types);
str(mat);

idxsN <- which(mat[, "types"] == "normal");

cmtN <- CalMaTeCalibration(dsList, tags=c("*", "normalReferences"),
references=idxsN);
print(cmtN);

dsCNList <- process(cmtN, verbose=verbose);
print(dsCNList);

extractSignals <- function(dsList, sampleName, reference=c("none",
"median"), refIdxs=NULL, ..., verbose=FALSE) {
    reference <- match.arg(reference);
      idx <- indexOf(dsList$total, sampleName);
      dfT <- getFile(dsList$total, idx);
      dfB <- getFile(dsList$fracB, idx);
      tcn <- extractRawCopyNumbers(dfT, logBase=NULL, ...,
verbose=verbose);
      baf <- extractRawAlleleBFractions(dfB, ..., verbose=verbose);
      if (reference == "median") {
        if (!is.null(refIdxs)) {
          dsR <- extract(dsList$total, refIdxs);
        } else {
          dsR <- dsList$total;
        }
        dfTR <- getAverageFile(dsR, verbose=verbose);
        tcnR <- extractRawCopyNumbers(dfTR, logBase=NULL, ...,
verbose=verbose);
        tcn <- divideBy(tcn, tcnR);
        tcn$y <- 2*tcn$y;
      }
    list(tcn=tcn, baf=baf);
 } # extractSignals()

sampleName <- "T10"
pch <- 19;
cex <- 0.8;
```

```
snT <- sampleName;
chr <-1
for (normalRefs in c(TRUE, FALSE)) {
  if (normalRefs) {
    figName <- sprintf("%s,Chr%02d,CalMaTe,normalReferences", snT, chr);
    dataT <- extractSignals(dsList, sampleName=snT, chromosome=chr,
reference="median", refIdxs=idxsN, verbose=verbose);
    dataTC <- extractSignals(dsCNList, sampleName=snT, chromosome=chr,
verbose=verbose);
  } else {
    figName <- sprintf("%s,Chr%02d,CalMaTe", snT, chr);
    dataT <- extractSignals(dsList, sampleName=snT, chromosome=chr,
reference="median", verbose=verbose);
    dataTC <- extractSignals(dsCList, sampleName=snT, chromosome=chr,
verbose=verbose);
  }


  toPNG(figName, width=1200, {  # (save to figures/)
    subplots(4, ncol=1);
    par(mar=c(2,5,1,1)+0.1, cex=cex, cex.lab=2.4, cex.axis=2.2);

    plot(dataT$tcn, ylim=c(0,4), pch=pch);
    plot(dataT$baf, pch=pch);
    plot(dataTC$tcn, ylim=c(0,4), pch=pch);
    plot(dataTC$baf, pch=pch);
  });
}
```

Tumorboost

```
library("aroma.cn");
log <- verbose <- Arguments$getVerbose(-8, timestamp=TRUE);
rootPath <- "totalAndFracBData";
rootPath <- Arguments$getReadablePath(rootPath);
dataSets <- c("SET13,10thPair,ACC,-XY,BPN,-XY,RMA,FLN,-XY");

dataSet <- dataSets[1];
```

```
fnt <- function(names, ...) {
  pattern <- "^(O10-[0-9]{2}-[0-9]{2})-([0-9]{1}[A-Z])[-]*(.*)";
  gsub(pattern, "\\1,\\2,\\3", names);
} # fnt()

ds <- AromaUnitFracBCnBinarySet$byName(dataSet, chipType="*",
paths=rootPath);
setFullNamesTranslator(ds, fnt);
print(ds);

sampleNames <- sort(unique(getNames(ds)));
sampleName <- sampleNames[1];

pair <- indexOf(ds, sampleName);
stopifnot(length(pair) == 2);

types <- sapply(extract(ds,pair), FUN=function(df) getTags(df)[1]);
o <- order(types);
types <- types[o];
pair <- pair[o];

dsPair <- extract(ds, pair);
dsT <- extract(dsPair, 1);
print(dsT);

dsN <- extract(dsPair, 2);
print(dsN);

rootPath <- "callData";
rootPath <- Arguments$getReadablePath(rootPath);

genotypeTag <- "NGC";
gsN <- AromaUnitGenotypeCallSet$byName(dataSet, tags=genotypeTag,
chipType="*");
setFullNamesTranslator(gsN, fnt);

types <- sapply(gsN, FUN=function(df) getTags(df)[1]);
types <- gsub("[A-Z]$", "", types);
keep <- which(is.element(types, c("2")));
gsN <- extract(gsN, keep);
print(gsN);

dsList <- list(normal=dsN, tumor=dsT, callsN=gsN);
```

```
sampleNames <- getNames(dsList$normal);
dsList <- lapply(dsList, FUN=function(ds) {
  idxs <- indexOf(ds, sampleNames);
  extract(ds, idxs);
});
print(dsList);

dummy <- lapply(dsList, FUN=function(ds) print(getFile(ds,1)));

tbn <- TumorBoostNormalization(dsList$tumor, dsList$normal,
gcN=dsList$callsN, tags=c("*", "NGC"));
dsTN <- process(tbn, verbose=log);
setFullNamesTranslator(dsTN, fnt);
print(dsTN);

dsList <- list(normal=dsN, tumor=dsT, tumorN=dsTN, callsN=gsN);
rm(dsN, dsT, dsTN, gsN);
dsList <- lapply(dsList, FUN=function(ds) {
  idxs <- indexOf(ds, getNames(dsList$normal));
  extract(ds, idxs);
});


figPath <- Arguments$getWritablePath("figures");
siteTag <- getTags(ds);
siteTag <- paste(siteTag[-1], collapse=",");
print(siteTag);
ugp <- getAromaUgpFile(dsList$tumor);

chromosome <- 1;
chrTag <- sprintf("Chr%02d", chromosome);
units <- getUnitsOnChromosome(ugp, chromosome=chromosome);

platform <- getPlatform(ugp);
if (platform == "Affymetrix") {
  require("aroma.affymetrix") || throw("Package not loaded:
aroma.affymetrix");
  snpPattern <- "^SNP";
} else if (platform == "Illumina") {
  snpPattern <- "^rs[0-9]";
} else {
  throw("Unknown platform: ", platform);
}
```

```
unf <- getUnitNamesFile(ugp);
unitNames <- getUnitNames(unf, units=units);

keep <- (regexpr(snpPattern, unitNames) != -1);
units <- units[keep];

pos <- getPositions(ugp, units=units);

kk <- 1;
dfList <- lapply(dsList, FUN=getFile, kk);
beta <- lapply(dfList, FUN=function(df) df[units,1,drop=TRUE]);
beta <- as.data.frame(beta);
beta <- as.matrix(beta);
names <- colnames(beta);
names[names == "tumorN"] <- "normalized tumor";

x <- pos/1e6;
xlim <- range(x, na.rm=TRUE);
xlab <- "Position (Mb)";

width <- 840;
width <- 1280;
aspect <- 0.6*1/3;

ylim <- c(-0.05,1.05);
ylim <- c(-0.1,1.1);
ylab <- "Allele B Fraction";
cols <- as.integer(beta[,"callsN"] != 1) + 1L;

for (cc in 1:3) {
  tag <- colnames(beta)[cc];
  name <- names[cc];
  figName <- sprintf("%s,%s,%s,%s,fracB", siteTag, sampleName, tag,
chrTag);
  pathname <- filePath(figPath, sprintf("%s.png", figName));
  if (!isFile(pathname)) {
    fig <- devNew("png", pathname, label=figName, width=width,
height=aspect*width);
    par(mar=c(2.7,2.5,1.1,1)+0.1, tcl=-0.3, mgp=c(1.4,0.4,0), cex=2);
    par(mar=c(1.7,2.5,1.1,1)+0.1);
    plot(NA, xlim=xlim, ylim=ylim, xlab=xlab, ylab=ylab, axes=FALSE);
    axis(side=1);
    axis(side=2, at=c(0,1/2,1));
```

```
    points(x, beta[,cc], pch=".", col=cols);
    label <- sprintf("%s (%s)", sampleName, name);
    stext(side=3, pos=0, label);
    stext(side=3, pos=1, chrTag);
    devDone();
  }
}

rootPath <- "totalAndFracBData";
rootPath <- Arguments$getReadablePath(rootPath);
dsC <- AromaUnitTotalCnBinarySet$byName(dataSet, chipType="*",
paths=rootPath);
setFullNamesTranslator(dsC, fnt);
print(dsC);

pairC <- indexOf(dsC, sampleName);
stopifnot(length(pairC) == 2);

types <- sapply(extract(dsC, pairC), FUN=function(df) getTags(df)[1]);
o <- order(types);
types <- types[o];
pairC <- pairC[o];

dsPairC <- extract(dsC, pairC);

C <- extractMatrix(dsPairC, units=units);
C <- 2*C[,1]/C[,2];

ylim <- c(0,6);
ylab <- "Copy number";
figName <- sprintf("%s,%s,%s,CN", siteTag, sampleName, chrTag);
pathname <- filePath(figPath, sprintf("%s.png", figName));
if (!isFile(pathname)) {
  fig <- devNew("png", pathname, label=figName, width=width,
height=aspect*width);
  par(mar=c(2.7,2.5,1.1,1)+0.1, tcl=-0.3, mgp=c(1.4,0.4,0), cex=2);
  par(mar=c(1.7,2.5,1.1,1)+0.1);
  plot(NA, xlim=xlim, ylim=ylim, xlab=xlab, ylab=ylab, axes=FALSE);
  axis(side=1);
  axis(side=2, at=c(0,2,4,6));
  points(x, C, pch=".");
  label <- sprintf("%s", sampleName);
  stext(side=3, pos=0, label);
```

```
 stext(side=3, pos=1, chrTag);
 devDone();}
```

Paired PSCBS

```
library("aroma.affymetrix");
verbose <- Arguments$getVerbose(-10, timestamp=TRUE);

dataSet <- "SET13,10thPair";
chipType <- "Mapping250K_Sty";

csR <- AffymetrixCelSet$byName(dataSet, chipType=chipType);


pair <- c(T="O10-010-010-1T", N="O10-010-010-2N");
csR <- extract(csR, indexOf(csR, pair));
print(csR);

res <- doASCRMAv2(csR, plm="RmaCnPlm", verbose=verbose);

data <- extractPSCNArray(res$total);
dimnames(data)[[3]] <- names(pair);
str(data);

CT <- 2 * (data[,"total","T"] / data[,"total","N"]);

betaT <- data[,"fracB","T"];

betaN <- data[,"fracB","N"];ugp <- getAromaUgpFile(res$total);
chromosome <- ugp[,1,drop=TRUE];
x <- ugp[,2,drop=TRUE];


df <- data.frame(chromosome=chromosome, x=x, CT=CT, betaT=betaT,
betaN=betaN);

library("PSCBS");

fit <- segmentByPairedPSCBS(df, verbose=verbose);

segs <- getSegments(fit);
```

```
print(segs);

pairName <- paste(pair, collapse="vs");
chrTag <- sprintf("Chr%s", seqToHumanReadable(getChromosomes(fit)));

toPNG(pairName, tags=c(chrTag, "PairedPSCBS"), width=840,
aspectRatio=0.6, {
  plotTracks(fit);});
```

VN Reference Algorithm

```
source("setupAromaExtensions.R");
setupAromaExtensions();

rm(updateSamplesFile.ChromosomeExplorer);
log <- verbose <- Arguments$getVerbose(-8, timestamp=TRUE);
options(digits=4)

settings<-aromaDefaultSettings();

doCRMAv2("allData", chipType="Mapping250K_Sty", plm="RmaCnPlm");

cesN<-
aromaPreprocessing(dataSet="allData",chipType="Mapping250K_Sty",verbos
e=verbose)

ces1<-extract(cesN,"arrays1");
print(ces1)
ces2<-extract(cesN,"arrays2");
print(ces2)

settings<-aromaDefaultSettings();

ref<-aromaGetReference(ces1,ces2,settings,verbose);

ces1<-extract(cesN,"arrays1,1");
print(ces1)
segModel<-aromaSegmentation(ces1,ref,settings,verbose);
res<-aromaReports(segModel,"T1",settings,verbose);
```

APPENDIX G

TOP 100 SIGNIFICANT SNPs, GENES AND PATHWAYS

Table G.1 Top 100 SNPs according unadjusted *p*-value.

Rank	Chr	rs ID	*p*- value	Rank	Chr	rs ID	*p*- value
1	16	rs6499861	4.90E-02	32	5	rs456648	0.001638
2	10	rs10884554	4.90E-02	33	7	rs1024455	0.00175
3	7	rs12154602	6.55E-02	34	14	rs17103831	0.00175
4	19	rs7249598	0.000158	35	6	rs9392131	0.00175
5	4	rs28716003	0.0002965	36	11	rs7942692	0.00175
6	15	rs16946687	0.0003873	37	2	rs4533500	0.00175
7	14	rs17120272	0.0003873	38	9	rs10816858	0.00175
8	4	rs6828250	0.0003873	39	14	rs17093516	0.002159
9	1	rs6657368	0.0003873	40	14	rs11846180	0.002159
10	1	rs10923183	0.0003873	41	10	rs12252255	0.002159
11	6	rs16896159	0.0003873	42	14	rs11846985	0.002293
12	18	rs17769881	0.0003873	43	10	rs11595306	0.002293
13	1	rs5744297	0.0003961	44	19	rs10411891	0.002293
14	4	rs3899794	0.0006022	45	8	rs2433147	0.002293
15	14	rs17753747	0.0006022	46	6	rs16890791	0.002293
16	16	rs17313499	0.0006022	47	2	rs2032765	0.002293
17	11	rs2958625	0.0006022	48	12	rs4766651	0.002293
18	15	rs12594008	0.0006022	49	9	rs2781542	0.002293
19	8	rs2614062	0.0006022	50	1	rs530599	0.002293
20	17	rs9895307	0.0006022	51	10	rs2050440	0.002293
21	4	rs3775261	0.0006022	52	18	rs16954169	0.002293
22	4	rs6845976	0.0006022	53	15	rs2654982	0.002293
23	17	rs7210007	0.0009721	54	2	rs13429411	0.002293
24	7	rs221770	0.0009721	55	9	rs10491538	0.002293
25	13	rs17051671	0.001359	56	8	rs10103191	0.002293
26	7	rs2686524	0.001359	57	10	rs1339634	0.002293
27	4	rs6840808	0.001638	58	9	rs4979455	0.002293
28	7	rs12701745	0.001638	59	17	rs7217422	0.003368
29	8	rs17200329	0.001638	60	8	rs7820388	0.003368
30	2	rs10196667	0.001638	61	8	rs6983190	0.003368
31	13	rs9554197	0.001638	62	5	rs17407759	0.003368

Table G.1 continued.

Rank	Chr	rs ID	*p*- value	Rank	Chr	rs ID	*p*- value
63	11	rs1491678	0.003368	82	4	rs16992080	0.004197
64	1	rs12049331	0.003368	83	5	rs6889096	0.004197
65	2	rs16866027	0.003368	84	3	rs11926292	0.004197
66	13	rs4470044	0.004197	85	13	rs12429798	0.004197
67	20	rs12625716	0.004197	86	1	rs11210477	0.004197
68	20	rs6066932	0.004197	87	19	rs4805272	0.004197
69	15	rs4887181	0.004197	88	10	rs17558323	0.004197
70	2	rs6760491	0.004197	89	9	rs1322307	0.004197
71	10	rs7915493	0.004197	90	16	rs7192189	0.004197
72	11	rs10768884	0.004197	91	11	rs11042886	0.004197
73	16	rs7200928	0.004197	92	3	rs12053924	0.004197
74	1	rs4271231	0.004197	93	14	rs1018508	0.004197
75	18	rs17657594	0.004197	94	8	rs7016551	0.004197
76	15	rs1797230	0.004197	95	22	rs5751704	0.004197
77	20	rs6129968	0.004197	96	17	rs8064530	0.004197
78	10	rs1236903	0.004197	97	11	rs7941147	0.004197
79	9	rs10739209	0.004197	98	4	rs7659534	0.004197
80	3	rs10936081	0.004197	99	7	rs17166288	0.004197
81	10	rs11817109	0.004197	100	3	rs6441567	0.004197

Table G.2 Top 100 genes according combined *p*-value.

Entrez Gene ID	Full Name	Location	P value
5552	Serglycin	10q22.1	2.10E-04
4487	Msh Homeobox 1	4p16.3-p16.1	6.02E-04
10248	Processing Of Precursor 7 Ribonuclease P/MRP Subunit (S. Cerevisiae)	7q22	9.72E-04
8395	Phosphatidylinositol-4-Phosphate 5-Kinase Type I Beta	9q13	0.001164378
6310	Ataxin 1	6p23	0.00150692
5836	Phosphorylase Glycogen Liver	14q21-q22	0.002087102
401124	Null	4p14	0.002141294
55676	Solute Carrier Family 30 (Zinc Transporter) Member 6	2p22.3	0.002280361
8110	D4 Zinc And Double PHD Fingers Family 3	14q24.3-q31.1	0.00261988
93649	Myocardin	17p11.2	0.003978406

Table G.2 continued.

729767	Carcinoembryonic Antigen-Related Cell Adhesion Molecule 18	19q13.41	0.004197
25791	Neuronal Guanine Nucleotide Exchange Factor	2q37	0.004270978
400954	Echinoderm Microtubule Associated Protein Like 6	2p16.2-p16.1	0.004291093
128272	Rho Guanine Nucleotide Exchange Factor (GEF) 19	1p36.13	0.004975
65055	Receptor Accessory Protein 1	2p11.2	0.005249727
1179	Chloride Channel Accessory 1	1p31-p22	0.005948452
220136	Coiled-Coil Domain Containing 11	18q21.1	0.006052
491	Atpase Ca++ Transporting Plasma Membrane 2	3p25.3	0.006162748
5738	Prostaglandin F2 Receptor Negative Regulator	1p13.1	0.006260145
57537	Sortilin-Related VPS10 Domain Containing Receptor 2	4p16.1	0.00655404
55357	TBC1 Domain Family Member 2	9q22.33	0.00668154
9223	Membrane Associated Guanylate Kinase WW And PDZ Domain Containing 1	3p14.1	0.006934604
23566	Lysophosphatidic Acid Receptor 3	1p22.3	0.007499123
128989	Chromosome 22 Open Reading Frame 25	22q11.21	0.007706537
4883	Natriuretic Peptide Receptor C/Guanylate Cyclase C (Atrionatriuretic Peptide Receptor C)	5p14-p13	0.007970144
5784	Protein Tyrosine Phosphatase Non-Receptor Type 14	1q32.2	0.008127025
55714	Odz Odd Oz/Ten-M Homolog 3 (Drosophila)	4q35.1	0.008342957
27122	Dickkopf Homolog 3 (Xenopus Laevis)	11p15.2	0.008359653
113691	Null	22q11.21	0.008737
26047	Contactin Associated Protein-Like 2	7q35-q36	0.009041033
594	Branched Chain Keto Acid Dehydrogenase E1 Beta Polypeptide	6q13-q15	0.009214765
117177	RAB3A İnteracting Protein (Rabin3)	12q14.3	0.009357431
23266	Latrophilin 2	1p31.1	0.009590199
1871	E2F Transcription Factor 3	6p22	0.009599635
2001	E74-Like Factor 5 (Ets Domain Transcription Factor)	11p13-p12	0.010094614
57184	Chromosome 15 Open Reading Frame 17	15q24.1	0.01023
29958	Dimethylglycine Dehydrogenase	5q14.1	0.01039492
1012	Cadherin 13 H-Cadherin (Heart)	16q24.2-q24.3	0.010681994
11181	Trehalase (Brush-Border Membrane Glycoprotein)	11q23.3	0.01073832
345051	Null	4q31.22	0.01073832
463	Zinc Finger Homeobox 3	16q22.3-q23.1	0.010928557
5700	Proteasome (Prosome Macropain) 26S Subunit Atpase 1	14q32.11	0.01093
1007	Cadherin 9 Type 2 (T1-Cadherin)	5p14	0.01093
401494	Protein Tyrosine Phosphatase-Like A Domain Containing 2	9p21.3	0.01093
390442	Olfactory Receptor Family 11 Subfamily H Member 4	14q11.2	0.01093

Table G.2 continued.

9757	Null	19q13.1	0.01093
390059	Olfactory Receptor Family 51 Subfamily M Member 1	11p15.4	0.01093
26526	Tetraspanin 16	19p13.2	0.011586749
421	Armadillo Repeat Gene Deletes İn Velocardiofacial Syndrome	22q11.21	0.01165437
84457	Phytanoyl-Coa 2-Hydroxylase İnteracting Protein-Like	10q11	0.01175
10128	Leucine-Rich PPR-Motif Containing	2p21	0.012744497
320	Amyloid Beta (A4) Precursor Protein-Binding Family A Member 1	9q13-q21.1	0.012858354
401265	Kelch-Like 31 (Drosophila)	6p12.1	0.01299
4680	Carcinoembryonic Antigen-Related Cell Adhesion Molecule 6 (Non-Specific Cross Reacting Antigen)	19q13.2	0.01299
245939	Defensin Beta 128	20p13	0.01299
8654	Phosphodiesterase 5A Cgmp-Specific	4q25-q27	0.013094324
134121	Chromosome 5 Open Reading Frame 49	5p15.31	0.013180766
8874	Rho Guanine Nucleotide Exchange Factor (GEF) 7	13q34	0.013325383
636	Bicaudal D Homolog 1 (Drosophila)	12p11.2-p11.1	0.013739522
117584	Ring Finger And FYVE-Like Domain Containing 1	17q12	0.014405103
10558	Serine Palmitoyltransferase Long Chain Base Subunit 1	9q22.2	0.014443981
1804	Dipeptidyl-Peptidase 6	7q36.2	0.014459985
161502	Chromosome 15 Open Reading Frame 26	15q25.1	0.014642854
4329	Aldehyde Dehydrogenase 6 Family Member A1	14q24.3	0.014667132
84230	Leucine Rich Repeat Containing 8 Family Member C	1p22.2	0.014801416
444	Aspartate Beta-Hydroxylase	8q12.1	0.014845056
58189	WAP Four-Disulfide Core Domain 1	16q24.3	0.014883839
8459	Tyrosylprotein Sulfotransferase 2	22q12.1	0.014918016
26084	Null	3q25.2	0.01493369
161357	MAM Domain Containing Glycosylphosphatidylinositol Anchor 2	14q21.3	0.014973053
1519	Cathepsin O	4q31-q32	0.015290047
153241	Centrosomal Protein 120kda	5q23.2	0.015346662
8685	Macrophage Receptor With Collagenous Structure	2q12-q13	0.01535449
51316	Placenta-Specific 8	4q21.22	0.01543775
727	Complement Component 5	9q33-q34	0.015745973
169026	Solute Carrier Family 30 (Zinc Transporter) Member 8	8q24.11	0.01592
112869	Coiled-Coil Domain Containing 101	16p11.2	0.01592
79648	Microcephalin 1	8p23.1	0.016092884
124152	IQ Motif Containing K	16p12.3	0.016134387
54715	Null	16p13.3	0.016189113
10564	ADP-Ribosylation Factor Guanine NEF2	20q13.13	0.016454267

Table G.2 continued.

84649	Diacylglycerol O-Acyltransferase Homolog 2 (Mouse)	11q13.5	0.017060892
919	CD247 Molecule	1q22-q23	0.017088148
7042	Transforming Growth Factor Beta 2	1q41	0.017527573
3559	İnterleukin 2 Receptor Alpha	10p15-p14	0.017785721
8476	CDC42 Binding Protein Kinase Alpha (DMPK-Like)	1q42.11	0.017867286
50615	İnterleukin 21 Receptor	16p11	0.018036102
9914	Atpaseca++ Transporting Type 2C Member 2	16q24.1	0.018159861
27347	Serine Threonine Kinase 39 (STE20/SPS1 Homolog Yeast)	2q24.3	0.018243434
51761	Atpase Aminophospholipid Transporter-Like Class I Type 8A Member 2	13q12	0.018437336
256536	Transcription Elongation Regulator 1-Like	10q26.3	0.018792204
771	Carbonic Anhydrase XII	15q22	0.018836241
51086	TNNI3 İnteracting Kinase	1p31.1	0.019480726
3157	3-Hydroxy-3-Methylglutaryl-Coenzyme A Synthase 1 (Soluble)	5p14-p13	0.019763618
7798	Leucine Zipper Protein 1	1p36	0.019972435
54988	Acyl-Coa Synthetase Medium-Chain Family Member 5	16p12.3	0.020165734
57161	Pellino Homolog 2 (Drosophila)	14q21	0.020511019
148	Adrenergic Alpha-1A- Receptor	8p21-p11.2	0.020542565
9313	Matrix Metallopeptidase 20	11q22.3	0.020697323
7520	X-Ray Repair Complementing Defective Repair İn Chinese Hamster Cells 5 (Double-Strand-Break Rejoining)	2q35	0.020897747

Table G.3 Top 100 pathways according combined *p*-value.

Pathway ID	Pathway System	Pathway Title	Gene Count	P-Value
GO:0004758	GO Function	Serine C-Palmitoyltransferase Activity	2	2.35E-04
GO:0006812	GO Process	Cation Transport	43	4.43E-04
GO:0048103	GO Process	Somatic Stem Cell Division	3	6.95E-04
GO:0016787	GO Function	Hydrolase Activity Acting On Acid Anhydrides	224	0.001368
ASPARTATE-DEG1-PWY	BioCyc	Aspartate Degradation I	4	0.001369
LCTACACAT-PWY	BioCyc	Lactate Oxidation	4	0.001369
P41-PWY	BioCyc	Pyruvate Fermentation To Acetate And Lactate I	4	0.001369

Table G.3 continued.

hsa00272	KEGG	Cysteine Metabolism	17	0.001967
ASPARAGINE-DEG1-PWY	BioCyc	Asparagine Degradation I	5	0.002247
GO:0004459	GO Function	L-Lactate Dehydrogenase Activity	5	0.002247
GO:0019642	GO Process	Anaerobic Glycolysis	5	0.002247
GO:0005089	GO Function	Rho Guanyl-Nucleotide Exchange Factor Activity	66	0.002902
GO:0035023	GO Process	Regulation Of Rho Protein Signal Transduction	68	0.003284
GO:0005436	GO Function	Sodium:Phosphate Symporter Activity	6	0.003318
GO:0006100	GO Process	Tricarboxylic Acid Cycle İntermediate Metabolic Process	6	0.003318
wiki_37	WikiPathways	Glycolysis And Gluconeogenesis	43	0.003701
GO:0016020	GO Component	Membrane	253	0.003941
GO:0006032	GO Process	Chitin Catabolic Process	7	0.004575
GO:0016740	GO Function	Transferase Activity Transferring Nitrogenous Groups	220	0.004575
GO:0001501	GO Process	Skeletal Development	106	0.004628
GO:0004568	GO Function	Chitinase Activity	8	0.006007
hsa04310	KEGG	Wnt Signaling Pathway	149	0.006068
GO:0030170	GO Function	Pyridoxal Phosphate Binding	53	0.007539
GO:0001654	GO Process	Eye Development	9	0.007606
GO:0005388	GO Function	Calcium-Transporting Atpase Activity	9	0.007606
wiki_95	WikiPathways	Steroid Biosynthesis	9	0.007606
GO:0048503	GO Function	GPI Anchor Binding	120	0.008014
GO:0045595	GO Process	Regulation Of Cell Differentiation	11	0.011269
GO:0046658	GO Component	Anchored To Plasma Membrane	11	0.011269
GO:0007166	GO Process	Cell Surface Receptor Linked Signal Transduction	134	0.012733
GO:0007268	GO Process	Synaptic Transmission	176	0.013224
hsa00640	KEGG	Propanoate Metabolism	34	0.013345
GO:0001532	GO Function	İnterleukin-21 Receptor Activity	1	0.015374
GO:0001544	GO Process	İnitiation Of Primordial Ovarian Follicle Growth	1	0.015374
GO:0001547	GO Process	Antral Ovarian Follicle Growth	1	0.015374
GO:0004057	GO Function	Arginyltransferase Activity	1	0.015374
GO:0004484	GO Function	Mrna Guanylyltransferase Activity	1	0.015374
GO:0004491	GO Function	Methylmalonate-Semialdehyde Dehydrogenase (Acylating) Activity	1	0.015374
GO:0004877	GO Function	Complement Component C3b Receptor Activity	1	0.015374
GO:0005915	GO Component	Zonula Adherens	1	0.015374
GO:0005924	GO Component	Cell-Substrate Adherens Junction	1	0.015374
GO:0005991	GO Process	Trehalose Metabolic Process	1	0.015374

Table G.3 continued.

GO:0005993	GO Process	Trehalose Catabolic Process	1	0.015374
GO:0006037	GO Process	Cell Wall Chitin Metabolic Process	1	0.015374
GO:0007258	GO Process	JUN Phosphorylation	1	0.015374
GO:0009593	GO Process	Detection Of Chemical Stimulus	1	0.015374
GO:0016598	GO Process	Protein Arginylation	1	0.015374
GO:0018478	GO Function	Malonate-Semialdehyde Dehydrogenase (Acetylating) Activity	1	0.015374
GO:0019859	GO Process	Thymine Metabolic Process	1	0.015374
GO:0021522	GO Process	Spinal Cord Motor Neuron Differentiation	1	0.015374
GO:0031648	GO Process	Protein Destabilization	1	0.015374
GO:0033364	GO Process	Mast Cell Secretory Granule Organization And Biogenesis	1	0.015374
GO:0033371	GO Process	T Cell Secretory Granule Organization And Biogenesis	1	0.015374
GO:0033373	GO Process	Maintenance Of Protease Localization İn Mast Cell Secretory Granule	1	0.015374
GO:0033382	GO Process	Maintenance Of Granzyme B Localization İn T Cell Secretory Granule	1	0.015374
GO:0042416	GO Process	Dopamine Biosynthetic Process	1	0.015374
GO:0042555	GO Component	MCM Complex	1	0.015374
GO:0043149	GO Process	Stress Fiber Formation	1	0.015374
GO:0045046	GO Process	Protein İmport İnto Peroxisome Membrane	1	0.015374
GO:0045810	GO Process	Negative Regulation Of Frizzled Signaling Pathway	1	0.015374
GO:0045823	GO Process	Positive Regulation Of Heart Contraction	1	0.015374
GO:0046327	GO Process	Glycerol Biosynthetic Process From Pyruvate	1	0.015374
GO:0047045	GO Function	Testosterone 17-Beta-Dehydrogenase (NADP+) Activity	1	0.015374
GO:0051795	GO Process	Positive Regulation Of Catagen	1	0.015374
GO:0007268	GO Process	Synaptic Transmission Glycinergic	176	0.015374
GO:0060013	GO Process	Righting Reflex	1	0.015374
GO:0060038	GO Process	Cardiac Muscle Cell Proliferation	1	0.015374
h_stathmin Pathway	BioCarta	Stathmin And Breast Cancer Resistance To Antimicrotubule Agents	13	0.015498
VALDEG PWY	BioCyc	Valine Degradation I	13	0.015498
GO:0016301	GO Function	Kinase Activity	186	0.016840
hsa00600	KEGG	Sphingolipid Metabolism	38	0.017696
GO:0016021	GO Component	İntegral To Membrane	233	0.017703
GO:0003824	GO Function	Catalytic Activity	146	0.017919
GO:0016779	GO Function	Nucleotidyltransferase Activity	15	0.020231
h_tob1Pathway	BioCarta	Role Of Tob İn T-Cell Activation	15	0.020231

Table G.3 continued.

GO:0008219	GO Process	Cell Death	42	0.022650
hsa00620	KEGG	Pyruvate Metabolism	42	0.022650
GO:0006606	GO Process	Protein İmport İnto Nucleus Docking	13	0.022650
GO:0008324	GO Function	Cation Transmembrane Transporter Activity	17	0.025412
Leukocyte signaling	WFINFLAM	Leukocyte Signaling	121	0.027952
GO:0016310	GO Process	Phosphorylation	18	0.028152
GO:0016491	GO Function	Oxidoreductase Activity	211	0.028233
GO:0006396	GO Process	RNA Processing	47	0.029634
GO:0007155	GO Process	Cell Adhesion	214	0.029818
ASPARTATE SYN-PWY	BioCyc	Aspartate Biosynthesis I	2	0.030278
GO:0000103	GO Process	Sulfate Assimilation	2	0.030278
GO:0000272	GO Process	Polysaccharide Catabolic Process	2	0.030278
GO:0001101	GO Process	Response To Acid	2	0.030278
GO:0001729	GO Function	Ceramide Kinase Activity	2	0.030278
GO:0001964	GO Process	Startle Response	2	0.030278
GO:0003863	GO Function	3-Methyl-2-Oxobutanoate Dehydrogenase (2-Methylpropanoyl-Transferring) Activity	2	0.030278
GO:0004020	GO Function	Adenylylsulfate Kinase Activity	2	0.030278
GO:0004069	GO Function	Aspartate Transaminase Activity	2	0.030278
GO:0004421	GO Function	Hydroxymethylglutaryl-Coa Synthase Activity	2	0.030278
GO:0004613	GO Function	Phosphoenolpyruvate Carboxykinase (GTP) Activity	2	0.030278
GO:0004702	GO Function	Receptor Signaling Protein Serine/Threonine Kinase Activity	2	0.030278

APPENDIX H

CNA AND LOH RESULTS OBTAINED BY THREE METHODS

CNA and LOH results obtained by three methods for all samples are given in following tables, Table G.1, Table G.2 and Table G.3.

Table H.1 CNA segments obtained by CalMaTe

Sample	Chr	Start	Stop	Mean	Count	Call
T1	5	72452355	72452619	-1.816	3	Loss
T1	7	104403259	104403270	1.733	2	Gain
T1	10	126968973	126980389	1.627	2	Gain
T2	7	104403259	104403270	1.836	2	Gain
T5	4	18922754	18923066	1.367	3	Gain
T6	1	176756854	176775678	1.247	3	Gain
T6	1	231982016	231985728	1.15	5	Gain
T6	7	34905983	34905998	4.121	2	Gain
T6	7	137275430	137284359	1.618	2	Gain
T7	17	3288535	3288545	-3.538	2	Loss
T7	18	44170385	44170471	-4.166	3	Loss
T8	1	70887099	70911359	3.208	3	Gain
T8	2	53514150	53540600	-1.664	3	Loss
T8	6	32448098	32563997	1.662	3	Gain
T8	7	32629579	32629640	1.885	2	Gain
T8	7	34905983	34905998	5.575	2	Gain
T8	8	126766585	126808810	-1.146	6	Loss
T8	11	101668066	101668330	1.899	2	Gain
T8	12	23185575	23218803	1.198	6	Gain
T8	17	42687933	42687961	-1.753	2	Loss
T8	17	61227637	61227768	-1.946	2	Loss
T8	18	25353854	25354031	1.463	3	Gain
T9	18	44170285	44170471	-4.594	4	Loss
T10	2	105403749	105403844	2.094	2	Gain

Table H.2 CNA segments obtained by VN algorithm

Sample	Chr	Start	Stop	Mean	Count	Call
T1	1	116991956	116996427	1.031	4	Gain
T1	8	6466608	6498174	1.419	3	Gain
T1	10	126947089	126980613	1.116	4	Gain
T1	12	106492887	106515992	1.053	5	Gain
T1	19	56156447	56180968	1.457	3	Gain
T1	21	34666160	34666492	1.143	3	Gain
T1	6	133726742	133727421	-1.088	3	Loss
T2	1	3665867	3684185	1.194	3	Gain
T2	1	54519107	54566190	1.28	3	Gain
T2	1	246715196	246801648	1.32	3	Gain
T2	3	124281604	124290320	1.096	3	Gain
T2	7	151417368	151431568	1.191	3	Gain
T2	10	73036391	73046177	1.21	3	Gain
T2	10	77205684	77225753	1.105	3	Gain
T2	11	825777	898518	1.226	3	Gain
T2	19	11227554	11233777	1.077	3	Gain
T2	19	56156447	56180968	1.367	3	Gain
T3	2	226090491	226101783	1.062	3	Gain
T3	9	132071153	132074905	1.227	3	Gain
T3	10	98916782	98930061	1.104	3	Gain
T3	12	50277412	50281414	1.159	3	Gain
T3	16	27444030	27458900	1.085	4	Gain
T3	19	56156447	56180968	1.117	3	Gain
T3	22	29554183	29566978	1.122	3	Gain
T3	1	64577482	64596766	-1.349	3	Loss
T3	1	87711620	87711755	-1.085	3	Loss
T4	2	102968075	102991369	1.191	6	Gain
T4	2	183277721	183380753	1.16	4	Gain
T4	2	241200202	241205113	1.401	3	Gain
T4	3	131884654	131889272	1.157	4	Gain
T4	3	189336370	189340241	1.095	5	Gain
T4	4	18922754	18923066	1.518	3	Gain
T4	4	88823007	88823643	1.169	4	Gain
T4	7	105473360	105478527	1.291	3	Gain
T4	7	131923644	131932880	1.134	5	Gain
T4	8	3819578	3820138	1.146	4	Gain
T4	8	5400160	5401120	1.131	4	Gain

Table H.2 continued.

T4	8	83903814	83955796	1.45	3	Gain
T4	9	8302233	8332726	1.048	5	Gain
T4	9	16306523	16306719	1.261	5	Gain
T4	10	52771341	52771503	1.304	3	Gain
T4	10	92817279	92818423	1.06	5	Gain
T4	12	84991845	84993571	1.264	5	Gain
T4	13	71871468	71881149	1.095	5	Gain
T4	13	75796834	75814878	1.006	5	Gain
T4	13	111146009	111152191	1.231	5	Gain
T4	14	100275413	100283248	1.402	3	Gain
T4	17	47345847	47389935	1.184	4	Gain
T4	18	13870705	13915442	1.136	4	Gain
T4	20	11730158	11764040	1.453	3	Gain
T4	21	37301261	37328106	1.068	6	Gain
T4	23	151017674	151022400	1.284	3	Gain
T4	1	242798937	242799035	-1.26	3	Loss
T4	13	87068836	87069220	-1.262	3	Loss
T5	2	79608891	79609002	1.386	3	Gain
T5	2	183277721	183380753	1.015	4	Gain
T5	2	232560984	232598197	1.286	3	Gain
T5	3	173400740	173406949	1.102	3	Gain
T5	4	18922754	18923066	1.467	3	Gain
T5	4	88823007	88823643	1.007	4	Gain
T5	7	132782294	132805848	1.185	3	Gain
T5	9	651333	651461	1.126	3	Gain
T5	10	52767174	52771503	1.151	4	Gain
T5	12	84991845	84993571	1.012	5	Gain
T5	14	100669440	100748844	1.189	3	Gain
T5	16	7294639	7295167	1.161	3	Gain
T5	17	47345847	47389935	1.081	4	Gain
T5	20	1659992	1660099	1.078	4	Gain
T5	9	20032707	20075009	-1.241	3	Loss
T5	23	150415723	150415979	-1.482	3	Loss
T6	1	71166556	71200049	1.564	9	Gain
T6	1	201422201	201423940	1.085	3	Gain
T6	1	206617766	206621927	1.06	3	Gain
T6	1	214347831	214361741	1.098	5	Gain
T6	1	220970441	220971339	1.012	4	Gain

Table H.2 continued.

T6	1	231982016	231998863	1.419	7	Gain
T6	3	64961866	64962082	1.364	3	Gain
T6	4	14132572	14137964	1.116	4	Gain
T6	9	24653566	24709463	1.153	3	Gain
T6	16	65468615	65482887	1.402	3	Gain
T6	19	34224816	34231709	1.271	4	Gain
T6	22	35337732	35337888	1.352	3	Gain
T7	1	59471986	59484502	1.434	5	Gain
T7	1	211168634	211168759	2.32	3	Gain
T7	2	9902902	9908901	1.587	4	Gain
T7	5	151738612	151751433	2.146	3	Gain
T7	6	57632711	57637260	1.698	4	Gain
T7	11	100449118	100449303	2.02	3	Gain
T7	13	20170110	20239913	1.929	3	Gain
T7	18	68800312	68809883	1.635	3	Gain
T7	20	37634664	37635714	2.117	3	Gain
T7	1	15542073	15545010	-1.843	3	Loss
T7	1	39163040	39165402	-1.74	3	Loss
T7	1	49113247	49130683	-1.013	9	Loss
T7	1	96504265	96595179	-2.387	3	Loss
T7	1	117475715	117479371	-1.666	3	Loss
T7	1	162193738	162195738	-1.922	3	Loss
T7	1	200970810	200999131	-1.802	4	Loss
T7	1	234621341	234635899	-1.759	3	Loss
T7	2	7520946	7541002	-1.806	3	Loss
T7	2	35659948	35744997	-1.287	5	Loss
T7	2	36612334	36623048	-2.562	4	Loss
T7	2	206086949	206141889	-1.295	15	Loss
T7	2	217348765	217427625	-1.271	7	Loss
T7	2	227448975	227506468	-1.376	6	Loss
T7	2	230256586	230368406	-1.201	9	Loss
T7	2	232408269	232423757	-1.92	3	Loss
T7	3	6795637	6821304	-1.282	8	Loss
T7	3	28756403	28762058	-1.654	3	Loss
T7	3	179400345	179422594	-2.004	4	Loss
T7	3	183210854	183210991	-1.884	3	Loss
T7	4	14381367	14466699	-1.069	19	Loss
T7	4	75199609	75253120	-1.663	5	Loss

Table H.2 continued.

T7	4	125656446	125779405	-1.932	3	Loss
T7	5	18394421	18394867	-2.006	3	Loss
T7	5	61010938	61011200	-3.12	3	Loss
T7	5	62658671	62658859	-1.521	4	Loss
T7	5	101487818	101553265	-1.938	3	Loss
T7	5	147548328	147558546	-1.894	3	Loss
T7	5	152935392	152935428	-1.833	3	Loss
T7	5	155081309	155087347	-1.798	3	Loss
T7	6	2059110	2129823	-1.251	6	Loss
T7	6	6319379	6319596	-1.541	3	Loss
T7	6	19012921	19013390	-2.494	3	Loss
T7	6	74999620	75030616	-2.125	3	Loss
T7	6	82221641	82277416	-1.56	6	Loss
T7	6	165425309	165425517	-2.101	3	Loss
T7	6	167019126	167030620	-2.572	3	Loss
T7	6	169564874	169565086	-3.273	3	Loss
T7	7	11757441	11778975	-1.727	4	Loss
T7	7	47105926	47128635	-2.26	4	Loss
T7	7	53593511	53628253	-1.539	4	Loss
T7	7	71056372	71076106	-1.308	6	Loss
T7	7	136025213	136025735	-1.862	3	Loss
T7	8	2929122	2929820	-1.912	3	Loss
T7	8	5425042	5467543	-1.436	5	Loss
T7	8	11458272	11463230	-1.138	7	Loss
T7	8	126674995	126692618	-2.16	3	Loss
T7	8	138526638	138530267	-1.832	4	Loss
T7	9	16306523	16306719	-1.754	5	Loss
T7	9	21114237	21131627	-1.861	3	Loss
T7	10	15216419	15246346	-1.852	5	Loss
T7	10	30898870	30927396	-1.533	5	Loss
T7	10	77225753	77227763	-1.709	3	Loss
T7	10	79842921	79876354	-1.786	4	Loss
T7	10	80212839	80213063	-1.945	3	Loss
T7	10	119048752	119072518	-1.552	4	Loss
T7	11	7776954	7784484	-2.793	3	Loss
T7	11	11204926	11211145	-1.813	5	Loss
T7	11	15949943	15952543	-2.392	3	Loss
T7	11	35321749	35334083	-1.912	4	Loss

Table H.2 continued.

T7	11	83341015	83358629	-1.79	3	Loss
T7	11	98631400	98658149	-1.656	4	Loss
T7	11	133150973	133184342	-1.657	4	Loss
T7	12	6208618	6238593	-1.984	3	Loss
T7	12		43933484	-1.753	3	Loss
T7	12	114879451	114879632	-2.178	3	Loss
T7	12	129507348	129517157	-2.481	3	Loss
T7	13	50443418	50443558	-2.312	3	Loss
T7	13	72619525	72619751	-1.89	3	Loss
T7	13	76345298	76363655	-1.786	3	Loss
T7	14	32671106	32689556	-2.437	3	Loss
T7	14	85507167	85591072	-1.876	6	Loss
T7	15	50590794	50650923	-1.312	7	Loss
T7	15	71622786	71628859	-1.926	3	Loss
T7	15	90908781	90972441	-1.383	3	Loss
T7	16	5608413	5608643	-1.794	4	Loss
T7	16	8304390	8305106	-2.439	4	Loss
T7	16	23939438	23950185	-1.767	3	Loss
T7	18	44170285	44170471	-4.581	4	Loss
T7	18	55144264	55144795	-1.872	3	Loss
T7	19	20110319	20111994	-1.587	3	Loss
T7	20	23814432	23893840	-1.377	7	Loss
T7	21	24215799	24365979	-2.06	4	Loss
T7	21	37321534	37321567	-2.928	3	Loss
T7	21	44273163	44295957	-2.177	3	Loss
T7	22	17317346	17387645	-1.301	6	Loss
T7	22	25593658	25668784	-1.422	13	Loss
T7	23	150219629	150238398	-1.092	8	Loss
T8	1	20544835	20572687	1.206	6	Gain
T8	1	55119144	55119515	1.598	3	Gain
T8	1	71188799	71200049	3.239	3	Gain
T8	1	108227361	108245615	1.246	5	Gain
T8	1	179994628	180080313	1.304	5	Gain
T8	3	162246284	162261021	1.185	5	Gain
T8	5	10726281	10743433	1.557	6	Gain
T8	6	32448098	32563997	1.539	3	Gain
T8	7	36848291	36871201	1.08	8	Gain
T8	7	111199127	111199356	1.549	3	Gain

Table H.2 continued.

T8	8	19303736	19313912	1.123	5	Gain
T8	12	13155166	13179236	1.206	4	Gain
T8	13	70653902	70882841	1.03	12	Gain
T8	13	112018391	112024683	1.909	3	Gain
T8	14	26395815	26396098	1.577	3	Gain
T8	17	39178221	39189921	1.059	5	Gain
T8	19	6896483	6906850	1.346	4	Gain
T8	20	47555943	47626736	1.022	10	Gain
T8	21	18195168	18213225	1.116	4	Gain
T8	1	58482341	58484277	-1.203	5	Loss
T8	2	53514150	53540600	-1.439	3	Loss
T8	3	60012350	60337913	-1.011	65	Loss
T8	3	62113277	62188720	-1.187	8	Loss
T8	4	181439874	181445182	-1.401	3	Loss
T8	6	19012921	19013390	-1.468	3	Loss
T8	6	169564874	169565086	-1.476	3	Loss
T8	7	78456148	78470511	-1.342	3	Loss
T8	8	27891564	27892514	-1.234	4	Loss
T8	8	126766585	126805422	-1.444	5	Loss
T8	10	124029259	124029471	-1.431	3	Loss
T8	15	99946503	99949289	-1.488	3	Loss
T8	16	75667791	75692964	-1.528	3	Loss
T8	16	79021610	79043240	-1.305	3	Loss
T9	2	124913517	124913598	3.081	3	Gain
T9	2	205186251	205186529	2.435	4	Gain
T9	3	127991149	127991938	2.463	3	Gain
T9	7	150831289	150892599	1.536	8	Gain
T9	10	80928793	80935083	1.783	6	Gain
T9	12	120210270	120210394	2.403	3	Gain
T9	14	23102784	23103040	2.464	4	Gain
T9	16	81639190	81639703	2.223	3	Gain
T9	22	49020281	49045663	1.864	6	Gain
T9	2	36612334	36624187	-1.885	5	Loss
T9	2	188356689	188436981	-1.999	4	Loss
T9	2	232408269	232423757	-2.172	3	Loss
T9	4	42902654	42902929	-2.678	3	Loss
T9	4	125656446	125779405	-2.236	3	Loss
T9	5	61010938	61011200	-2.885	3	Loss

Table H.2 continued.

T9	6	67965720	67968795	-2.372	3	Loss
T9	6	169564874	169565086	-2.609	3	Loss
T9	8	126674995	126692618	-2.469	3	Loss
T9	9	16306523	16306614	-2.281	4	Loss
T9	9	120964999	120991462	-2.253	3	Loss
T9	10	2235509	2290492	-1.425	7	Loss
T9	10	21159236	21174672	-1.695	5	Loss
T9	10	72263690	72279151	-2.462	3	Loss
T9	10	80212839	80213063	-2.321	3	Loss
T9	10	121884070	121884166	-3.337	3	Loss
T9	11	93594101	93598595	-1.83	4	Loss
T9	11	133739685	133765127	-2.226	3	Loss
T9	12	28738933	28739024	-2.499	3	Loss
T9	12	54977871	54997372	-1.4	8	Loss
T9	12	108425202	108472527	-1.452	8	Loss
T9	13	50443418	50443558	-2.522	3	Loss
T9	14	32671106	32689556	-2.429	3	Loss
T9	14	85507167	85593694	-1.328	7	Loss
T9	14	100022124	100022864	-1.561	6	Loss
T9	15	71622786	71628859	-2.148	3	Loss
T9	15	78506245	78527616	-2.16	3	Loss
T9	15	82387776	82409917	-1.824	5	Loss
T9	17	54008626	54035820	-2.89	3	Loss
T9	18	44170285	44170471	-4.631	4	Loss
T9	21	24215799	24365979	-2.13	4	Loss
T9	21	37321534	37321567	-2.796	3	Loss
T10	1	26020784	26032034	1.105	4	Gain
T10	1	31005669	31007037	1.05	4	Gain
T10	1	61617180	61620604	1.032	4	Gain
T10	1	62674507	62674547	1.134	3	Gain
T10	1	77287551	77324865	1.169	4	Gain
T10	1	80514773	80589958	1.036	4	Gain
T10	1	95024183	95028801	1.145	3	Gain
T10	1	162003746	162027992	1.006	5	Gain
T10	2	102968075	102991369	1.006	6	Gain
T10	2	134420944	134437471	1.054	4	Gain
T10	2	183277721	183380753	1.073	4	Gain
T10	3	131884654	131889272	1.068	4	Gain

Table H.2 continued.

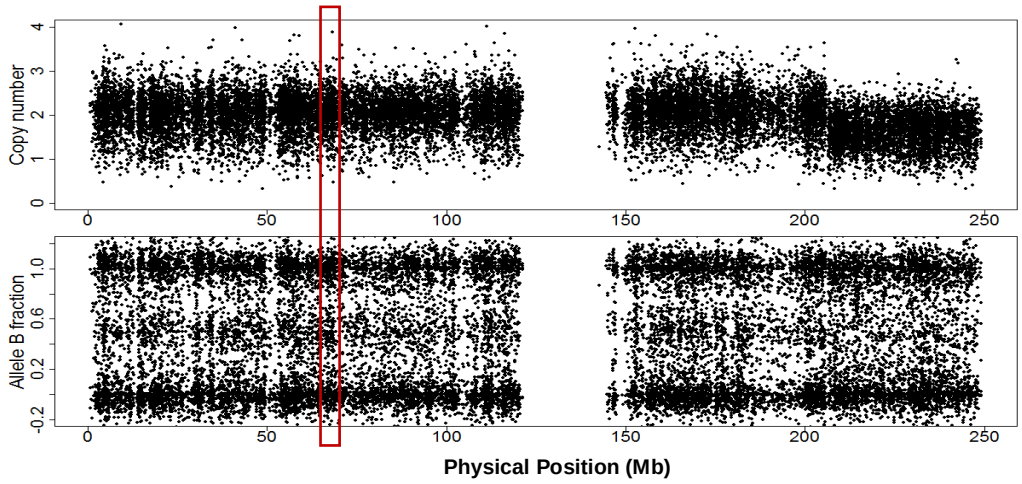
T10	3	189336370	189340241	1.087	5	Gain
T10	4	18922754	18923066	1.548	3	Gain
T10	4	31037602	31077197	1.164	3	Gain
T10	5	2902488	2902999	1.205	3	Gain
T10	5	9837791	9838122	1.01	4	Gain
T10	5	107657072	107688854	1.181	3	Gain
T10	5	152935392	152935428	1.148	3	Gain
T10	6	90497311	90586725	1.111	3	Gain
T10	7	105473360	105478527	1.11	3	Gain
T10	7	111637240	111650835	1.111	3	Gain
T10	7	131923644	131932880	1.101	5	Gain
T10	7	132782294	132805848	1.211	3	Gain
T10	7	140014425	140015104	1.163	3	Gain
T10	8	3819578	3820138	1.024	4	Gain
T10	8	11458343	11463044	1.061	5	Gain
T10	10	52771341	52771503	1.02	3	Gain
T10	10	67598906	67600010	1.238	3	Gain
T10	11	133669650	133682972	1.18	3	Gain
T10	13	22522198	22525954	1.108	4	Gain
T10	13	28892892	28893642	1.009	4	Gain
T10	14	100275413	100283248	1.219	3	Gain
T10	15	87252779	87256321	1.194	3	Gain
T10	16	8060441	8064041	1.162	3	Gain
T10	17	47345847	47389935	1.11	4	Gain
T10	18	40280407	40381039	1.048	4	Gain
T10	20	1659992	1660099	1.155	4	Gain
T10	20	61702029	61706149	1.243	3	Gain
T10	21	32344644	32362409	1.132	4	Gain
T10	23	97833775	97833958	-1.251	3	Loss

APPENDIX I

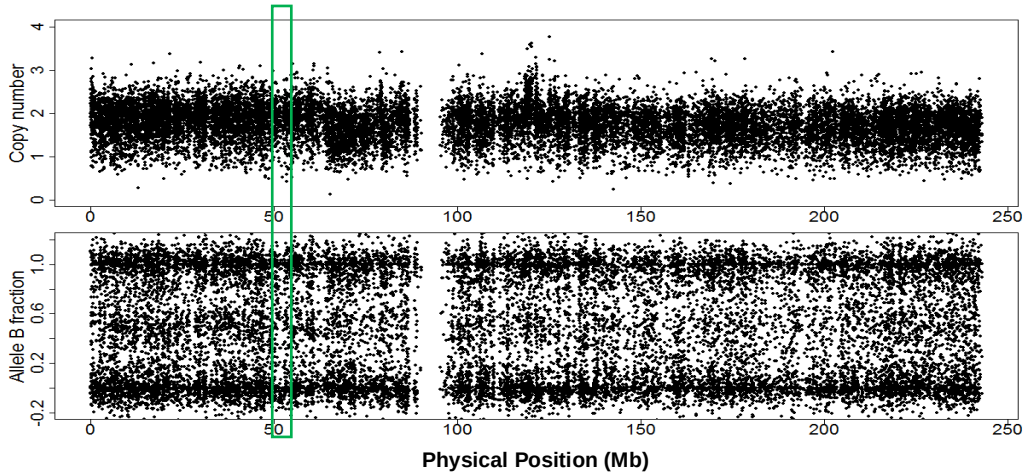
RAW CN AND BAF VS GENOMIC POSITION GRAPHS OF CHROMOSOMES WITH CNA FOR SAMPLE T8

Results of CalMaTe

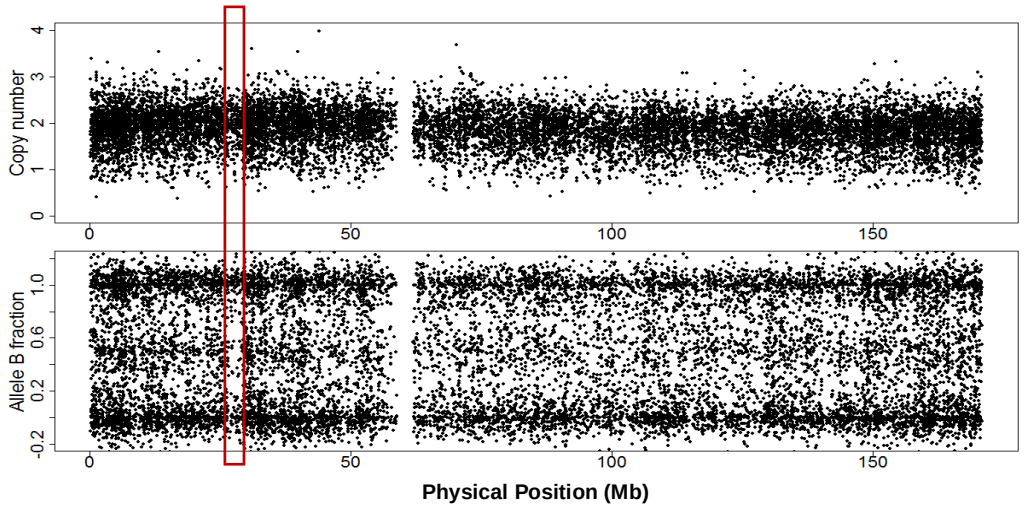
a) Chromomosome 1



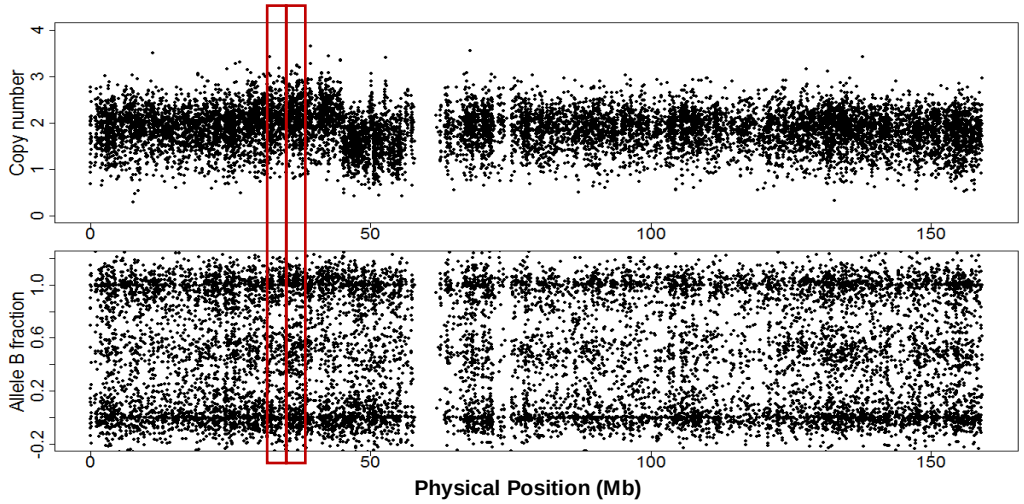
b) Chromosome 2



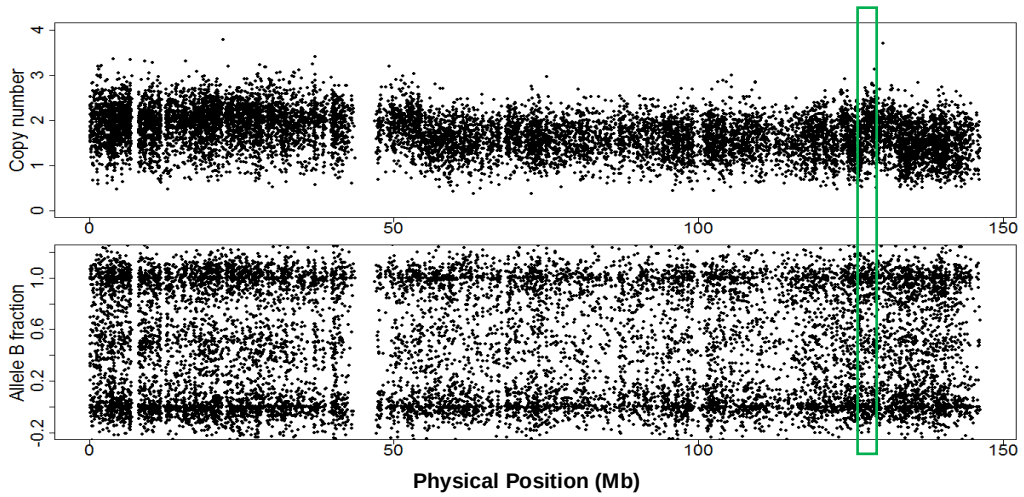
c) Chromosome 6



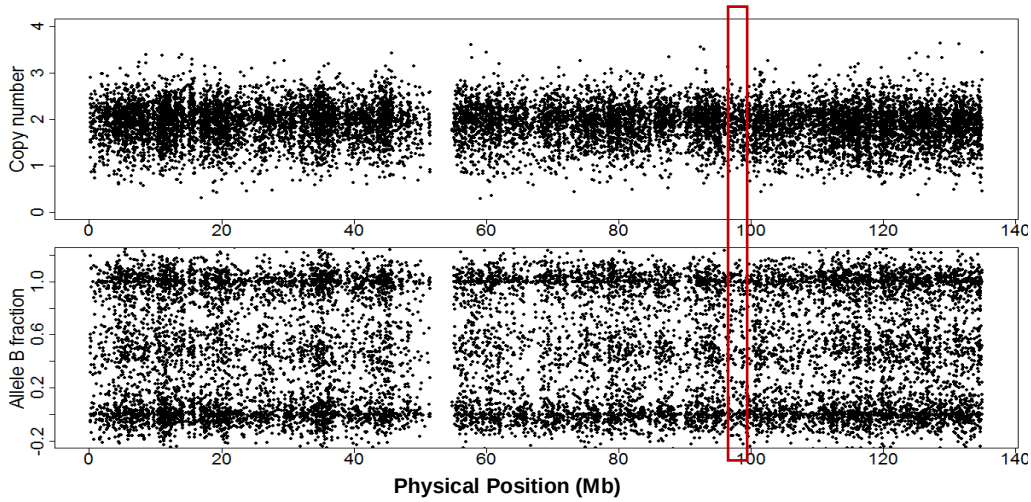
d) Chromosome 7



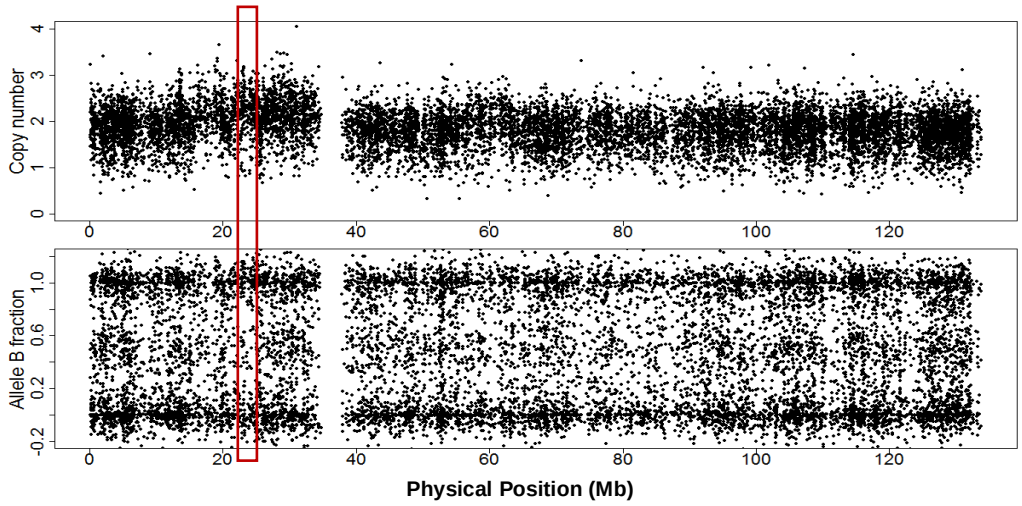
e) Chromosome 8



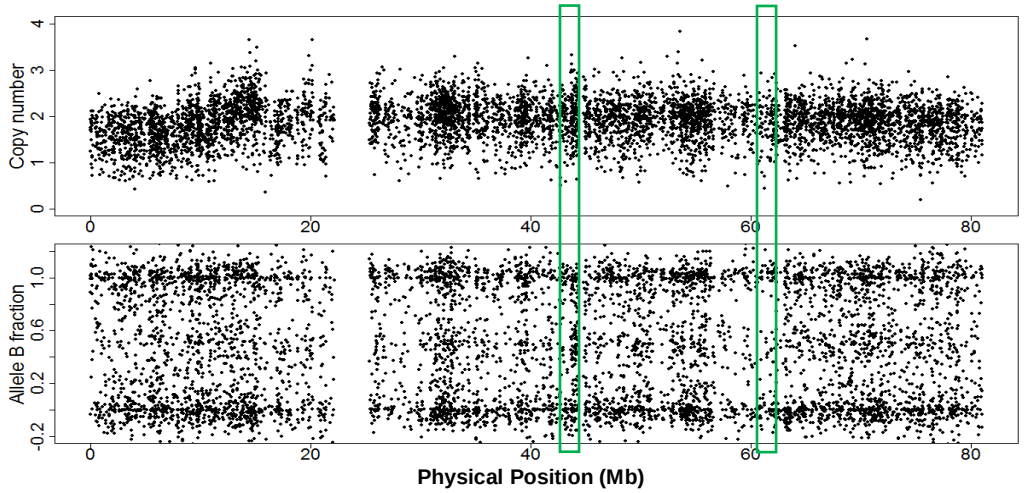
f) Chromosome 11



g) Chromosome 12



h) Chromosome 17



i) Chromosome 18

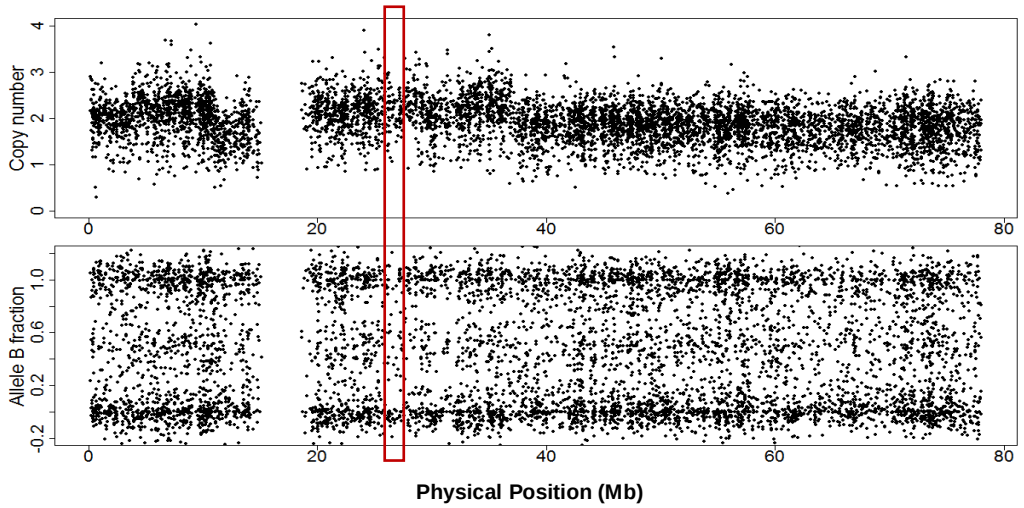


Figure I.1 The panels for a) Chr1 b) Chr2 c) Chr6 d) Chr7 e) Chr8 f) Chr11 g) Chr12 h) Chr17 i) Chr18 are corresponding to raw CN and BAF graphs drawn vs physical position. Black dots: locus-level estimates; red rectangle: gain; green rectangle: loss.

Results Of PSCBS

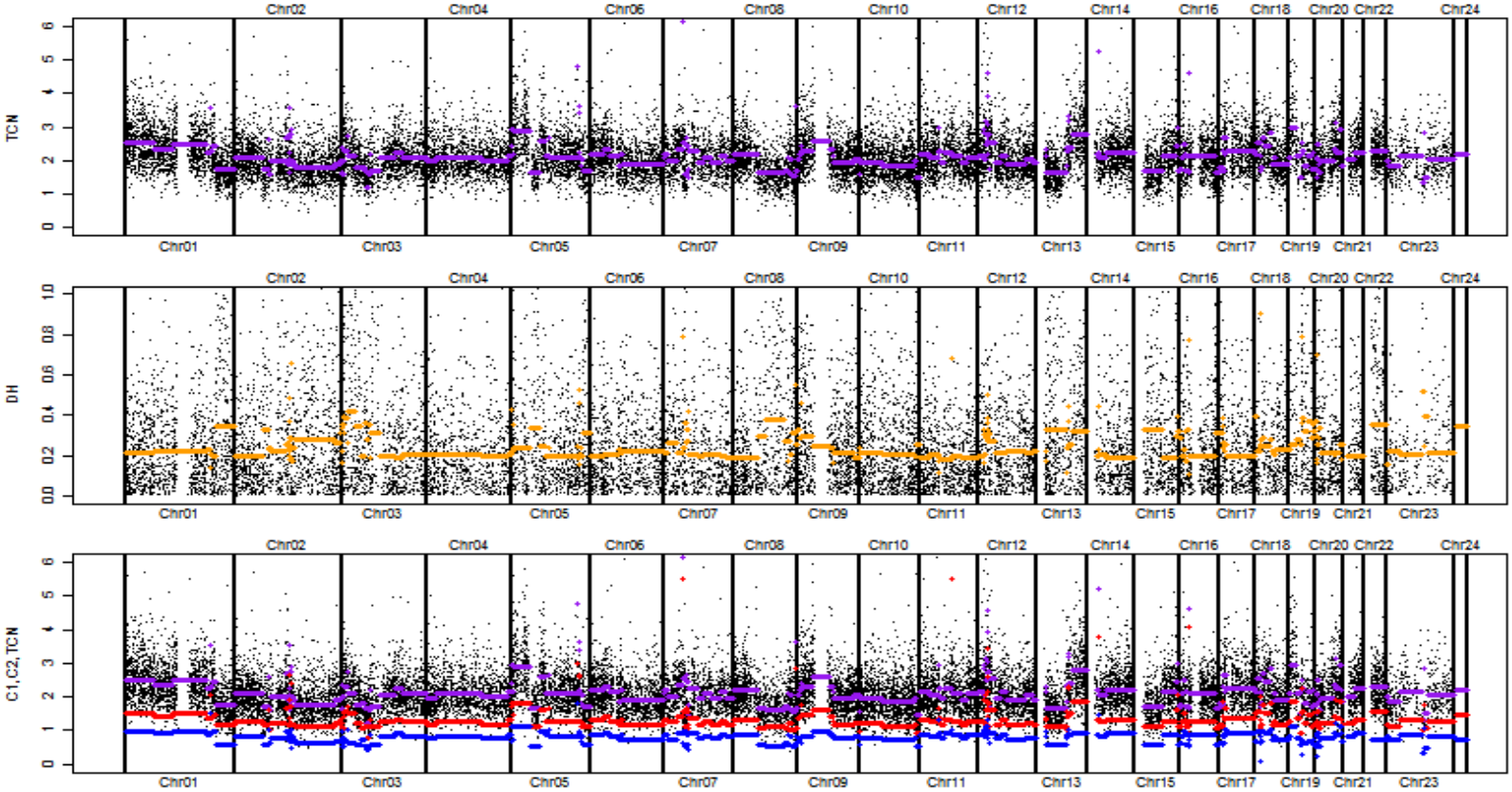


Figure I.2 Segmentation results of sample T8 over all chromosomes. TCN: Total copy number (purple segments in the upper and bottom panel); DH: Decrease in heterozygosity (orange segments in the middle panel; C1: Minor copy number (blue segments in the bottom panel); C2: Major copy number (red segments in the bottom panel).