

**GENOME-WIDE ANALYSIS OF LOSS OF  
HETEROZYGOSITY AND DISCOVERY OF NOVEL  
TUMOR SUPPRESSOR GENES IN GASTRIC  
CANCER**

**WU YINGTING**

*(B. Sc. Shanghai Jiao Tong University)*

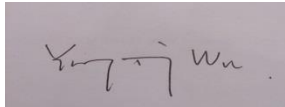
**A THESIS SUBMITTED  
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY  
IN COMPUTATION AND SYSTEMS BIOLOGY (CSB)  
SINGAPORE-MIT ALLIANCE  
NATIONAL UNIVERSITY OF SINGAPORE**

**2012**

## DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

A rectangular box containing a handwritten signature in black ink. The signature appears to be 'Wu Yingting' written in a cursive style.

---

Wu Yingting

25/07/2012

## ACKNOWLEDGEMENTS

First and foremost I would like to offer my sincerest gratitude to my supervisors, A/P Steve Rozen and Professor Roy Welsh. Without their patient guidance and strong support, this dissertation would not have been possible. Steve's perpetual energy and enthusiasm in research had motivated me in all my PhD candidature period. He is such a good supervisor that inspired my creative job and is always willing to help. Therefore, I really underwent a smooth and rewarding research life during these four years. Roy, although based in United States, instructed me with his profound knowledge and visited Singapore frequently for project discussions. His rigorous academic attitude deeply impressed me.

Secondly, I am thankful to A/P Patrick Tan and Assistant Professor Goh Liang Kee. Patrick kindly provides me the original SNP array data for analysis and I also receive the support of lab resources and techniques from him and his lab. His knowledge of cancer research is also valuable to me to solve problems in the project. Dr Goh is an expert in microarray data analysis and provided valuable comments and opinions to me.

Thirdly, I want to express my thanks to the lab members in A/P Steve Rozen, Assistant Professor Goh Liang Kee and A/P Patrick Tan's labs. Niantao Deng and Gengbo Chen helped in research design and programming. Dr. Zhengdeng Lei and Dr. John McPherson supported technical issues. Dr. Yew Chung Tang step-by-step demonstrated and instructed the experiments on cell lines. Yansong Zhu kindly offer help to my cell proliferation experiments. I will not list all the names here, but their assistance and friendship will be a memorable treasure in my whole life.

In addition, I wish to extend my thanks to the Singapore-MIT Alliance-Computation and Systems Biology (SMA-CSB) program for providing me with the abundant funding and administrative support and providing me a remarkable graduate school experience.

Finally, I want to express my special thanks to my husband, my parents, and all my friends, who like me and have fun with me. Without their encouragement, I would never have made it. This thesis is dedicated to all of them.

## TABLE OF CONTENTS

DECLARATION .....	i
ACKNOWLEDGEMENTS.....	ii
TABLE OF CONTENTS .....	iv
SUMMARY.....	viii
TABLE OF ABBREVIATIONS .....	x
LIST OF TABLES.....	xi
LIST OF FIGURES .....	xii
CHAPTER 1 INTRODUCTION AND LITERATURE REVIEW .....	1
1.1 Aims and Outlines.....	1
1.2 General Introduction of Gastric Cancer .....	3
1.2.1 Epidemiology.....	3
1.2.2 Diagnosis .....	3
1.2.3 Histology .....	4
1.2.4 Treatment.....	6
1.2.5 Etiology .....	7
Helicobacter Pylori Infection.....	7
Diet, Smoking and Alcohol.....	7
Genetic Susceptibility .....	8
1.2.6 Molecular Pathogenesis.....	9
1.2.6.1 Cancer-Related Pathways .....	10
TP53 Pathway .....	10
NF- $\kappa$ B Pathway.....	11
Wnt Signaling Pathway .....	12
1.2.6.2 Genetic and Epigenetic Alterations .....	12
Gene Mutations .....	13
Copy Number and Gene Expression Alterations .....	14
Microsatellite Instability (MSI).....	15
Epigenetic Modifications .....	16
1.2.7 Gastric Cancer Biomarkers.....	17
1.3 Loss of Heterozygosity.....	18
1.3.1 General Introduction to LOH .....	18

1.3.2 LOH and TSGs in Cancers .....	19
1.3.3 LOH in Gastric Cancer .....	20
1.4 Genome-Wide SNP Array Application on CNA and LOH Analysis ....	21
1.4.1 Genome-wide SNP Array .....	21
1.4.2 Application of SNP Arrays on Copy Number and LOH Analysis ..	26
<b>Chapter 2 COMPARISON OF SOFTWARE FOR IDENTIFYING COPY NUMBER ALTERATIONS AND LOSS OF HETEROZYGOSITY FOR AFFYMETRIX SNP 6.0 ARRAYS .....</b>	<b>28</b>
2.1 Abstract.....	29
2.2 Introduction.....	30
2.3 Materials and Methods .....	35
2.3.1 Lung Cancer Cell lines .....	35
2.3.2 CNA and LOH Analysis .....	35
CNAG .....	35
Birdsuite.....	35
GAP.....	35
PennCNV .....	36
PICNIC .....	36
ASCAT .....	36
TAPS.....	37
Paired-PSCBS .....	37
2.4 Results and Discussion .....	38
2.4.1 Data Pre-processing.....	38
2.4.2 Genotyping .....	40
2.4.3 LRR, BAF and Decrease in Heterozygosity.....	41
2.4.4 Segmentation .....	42
2.4.5 CNA and LOH Calls in Programs that Use HMMs .....	43
2.4.6 CNA and LOH Calls in Programs that Use Segmentation.....	44
2.4.7. Evaluation and Comparison of Eight Programs in Data from a Dilution Series .....	46
2.5 Conclusions .....	48
2.6 Figures.....	49

CHAPTER 3 GLOBAL ANALYSIS OF LOSS OF HETEROZYGOSITY AND GENOMIC COPY NUMBER ALTERATIONS IN GASTRIC CANCER IMPLICATES KNOWN AND NOVEL CANCER GENES.....	58
3.1 Abstract .....	59
3.2 Introduction .....	60
3.3 Materials and Methods .....	62
3.3.1 Patients and Samples .....	62
3.3.2 DNA Extraction and Hybridization .....	62
3.3.3 SNP Array Data Pre-processing .....	62
3.3.4 ASCAT Profiling of Allele-Specific Copy Numbers .....	63
3.3.5 Cell Culture.....	64
3.3.6 Preparation and Transfection of siRNA .....	64
3.3.7 Western Blot Analysis .....	65
3.3.8 Cell proliferation assays .....	65
3.4 Results .....	66
LOH Landscape in Gastric Cancer. ....	66
Recurrent Somatic CNAs in Gastric Cancer .....	67
Relation between Genomic Alterations and Tumor Characteristics .....	68
LOH and <i>TP53</i> mutations .....	68
3.5 Discussion .....	70
Limitations.....	70
Comparison to previous findings on CNA in gastric cancer .....	71
Comparison to previous findings on LOH in gastric cancer .....	71
Candidate tumor-suppressor genes subject to frequent LOH.....	72
3.6 Conclusions .....	74
Acknowledgements.....	74
Author Contributions .....	74
3.7 Figures .....	75
3.8 Tables .....	85
3.9 Supplementary Table.....	93
3.10 Supplementary Figure: .....	96
Chapter 4 LOH ANALYSIS IN CANCER RESEARCH .....	98
4.1 Results .....	98

4.1.1 LOH Analysis of Candidate TSG detected by Whole-Exome Sequencing in Gastric Cancer.....	98
4.1.2 CNA and LOH Analysis in Both SNP Array and Next-Generation Sequence Data .....	99
4.2 Figures .....	101
CHAPTER 5 CONCLUDING REMARKS .....	104
REFERENCES .....	106



## SUMMARY

Gastric cancer is the second leading cause of cancer death worldwide, but has been little studied compared with many other tumors. Copy number alteration (CNA) and loss of heterozygosity (LOH) are two common events that are related to the tumorigenesis in gastric cancer. LOH is a genetic abnormality that causes the loss of one normal allele of a specific gene when the other allele has already been mutated. LOH can result in the inactivation of tumor suppressor genes (TSGs), and therefore, regions that are frequently and independently subject to LOH often harbor TSGs. So far, however, there have been few systematic, genome-wide studies of LOH in gastric cancer. Here we report the results of genome-wide assessments of CNAs and LOH in 45 gastric tumors assayed by Affymetrix SNP 6.0 arrays, each with matched non-malignant DNA. Analysis of regions that frequently undergo LOH in these 45 tumors implicates TSGs already known to contribute to gastric carcinogenesis; these include *TP53* (80%), *CDKN2A* (67%) and *APC* (53%). This analysis also implicates several candidate TSGs that, to our knowledge, have not been previously linked to gastric carcinogenesis. These genes include *PTPRD* and *DOCK8* on chromosome 9p. In addition, we unexpectedly found that the extent of LOH in tumors is highly correlated with gender and with tumor subtypes. These correlations may reflect underlying differences in the mechanisms of cancer development and progression.

Because the accuracy of inference of copy number alteration, LOH, and allelic imbalance from SNP arrays mainly depends on the software applied, we compared eight commonly used free programs with respect to

their sensitivities and specificities. We concluded that ASCAT and CNAG outperformed the other methods.

Finally, our analysis of LOH facilitated and supported the discoveries of novel TSGs in gastric carcinoma and cholangiocarcinoma (bile duct cancer) by next-generation whole-exome sequencing. We also show that one of the analytical methods that we studied, ASCAT, can be applied not only to SNP-array data, but also to next-generation sequencing data.

## TABLE OF ABBREVIATIONS

**GC:** gastric cancer

**CNA:** copy number alteration

**TSG:** tumor suppressor gene

**LOH:** loss of heterozygosity

**BAF:** B allele frequency

**LRR:** log R ratio

**GTC:** Affymetrix Genotyping Console

**APT:** Affymetrix Power Tools

**CNAG:** Copy Number Analyser for Genechip

**GAP:** Genome Alteration Print

**PSCN:** Parent-Specific Copy Number (PSCN)

**TAPs:** Tumor Aberration Prediction Suite

**ASCAT:** Allele-Specific Copy Number analysis of Tumors

**GISTIC:** Genomic Identification of Significant Targets in Cancer

**HDGC:** hereditary diffuse gastric cancer

**RDAAC:** Read Depth and Allele Count, a modified version of ASCAT

applied to exome sequencing data

## LIST OF TABLES

TABLE 1. Several known TSGs that undergo LOH in various cancers. ....	20
TABLE 2. Regions that undergo frequent LOH in gastric cancer.....	21
TABLE 3. Clinical and pathological information on tumors studied.....	86
TABLE 4. Regions with LOH in $\geq 35\%$ of gastric adenocarcinomas.....	87
TABLE 5. Summary of frequently deleted regions.....	88
TABLE 6. Summary of frequently amplified regions.....	89
TABLE 7. Summary of regions with homozygous deletions in more than one sample.....	90
TABLE 8. Strong association between mutations in <i>TP53</i> hotspots and LOH at <i>TP53</i> .....	91
TABLE 9. Cox proportional hazards analysis provides no evidence that LOH proportion influences prognosis.....	92
TABLE 10. Values of different parameters used in ASCAT analysis.....	93
TABLE 11. Tumors for which ASCAT was unable to estimate allele-specific copy numbers.....	95

## LIST OF FIGURES

FIGURE 1. Epidemiology of gastric cancer. ....	4
FIGURE 2. The TP53 pathway.....	10
FIGURE 3. NF- $\kappa$ B pathway activation induced by <i>H. pylori</i> infection. ....	13
FIGURE 4. The canonical Wnt signaling pathway. ....	13
FIGURE 5. Different genetic mechanisms that cause LOH. ....	19
FIGURE 6. The overview of the flow of a Affymetrix Genomewide SNP array. .....	24
FIGURE 7. Flowchart of SNP array data processing procedure for CNA and LOH detection.....	49
FIGURE 8. The clustering and pattern recognition algorithms used by GAP and TAPS. ....	50
FIGURE 9. Comparison of CNA and LOH detection across the genome for different methods at different proportions of tumor and non-malignant cells. 52	
FIGURE 10. LOH on chromosome 1 at varying proportions of tumor DNA as inferred by six programs. ....	54
FIGURE 11. Comparison of sensitivities (A, B, C) and specificities (D)for different methods. ....	55
FIGURE 12. Example ASCAT profile and allele-specific copy numbers. ....	75
FIGURE 13. Examples of tumor and non-malignant pairs that ASCAT was unable to analyze.....	77
FIGURE 14. Relationship between tumor content and ASCAT's ability to generate an allele-specific-copy-number model. ....	78
FIGURE 15. Frequencies of LOH and CNA across 45 gastric tumors. ....	78
FIGURE 16. Identification of significant somatic copy number alterations across gastric cancer by GISTIC.....	79
FIGURE 17. LOH and CNA proportions in males and females.....	81
FIGURE 18. Comparisons of proportions of LOH and CNA in gastric tumors according to the Lauren histological subtypes.....	82
FIGURE 19. Relationship between TP53 mutation and proportion of genome subject to CNA.....	82
FIGURE 20. Relationship between standard deviations of segmented BAF and segmented LRR.....	84
FIGURE 21. Kaplan-Meier survival analysis comparing outcomes by proportion of LOH and average ploidy.....	85
Figure 22. Western blot of proteins PTPRD and DOCK8 using various cell lines. ....	96
FIGURE 23. DOCK8 and PTPRD siRNA knock-down analysis show no significant effect of these two genes on cell proliferation. ....	97

FIGURE 24. The ASCAT profile of two gastric tumors assayed by Affymetrix SNP 6.0. .... 101

FIGURE 25. Comparison of RDAAC analysis using next-generation sequencing data and ASCAT analysis using Affymetrix SNP 6.0 data. .... 103

## **CHAPTER 1 INTRODUCTION AND LITERATURE REVIEW**

### **1.1 Aims and Outlines**

Gastric cancer is a complex disease with high mortality. It is the second most common cause of cancer death worldwide [1], and the five year survival rate for patients with the late stage of cancer is  $\sim 4\%$  [2]. The effects of treatment are very limited, partially due to the intrinsic heterogeneity of the cancer. Therefore, it is important to understand the mechanisms of gastric cancer comprehensively in order to develop more effective therapy.

Tumorigenesis in gastric cancer is caused primarily by genetic alterations [3]. Tumorigenesis is marked by the aberrant regulation of genes that are involved in different signaling pathways such as cell proliferation and apoptosis. The aberrations include genome copy number changes, chromosomal translocations, single nucleotide substitutions, epigenetic modifications, insertions-deletions, and loss of heterozygosity (LOH).

LOH is a type of genetic alteration that often provides the second hit of tumorigenesis in the Knudson two-hit model of tumorigenesis. In the model, the first hit is a mutation on one allele of a gene, and LOH will cost the loss of the other wild-type allele [4]. The detection of LOH in tumors facilitates the discovery of tumor suppressor genes (TSGs) since TSGs are often located in regions that recurrently undergo LOH. In general, at the cellular level, mutations in tumor suppressor genes are recessive, and cells that contain one normal and one mutated gene copy still behave normally. However, LOH causes cells to lose the remaining normal gene and thus to develop into a tumor. TSGs play a key role in carcinogenesis and tumor progression, but our understanding of

TSGs is still limited. While some commonly mutated TSGs have been well studied, there is evidence that more remain to be discovered. The development of SNP arrays and the analytical approaches facilitate the discovery of regions that frequently undergo LOH in gastric cancer, which can lead to the discovery of regions containing TSGs. Therefore, it would be of great significance to understand the pattern of recurrent LOH in gastric cancer.

Since few studies of gastric cancer LOH have been undertaken, our aim is to understand the effect of LOH in tumorigenesis through a comprehensive genome-wide study, and a detailed study of LOH in gastric cancer may present potential biomarkers for prognosis and treatment. In addition, we aim to find novel tumor suppressor genes in LOH regions and to understand the effect of these genes on gastric cancer development.

In the current chapter, we provide the background information regarding gastric cancer epidemiology and pathology, LOH and the Affymetrix SNP 6.0 array platform that we used for whole-genome copy number and LOH analysis.

In Chapter 2, we focus on the analysis of data of gastric cancer tumor-normal pairs with the Affymetrix Genomewide SNP 6.0 platform. We present for the first time a whole-genome LOH map of gastric cancer and evidence for several novel tumor suppressor genes that may play a key role in tumorigenesis. We also showed a significant correlation between LOH and other genetic alterations.

In Chapter 3, we investigate the application of diverse approaches to copy number and LOH analysis and discussed the sensitivity and specificity of these approaches as affected by different tumor content.



In addition to the genomic alteration analysis in gastric cancer, in Chapter 4, we also present the application of LOH analysis to other types of cancers, which aids the discoveries of novel TSGs in the respective cancer types. We also show that the analytical methods of LOH on SNP 6.0 arrays can be adapted to next-generation sequencing data.

## **1.2 General Introduction of Gastric Cancer**

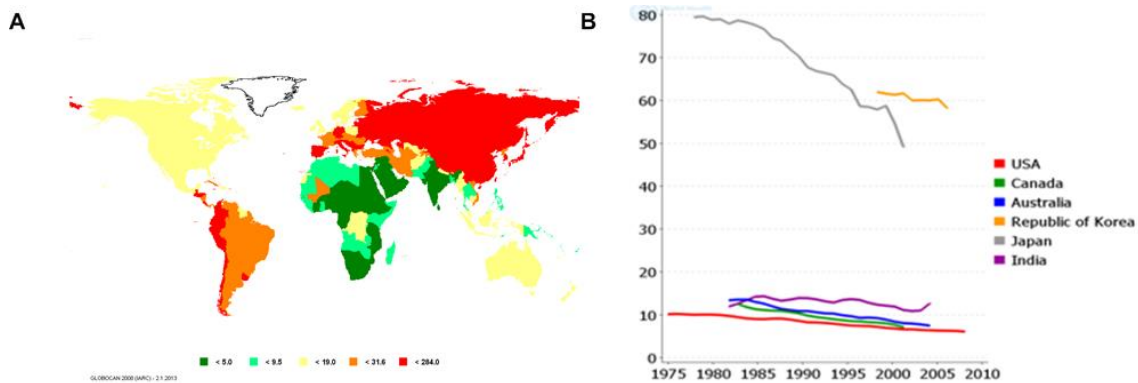
### **1.2.1 Epidemiology**

Gastric Cancer was the fourth most common cancer in the world, with an estimated 989,600 new cases [1]. GC occurrence varies globally and is more prevalent in East Asia (extremely high in South Korea, Japan and China) and South America (Figure1). With 72% of all new cases occurring in male patients, the incidence rate is twice as high in males as in females. Gastric cancer has a high mortality rate and is the second leading cause of cancer death worldwide, accounting for over 700,000 deaths annually [1]. In Singapore, the incidence rate was ~ 22.3 per 100,000 in 2002, with a mortality rate of 17.8 per 100,000 [5]. Despite the steady decline of gastric cancer incidence rates over the 30 years (Figure 1B) presumably due to the diet changes, improved sanitation and increased screening (especially in Japan), the absolute incidence rate has risen because of the aging of the world population.

### **1.2.2 Diagnosis**

Early stage gastric cancer is often asymptomatic and thus can seldom be detected. Therefore, in countries with high incidence rates of gastric cancer such as Japan, mass endoscopic screening programs are conducted for early diagnosis and treatment [6]. Usually, a double-contrast barium x-ray followed by an upper endoscopy (EGD) is the

main procedure to diagnose gastric cancer on patients who present with the symptoms such as weight loss, abdominal pain, nausea and vomiting or those with multiple risk factors [7]. A biopsy is required if any abnormality is seen by EGD. Further tests, including endoscopic ultrasound, computed tomography scan and positron emission tomography scan are necessary after the initial diagnosis of gastric cancer to determine treatment options.



**FIGURE 1. Epidemiology of gastric cancer.**

(A) Age adjusted incidence rates of gastric cancer per 100,000 in 2008. (B) Trends in age adjusted incidence rate of stomach cancer per 100,000 men in selected countries. (Reproduced from Globocan 2008. <http://globocan.iarc.fr/>)

Tumor stage is an important indicator for both diagnosis and treatment of gastric cancer. There are two major staging systems for gastric cancer: the Japanese Classification of Gastric Cancer (JCGC) [8] and the International Union Against Cancer's tumor-node-metastasis (TNM) system [9]. The 5-year survival rates for gastric cancer vary significantly for different stages. The 5-year survival rate for stage I can reach up to 60% while the survival rate for stage IV is only around 4% [10].

### 1.2.3 Histology

Two main systems of histological classification are widely used: the Laurén classification and the WHO classification.

The Laurén system classifies according to the pathological criteria and consists of two main types: the intestinal type and the diffuse type [11]. Intestinal type gastric tumors form irregular tubular or papillary structures and are normally well differentiated. Diffuse type gastric tumors have structures that are inconspicuous, may have signet-ring cells, and are undifferentiated or poorly differentiated. Adenocarcinomas of this type tend to aggressively invade the gastric wall. The intestinal type carcinomas frequently occur in old men while the diffuse type is more prevalent in young women [12]. A third type, termed "mixed", contains both intestinal and diffuse histological features [13]. The mixed type is more aggressive and tends to have larger sized tumors, deeper invasion, and more common lymph node metastasis compared to the other two types [12].

The Laurén classification is roughly comparable with the WHO classification [14]. The WHO classification classifies gastric cancer into four main categories: Tubular, papillary, mucinous, and poorly cohesive based on the descriptive criteria [15]. The papillary, tubular and mucinous subtypes of gastric cancer are usually classified as intestinal according to the Laurén system, while poorly cohesive tumors are always classified as diffuse in the Laurén system.

Genetic and epigenetic alterations vary significantly between intestinal and diffuse gastric cancer [16], which has led to the hypothesis that the two subtypes have different etiologies. The "Correa Model" [17] posits that an intestinal tumor progresses through a number of sequential steps, which usually start from the gastritis caused by *H. pylori* infection. Then it progresses subsequently from atrophic gastritis to carcinoma.

However, diffuse gastric cancer is not included in this model, and no premalignant lesion is known. Thus this model suggests that these two subtypes differ significantly in their molecular pathways of tumorigenesis.

#### **1.2.4 Treatment**

Surgery is the most common treatment for all stages of gastric cancer. The basic goal is to remove all cancer and a margin of normal tissue, but the effects depend on the extent of invasion and the location of the tumor.

Besides surgery, radiation therapy and chemotherapy are often applied to treat gastric cancer. Radiation therapy kills cancer cells by high-energy x-rays and chemotherapy uses drugs to stop the growth and division of cancer cells. Most chemotherapy treatments apply the combination of at least two drugs, such as fluorouracil (5-FU, Adrucil) and cisplatin (Platinol) [18, 19]. Our recent study [20] identified three robust subtypes in gastric tumors: "invasive", "proliferative" and "metabolic" based on gene expression profiling, and we found that metabolic-subtype tumors were preferentially sensitive to 5-FU treatment, while invasive-subtype may be more sensitive to PI3K/AKT/mTOR pathway inhibitors. Another study found that patients with higher EGFR expression benefit from the chemotherapy using the combination of 5-FU and cisplatin. [21].

No standard of care has been established for gastric cancer yet, because gastric cancer is a heterogeneous disease and the relative benefits of drugs are unclear [22]. However, the development of targeted therapy casts light on the treatment of gastric cancer. Targeted therapy interferes with specific molecules that are involved in tumor growth and progression in order to inhibit the growth of cancer. It can improve clinical

outcomes and generally does not have the same types of severe side effects as standard chemotherapy. The recent ToGA (Trastuzumab for Gastric Cancer) trial has shown that Trastuzumab, a HER2 inhibitor, when combined with chemotherapy, improved the overall survival of patients with HER2-positive gastric tumors [22]. We have also reported that FGFR2-amplified tumors show sensitivity to dovitinib, a FGFR/VEGFR targeting agent [23].

### **1.2.5 Etiology**

There is strong evidence that environmental factors, which presumably lead to somatic genetic and epigenetic alterations, play a major role in gastric carcinogenesis. In some cases, there are also known inherited genetic risk factors for gastric adenocarcinoma. Chronic inflammation, exposure to carcinogens and genetic susceptibility significantly increase the risk of gastric cancer [24, 25].

#### **Helicobacter Pylori Infection**

*H. pylori*, a bacterium that colonizes the gastric epithelium, is the strongest known risk factor for gastric cancer, especially cancers in the lower part of the stomach. *H. pylori* is estimated to contribute ~75% of the risk of gastric cancer [26]. *H. pylori* carcinogenesis includes several mechanisms. *H. pylori* infection may lead to chronic gastritis, gastric atrophy and intestinal metaplasia [27-31], which constitute progression towards intestinal-type gastric cancer [17]. Strain-specific bacterial virulence factors, such as the vacuolating cytotoxin VacA and CagA of the cag pathogenicity island (cag-PAI), also play a key role in disease outcome [32, 33].

#### **Diet, Smoking and Alcohol**

Evidence has shown that consumption of salted meat and fish, smoked foods and N-nitroso compounds, together with a low intake of fresh fruits and vegetables, increases the risk of developing gastric cancer [34-36]. Animal experiments indicated that ingestion of salt can cause gastritis and enhance pathogenic response to *H. pylori* infection [37]. A questionnaire study on 2112 Welsh men with 13.8 years follow-up revealed a significant decrease in cancer risk by the consumption of fresh fruits and vegetables [38]. The same result was also observed in a large scale cohort study of 265,118 adults in Japan from 1966 to 1982.

In addition, studies have shown that gastric cancer is associated with smoking and alcohol consumption. The study of 19,657 men from 1990 to 1999 revealed that smoking increased the risk of the differentiated type gastric cancer [39]. Another study between 1974 to 1992 on the cohort of 32,906 people showed that the relationship of gastric cancer with smoking is dose-dependent [40]. Cigarette smoke promotes gastric tumor growth [41] and is a more pronounced risk factor in cardia gastric cancer (gastric cancer in the upper part of the stomach). Another case-control study from Russian based on 448 cases and 610 controls indicated the relationship between hard liquor drinking and cardia gastric cancer in men [42].

### **Genetic Susceptibility**

Several known inherited factors attribute to risk for gastric cancer, including inherited cancer predisposition syndromes, genetic polymorphisms, and germline mutations.

Individuals with inherited cancer predisposition syndromes such as Lynch syndrome, Li-Fraumeni syndrome, and hereditary diffuse gastric cancer (HDGC)

syndrome, have a higher risk of developing gastric cancer [43, 44]. Lynch syndrome, also known as hereditary non-polyposis colorectal cancer, is an autosomal dominant inherited medical condition that has an increased risk of gastric cancer as well as other cancers. Lynch syndrome is caused by the defects in DNA mismatch repair genes such as *MSH2*, *MLH1*, *MSH6*, and *MLH3*, which lead to microsatellite instability [45-48]. Li-Fraumeni syndrome is a hereditary disorder caused by germline mutations in *TP53* [49]. It is characterized by first appearance of cancer at a young age and recurrence over the whole life span. HDGC syndrome is characterized by susceptibility for diffuse gastric cancer and the majority of HDGC patients possess germline mutation of E-cadherin (*CDH1*) [50-53].

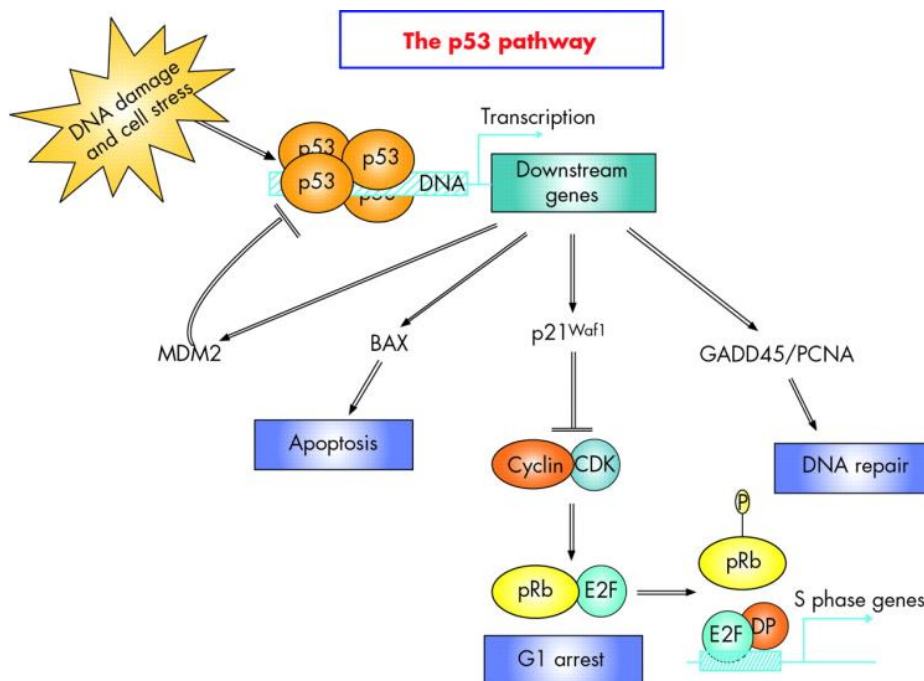
Germline mutations of certain genes were also observed in gastric cancer.  $\beta$ -catenin and *APC* mutations were frequently observed in intestinal type gastric cancer [54, 55]. In addition, *BRCA2* germline mutations were found in 21% of HDGC patients [56].

Polymorphisms in the human interleukin (IL)-1 beta gene and IL-1 receptor antagonist gene are associated with an increased risk of gastric cancer due to *H. pylori* infection [57], and the pro-inflammatory polymorphisms of cytokines TNF- $\alpha$  and IL-10 also altered the risk of noncardia gastric cancer (gastric cancer in all other areas of the stomach other than the top portion) [58]. In addition, recent genome-wide association study of gastric adenocarcinoma tested over 500,000 single nucleotide polymorphisms (SNPs) for association with gastric cancer in 2,240 cases and reported that polymorphism of a SNP in *PLCE1* at 10q23 showed significant relationship with the susceptibility of cardia gastric cancer [59].

### **1.2.6 Molecular Pathogenesis**

Gastric cancer is triggered by multiple somatic alterations, genetic or epigenetic, involving a number of oncogenes, tumor-suppressor genes, and DNA-repair genes. These prominent aberrations include somatic mutations, genomic copy number alterations, LOH, and DNA methylation, histone acetylation/methylation. Accumulated genomic damage eventually affects different cellular pathways and causes them to sustain proliferative signaling, evade growth suppressors, resist cell death, enable replicative immortality, induce angiogenesis, and activate invasion and metastasis [60].

### 1.2.6.1 Cancer-Related Pathways



**FIGURE 2. The TP53 pathway.**  
Reproduced from [61].

### TP53 Pathway

Frequent LOH and mutations in *TP53* are well-known mechanisms of carcinogenesis [62-65]. The *TP53* tumor suppressor gene plays a vital role in the response to environmental and intracellular stresses.



In normal cells, *MDM2*, an E3 ubiquitin ligase, forms a complex with *TP53* to regulate its degradation. Inhibition of this process causes the activation and accumulation of *TP53*. The *TP53* gene then directly regulates the downstream genes to modulate growth arrest (*p21*), apoptosis (*BAX*), DNA repair (*GADD45*) or protein degradation (*MDM2*) (Fig 3).

### **NF- $\kappa$ B Pathway**

Inflammation caused by *H. pylori* infection is strongly associated with gastric carcinogenesis, and the activation of NF- $\kappa$ B is a critical regulator of genes involved in immune and inflammatory responses [66]. The general pathway is described in Fig 4. Without stimulation, NF- $\kappa$ B dimers interact with the inhibitors of NF- $\kappa$ B (I $\kappa$ Bs) and remain inactive in the cytoplasm [66]. The activation and translocation of NF- $\kappa$ B into the nucleus is controlled by the degradation of I $\kappa$ Bs. I $\kappa$ Bs are phosphorylated by the I $\kappa$ B kinases (IKKs) and undergo proteasome-dependent degradation in response to a variety of extracellular stimuli following *H. pylori* infection [67]. *H. pylori* delivers cytotoxin-associated gene A (CagA), a cag-PAI encoded protein, into the epithelial cell cytosol, which is phosphorylated and binds to tyrosine phosphatase to trigger the NF- $\kappa$ B activation cascade. In addition, the activation of the NF- $\kappa$ B pathway is triggered by various pro-inflammatory cytokines such as IL-1 $\beta$  and TNF- $\alpha$ . Lipopolysaccharide (LPS), which target spanning-membrane receptors IL-1R, TNFR and TLR4 (Toll-like receptor 4) respectively, is a major outer membrane component activated by the NF- $\kappa$ B pathway. In tumor cells, the activation of NF- $\kappa$ B has impaired regulation [68-70].

Subsequently, The activation of NF- $\kappa$ B induces inflammatory and tissue-repair genes

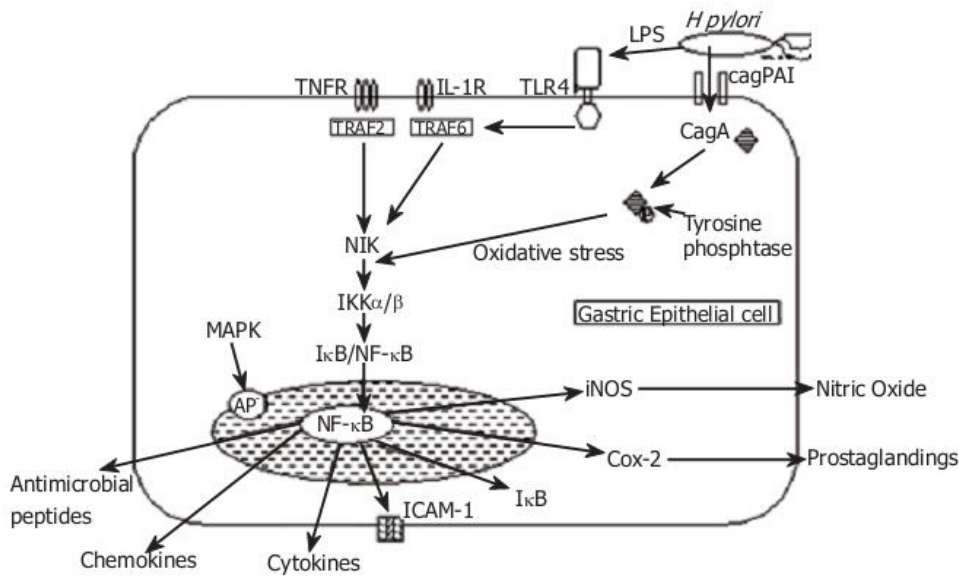
including *MIP-2*, *MMP3*, and *VEGF* and also increases the expression of ICAM-1, a cell adhesion molecule [67].

### **Wnt Signaling Pathway**

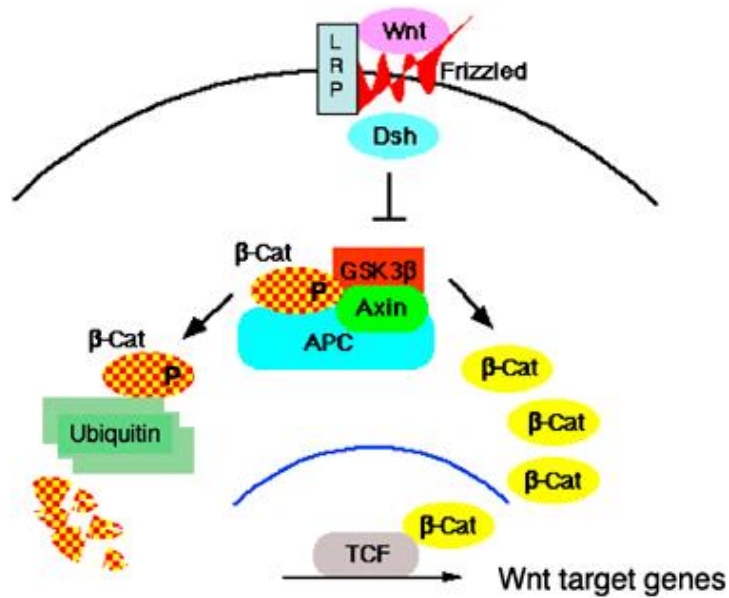
Wnt signaling plays an important role in regulating cell proliferation and differentiation, and the pathway involves multiple interacting factors. Mutations in pathway-related genes including *APC*,  $\beta$ -catenin (*CTNNB1*) and *AXIN*, have been reported in gastric cancers [52, 71, 72]. In the absence of the extracellular Wnt ligands, the APC forms a complex with AXIN and GSK3  $\beta$  to phosphorylate  $\beta$ -catenin, an intracellular signaling molecule, which leads to the degradation of  $\beta$ -catenin. With the activation of Wnt, it binds to the cell-surface receptors of the Fizzled family and activates disheveled family protein (DSH), DSH then inhibits the activity of the APC-Axin-GSK3  $\beta$  -  $\beta$ -catenin complex, resulting in the accumulation of  $\beta$ -catenin in the cytoplasm.  $\beta$ -catenin binds to TCF reporters to activate Wnt target genes involved in cell cycle, apoptosis, cell growth and cell adhesion (Fig 5).

#### **1.2.6.2 Genetic and Epigenetic Alterations**

Gastric cancer is a complex disease and undergoes various somatic genetic and epigenetic alterations, including single nucleotide substitution, genomic copy number changes, loss of heterozygosity, and epigenetic modifications.



**FIGURE 3. NF-κB pathway activation induced by *H. pylori* infection.**  
 Reproduced from [73].



**FIGURE 4. The canonical Wnt signaling pathway.**  
 Reproduced from [74].

**Gene Mutations**

Somatic mutations have long been studied in gastric cancer [52, 55, 75]. The initial studies of somatic mutations in primary gastric cancer focused on *TP53*, a well-known tumor suppressor gene, and discovered a range of non-synonymous mutations that frequently occur in gastric cancer, especially in early-stage tumors [63, 64]. Subsequently, studies have revealed that somatic mutations in *APC* and  $\beta$ -catenin play key roles in tumor progression [52, 55]. *KRAS*, an important gene in the mitogen-activated protein kinase (MAPK) cascade, is also frequently activated due to mutations in the caused progression of gastrointestinal malignancies [76, 77]. In addition, *TTFII*, a candidate tumor-suppressor gene that provides a physical barrier at the gastric mucosa against various noxious agents, was found to lose its expression due to mutations [75].

The development of next-generation sequencing has vigorously boosted the genome-wide mutational analysis of gastric cancer with lower cost compared to previously available sequencing methods. Recent studies using exome-sequencing have reported several novel cancer-related genes that are frequently mutated in gastric cancer, including *ARIDIA* and *FAT4* [78, 79]. Somatic inactivation of these genes by mutations likely contributes to gastric tumorigenesis based on analysis in large series of tumors and based on experiments in cell lines.

### **Copy Number and Gene Expression Alterations**

Genome-wide analyses of gastric cancer have revealed several regions of recurrent changes in DNA copy number, which indicate the possible location of oncogenes and tumor suppressor genes involved in gastric tumorigenesis. Copy number alterations such as recurrent genomic amplifications on chromosome arms 1p, 6p, 7q, 8p, 11q, 16q, 17q, and 20q and deletions on chromosome arms 3p, 4q, 5q, 9p, 16p, 17p and

18q are common in gastric cancer [80-83]. Copy number changes are often accompanied by expression alterations of genes in the corresponding regions. Many gastric cancer-related oncogenes are up-regulated by DNA amplification, including *C-MET*, *K-SAM* and *ERBB2* [83-87]. In addition, gastric cancers sometimes show down-regulation of tumor-suppressor genes (TSGs), including *RUNX3* and *FHIT*, by copy number loss, leading to the loss of functions of these two tumor suppressor genes [88, 89].

The global gene expression profiles of gastric cancer have provided distinct gene signatures for diagnosis and treatment [20, 90-92]. A diversity of gene expression patterns in gastric cancer reflects variations in intrinsic properties of tumor and normal cells and variations in the cellular composition of gastric cancer [93]. Moreover, distinct gene expression profiles were also found between diffuse and intestinal gastric cancer [94]. Intestinal gastric tumors show overexpression of genes involved in cell proliferation, such as *CDX1*, *MYO1A*, *MTP*, and down-regulation of genes that are associated with epithelial differentiation. In contrast, in the diffuse type, genes encoding extracellular proteins are up-regulated, which is accompanied by the down-expression of e-cadherin (*CDH1*). In addition, amplification of *ERBB2* is especially common in intestinal gastric adenocarcinomas, while *K-SAM* and *C-MET* overexpression is more common in diffuse gastric tumors [95, 96].

### **Microsatellite Instability (MSI)**

Microsatellites are short, repetitive DNA sequences that are widely and randomly distributed throughout the human genome. Microsatellite Instability (MSI) is characterized by novel-sized alleles detected in microsatellite sequences that are only in tumor tissues. MSI has been reported in many sporadic gastric cancers [63, 97, 98] and is

sometimes associated with germ-line mutations of the DNA mismatch repair (MMR) genes such as *MSH2* and *MLH1*, which are involved in base-base MMR during DNA replication [99, 100]. Loss of MMR leads to an accumulation of DNA replication errors in cell proliferations, especially in short repetitive nucleotide sequences, which thus leads to MSI. Several studies have found that gastric cancers with high-frequency MSI (MSI-H) show specific clinical phenotypes compared to low-frequency MSI (MSI-L) and microsatellite stable (MSS) tumors [101, 102]. MSI-H tumors tend to be of the intestinal subtype and have higher survival rates. MSI status also plays an important role in characterizing tumors and predicting prognosis, with MSI-L/MSS group showing better response to 5-FU treatment [103].

### **Epigenetic Modifications**

Epigenetic modifications, such as DNA methylation and histone acetylation or methylation, are important alterations in gastric cancer. DNA hyper- and hypomethylation at CG dinucleotides (CpGs), were discovered to have a correlation with tumor suppressor gene silencing and oncogene overexpression, respectively [104, 105], and hypermethylation of CpG islands (CGIs, regions of high CpG density) in gene promoters is widely associated with transcriptional silencing in cancer [105]. Numerous studies have investigated the role of DNA methylation in gastric cancer development, identifying genes frequently hypermethylated in gastric tumors such as *MLH1* and *CDKN2A* [55, 98, 106]. Methylation is also a common second hit to the *CDH1* gene subsequent to the germ-line mutation of the first allele, which is a major cause for hereditary diffuse gastric cancer [107]. In eukaryotic cells, five different histone proteins exist, termed H1, H2A, H2B, H3 and H4. Certain histone modifications, such as

methylation, and acetylation can lead to the change of the structure of the chromatin, which may contribute to gene activation or silencing. Hypo-acetylation of histones H3 and H4 in the *p21* promoter region is frequently observed in gastric cancer [108]. While conducting global acetylation analysis, the level of acetylated histone H4 is much lower in gastric cancers compared with that in non-neoplastic mucosa, which indicates a strong correlation between the reduced histone H4 acetylation and the tumor progression [109].

### **1.2.7 Gastric Cancer Biomarkers**

Gastric cancer is a heterogeneous disease comprising multiple intrinsic subtypes and is naturally resistant to many anticancer drugs. The discoveries of biomarkers to this cancer may make personalized treatment possible, which may reduce the mortality and improve the effectiveness of therapies. Tan *et al* used gene expression profiles to reveal the distinct biological properties of gastric cancer groups, and found two intrinsic genomic subtypes (G-INT and G-DIF) that had different response to 5-FU and oxaliplatin treatment [110]. Another recent study showed three robust subtypes ("invasive", "proliferative" and "metabolic") from the study of two large gastric cancer cohorts and discussed their differences in therapeutic vulnerabilities [20]. In addition, a comprehensive bench-to-bedside model for personalized treatment of gastric cancer considering both genomic markers and environmental effects has been developed and proposed [111]. Patients can be classified into low and high metastatic risk groups according to prognostic gene signatures and low-risk patients can avoid chemotherapy toxicity by applying surgery alone. Although gastric cancer treatment remains a major challenge, an increasing number of studies reveals that new, robust biomarkers may significantly improve the survival rates of patients.

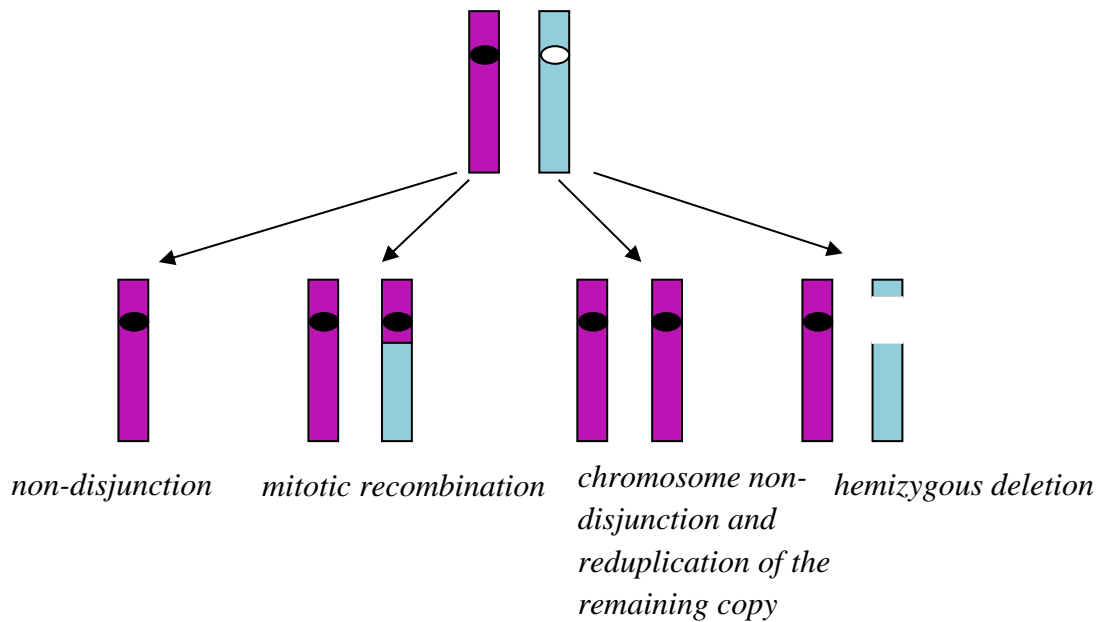
## **1.3 Loss of Heterozygosity**

### **1.3.1 General Introduction to LOH**

Loss of heterozygosity (LOH) is a genetic abnormality that causes the loss of one normal allele of a specific gene when the other allele has already been mutated. In other words, LOH is a constitutional genotype change from heterozygous to homozygous in somatic cells. According to Knudson's "two hit" hypothesis of tumorigenesis raised in 1971 [4], the first hit is usually a point mutation that inactivates one copy of a tumor suppressor gene (TSG), and individuals will not develop cancer at this point. Instead, they develop cancer only when the second hit occurs, which causes the loss of the remaining functional TSG allele. LOH occurs at a higher frequency than single nucleotide substitutions and the still-intact copy of the gene is far more likely to be lost through LOH than by a second point mutation. Hence, LOH is a common second hit in carcinogenesis. As a result, detection of LOH has been widely used to identify genomic regions that harbor TSGs and to characterize different tumor types, pathological stages and progression [112, 113].

LOH may occur due to non-disjunction, mitotic recombination, deletion, chromosome non-disjunction and reduplication, or gene conversion (Fig 6). Regions subject to hemizygous loss of DNA copy number exhibit LOH, but the converse is not always true. LOH without copy number changes, or Copy Number Neutral Loss of Heterozygosity (CNNLOH), is caused by duplication of the chromosome containing the mutated allele and loss of the chromosome containing the normal allele. It is also common in gastric cancer.





**FIGURE 5. Different genetic mechanisms that cause LOH.**

(White circle: normal allele, black circle: mutated allele.)

### 1.3.2 LOH and TSGs in Cancers

LOH is a common genetic alteration that is observed in various solid tumors. LOH was found on chromosome 5 in 20% of the colorectal cancers [114]. Subsequent studies showed that the mutational inactivation of TSGs caused by LOH predominates in colorectal cancer, including LOH of alleles at chromosomal regions 5q (*APC*), 17p (*TP53*) and 18q [115]. A significant level of LOH has been found at several sites in ovarian cancer, including 3p, 6q, 11p, 17q [116-118]. LOH occurs most frequently on chromosomes 3p, 13q and 17p in lung cancer, likely representing the inactivation of key tumor suppressor genes including *FHIT*, *TP53* and *RB* [119-121]. In such cancers as lung,

ovarian, and colorectal cancers, LOH is found at an early stage of tumor progression [122-124].

These findings, accompanied by the discoveries of gene mutations in region of LOH, reveal many TSGs. Inactivation of *TP53* due to loss of heterozygosity has been demonstrated in a variety of cancers [125-127]. *TP53* plays a key role in apoptosis, genomic stability, and inhibition of angiogenesis. The *p16* gene encodes a protein that can inhibit the ability of CDK4 and CDK6 to phosphorylate the retinoblastoma protein and the inactivation of this protein may lead to uncontrolled cell cycling and growth. The region containing *CDKN2A* undergoes frequent allelic loss in multiple cancers [128-131]. Several well-studied known TSGs are summarized in table 1.

### 1.3.3 LOH in Gastric Cancer

LOH plays an important role in gastric tumorigenesis due to its ability to inactivate TSGs. LOH can be detected in up to 80% of gastric tumors, and LOH frequency increases during tumor progression [132]. Many studies have been conducted to comprehensively analyze LOH in gastric cancer and have revealed several chromosome arms that frequently undergo LOH, including 1q, 3p, 4p, 5q, 7p, 8p, 9p, 11q, 12q, 13q, 17p, 18q, 21q, and 22q [133-135]. LOH analysis also identified several arms and regions along the genome that contain TSGs important in gastric tumorigenesis, such as 17p (*TP53*) [134], 5q (*APC*), and 18q (*DCC* and *SMAD4*). Table 2 summarizes the chromosomal regions that frequently undergo LOH in gastric cancer and the tumor suppressor genes (if known) that are targeted by LOH events in these regions.

**TABLE 1. Several known TSGs that undergo LOH in various cancers.**

Gene	Chromosomal Location	Sporadic Cancers	Function of Proteins	Reference
------	----------------------	------------------	----------------------	-----------

<i>RUNX3</i>	1p36	gastric cancer	transcription factor (TF) co-factor	[136]
<i>FHIT</i>	3p14.2	many types	diadenosine triphosphate hydrolase	[137-139]
<i>APC</i>	5p21	colorectal, pancreatic, and stomach carcinomas; prostate carcinoma	$\beta$ -catenin degradation	[140-143]
<i>CDKN2A (p16)</i>	9p21	many types	CDK inhibitor	[128-131]
<i>PTEN</i>	10q23.3	glioblastoma; prostate, breast, and thyroid carcinomas	PIP <sub>3</sub> phosphatase	[144, 145]
<i>RB</i>	13q14	retinoblastoma; sarcomas; bladder, breast, esophageal, and lung carcinomas	transcriptional repression; control of E2Fs	[146-149]
<i>CDH1</i>	16q22.1	invasive cancers	cell-cell adhesion	[150, 151]
<i>TP53</i>	17p13.1	many types, up to 50% of all tumors	transcription factor	[125-127]

**TABLE 2. Regions that undergo frequent LOH in gastric cancer.**

Chromosomal regions	TSGs	Frequency	References
1q		50%-67%	[134, 152]
3p	<i>FHIT</i>	32.4%	[153]
5q	<i>APC, MCC</i>	34%-60%	[134, 154]
7q	<i>TES</i>	32%-43%	[154-156]
8p		44%	[157]
9p		36.4%	[158]
10p	<i>KLF6</i>	53%	[159]
12q		55%	[152]
13q		38.1%-41%	[158, 160]
17p	<i>TP53</i>	37.5%-68%	[134, 156, 158, 161]
18q	<i>SMAD4, DCC, BCL2</i>	29%-61%	[158, 161-163]
21q		40%-43%	[154, 156]

## 1.4 Genome-Wide SNP Array Application on CNA and LOH Analysis

### 1.4.1 Genome-wide SNP Array

Traditional methods to study CNA and LOH such as restriction fragment length polymorphism (RFLP), molecular cytogenetic analysis, fluorescence in situ hybridization

(FISH) and microsatellite analysis, require a lot of time and effort. The instability of the markers and the difficulty of automating PCR based analysis make their usage for genome-wide studies unpractical.

The invention of microarrays makes possible the high-throughput analysis of CNA and LOH on a genome-wide scale. Array-based comparative genomic hybridization (aCGH) is a technology with probes detecting total copy number of genomic sites. Because aCGH is unable to detect allelic states of SNPs, it is only applicable to analyze genome-wide DNA CNAs [164-166]. In addition, the number of probes and thus the genomic resolution in this older array technology was lower than that of the chips now available [167].

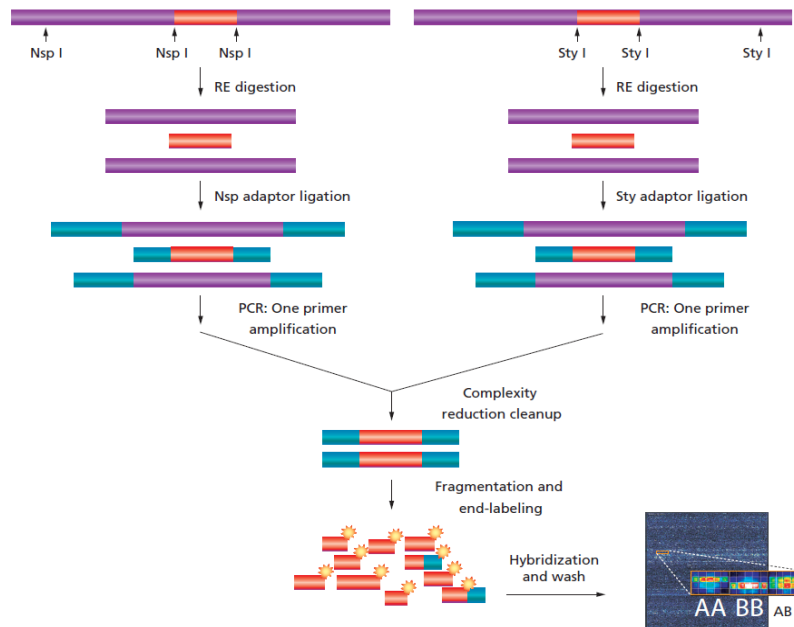
The development of single nucleotide polymorphism (SNP) arrays provides a major advance. SNP arrays offer the advantage of providing high resolution genotype information in addition to copy number variation information in a single experiment with a very high density of genomic coverage. With genotype information, it is possible not only to determine the de novo CNA in patients, but also to decide the origins of chromosomes, check for sample mix-up, and study copy number neutral genomic variation such as uniparental disomy and copy number neutral loss of heterozygosity. Advances in computational methodology have been critical in facilitating application of this technology to molecular genetics.

Affymetrix and Illumina are two major manufacturers of SNP array platforms that are widely used. The differences between them are the underlying technologies. Affymetrix GeneChip assays are based on the hybridization of genomic DNA to assays. Oligonucleotides of probes are printed directly on the chips. The probes are 25-mer

sequences targeting both alleles of the SNPs. A DNA segment with complete complementary sequences will be hybridized to the chip and bind to the specific probe more efficiently than a probe with a mismatch, which is represented by a higher fluorescence signal that can be detected. However, Illumina Infinium II assays use a single base extension method to obtain the allelic information for SNPs. Beads with DNA probes sticking out of them are randomly deposited into microwells on a substrate. Each bead contains a 29-base unique sequence to allow the identification of probes. The probe sequences are 50 bases long SNP locus-specific primers and are complementary to the sequences adjacent to the SNP sites. After DNA fragmentation and hybridization to probes, they are extended with hapten-labelled nucleotides and the incorporated nucleotides are detected by fluorescence-labelled antibodies for further analysis. Because of these differences, data processing procedures of the platforms from these two manufacturers are slightly diversified. Our study only uses Affymetrix SNP 6.0 arrays and the details will be described.

The Affymetrix Genome-Wide Human SNP Array 6.0 allows us to look simultaneously at more than 1.8 million probe sites on a single array with an inter-marker distance of 696 base pairs [168]. The chip includes 906,600 SNP probes and 946,000 non-polymorphic probes, with the latter targeting non-SNP loci. For each locus, the chip contains three or four replicated pairs of 25-nt perfect-match (PM) probes quantifying the amount of DNA target to optimize the accuracy of estimation. The two alleles of a SNP are arbitrarily labeled as "allele A" and "allele B". Therefore, the genotype of a SNP from a diploid sample is typically AA, AB or BB.

The basic flow of processing for SNP 6.0 arrays is described in Fig 8. In short, 500 ng of total genomic DNA is digested with Nsp I and Sty I restriction enzymes. Fragments from restriction enzyme digestion, regardless of size, are then ligated to universal adaptors. A generic PCR primer that recognizes the adaptor sequence is used to amplify these adaptor-ligated DNA fragments, and PCR products of 200 to 1,100 bp in size range are preferred by PCR conditions. These PCR amplified products are combined and purified, and then they are fragmented, labeled, and hybridized to a SNP 6.0 Array (Figure 7). The probe-level fluorescence signal intensities are then detected and processed.



**FIGURE 6. The overview of the flow of a Affymetrix Genomewide SNP array.** Reproduced from [www.affymetrix.com](http://www.affymetrix.com).

Each allele of a SNP has three or four replicated perfect match probes that are completely complementary to the sequences of the allele. These probes are printed directly on the chip. As shown in Figure 7, six rectangles comprises the probes for a SNP,

with the top three target one allele and the bottom three target the other allele, respectively. If there are equal amounts of DNA with each genotype, the post-hybridization intensities of the top three rectangles will be similar to those of the bottom four rectangles. If only one allele is present in the DNA, e.g. the allele complementary to the top rectangles, then the top three rectangles will be much brighter than the bottom four rectangles. If there is allelic imbalance, which can occur in cancer, then the two alleles might be present, for example, in a ratio of 2:1. In this case, if the more abundant allele is detected by the top rectangles, these will be approximately twice as bright as the bottom rectangles. Therefore, from the signal intensities of the rectangles detected, we can infer the allele specific copy numbers.

After hybridization, each SNP array slide is scanned and the array probe signal intensities are obtained. A number of preprocessing steps are required to convert raw intensity measurements into biological inferences, and these steps can significantly influence the quality of the ultimate measurements. The underlying principle is that the signal intensities mainly depend on the amount of target DNA in the sample. The intensities of the fluorescence signals may be affected by various systematic variations such as the array manufacturing process and the affinity between targets and probes. To accurately estimate the true copy number of each allele, the raw data need to be normalized considering these systematic errors. Many algorithms have been developed to normalize the raw intensities [169-171], such as quantile normalization [172] and Copy-number estimation using Robust Multichip Analysis (CRMA) v2 [173]. These algorithms take the raw intensity image of SNP arrays as inputs and derive the signal intensities for

the A and B alleles through several normalization steps. The former algorithm is applied in analyzing multiple arrays, while the latter one is suitable for single array processing.

Genotyping is also crucial to obtain information out of the raw intensities. For the early series of SNP data, the partitioning around medoids (PAM)-based algorithm [174] and a dynamic model algorithm [175] are utilized. With the evolution of the platforms, more and more new algorithms have been developed and are proven more accurate in genotyping the SNP arrays, such as BRLMM-P [176], which models the log-transformed intensities as a stochastic function, and Birdseed [177], which uses an expectation-maximization (EM) procedure to fit the signals from the test samples to a two-dimensional Gaussian mixture model with a priori.

The preprocessed data are further analyzed for various applications such as genome-wide association studies, copy number and LOH analysis. The latter two analyses will be described in the next part.

#### **1.4.2 Application of SNP Arrays on Copy Number and LOH Analysis**

The SNP array offers the ability to define CNA and LOH in a tumor simultaneously and is a powerful platform for oncogene and TSG discoveries. Genome-wide LOH analysis using SNP arrays is typically performed by comparison of tumor and adjacent normal genomic DNA from the same individual. LOH can be discovered by a change from a heterozygous state in the normal sample to a homozygous state in the tumor sample. If a matched normal is unavailable, several algorithms are also available to analyze the SNP array using pooled references [178, 179].

Many free or commercial tools have been developed for the SNP array data analysis, including Affymetrix Genotyping Console (GTC), Affymetrix Power Tools



(APT), Nexus (BioDiscovery, <http://www.biodiscovery.com/software/nexus-copy-number/>), Copy Number Analyser for Genechip (CNAG) [179], dchip [180], PennCNV [163], PICNIC [181], QuantiSNP [182], Genome Alteration Print (GAP) [183], Parent-Specific Copy Number (PSCN) [184], Tumor Aberration Prediction Suite (TAPs) [185] and Allele-Specific Copy Number analysis of Tumors (ASCAT) [186]. Details will be described in Chapter 2.

**CHAPTER 2 COMPARISON OF SOFTWARE FOR IDENTIFYING COPY  
NUMBER ALTERATIONS AND LOSS OF HETEROZYGOSITY FOR  
AFFYMETRIX SNP 6.0 ARRAYS**

Yingting Wu<sup>1,2</sup>, Roy Welsch<sup>3</sup>, Steve Rozen<sup>1,2</sup>

<sup>1</sup> Computation and Systems Biology, Singapore-MIT Alliance

<sup>2</sup>Neuroscience & Behavioral Disorder program, Duke-NUS Graduate Medical School,  
Singapore

<sup>3</sup>Sloan School of Management and Engineering System Division, Massachusetts Institute  
of Technology, MA, USA

**\*Author Contributions:**

Yingting Wu carried out all analyses and writing with supervision from her thesis  
advisors.

## **2.1 ABSTRACT**

We evaluated several publicly available software packages for analysis of copy number analysis (CNA) and loss of heterozygosity (LOH). We evaluated their performance on a previously published data set [185] that consists of a dilution series of cancer-cell-line DNA mixed with matched germ-line DNA. The dilution series was assayed on Affymetrix SNP 6.0 arrays. Here we describe, compare, and evaluate the performance of the algorithms utilized by these methods in each step of analyzing SNP array data. ASCAT and CNAG outperformed the other methods in inference of CNA and LOH.

## 2.2 INTRODUCTION

CNA and LOH are important types of genetic alteration in cancers [187], and characterization of these alterations plays a key role in both diagnosis and drug development. Single nucleotide polymorphism (SNP) arrays provide a genome-wide high resolution view of these alterations. Several high-throughput studies have applied SNP arrays to characterize CNA and LOH in various cancers [188-191]. The power of SNP arrays for this application depends on sophisticated computational methods.

Affymetrix Genome-Wide Human SNP Array 6.0 arrays interrogate > 1.8 million genomic sites, including 906,600 SNPs and 946,000 non-polymorphic sites, at the latter of which it assesses copy number. For this array design, average inter-marker distance is 696 base pairs [168]. For each SNP site, the chip contains three or four overlapping pairs of 25-nt probes. Each probe is perfectly complementary to one of the two alleles at the SNP. The two alleles of a SNP are by convention labeled "allele A" and "allele B" regardless of the actual bases. Thus, each SNP probe is associated with either "allele A" or "allele B", and the genotype of a SNP is denoted AA, AB or BB. To estimate copy number and allelic imbalance at each SNP one uses two values: the log R ratio (LRR) and the B allele frequency (BAF). LRR is the log of the ratio of observed tumor probe intensities to reference normal intensities, and deviations of LRR from zero are evidence for CNA. BAF is the proportion of the B allele in the two-allele mixture. Deviations of BAF values from the expected 1:1 ratio at heterozygous sites constitutes allelic imbalance and indicates aberrant copy numbers of at least one of the two homologous chromosomes at that site.

Complications in estimation of CNA and LOH from LRR and BAF arise from several sources, including prominently (1) tumor aneuploidy and polyploidy and (2) admixture of DNA from non-malignant genomes.

Tumor aneuploidy is the chromosomal instability that reflects the defects in mitotic segregation in cancer cells. The total amount of DNA in an aneuploid tumor sample can differ significantly from the diploid normal sample. Due to the restriction of the technique, the protocol for SNP arrays constrains the amount of DNA other than the number of cells to be the same for each assay. Therefore, a  $2n$  segment in a triploid tumor sample will show smaller signal intensities compared to the same  $2n$  segment in the diploid normal sample ( $LRR < 0$ ), and without adjusting the ploidy state, the data will be similar as a hemizygous deleted segment in a diploid tumor sample. The reason is that the zero baseline of the LRR does not represent a normal diploid copy number but an average copy number of the tumor sample.

The second problem in mining SNP array data arises from the admixture of non-tumor cells in the tissue sample from which the DNA sample is extracted. The presence of a normal DNA dilutes the amplitude of the signal changes that reflect the genomic alterations in the tumor DNA. Thus, using fixed thresholds to detect the copy number variations may fail due to this admixture of non-tumor DNA [192, 193], and considering the proportion of tumor DNA increases the accuracy of copy number estimation [185, 186].

The basic steps to analyze data generated by SNP arrays include: (1) normalization, (2) genotype calling, (3) LRR and BAF calculation, (4) segmentation, and

(5) CNV and LOH calls. Figure 20 provides a comprehensive overview of the algorithms used by each tool for each step.

Many free or commercial tools have been developed to estimate CNA and LOH from SNP array data, and they vary in their approaches to each of the basic steps.

Several studies have evaluated and compared methods for CNA detection. Winchester *et al.* [194] compared the performance of five methods, including Birdsuite [177], CNAT (Copy Number Analysis Tool) [195], GADA (Genome Alteration Detection Analysis) [196], PennCNV [163], and QuantiSNP [182], on data generated by both Illumina 1M Duo and Affymetrix SNP 6.0 platforms. This study suggested the use of any two programs on a single dataset in order to utilize the advantages of each software package to improve sensitivity and specificity. This study also recommended the use of software that is specially designed for the platform to be used, such as QuantiSNP for Illumina SNP array data. In another study, Dellinger *et al.* [197] evaluated seven methods, including CBS (circular binary segmentation) [198], CNVFinder [199], cnvPartition [200], GLAD (Gain and Loss Analysis of DNA) [201], Nexus [202], PennCNV and QuantiSNP, in various processing steps on data generated by both the Illumina HumHap 550 and Affymetrix SNP 6.0 platforms. This study recommended determining the optimal parameters using a subset of samples with high-quality genotype call rates before analyzing the whole dataset. In yet another study, Eckel-Passow *et al.* [203] focused on comparing the locus-level copy number estimates generated by four different tools, including APT (Affymetrix Power Tools) [204], Aroma. Affymetrix [205], PennCNV and CRLMM (Corrected Robust Linear Model with Maximum Likelihood Distance) [206], on data from Affymetrix SNP 6.0 arrays. They found that PennCNV had

a better performance and a more user-friendly interface and detected CNAs with smaller bias and variability of locus-level copy number data. These studies all showed that various methods have diverse advantages and need to be carefully chosen to meet specific requirements.

However, these studies only focused on CNA analysis and did not carry out any comparison of the performance of LOH analysis, even when the packages evaluated offered this capability. In addition, all these studies compared segmentation-based algorithms and HMM-based algorithms to identify CNAs, but the performance of recently developed pattern recognition algorithms have not been investigated. The differences between the types of algorithms will be discussed later. An additional limitation of these studies is that they are all based on unpaired analysis, despite the fact that paired analysis of CNA and LOH by comparing the tumor sample to its matched normal samples has been effective in avoiding the miscall of germ-line copy number polymorphisms. Therefore, the present study focuses on the performance of different methods that simultaneously carry out both CNA and LOH detection.

Here, we evaluate eight programs, including GAP (Genome Alteration Print) [183], Birdsuite, PennCNV, CNAG (Copy Number Analyzer for GeneChip) [179], PICNIC (Predicting Integral Copy Numbers In Cancer) [207], paired PSCBS (Parent-Specific Copy-number Segmentation) [208], TAPS (Tumor Aberration Prediction Suite) [185] and ASCAT (Allele-Specific Copy Number analysis of Tumors) [186]. Although some of the programs have been already evaluated, we assessed the performances of both CNA and LOH analysis by these programs. We apply each of these methods to data from a dilution series of a single tumor cell line (NCI-H1395) mixed with increasing

proportions of germ-line DNA from the same donor, as reported in [185]. Spectral karyotyping indicates that the cell is approximately triploid [209]. The dilutions were assayed on Affymetrix Genome-wide SNP 6.0 arrays, and a spectral karyotype of the tumor was available to provide information on copy number independent of the Affymetrix data. The analysis provides a detailed picture of the performance of these eight programs at varying mixtures of tumor and normal DNA.



## **2.3 MATERIALS AND METHODS**

### **2.3.1 Lung Cancer Cell lines**

The Affymetrix Human Genome-wide SNP 6.0 data of lung cancer cell line NCI-H1395 with different tumor content (100%, 70%, 50% and 30%) and its patient-matched blood cell line NCI-BL1395 were obtained as .CEL file from GEO accessions GSE29172 and GSM645856 ([www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/)). In the dilution series, DNA from normal blood cell line NCI-BL1395 was mixed with DNA from the lung cancer cell line NCI-H1395, and the DNA ratio was adjusted to compensate for NCI-H1395 being nearly triploid, so the proportions of tumor DNA were 100%, 80%, 65%, and 42%.

### **2.3.2 CNA and LOH Analysis**

#### **CNAG**

CNA and LOH analysis was performed by CNAG (version 3.3.0.1 beta) using the default parameters. Paired samples with their references are matched in the data extraction. We chose “non-allele-specific analysis” with self-reference only. CN gains, losses and LOH were defined according to the default.

#### **Birdsuite**

CNA and LOH analysis was performed by PennCNV according to the online instructions (<http://www.broadinstitute.org/science/programs/medical-and-population-genetics/birdsuite/birdsuite-manual>). Affymetrix Power Tools were utilized for data normalization before Birdsuite analysis. Birdseed was used for SNP genotyping and Canary was applied to genotype the known CNPs. The default settings by Birdsuite were used to run the programs.

#### **GAP**

CNA and LOH analysis was performed by GAP according to the online instructions and default settings ([http://bioinfo-out.curie.fr/projects/snp\\_gap/](http://bioinfo-out.curie.fr/projects/snp_gap/)). Allelic difference data were output from Affymetrix Genotyping Console 4.0 and were directly utilized by GAP as the inputs.

### **PennCNV**

CNA and LOH analysis was performed by PennCNV according to the online instructions ([http://www.openbioinformatics.org/penncnv/penncnv\\_tutorial\\_affy\\_gw6.html](http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affy_gw6.html)). Quantile normalization and birdseed-v2 calling algorithm were used by Affymetrix Power Tools to generate the signal intensities and genotyping calls from raw CEL files. Allele-specific signals were then extracted to calculate LRR and BAF values. Copy number variation and LOH were calculated based on the Hidden Markov Model (HMM).

### **PICNIC**

CNA and LOH analysis was performed by PICNIC according to the online manual ([ftp://ftp.sanger.ac.uk/pub/cancer/picnic\\_software/picnic\\_src/PICNIC\\_implementation\\_guide.pdf](ftp://ftp.sanger.ac.uk/pub/cancer/picnic_software/picnic_src/PICNIC_implementation_guide.pdf)). Matlab is used to execute the program. In general, we applied the default settings of PICNIC to normalized the raw signal intensities and generate contamination fraction and ploidy estimations. Then such information was the premise to infer CNA and LOH by HMM.

### **ASCAT**

CNA and LOH analysis was performed by ASCAT. We preprocessed the raw SNP 6.0 array fluorescence signal with CRMA v2 and TumorBoost to get the normalized

LRR and BAF data. Then we applied ASCAT to generate allele-specific copy numbers from the pre-processed data. We used the following parameter settings: median smoothing; minimum segment length ( $k_{\min}$ ) = 100 SNPs; LRR compaction factor ( $\gamma$ ) = 0.5. In addition, we modified ASCAT to use a new parameter,  $\alpha$ , which represents a BAF compaction factor analogous to the LRR compaction factor  $\gamma$ . We used  $\alpha = 0.6$  for our analysis of Affymetrix SNP 6.0 arrays. We also modified the procedure whereby ASCAT's segmentation algorithm determines whether there is allelic imbalance at a particular segment. All the programs are written in R.

### **TAPS**

CNA and LOH analysis was performed by TAPS. We used CRMA v2 to preprocess the raw data, which is the same as described in ASCAT preprocessing. Segmentation of LRR is conducted by CBS. TAPS were then performed using the default settings.

### **Paired-PSCBS**

CNA and LOH analysis was performed by paired-PSCBS according to the online instructions (<http://www.aroma-project.org/vignettes/PairedPSCBS-lowlevel>). We used CRMA v2 to preprocess the raw data, which is the same as described in ASCAT preprocessing. The R package "Paired PSCBS" was utilized to run the programs. LOH was inferred by the LOH calling algorithm integrated into the package. We calculated the standard deviations by comparing the observed LRR to the estimated LRR. We also calculated the median estimated LRR, and the positive value of the median estimated LRR was set as the arbitrary sample-adaptive threshold for CN gains and the negative value was the threshold for CN losses.

## **2.4 RESULTS AND DISCUSSION**

Using the dilution series data generated from the lung cancer cell line NCI-H1395 and its matched blood cell line NCI-BL1395 from [185], as described above, we evaluated the performance of eight programs commonly used for analyzing CNA and LOH. These tools apply various algorithms, as shown in Figure 20. A summarization of the features of these tools is also presented in Tables 10 and 11.

### **2.4.1 Data Pre-processing**

The intensities of the fluorescence signals obtained by scanning the image of the chips are affected by various factors including both biological and non-biological variation. Several preprocessing steps are required to convert raw intensity measurements into biological inferences, and these steps can significantly influence the qualities of the ultimate inferences. Several preprocessing algorithms have been proposed to reduce unwanted non-biological variability and obtain the accurate CN information for each allele.

GAP utilizes the CN5 algorithm [210], which is implemented on the Affymetrix Genotyping Console (GTC) to normalize signal intensities and estimate raw CNs. CN5 applies adaptive background correction to address issues of optical background noise, the effects of non-specific hybridization, to probe-specific variation in intensity. To normalize across arrays, it uses sketch quantile normalization as the across-array normalization algorithm to address the effects of non-biological variation due to chip manufacturing and experimental procedures. Quantile normalization is a scaling based algorithm that makes use of a baseline array (usually with the highest genotype calls). The algorithm assumes that the distribution of signal intensity will not change and tries to

make the test array distribution and the baseline array distribution identical in statistical properties.

Birdsuite and PennCNV employ Affymetrix Power Tools (APT) for preprocessing. APT uses RMA (Robust Multi-array Analysis) for background correction, sketch quantile normalization for inter-array variation removal, and PLIER (Probe Logarithmic Intensity Error) for probe intensity summarization.

PICNIC [207] first constructs a training group in the preprocessing step with 461 normal samples from the Affymetrix SNP 6.0 array. It then applies a Bayesian approach to fit the allelic signal intensities of the tumor sample to the three clusters observed in the training group and uses the maximum posterior estimation to obtain normalized allele-specific signals that map the three clusters. During this process, aneuploidy samples are automatically adjusted. It also estimates parameters required for segmentation, including tumor content and tumor ploidy.

GAP, Birdsuite, PennCNV, and PICNIC all normalize the signal intensities of the tumor samples to pooled reference samples. The former three methods use the 270 HapMap samples from the International HapMap Project [211] and PICNIC uses the 461 normal samples in the training group from [207]. Normalization strategies based on a pooled reference sample do not in practice distinguish somatic copy number alterations from germ-line copy number variation, because the algorithms do not examine the non-tumor reference samples for copy number variation.

CNAG [179] corrects the raw signal intensities of a sample by compensating for varied enzyme-digested fragment lengths and for GC content. CNAG addresses the issue of aneuploidy by accepting from the user an indication of regions that are a-priori thought

to be diploid. In CNAG, a paired normal sample is used as the reference for the corresponding tumor sample.

CRMA v2 (Copy-number estimation using Robust Multichip Analysis) [173] is a single-array method to remove chip background noise, crosstalk between alleles, the effects of probe sequence composition on the stability of hybridization, and effects due to the varying lengths of the restriction fragments that are hybridized to the probes (and that vary systematically from SNP to SNP). Unlike Birdsuite and PennCNV that utilize quantile normalization and process multiple samples together, CRMA v2 can process each array independently, and for the paired analysis, only the tumor sample with its matched normal sample are required. Thus, it is easy to apply CRMA v2 to new arrays as they are produced without having to reprocess the old arrays. ASCAT, paired PSCBS and TAPS all use CRMA v2 for preprocessing.

#### **2.4.2 Genotyping**

Genotyping is a crucial step in the analysis of SNP array data. Accurate genotyping helps to exclude non-informative SNPs (those that are homozygous in the normal sample). However, because of tumor aneuploidy and because DNA from solid tumors usually also contains DNA from non-malignant cells, genotyping SNPs in the tumor sample is a very different problem from, and much more difficult than, genotyping SNPs in non-malignant DNA.

GAP uses BRLMM-P (Bayesian Robust Linear Model with Mahalanobis distance classifier) [176] conducted on GTC for genotyping. BRLMM-P first calculates the initial genotype for each SNP using the DM algorithm [212]. It then random selects a subset of non-monomorphic SNPs to estimate the prior information of cluster centers and variance-

covariance matrices. Combining the initial genotypes and the prior information, the algorithm applied a Bayesian procedure to get the posterior estimation of the three cluster centers as well as the variance matrices. Finally, SNP genotypes are delineated based on the Mahalanobis distance between SNPs and the three cluster centers.

Birdsuite and PennCNV rely on Birdseed v2 [177] for genotyping. Birdseed v2 uses an expectation-maximization procedure to fit the signals from the test samples to a two-dimensional Gaussian mixture model with a priori estimation. PICNIC genotypes SNPs simply based on the three clusters generated in the preprocessing step, and these clusters have already been adjusted by the aneuploidy information it estimates.

CNAG utilized the WGSA (whole-genome sampling analysis) algorithm [213] for genotyping in order to remove non-informative SNPs from further analysis. WGSA derives clusters of genotypes from a fix set of 108 non-malignant training samples. The algorithm then partitions around medoids of clusters to get the genotype for each SNP. Only three genotypes are called: AA, AB and BB.

ASCAT, paired PSCBS, PICNIC, and TAPS employ the “naive genotyping algorithm” [214] to identify informative SNPs. This algorithm simply calculates the two local minima of the empirical density of BAF data in the normal sample and it sets the threshold based on the normal BAFs to call the genotypes.

### **2.4.3 LRR, BAF and Decrease in Heterozygosity**

All the programs that we evaluated rely on LRR and BAF as key values from which to infer tumor genomic copy number and LOH state at each SNP. LRR and BAF are calculated based on the normalized signal intensities of each SNPs. LRR can be used to estimate total copy number, while BAF quantifies the imbalance between two alleles.

SNP arrays provide both total and allele-specific signals at SNP loci and only total copy number estimates at non-polymorphic loci. LRR at locus  $j$ , denoted  $R_j$ , is defined as to increase the signal-to-noise ratio, and the normalized decrease in heterozygosity can be used by ASCAT, paired PSCBS, and TAPS.

To assess the accuracy of the five preprocessing approaches (CN5, Quantile, CRMA v2, CNAG, and PICNIC), we compared the median LRR and the median absolute deviation (MAD) of LRR across all loci (both polymorphic and non-polymorphic) on chromosome 17 as calculated from data from the lung cancer cell line NCI-H1395 (Table 12). The spectral karyotype of this cell line indicates that all of chromosome 17 is triploid [209]. We found that CRMA v2 produced the minimum MAD when compared to other preprocessing algorithms. The models used by CNAG and PICNIC can account for aneuploidy and polyploidy in the tumor in calculation of LRR, and they have done so in the case. As a consequence, they both estimated higher LRRs than the other three methods, which implicitly assume that the tumor samples are approximately diploid when calculating LRRs. Preprocessing algorithms that use paired normal samples as references (CRMA v2, CNAG, and PICNIC) obtained lower MADs than preprocessing algorithms using pooled normal references (CN5 and Quantile normalization).

#### **2.4.4 Segmentation**

The signal intensities that reflect copy numbers are noisy, and four of the eight tools that we evaluate (GAP, paired PSCBS, TAPS, ASCAT) smooth LRRs and BAFs by segmentation. Segmentation splits the genome into regions with equal copy numbers. Circular binary segmentation (CBS) [198] is the most widely used algorithm for LRR segmentation. Because both LRR and BAF may carry information regarding the location



of copy-number change-points, CBS is also sometimes used to segment BAF values (usually transformed to decrease in heterozygosity,  $\rho$ ) to identify regions with allelic imbalance and LOH. CBS assumes that the changes in total copy number or allele-specific copy number that underlie the changes in measured LRR or BAF are discrete and affect contiguous markers on the genome. In our tests of CBS we have found that it is sensitive to consecutive outliers and tends to segment noisy regions into small fragmentary pieces; we have been unable to correct this by adjusting the smoothing parameters available (data not shown).

GAP determines the breakpoints of LRR and BAF separately using CBS. TAPS and paired PSCBS apply a two-step segmentation strategy: they first detect change-points from the LRRs alone, and then further improve the change-points with the decrease in heterozygosity ( $\rho$ ) values. TAPS makes use of k-means clustering method to identify and remove segments with non-informative SNPs, while Paired PSCBS directly removes non-informative SNPs based on the naive genotype calls of paired normal samples.

ASCAT segments using the ASPCF (Allele-Specific Piecewise Constant Fitting) algorithm [186], which simultaneously fits piecewise constant functions to the LRR and  $\rho$ 's. ASPCF includes a penalty term for creation of each segment, and the sensitivity and specificity of ASPCF is significantly affected by the value of this term.

#### **2.4.5 CNA and LOH Calls in Programs that Use HMMs**

Four of the approaches we studied (CNAG, Birdsuite, PennCNV and PICNIC) do not have a separate segmentation step, but rather directly apply discrete state hidden Markov models (HMM) to LRRs and possibly BAFs to infer CNA and LOH. The HMMs aim to determine unobserved underlying states from a sequence of observed data points.

For these applications, the underlying states are taken from pre-specified, finite sets of total copy numbers and allele-specific copy numbers. The HMMs may fail to detect tumor-specific copy number variants and LOH due to the heterogeneity of tumor samples, which appear as fractional copy number changes rather than integer copy number states. Previous comparison of these algorithms with other methods have shown a loss of ability to detect copy number changes and LOH when tumors are diluted with normal cells [192], as is usually the case for solid tumors obtained by biopsy or surgery.

CNAG defines seven total-copy-number emission states (0-6 copies) and two LOH states (“present” or “absent”). PennCNV’s HMM has six emission states: loss of one copy, loss of two copies, normal state (diploid), normal state with LOH, single copy duplication or double copy duplication. However, PennCNV does not provide information in its output about the copy-neutral LOH state. Birdsuite considers only total copy numbers, and emits one state out of the five possible pre-defined copy number levels (0-5 copies); thus, Birdsuite cannot identify regions of LOH. PICNIC, on the other hand, considers allele-specific copy numbers, which allows it to identify regions of likely LOH.

#### **2.4.6 CNA and LOH Calls in Programs that Use Segmentation**

Four of the programs that we assessed (Paired PSCBS, GAP, TAPS, and ASCAT) have separate steps that segment LRR and BAF prior to inferring total and allele-specific copy numbers. Paired PSCBS calls CN gains and losses based on a sample-specific arbitrary threshold, which is the median estimated LRR plus or minus 0.25 standard deviations (of the segment, respectively). Segments with LRRs greater than the upper threshold are called CN gains and those LRRs below the lower threshold are CN losses.

GAP and TAPS employ clustering and pattern recognition to decide CN and LOH states. They create two-dimensional clusters based on LRR and a measure that captures BAF information, and then match these clusters to specific allele-specific copy numbers. Both methods address the problems of tumor aneuploidy and admixture of DNA from non-malignant cells. To capture BAF information, GAP uses a measure called allelic difference, which is the ratio of intensity differences between the each of the two alleles in the tumor and the reference. TAPS uses a measure called the allelic imbalance ratio, which is calculated as  $(B/(A + B))$ , where A and B are the normalized signal intensities of allele A and allele B. These two methods both utilize two-dimensional scatter plot with LRR by either allelic difference (GAP) or allelic imbalance ratio (TAPS) (Figure 21), and both scatter plots capture similar information. An advantage of these clustering-based approaches is that that they can detect heterogeneity of chromosomal aberrations within a tumor, that is, the situation in which there are different populations of cells with different genomics aberrations. However, the accuracy of these algorithms is significantly affected by the number clusters present in the sample. If there are only a few clusters, these methods tend to associate clusters with incorrect allele-specific copy numbers. Therefore, GAP and TAPS accept manual input regarding tumor ploidy, which can ameliorate this problem.

In estimating allele-specific copy numbers, ASCAT also estimates tumor ploidy and the proportion of non-malignant cells in the source of the DNA sample as parameters. ASCAT searches across possible values for average tumor ploidy and for possible proportions of non-tumor cells to find values that minimize the total distance between estimated allele-specific copy numbers and nearest nonnegative whole numbers.

Therefore, ASCAT is more robust in the face of tumor aneuploidy and low tumor content than other methods. However, because the ASCAT algorithm tends to rely heavily on BAF information, tumors with BAF values that are uniformly close to 0.5 often cannot be analyzed by ASCAT. ASCAT also generates a goodness-of-fit score for each identified chromosomal copy-number aberration. This score indicates ASCAT's confidence in the estimated allele-specific copy numbers. ASCAT often cannot estimate allele-specific copy numbers from data with very noisy LRR or from samples with extremely low tumor content.

#### **2.4.7. Evaluation and Comparison of Eight Programs in Data from a Dilution Series**

We investigated how well the eight programs delineated CNA and LOH across the genome as the proportion of non-malignant DNA increases (Figure 22). Birdsuite, PennCNV and GAP detected the fewest CN gains and GAP found the fewest regions of LOH even in the pure tumor sample. The observation might be explained partially by the unpaired normalization that Birdsuite, PennCNV, and GAP performed during the preprocessing step. The performance of paired PSCBS in detecting CN gains and losses indicates the weakness and limitations of this algorithm in tackling aneuploid tumor samples. CNAG, PICNIC, ASCAT and TAPS show similar performances in detecting CN alterations in the pure tumor. However, the proportion of copy number gains detected by PICNIC drops drastically as the proportion of non-malignant DNA increases. This result stems from the ability of PICNIC to consider mixture of non-tumor DNA in the sample, because the three methods that can account for this show much less loss of ability to identify CNAs as the proportion of tumor in the sample decreases. CNAG and TAPS

both successfully detected around 25% LOH in the pure tumor, and then the proportion of LOH detected by these two samples increased in 70% tumor sample (Figure 22C). This may be due to misinterpretation of allelic imbalance as LOH in some regions. The power of CNAG to infer LOH regions is kept steady while the TAPS detects almost no LOH when the tumor content at or below 50%: TAPS's performance degrades markedly in samples with <50% tumor content. ASCAT shows almost no loss in its ability to detect LOH in samples with very low proportions of tumor cells (Figure 22C); it thus may be the most suitable algorithm for detecting LOH in samples with low tumor content. Figure 23 shows details of LOH analyses on chromosome 1 by various methods. As BAF gradually shrinks to 0.5 with the decrease of tumor content, fewer LOH regions were called. However, CNAG, PICNIC and ASCAT still preserve some ability to detect LOH even when there is only 30% tumor content.

In addition, we tested the sensitivity and specificity of those methods by analyzing CNA and LOH using an approach first reported in [185] (Figure 24). This approach takes regions of CNA as real when they are called by five of the eight programs in the samples with 100% tumor content. Analogously, it takes regions of LOH as real if they are called by four out of the six programs that generate LOH estimates. The pattern of sensitivities of different methods is similar to the pattern of proportions of alterations found by these methods. With one exception, all the programs show high specificities in the analyses. The exception is paired PSCBS in analyzing 70% tumor, which over-calls LOH, possibly by mistaking allelic imbalance for LOH.

## 2.5 Conclusions

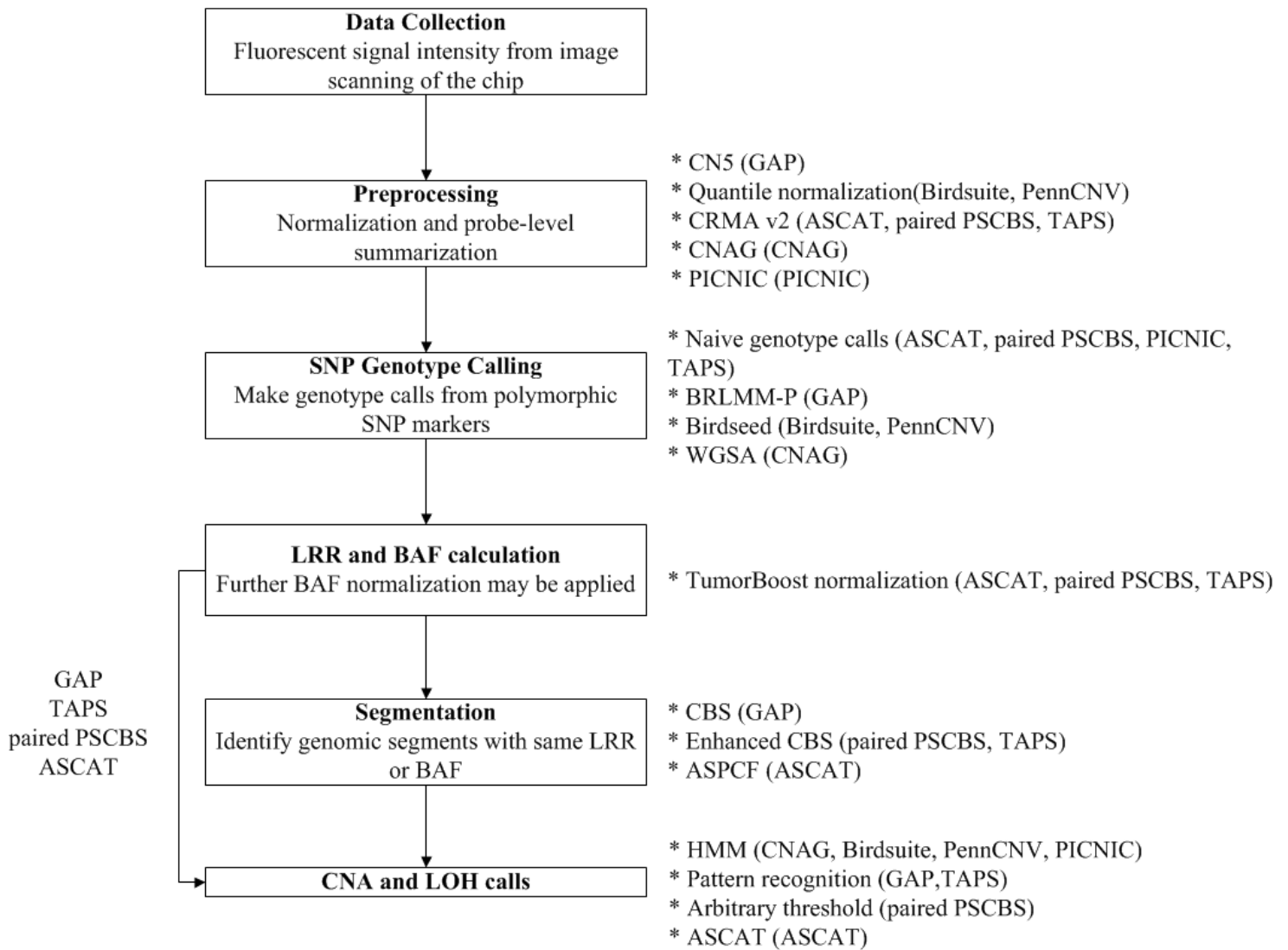
We compared eight commonly-used tools for analyzing CNA and LOH in tumors assayed by Affymetrix SNP 6.0 arrays. We evaluated their performances in a triploid tumor in a dilution series representing 100%, 70%, 50%, and 30% tumor cells mixed with non-malignant cells. A spectral karyotype of the tumor was available to provide information on copy number independently of the Affymetrix data.

We found that ASCAT performed the best with respect to sensitivity and specificity for detecting CNA and LOH. ASCAT is especially stable in calling LOH from samples with tumor content as low as 30% (Figure 24C). In addition, we found that CNAG, although it does not estimate and adjust the tumor content, nevertheless performs well in estimating CNA and LOH in samples with lower tumor proportion. We conclude that ASCAT and CNAG are the best two choices among the eight tools to obtain accurate estimates of CNA and LOH from paired tumor-normal samples assayed by Affymetrix SNP 6.0 arrays. ASCAT is written in R and its parameters can be changed by programmers, but not by end users. Therefore, it is more suitable for labs with computational expertise. CNAG has a more user-friendly interface and is more convenient to use.

## 2.6 Figures

**FIGURE 7. Flowchart of SNP array data processing procedure for CNA and LOH detection.**

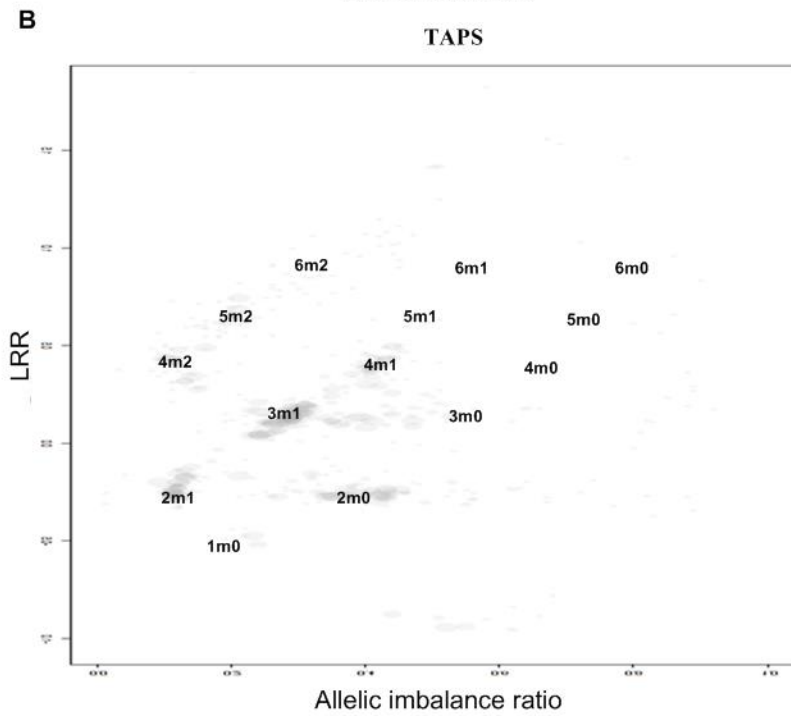
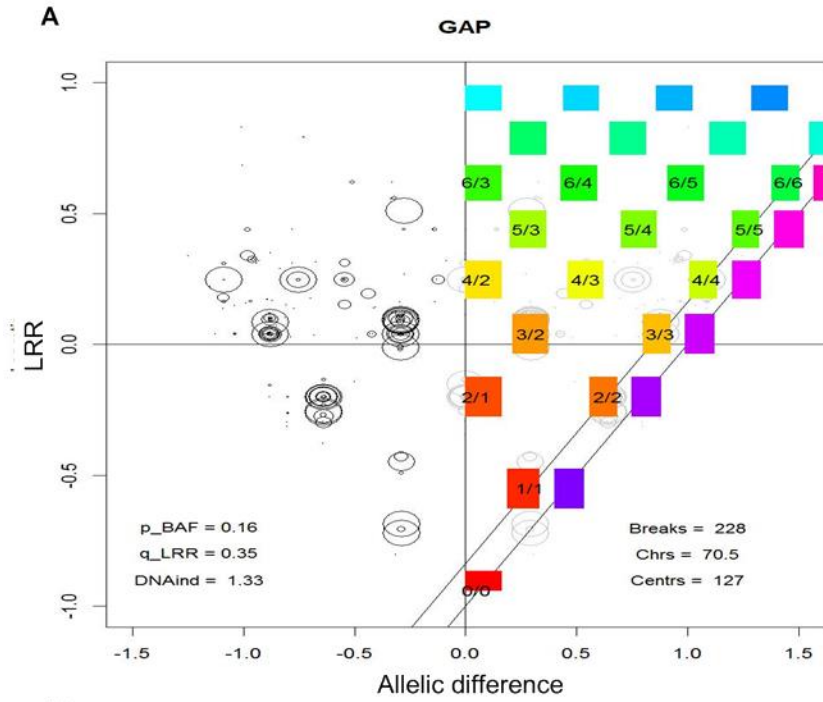
\* indicates algorithms applied in various tools for each step.



**FIGURE 8. The clustering and pattern recognition algorithms used by GAP and TAPS.**

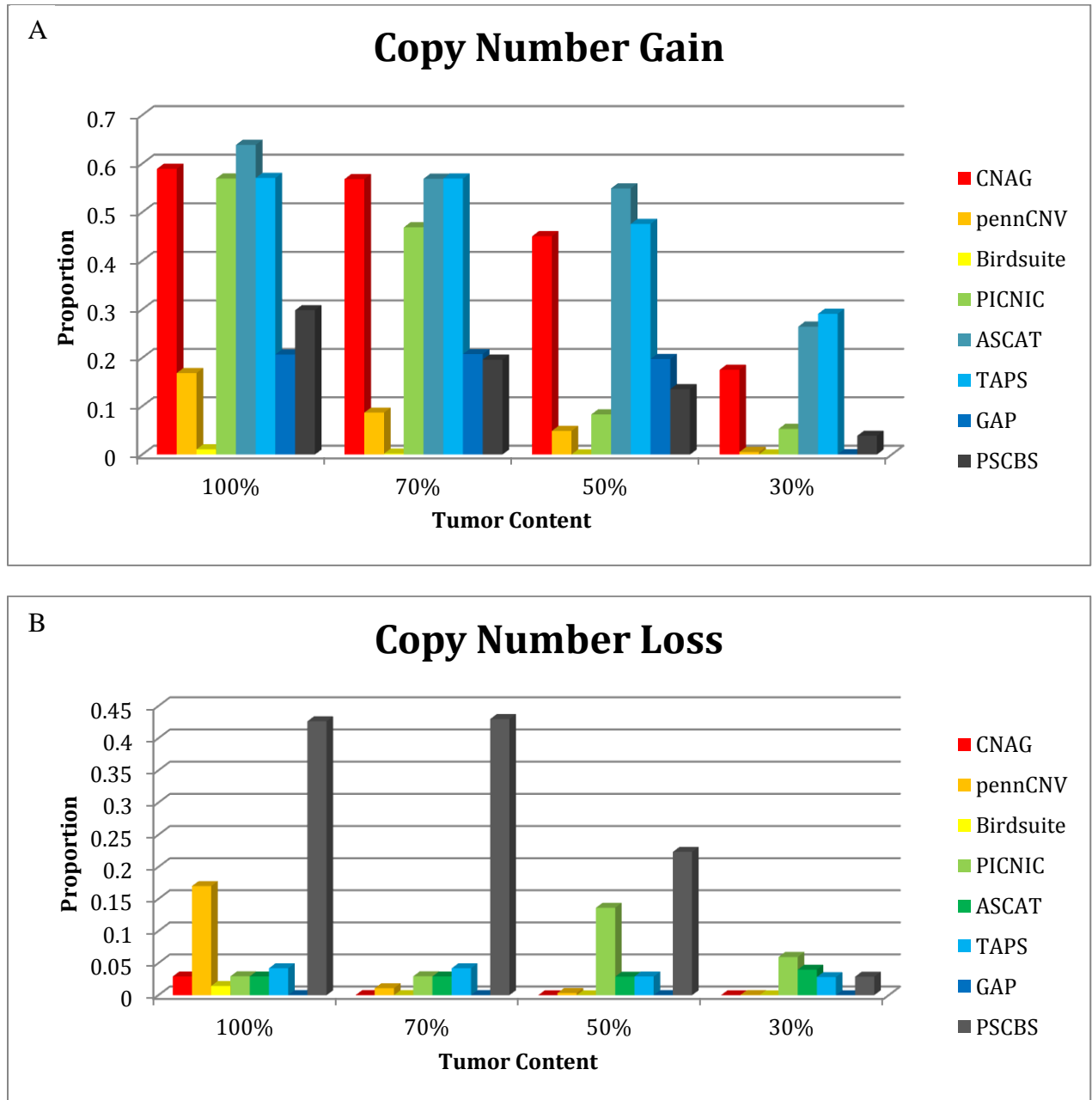
The graphs were generated from the SNP array data for the lung tumor cell line NCI-H1395 (100% tumor content). (A) GAP uses a scatter plot of LRR by allelic difference, with the best-fitting model that allows interpretation of the CN and LOH states for all clusters. The numbers in the colored boxes (e.g. 2/1) indicate that the total copy number for segments in the cluster is 2 and that the B alleles have 1 copy. (B) TAPS uses a scatter plot of LRR by allelic imbalance ratio. The graph shows allele-specific copy numbers determined for each cluster. For example, “3m1” indicates that a total copy number of 3 and a minor allele copy number of 1.

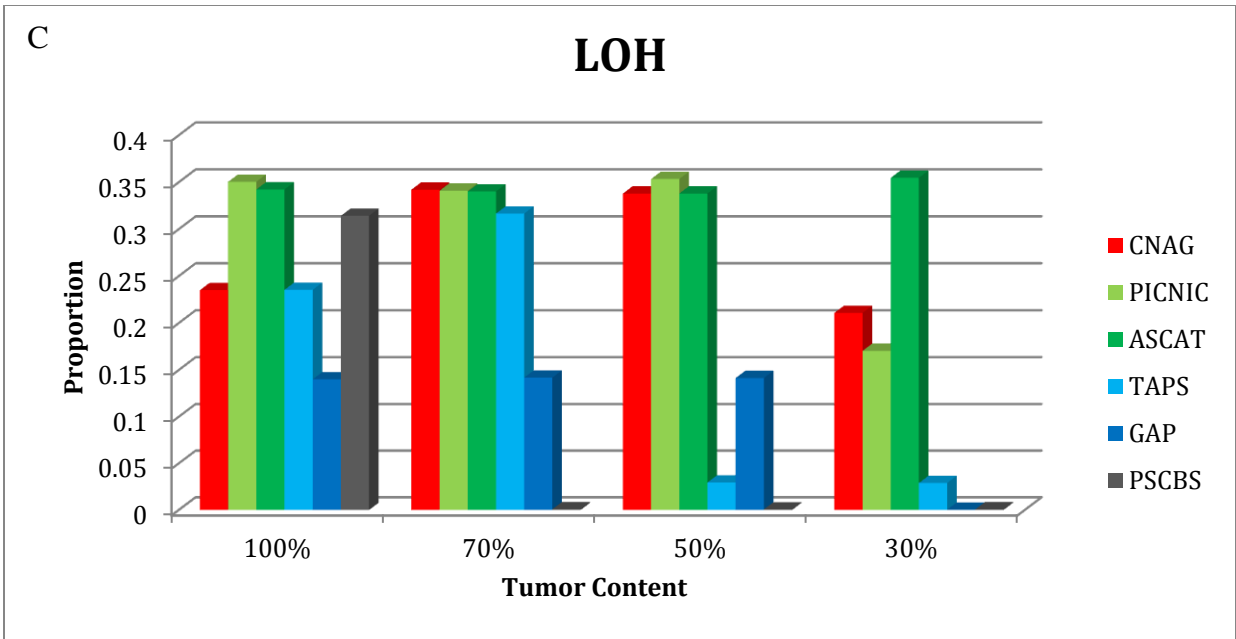




**FIGURE 9. Comparison of CNA and LOH detection across the genome for different methods at different proportions of tumor and non-malignant cells.**

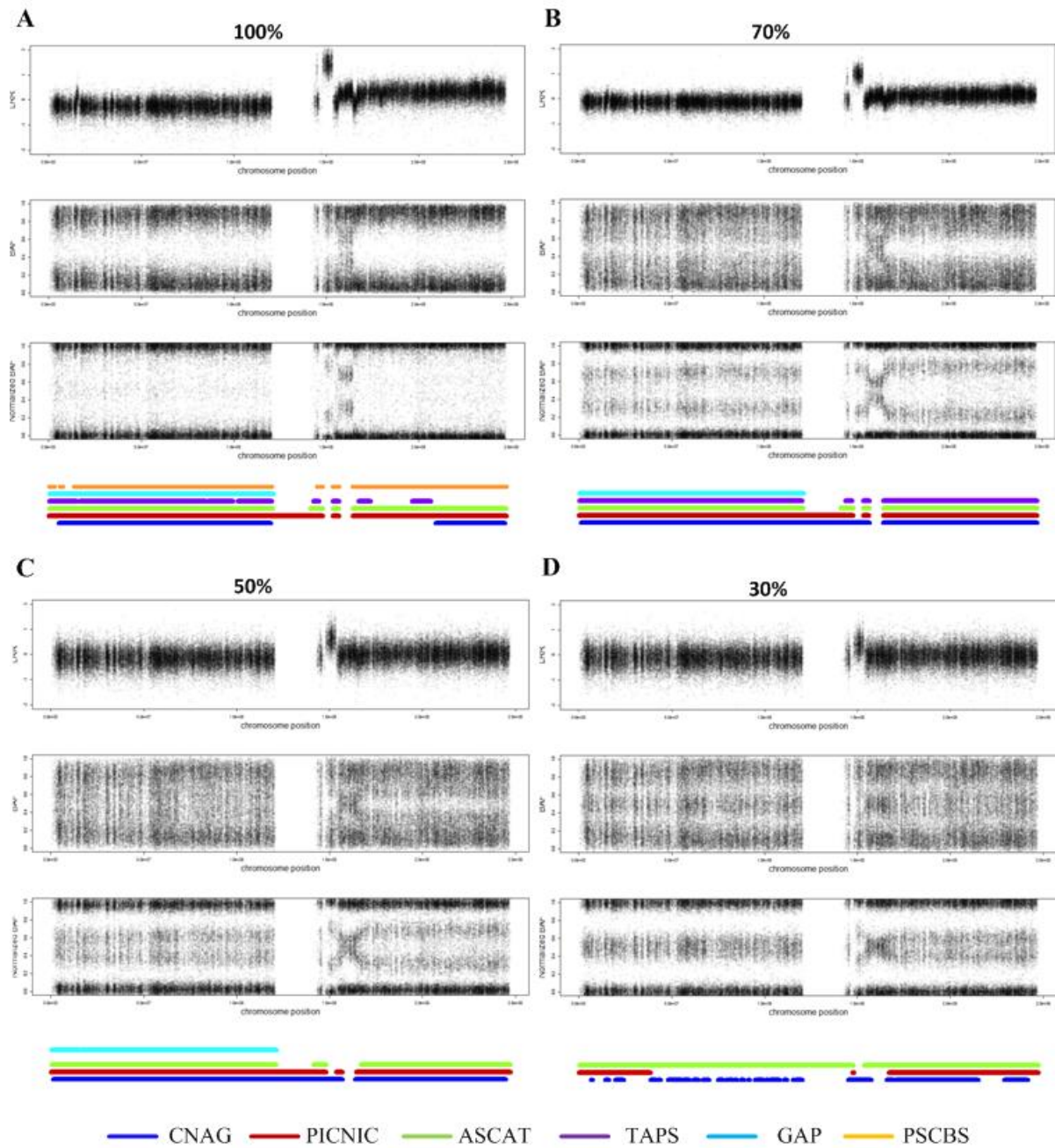
Proportions of CN gains, losses and LOH are based on the ratio of altered SNP number to the total SNP number. Copy number gain refers to regions with > 2 copies and copy number loss refers to regions with 1 or 0 copies.





**FIGURE 10. LOH on chromosome 1 at varying proportions of tumor DNA as inferred by six programs.**

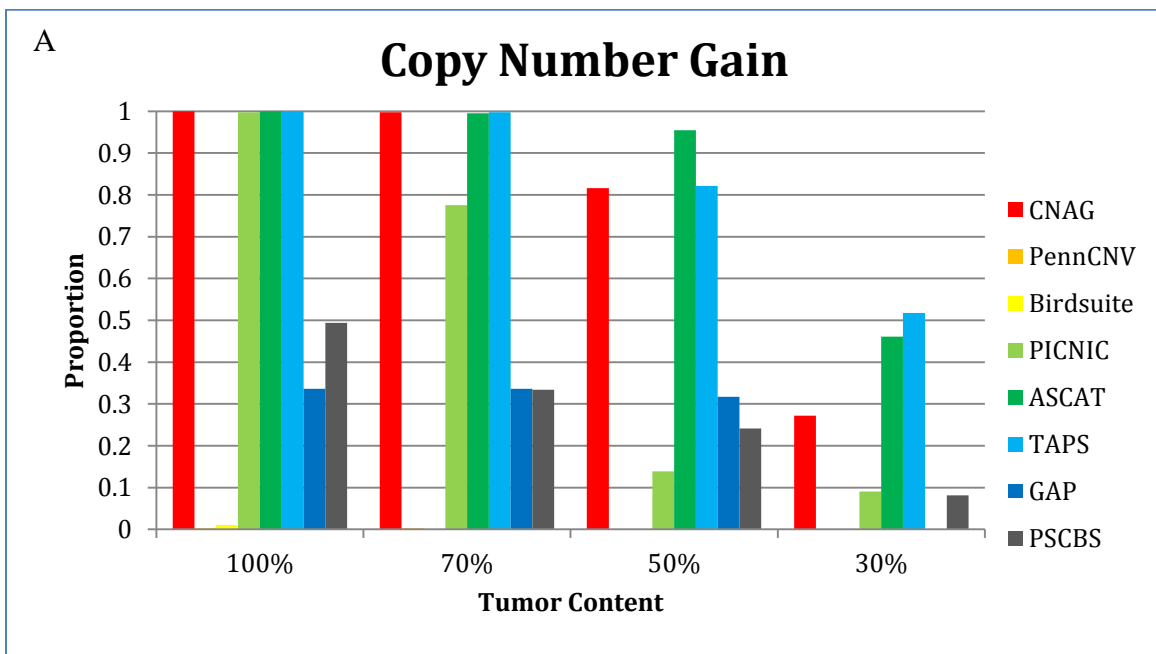
100%, 70%, 50%, and 30% indicate the proportion of tumor cells in the dilution series. Each graph contains four rows. The first row is the LRR normalized data along chromosome 1. The second row is the raw BAF data, and the third row is the BAF data after normalization by TumorBoost. The colored lines in the fourth rows represent the

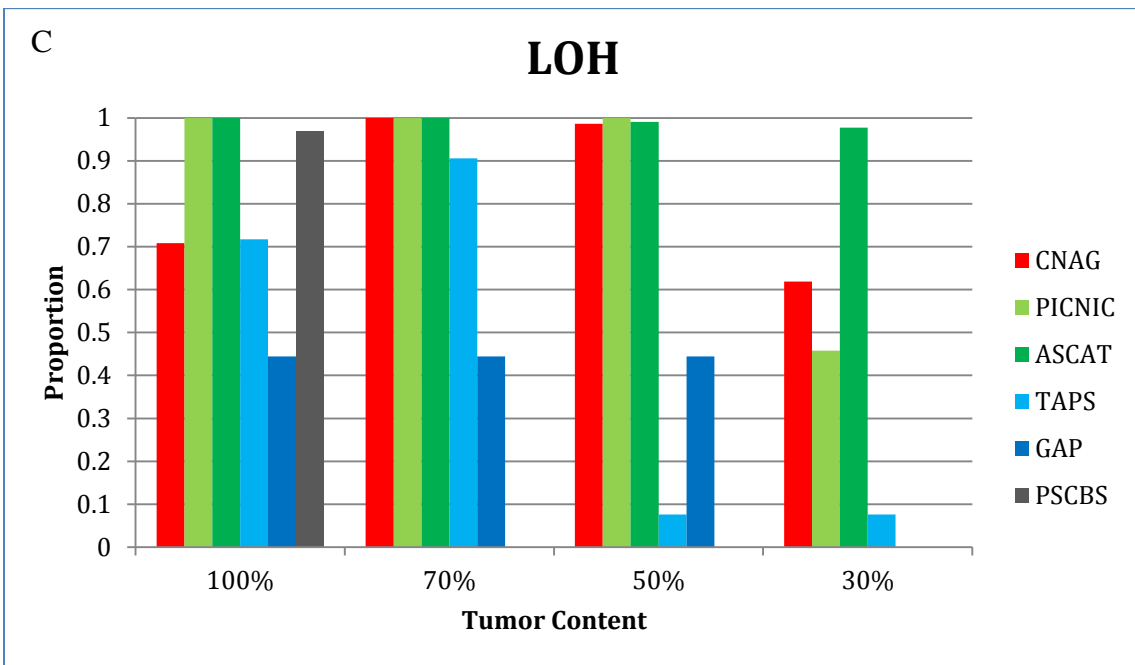
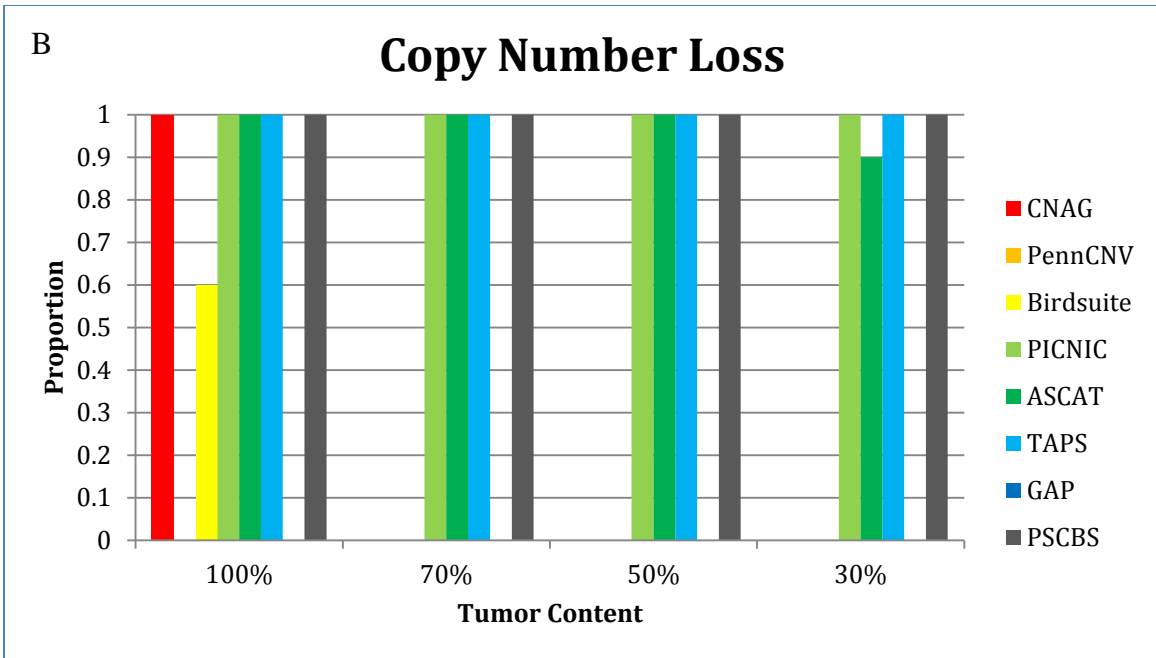


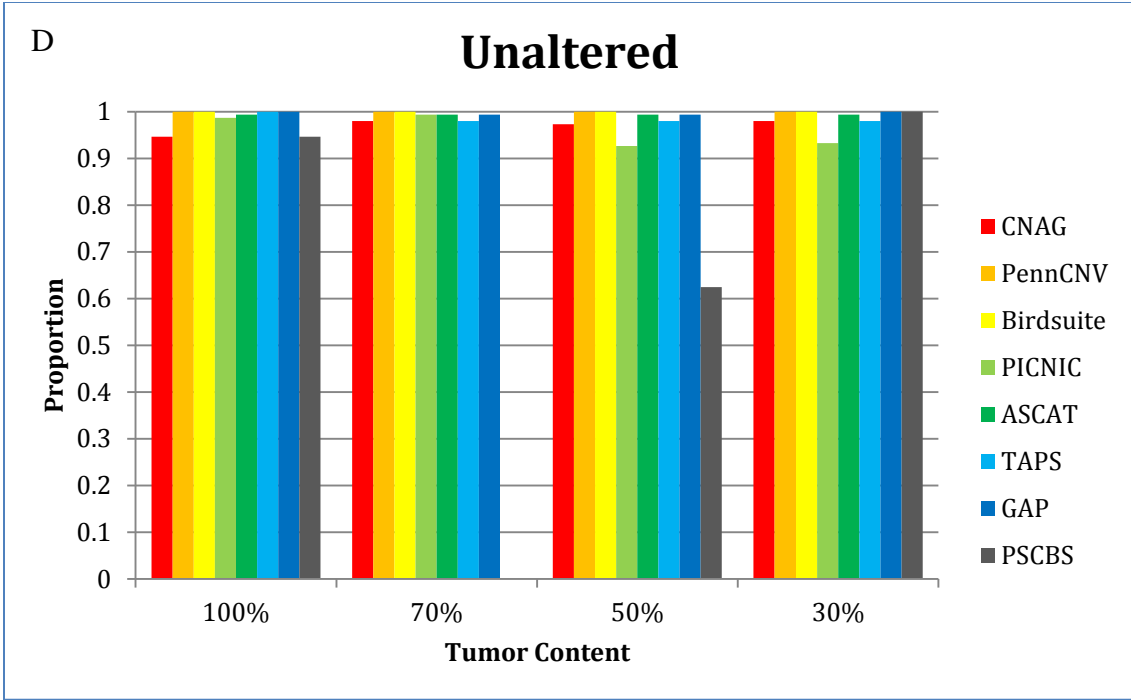
regions of LOH detected by different methods on chromosome 1.

**FIGURE 11. Comparison of sensitivities (A, B, C) and specificities (D) for different methods.**

For panels A, B, and C, the Y axes indicate the proportion of assumed real events detected (calculated by numbers of SNPs). For copy number alteration, sensitivity was calculated as the ratio of the number of altered SNPs detected by the given method to the number of SNPs detected by five out of eight of the programs. For panel D, the Y axis indicates the proportion of assumed real 2N regions without LOH that were determined as such by each method. Common unaltered regions were defined as the heterozygous copy number 2 called by at least five out of the eight methods and the specificity was calculated as the percentage of unaltered SNP number detected by the specific method to the common unaltered SNP number.







**CHAPTER 3 GLOBAL ANALYSIS OF LOSS OF HETEROZYGOSITY AND  
GENOMIC COPY NUMBER ALTERATIONS IN GASTRIC CANCER  
IMPLICATES KNOWN AND NOVEL CANCER GENES**

Yingting Wu<sup>1,2</sup>, Ioana Cutcutache<sup>2</sup>, Zhengdeng Lei<sup>2</sup>, Niantao Deng<sup>2,3</sup>, John Richard  
McPherson<sup>2</sup>, Roy Welsch<sup>4</sup>, Patrick Tan<sup>2</sup>, Steven G. Rozen<sup>1,2</sup>

<sup>1</sup> Computation and Systems Biology, Singapore-MIT Alliance

<sup>2</sup> Cancer and Stem Cell Biology program, Duke-NUS Graduate Medical School,  
Singapore

<sup>3</sup>NUS Graduate School for Integrative Science and Engineering, National University of  
Singapore, Singapore

<sup>4</sup>Engineering System Division, Massachusetts Institute of Technology, MA, USA

Correspondence: gmstanp@duke-nus.edu.sg, steve.rozen@duke-nus.edu.sg



### 3.1 Abstract

Gastric cancer is the second leading cause of cancer death worldwide but has been little studied compared with other cancers that impose similar burdens on public health. Gastric cancers often undergo loss of heterozygosity (LOH), sometimes resulting in the inactivation of tumor suppressor genes. Therefore, regions that are frequently and independently subject to LOH are likely to harbor tumor suppressors. However, patterns of LOH across gastric tumors have yet to be comprehensively studied by the most sensitive method currently available: high-density single-nucleotide polymorphism arrays. Here we report the results of genome-wide assessments of LOH and copy number alterations in 77 gastric adenocarcinomas. We used an array that assays genotype and allelic copy number at 906,600 single nucleotide polymorphisms. LOH is prevalent; on average 27% of each tumor genome was subject to LOH. The analysis of LOH implicates well-known tumor suppressors, including *TP53* (61% of the tumors), *CDKN2A* (58%) and *APC* (42%). This analysis also implicates a candidate tumor suppressor, *DOCK8*, which, to our knowledge, has not been previously linked to gastric carcinogenesis. We also found that *TP53* mutations occur almost exclusively in samples with LOH at that gene, confirming the important role of LOH in its inactivation. In addition, our analysis was able to detect somatic homozygous deletions in a mixture of tumor and non-malignant DNA, which allowed us to survey these deletions as well. The systematic and broad characterization of LOH and homozygous deletions presented here can serve as resource to aid in the future identification of new driver genes in gastric adenocarcinomas.

### 3.2 Introduction

Gastric cancer is the fourth most common cancer in the world and the second most common cause of cancer death [215]. In 2008, it caused 738,000 deaths (10% of all cancer-related deaths) [1]. Gastric cancer is especially prevalent in East Asia, Eastern Europe, and parts of Central and South America [1]. Current treatments have only slight survival benefits. Except in Japan, where endoscopic screening is common, the overall five year survival rate is only 20-25% [216].

Although there have been many studies of LOH in gastric cancer, due to the limitations of previously available technologies, most have focused on small chromosomal regions or have assayed only a few markers on each chromosome arm. In particular, there were many studies based on restriction-fragment length polymorphism and microsatellite markers in specific genomic regions, which found that several tumor suppressor genes (TSGs), including *TP53*, *APC*, and *DCC* are often affected by LOH in gastric cancer [134, 155, 157, 162, 217-219]. In addition, surveys that sampled a few loci on each chromosome arms have identified LOH affecting every chromosome arm [135, 154]. Nevertheless, genome-wide patterns of LOH across multiple gastric tumors have yet to be systematically studied by the most thorough and sensitive method currently available: high-density single-nucleotide-polymorphism (SNP) arrays.

In addition to LOH, gastric cancers also commonly possess genomic copy-number alterations (CNAs), i.e. genomic amplifications and deletions. The most recent studies used comparative genomic hybridization (CGH), array CGH, or SNP arrays and have provided an extensive and detailed view of CNA in gastric cancer [23, 80, 81, 217, 220]. Some CNAs have clinical implications. Amplification and over-expression of *MET*

and *ERBB2* are associated with poor survival [84, 86, 87], as is *KRAS* amplification [23]. In addition, there is evidence that patients with amplifications of *FGFR2* benefit from dovitinib treatment [23].

Here we delineate a high-resolution, comprehensive view of genomic regions subject to LOH and CNA, including homozygous deletion, based on microarray assays of ~906,600 SNPs in gastric adenocarcinomas and matched non-malignant tissue.

### **3.3 Materials and Methods**

#### **3.3.1 Patients and Samples**

Primary gastric adenocarcinomas and matched non-malignant tissue samples were obtained from Singapore Health Services with approval from the institutional review board. All samples were obtained with signed informed consent. Table 3 summarizes histopathological and patient characteristics.

#### **3.3.2 DNA Extraction and Hybridization**

Genomic DNAs from snap-frozen gastric tumors and matched non-malignant gastric tissues was extracted using a Qiagen genomic DNA extraction kit. The DNA was then hybridized to Affymetrix Human Mapping SNP 6.0 arrays (Affymetrix, Santa Clara, CA) according to the manufacturer's protocol. The chips were scanned with a GeneChip scanner using the Affymetrix GeneChip Operating Software. SNP positions were represented according to the hg18 (build 36) version of the human genome sequence. Some of the array data were previously published in [23]. However, these data were not previously analyzed for LOH, allelic imbalance, or homozygous deletions.

#### **3.3.3 SNP Array Data Pre-processing**

We used CRMA-v2 (Copy-number estimation using Robust Multichip Analysis version 2) [173] to extract intensity values for both alleles of each SNP from the SNP array data in the .CEL files. In this process, CRMA attempts to account for (1) crosstalk between alleles, (2) probe sequence effects, and (3) the effects of the varying sizes of fragments generated by restriction enzyme digestion prior to hybridization. We then processed each tumor-and-non-malignant pair with TumorBoost [214] to increase the signal-to-noise ratio of the allele-specific signals. This made it substantially easier to

detect allelic imbalance. Matched non-malignant samples were used as the reference for generating log<sub>2</sub> relative ratios (LRRs) and B-allele frequencies (BAFs) for the SNPs. The LRR of a SNP is the log<sub>2</sub> of the signal intensity at that SNP (summed over both alleles) in the tumor sample divided by the signal intensity in the matched non-malignant sample. The BAF of a SNP is the proportion of the total signal in the tumor deriving from the non-reference allele. (The non-reference allele is designated the B allele, whence the term B-allele frequency.)

### **3.3.4 ASCAT Profiling of Allele-Specific Copy Numbers**

We used ASCAT (Allele-Specific Copy number Analysis of Tumors) [186] versions 2.0 and 2.1 to estimate allele-specific copy numbers from the LRRs and BAFs while accounting for the effects of cancer-cell polyploidy and aneuploidy and the admixture of DNA from non-malignant cells (FIGURE 12). We selected ASCAT after evaluating several other analytical software packages, including CNAG (Copy Number Analyzer for GeneChip) [179] and GAP (Genome Alteration Print) [183]. For evaluation we used published data from a dilution series of cancer cell-line DNA mixed with DNA from non-malignant tissue from the same person [185]. We evaluated the software packages based on their (1) ability to detect LOH, allelic imbalance, and CNA in samples with a low proportion of tumor cells and (2) ability to be used in semi-automated fashion from the command line. Details of the evaluation are presented in Chapter 2. We made several modifications to ASCAT to work effectively with Affymetrix Human Mapping SNP 6.0 arrays, as detailed in Table 10.

The main inputs to ASCAT are LRRs and BAFs computed from DNA samples from a tumor and matched non-malignant tissue as described above (Figure 12A,B).

ASCAT analyses these for SNPs that are heterozygous in the non-malignant sample, and therefore informative with respect to LOH and allelic imbalance. ASCAT smoothes random SNP-to-SNP variation in LRRs and BAFs by segmentation. The green dots in Figure 12A and B show the segmented LRR and BAF values, superimposed on the original, unsegmented values, which are indicated by the red dots. After segmentation, ASCAT generates genome-wide allele-specific copy number profiles (Figure 12D). The profiles estimate (1) the proportion of tumor and non-tumor cells in the tumor sample (“aberrant cell fraction” in Figure 12D), (2) allele-specific copy numbers of chromosomal segments across the genome, and (3) reliability scores for these estimates (Figure 12E). ASCAT also provides an average ploidy for each tumor sample, which is the average of the copy numbers of informative SNPs across the genome (“Ploidy” in Figure 12D).

### **3.3.5 Cell Culture**

Six gastric cancer cell lines (YCC10, SCH, NUGC3, IM95, N87 and YCC16) and one gastric mucous epithelium cell line (HFE145) were selected for studies of candidate TSGs *PTPRD* and *DOCK8*. Cell lines IM95 and HFE145 were cultured in Dulbecco's modified Eagle medium (DMEM) with 10% FBS (fetal bovine serum) and 10mg/L insulin. Cell lines YCC16 and YCC10 were cultured in Minimum Essential Medium Eagle medium (MEM) with 10% FBS. Cell lines SCH, NUGC3 and N87 were cultured in Roswell Park Memorial Institute medium (RPMI) with 10% FBS.

### **3.3.6 Preparation and Transfection of siRNA**

The siRNAs targeting *PTPRD* and *DOCK8* were purchased from Dharmacon ([www.dharmacon.com](http://www.dharmacon.com)). Cells were seeded onto 60mm culture plates at a density of  $2 \times 10^5$  cells/mL and cultured overnight to reach 80% confluency before transfection. Cells

were transfected with 5nM of siRNA using DharmaFECT reagent in opti-MEM serum-free medium at 37 °C in 5% CO<sub>2</sub> atmosphere for 24h. Then the medium was removed and replaced with the original culture medium.

### **3.3.7 Western Blot Analysis**

Cells were cultured until 80-90% confluency and proteins were extracted with lysis buffer (2% SDS). Proteins were separated on 6% SDS-PAGE, and then transferred to nitrocellulose membranes. After blotting with 5% non-fat dry milk and 0.1% Tween 20 in Tris-buffered saline, membranes were incubated separately with either rabbit polyclonal anti-PTPRD or rabbit polyclonal anti-DOCK8 antibodies (1:200 diluted in 1% non-fat dry milk). The membranes were subsequently incubated with goat anti-rabbit secondary antibodies and screened by Odyssey® Imager of Li-COR.

### **3.3.8 Cell proliferation assays**

Cell proliferation assays were performed with a CellTiter96 Aqueous Nonradioactive Cell Proliferation Assay kit (Promega) following the manufacturer's instructions. The 96-well plates were measured with a Perkin-Elmer plate reader.

### 3.4 Results

We analyzed 113 gastric tumors with their paired adjacent non-malignant tissues using ASCAT. For 77 of the 113 pairs, ASCAT was able to estimate allele-specific copy numbers across the genome. ASCAT was unable to estimate allele-specific copy numbers for the remaining pairs for the following reasons (Table 11): (1) excessively variable LRR data that ASCAT was unable to segment reasonably (12 tumors, Figure 13A); (2) BAF values that are almost uniformly 0.5, which could be due to a tumor genome completely lacking LOH or allelic imbalance or, alternatively, to a very low proportion of tumor cells contributing to the DNA sample (22 tumors, Figure 13B); or (3) apparently low tumor content as evidenced by very little variation in the segmented LRRs and few divergences of the BAFs from 0.5 (2 tumors). The samples that ASCAT was not able to analyze had lower pathologist-estimated tumor proportions compared to samples that ASCAT succeeded in analyzing ( $p = 0.034$ , one-sided Wilcoxon rank-sum test, Table 3, Figure 14A). Furthermore, the minimum tumor cell proportion that ASCAT was able to estimate was 18%, suggesting that it is not able to analyze samples with content lower than this (Figure 14B). These two observations implicate low tumor content as a possible major reason that ASCAT could not estimate allele-specific copy number for some samples.

#### **LOH Landscape in Gastric Cancer.**

LOH is prevalent in gastric cancer (Figure 15, Table S3): on average, 26.8% of each gastric cancer genome was subject to LOH (range 0.13% to 77.7%). LOH is also pervasive: 98% of the SNPs assayed were subject to LOH in at least 10% of tumors. We focused on the 10 regions that were subject to LOH in  $\geq 35\%$  of the tumors (Table 4).



Chromosome arm 17p, which contains *TP53*, is the region most frequently subject to LOH, and 61% of the tumors show LOH there. Chromosome arms 9p (58% of tumors, a region containing the well-known TSG *CDKN2A*) and 5q (42% of tumors, a region containing the well-known TSG *APC*) also frequently undergo LOH. Several other TSGs in regions subject to frequent LOH include *MAP2K4*, *NCOR1*, and *CDKN2A* (Table 4). Copy-number neutral LOH (CNNLOH) accounted for an average of 51% of LOH in each sample. However, in some LOH peaks, for example the one in 18q, more than half of the tumors with LOH have hemizygous deletions; 18q contains the likely TSG *DCC* (deleted in colorectal cancer).

### **Recurrent Somatic CNAs in Gastric Cancer**

Because ASACT estimates allele-specific copy number in order to identify regions of LOH, it also detects CNAs. For this analysis, we considered a region to be subject to CNA in a sample if the ASCAT-determined integral copy number was less than or greater than two. We observed a total of 2,037 somatic CNAs, (954 gains and 1,083 losses) across the 77 tumors; the mean number of copy number gains per sample was 12 (median=7) and the mean number of losses was 14 (median=6). With the ASCAT profile for each sample as input, GISTIC (Genomic Identification of Significant Targets in Cancer [221]) identified 40 regions (17 gains and 23 losses) that underwent significant CNA (Figure 16, Tables 5 and 6). The regions of significant CN gain and loss identified in this study are similar to those identified in a previous study using different methodologies in 193 tumor samples from the same collection [23].

Because the current analysis considers the effects of tumor aneuploidy and mixture of non-malignant genomes in the tumor samples (Figure 16), it can identify

homozygous deletions. The highest frequency of homozygous deletion involves a region on 3p that contains no well-characterized genes (6 tumors). Other recurrent homozygous deletions are in within chromosome arms 8p (4/77) and 18q (4/77) (Figure 15D, Table 7). The well-known TSGs *CDKN2A* (*p16*, *ARF*), *AXIN1*, and *SMAD4* were each homozygously deleted in 2 tumors.

### **Relation between Genomic Alterations and Tumor Characteristics**

In order to investigate the biological implications of LOH and CNA in gastric cancer, we compared LOH and CNA proportions among different clinical subgroups. We found that tumors from males tend to have higher proportions of LOH than those from females (Figure 17A,  $p=0.04$ , Wilcoxon rank sum test, not significant considering that we tested four hypotheses). There were no significant relationships between gender and any of CNNLOH proportion, CN loss proportion, or CN gain proportion. (Figure 17B-D).

We also evaluated the LOH proportions in intestinal compared to diffuse gastric tumors. We found no significant relationship between CNNLOH or CNA proportion and these two histological subtypes (Figure 18).

### **LOH and *TP53* mutations**

We also found evidence that *TP53* mutations are associated with higher levels genomic instability as detected by ASCAT. We previously reported results of screening for mutations in the *TP53* hot spots (exons 4, 5, 6, 7, 8 and 9) [78] in 56 of the 77 tumors that we analyzed. Among these 56 tumors, those with *TP53* mutations had a higher proportion of the genome affected by CNAs ( $p = 0.029$  by Wilcoxon rank-sum test, Figure 19). This result is consistent with previous studies that have linked genomic instability to mutations on *TP53* [222, 223].

In addition, we found that *TP53* mutations occurred almost exclusively in tumors with LOH at that gene: only three samples had *TP53* mutations but no LOH at that gene ( $p=4.75e-4$ , Fisher's exact test, Table 8). This is consistent with Knudson's model that LOH is a common second hit after a previous heterozygous loss of function mutation [4].

### 3.5 Discussion

#### Limitations

Genome-wide analyses of CNA and LOH are challenging due to the mixture of malignant and non-malignant genomes in tumor samples. No standard method has emerged as the most appropriate to accurately estimate allele-specific copy numbers in tumors with low proportions of malignant cells. ASCAT performed well in our evaluation of it in a dilution series in the face of a low proportion of malignant cells [224]. Nevertheless, it appears that ASCAT's ability to complete its analysis rapidly approaches nil when the proportion of malignant cells is  $< 20\%$  (Figure 14B). Indeed, it is impossible for any analytic approach to distinguish a tumor sample consisting almost entirely of non-malignant cells from a sample consisting entirely of malignant cells with normal genomes that completely lack large-scale genomic aberrations. ASCAT was not able to generate allele-specific-copy-number profiles for 22 tumor samples with flat BAFs (Figure 13B). If these included some tumors that completely lacked genomic aberrations, then we would have overestimated the proportion of gastric adenocarcinomas with LOH, allelic imbalance, and CNA. We note, however, that on the whole, earlier genome-wide surveys found similar or higher estimates of the proportions of tumors affected by LOH [135, 154] or CNA [81].

In addition, ASCAT was unable to complete its analysis of 12 tumors for which the LRRs were excessively variable (Figure 13A). In most tumors for which ASCAT was able to generate an allele-specific-copy-number profile, the standard deviations of the segmented LRRs and BAFs are tightly correlated (blue circles in Figure 20). For the samples with excessively variable BAFs, however, the standard deviations of the BAFs

are relatively lower (red squares in Figure 20). Indeed, the LRRs change frequently without corresponding changes in BAFs in these tumors. This then suggests the high variability of the LRRs represent an experimental artifact rather than a characteristic of these gastric cancer genomes.

### **Comparison to previous findings on CNA in gastric cancer**

Compared to the CN amplification pattern in 34 gastric cancer cell lines found by Tada *et al* [217], we found similar regions that are frequently amplified in gastric cancer, including 8q, 11p, and 20q. We also corroborated several other regions that were previously reported to frequently undergo CNA [81-83]. When comparing the CNA landscape in our analysis to a previous CN study using 193 tumor samples from the same collection [23], we found marked similarity in the CNA patterns, which confirmed several previous findings, such as the amplifications of *FGFR2* and *ERBB2*.

### **Comparison to previous findings on LOH in gastric cancer**

Several studies have found that LOH occurs more often in intestinal-type gastric tumors than diffuse-type tumors [225, 226]. However, this was not observed in other studies [132, 154, 156]. The present whole-genome analysis found no evidence of differences in the proportion of LOH between the two types (Figure 19A,B).

Previous univariate analyses indicated that patients with low levels of LOH had better survival than those with high levels of LOH or non-detectable LOH [101, 217]. To assess the prognostic impact of LOH in gastric cancer in our data, we conducted a Kaplan Meier survival analysis comparing patients with high versus low proportion of LOH. We found no evidence of differences between the two groups (Figure 21A). Multivariate analysis in a Cox proportional hazards model that treated the proportion of LOH as a

continuous variable likewise showed no evidence that survival was related to the proportion of LOH (Table 9). We also found no evidence of an effect of average ploidy on survival (Figure 21B).

### **Candidate tumor-suppressor genes subject to frequent LOH**

We found that the short arm of chromosome 9 (9p) is subject to frequent LOH in gastric cancer (58%). Several studies have reported similar observations, with frequencies ranging from 22% to 64%. The well-known TSG *CDKN2A* (*p16*) is located on 9p21, and is mutated in numerous tumor types [227-229]. This gene is frequently deleted or hypermethylated [230-235] in gastric cancer, and in our analysis this gene was homozygously deleted in 2 tumors (Figure 15D). However, it is possible that there are other TSGs in this region that contribute to gastric carcinogenesis. Two genes that look promising in this regard are *PTPRD* (protein tyrosine phosphatase, receptor type, D) and *DOCK8* (dedicator of cytokinesis 8).

*PTPRD* is inactivated by gene deletion or mutation in various cancers, including lung cancer, neuroblastoma, glioblastoma, melanoma, and squamous cell carcinoma [236-241] and was previously noted to undergo LOH in gastric cancer [217]. In addition, *PTPRD* was homozygously deleted in 2 out of 77 gastric tumors. *PTPRD* also interacts with MIM, a potential metastasis suppressor gene, in regulating cytoskeletal remodeling [242]. *PTPRD*'s ability to dephosphorylate STAT3 is abrogated by cancer-specific mutations in *PTPRD* [239]. The evidence above supports the hypothesis that *PTPRD* acts as a TSG.

*DOCK8*, another candidate TSG in 9p is a guanine nucleotide exchange factor (GEF) that is responsible for activation of Rho GTPases by exchanging bound GDP for

free GTP. Homozygous deletion and reduced expression of *DOCK8* were observed in lung cancer [243-245]. We also observed increased mRNA expression levels of *DOCK8* in tumors compared to non-malignant gastric epithelium (p=0.02, data are not shown). This may indicate an attempt by cells to up-regulate mutationally inactivated variants of the gene.

### **3.6 Conclusions**

This genome-wide survey of LOH, allelic imbalance, and CNA, including homozygous deletions, in 77 gastric adenocarcinomas provides a more systematic and detailed picture than previously available. Because the regions commonly affected LOH and CNA are broad and encompass many genes, information about these regions alone is not sufficient to unambiguously identify driver genes. In the future, when integrated with information on somatic point mutations and experimental studies, the results from this study may aid in the identification of new driver genes in gastric adenocarcinomas.

### **Acknowledgements**

We thank Gengbo Chen for assistance with plotting LOH proportions.

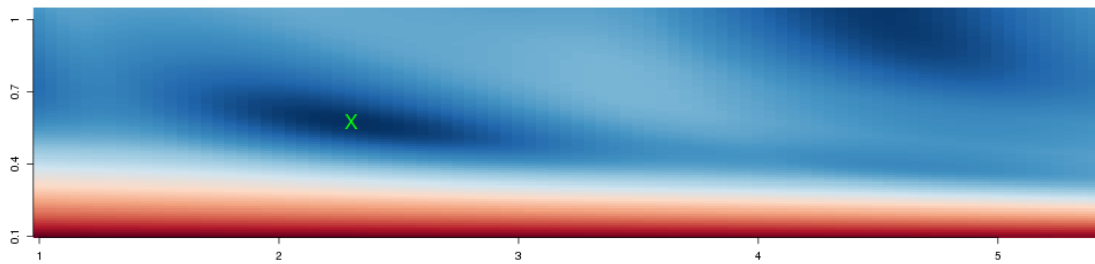
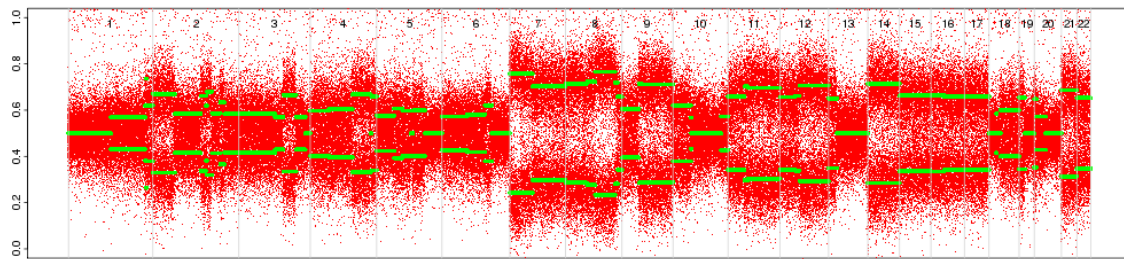
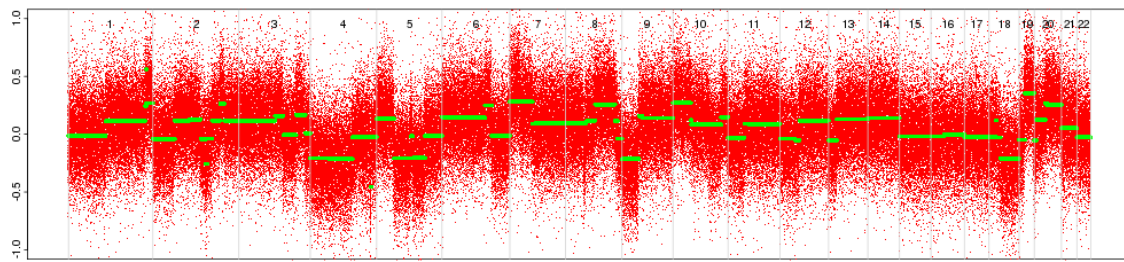
### **Author Contributions**

AYW Designed the study and carried all of the laboratory experiments and most of the bioinformatic analysis, and drafted the text and most of the figures. IC independently reviewed and checked the bioinformatic analysis and drafted additional figures. ZL, ND, and JRM assisted with bioinformatics analysis. RW reviewed and consulted on the study design and manuscript. PT supervised the study and organized generation of the data. SGR supervised the study and edited the manuscript and figures.



### 3.7 Figures

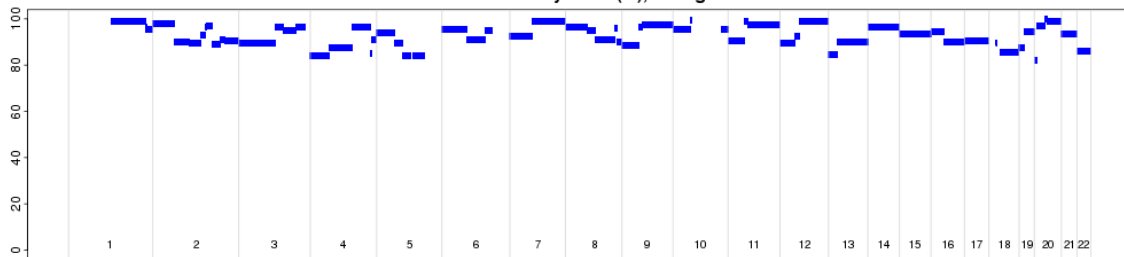
**FIGURE 12. Example ASCAT profile and allele-specific copy numbers.** (Data for tumor 97005.) (A) LRR. The x-axis shows the indices of autosomal SNPs that are heterozygous in the non-malignant sample. The y-axis shows the LRR value of SNPs in the tumor relative to the non-malignant sample. Green dots show ASCAT's segmentations. (B) BAFs for the SNPs plotted in A. Green lines show ASCAT's segmentation. (C) The solution space for the two parameters "ploidy" and "aberrant cell fraction", with the location of the chosen values marked with an "X". (D) ASCAT's model of allele-specific copy numbers. The y-axis indicates the estimated (integer) chromosomal copy number. Red and green lines indicate the more common and less common chromosomal haplotypes, respectively. The lines are vertically offset slightly to avoid superimposition. (E) The aberration reliability score, a measure of how well the calculated allele-specific copy-number model in panel D explains the observed LRRs and BAFs.



Ploidy: 2.57, aberrant cell fraction: 58%, goodness of fit: 91.7%

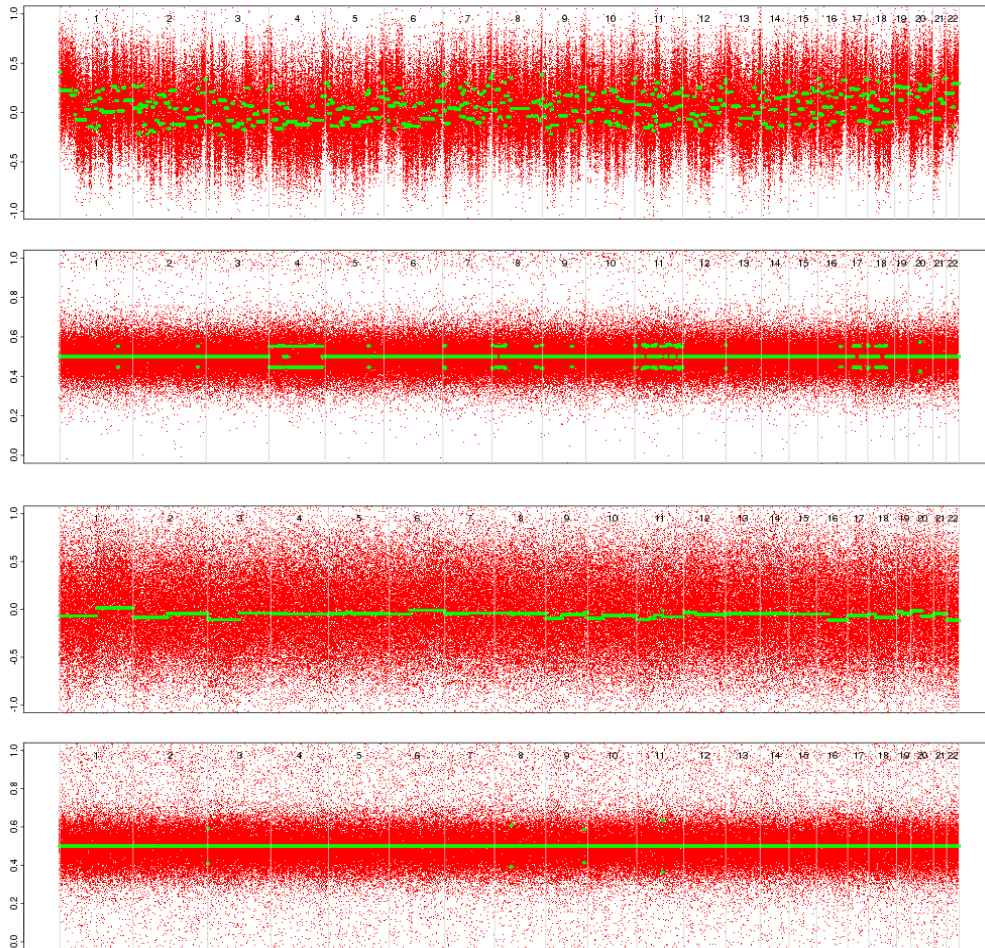


Aberration reliability score (%), average: 93%



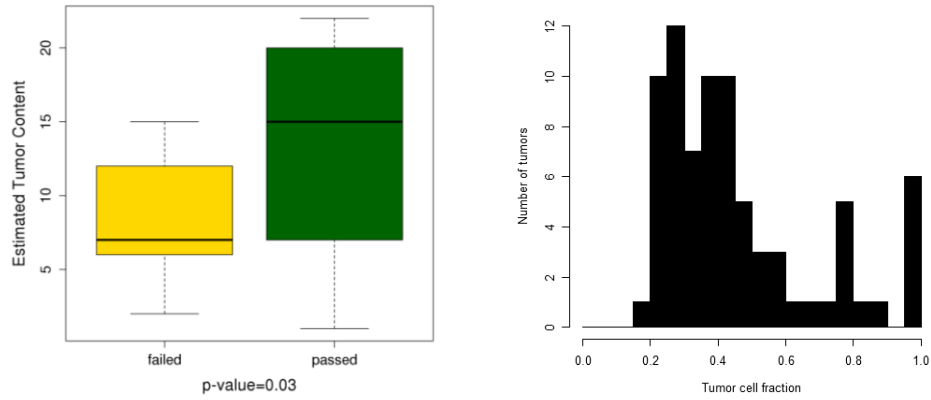
**FIGURE 13. Examples of tumor and non-malignant pairs that ASCAT was unable to analyze.**

(A) Excessively noisy LRR leading to breakdown in segmentation. (B) Flat BAF.

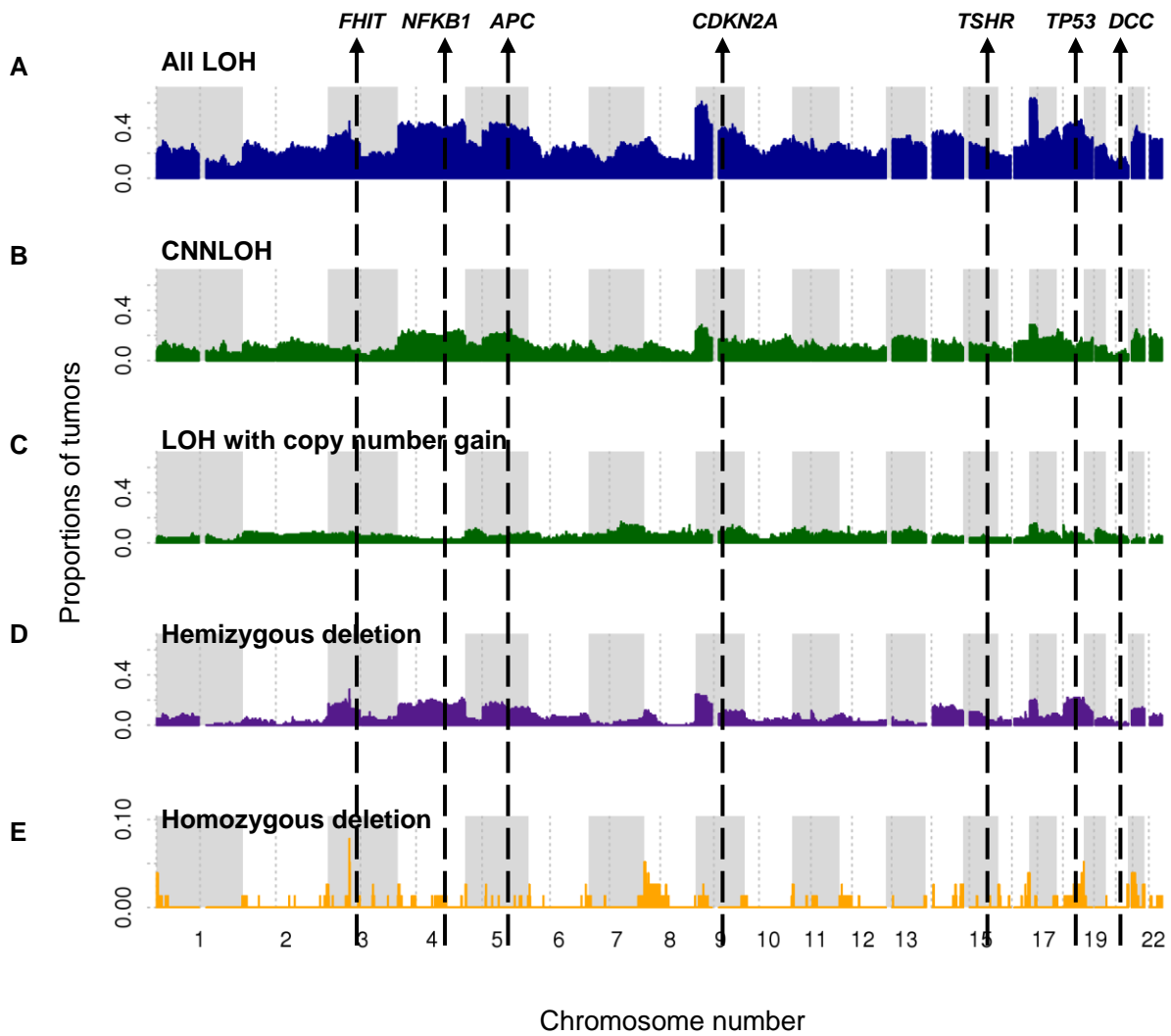


**FIGURE 14. Relationship between tumor content and ASCAT’s ability to generate an allele-specific-copy-number model.**

(A) Pathologist-estimated tumor proportions in samples for which ASCAT was able or unable to generate allele-specific-copy-number models. P value by one-sided Wilcoxon rank-sum test. (B) Distribution of ASCAT-estimated tumor-cell proportion for 77 samples.

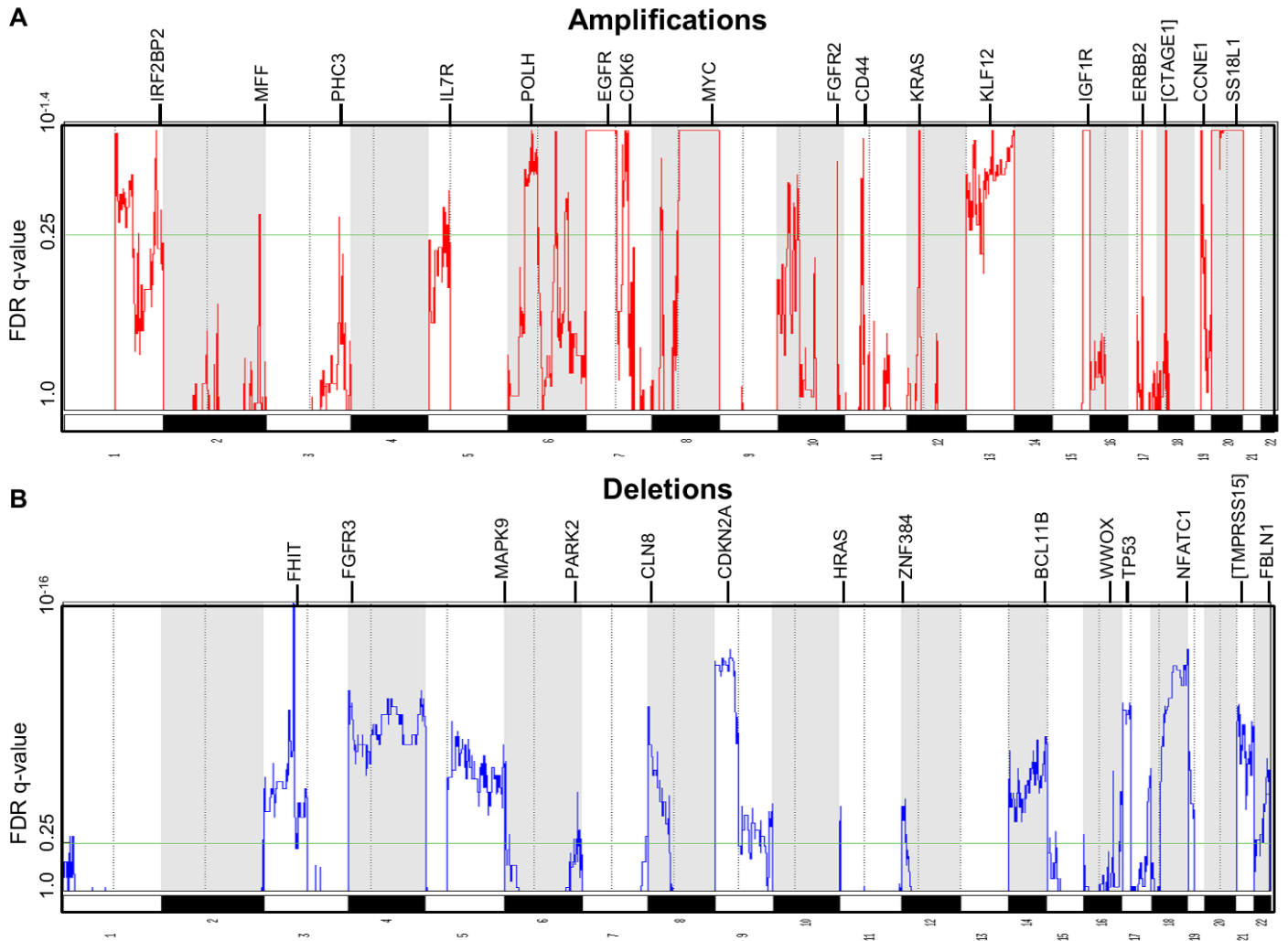


**FIGURE 15. Frequencies of LOH and CNA across 45 gastric tumors.**  
 Several known TSGs are indicated.

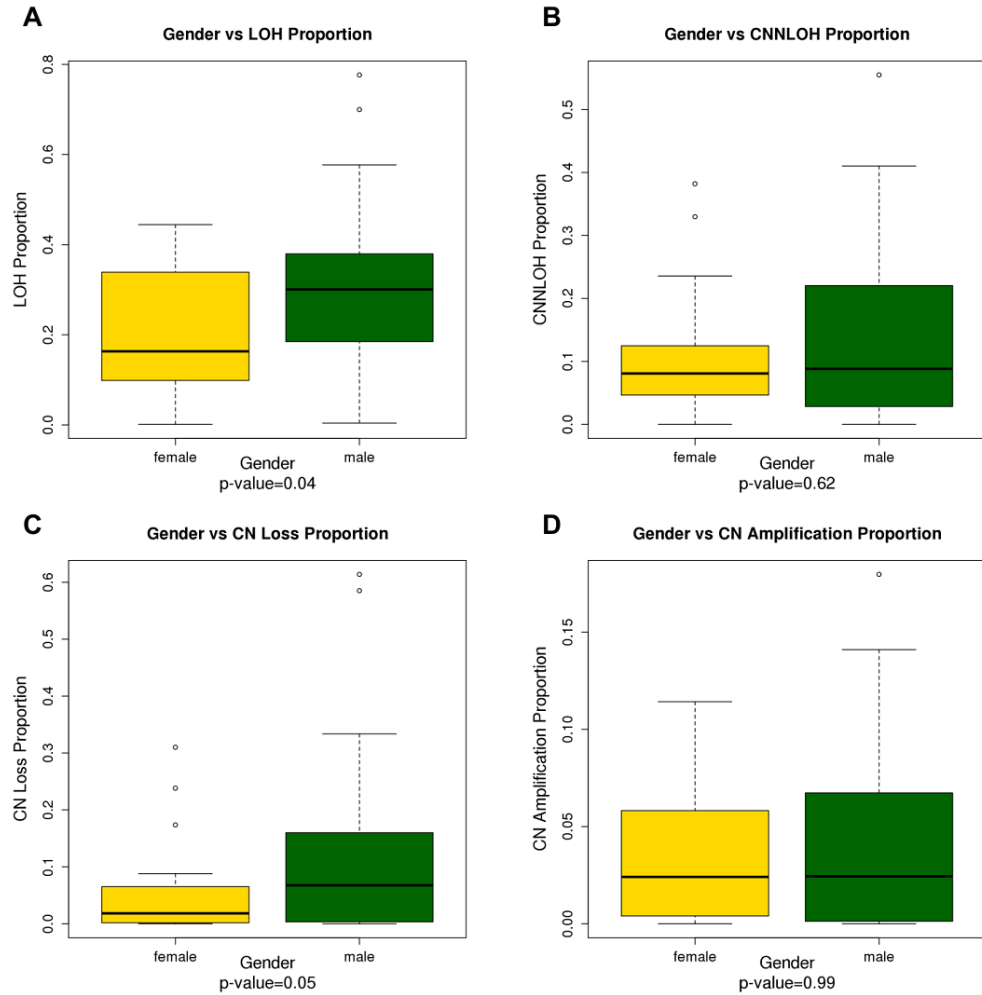


**FIGURE 16. Identification of significant somatic copy number alterations across gastric cancer by GISTIC.**

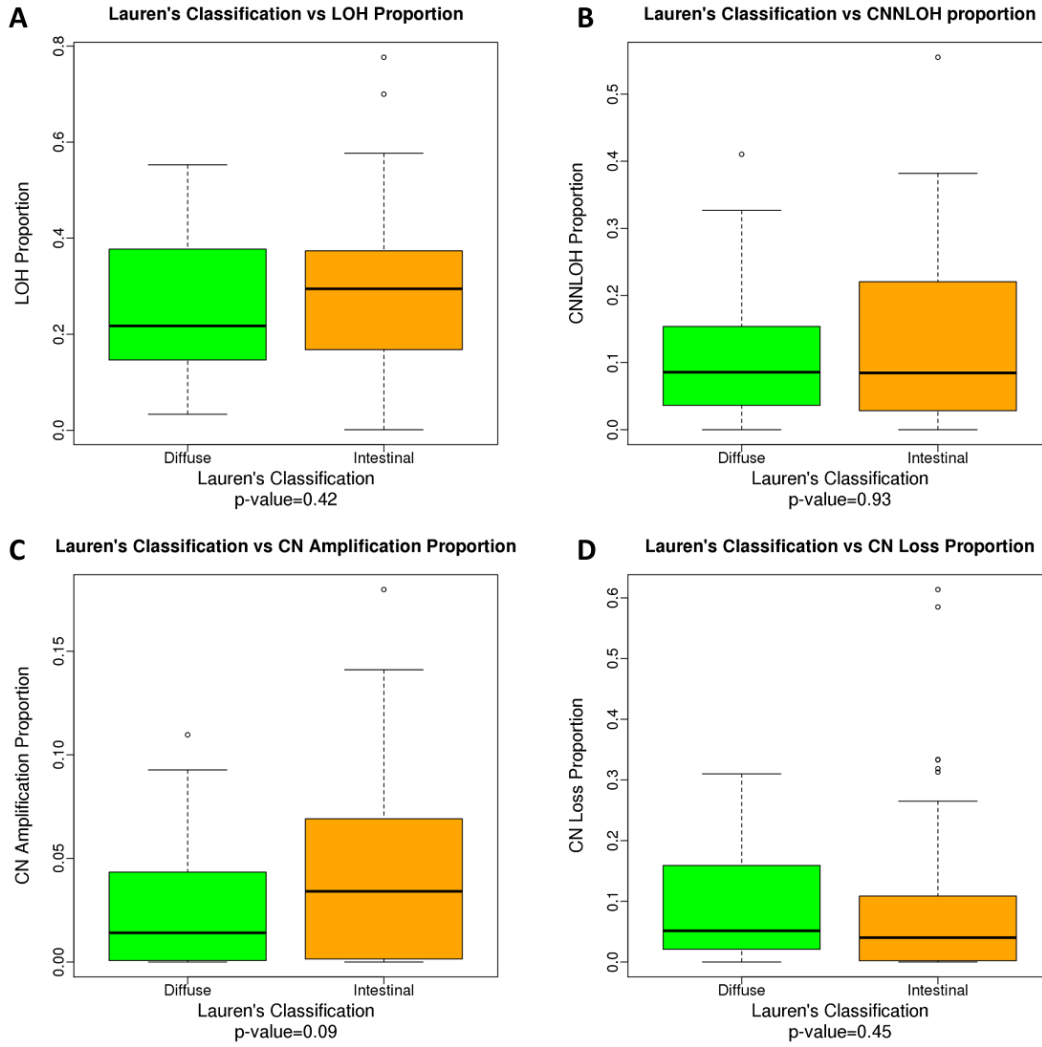
The y-axes are the positions across the autosomal genome, and the x-axes are the GISTIC q-values. Several known or putative gene cancer-related genes are indicated in peaks that passed the cut-off (q-value <0.25).



**FIGURE 17. LOH and CNA proportions in males and females.**



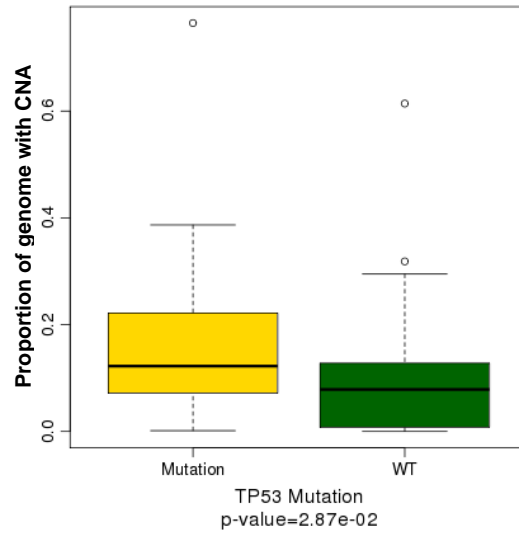
**FIGURE 18. Comparisons of proportions of LOH and CNA in gastric tumors according to the Lauren histological subtypes.**



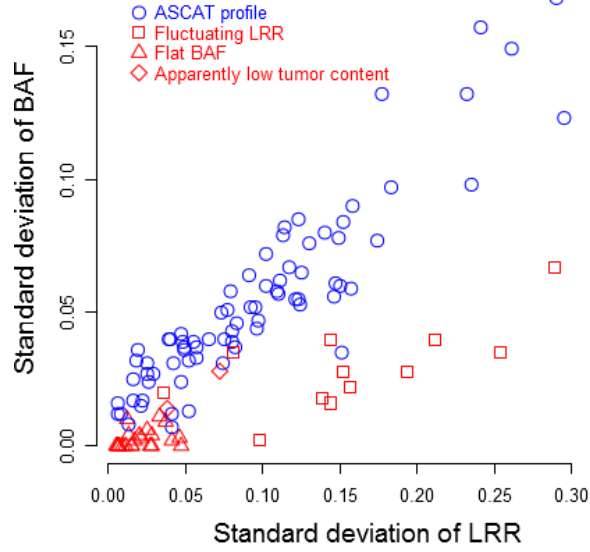


**FIGURE 19. Relationship between TP53 mutation and proportion of genome subject to CNA.**

P-value by Wilcoxon rank-sum test.

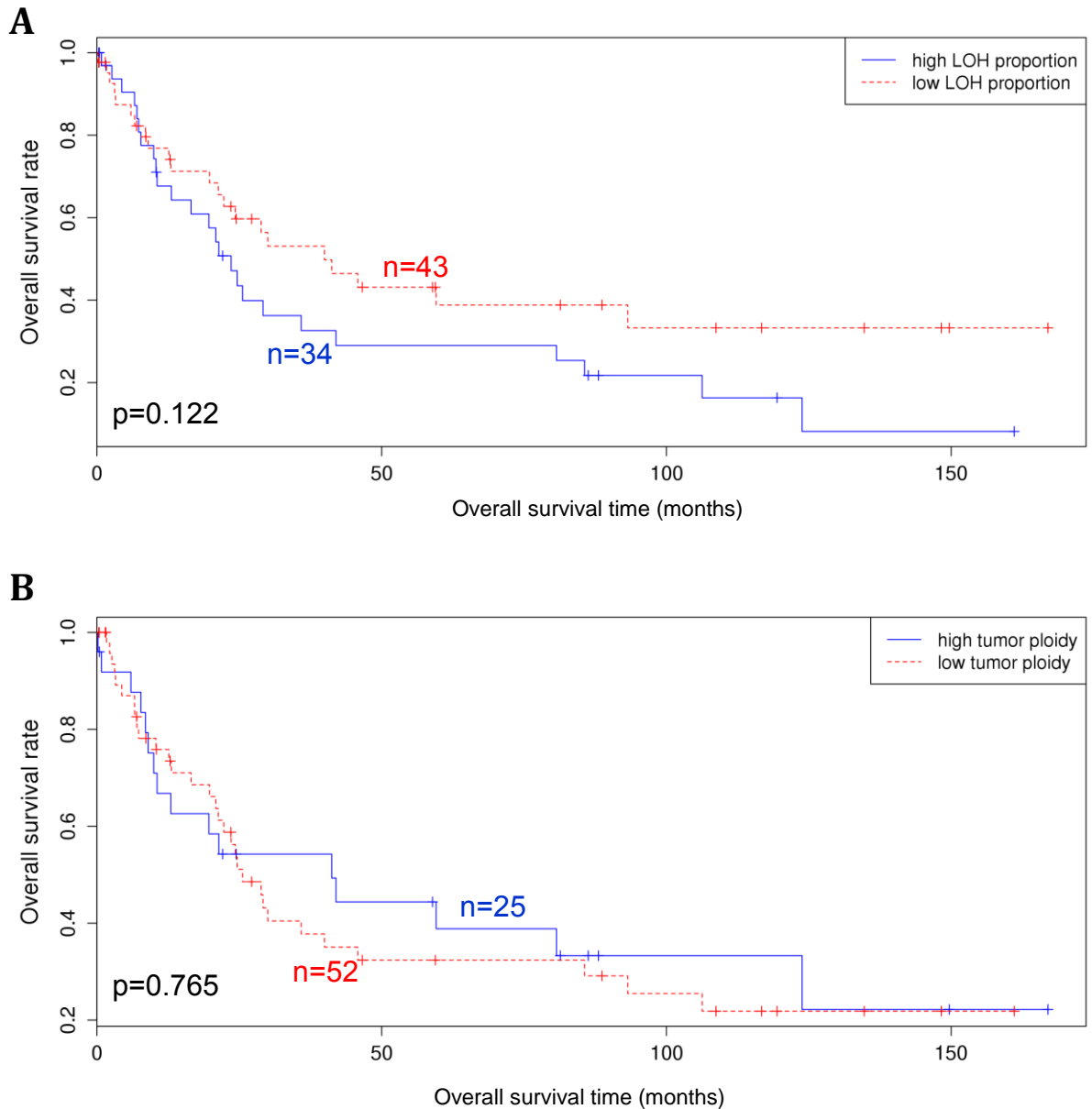


**FIGURE 20. Relationship between standard deviations of segmented BAF and segmented LRR.**



**FIGURE 21. Kaplan-Meier survival analysis comparing outcomes by proportion of LOH and average ploidy.**

(A) Patients with high proportions of LOH (24 patients, >30% of the genome) versus patients with low proportions of LOH (21 patients, <30%). (B) Patients with high tumor ploidy (21 patients, average ploidy >2.6) versus patients with low tumor ploidy (24 patients, average ploidy < 2.6). The outcome was overall survival.



### 3.8 Tables

**TABLE 3. Clinical and pathological information on tumors studied.**

Category	Subcategory	All tumors	Tumors with ASCAT estimate	Tumors without ASCAT estimate	P-value, with vs. without ASCAT estimate	Test
Age	Range	25-92yrs	25-92yrs	32-88yrs	0.39	t
	Median	67yrs	65yrs	68yrs		
Pathologist-estimated tumor content	Range	0%-90%	0%-90%	5%-60%	0.034	One-sided Wilcoxon rank-sum
	Median	35%	60%	20%		
Gender	Female	41	23	18	0.058	Fisher's exact
	Male	72	54	18		
Laur n classification	Diffuse	44	26	18	0.139	Fisher's exact (excluding mixed)
	Intestinal	62	46	16		
	Mixed	7	5	2		
TNM Stage	I	14	10	4	0.956	Chisq
	II	20	13	7		
	III	61	41	20		
	IV	18	13	5		
G-INT / G-DIF subtype	G-INT	57	43	14	0.087	Fisher's exact (excluding NC)
	G-DIF	45	26	19		
	NC	11	8	3		
Overall survival time (months)	Range	0.2-178.1	0.2-167.0	2.27-178.1	0.298	Kaplan-Meier log-rank <sup>b</sup>
	Median <sup>a</sup>	30.0	28.8	45.1		
Tumor grade	Poorly diff	69	45	24	0.81	Fisher's exact
	Moderately diff	40	29	11		
	Well diff	4	3	1		

T: t-test; G-INT: genomic intestinal [110]G-DIF: genomic diffuse [110]; NC: not classifiable; Chisq: chi-squared test; diff: differentiated.

<sup>a</sup> Medians and log-rank test calculated by the survfit and survdiff functions in the R package “survival” (<http://cran.r-project.org/web/packages/survival/index.html>) [246].

**TABLE 4. Regions with LOH in  $\geq 35\%$  of gastric adenocarcinomas.**

<b>Region</b>	<b>Start</b>	<b>End</b>	<b>Size (Mb)</b>	<b># tumors</b>	<b>Frequency</b>	<b>Cancer genes in region</b>
17p	6,689	19,117,656	19.1	47	0.61	<i>MINK1, NLRP1, <u>TP53</u>, MAP2K4</i>
9p23-p21.1	11,219,652	30,245,385	19.0	45	0.58	<i><u>CDKN2A (p16, ARF)</u></i>
9p24.3	36,431	597,816	0.5	43	0.56	
9p24.2-p23	3,935,361	11,218,369	7.3	43	0.56	
9p24.3-p24.2	597,901	3,934,989	3.3	42	0.55	
4p16.1-p15.33	5,374,196	36,322,279	9.7	33	0.43	<i>EVC2</i>
4q13.1-q24	58,284,649	187,905,789	75.2	33	0.43	<i>CDS1, <u>NFKB1</u></i>
18q12.1-q22.3	28,823,732	69,726,488	40.9	33	0.43	<i><u>DCC</u>, ELP2, SMAD2, SMAD4</i>
3p14.2	60,114,683	60,463,734	0.3	33	0.43	<i><u>FHIT</u></i>
5q13.2-q31.3	70,702,961	141,259,534	70.6	32	0.42	<i>PIK3R1, <u>APC</u>, CTNNA1</i>
9q31.1-q31.3	104,435,842	112,108,911	7.7	31	0.4	
18p11.32-p11.22	1,543	8,733,340	8.7	20	0.26	<i>ROCK1</i>
21q21.1-q21.2	20,978,977	24,744,845	3.8	30	0.39	
14q13.1-q32.2	31,265,961	98,535,237	67.2	27	0.35	<i>NEK9, <u>TSHR</u></i>

**TABLE 5. Summary of frequently deleted regions.**

<b>Location</b>	<b>Start</b>	<b>End</b>	<b>Size (Mb)</b>	<b>q-value</b>	<b>Genes in the region</b>
3p14.2	60,375,508	60,463,797	0.1	8.80E-17	<i>FHIT</i>
9p21.3-p21.2	19,929,120	26,948,438	7.0	1.80E-10	<i>CDKN2A, CDKN2B, MLLT3</i>
18q23	73,971,555	76,117,153	2.1	1.80E-10	
4p16.3-p16.1	1,521,326	9,443,993	7.9	6.40E-07	<i>FGFR3, WHSC1, STK32B, EVC2, POLN</i>
4q34.3-q35.1	181,383,219	182,941,584	1.6	6.40E-07	
17p13.3	1	8,692,440	8.7	4.30E-06	<i>CRK, PER1, TP53, YWHAE, USP6, RABEP1, NLRP1, MINK1, CYB5D2</i>
21q21.1	19,272,221	19,819,004	0.5	5.70E-06	
8p23.3	417,858	2,188,791	1.8	7.70E-06	
5q11.2-q12.1	58,333,900	59,098,226	0.8	5.00E-05	
19p13.3	1	5,106,307	5.1	5.00E-05	<i>GNA11, MAP2K2, SH3GL1, STK11, TCF3, FSTL3</i>
14q32.2-q32.33	98,542,686	103,995,378	5.5	2.80E-04	<i>HSP90AA1, CDC42BPB, BCL11B</i>
5q35.3	178,644,954	179,734,920	1.1	6.30E-04	<i>MAPK9</i>
14q11.2	21,418,624	22,057,861	0.6	1.03E-03	
22q13.31	43,195,272	46,946,850	3.8	2.06E-03	
6p25.3	1,846,495	2,122,022	0.3	2.62E-03	
16q23.1	77,106,158	77,279,261	0.2	5.58E-03	<i>WWOX</i>
1p36.32	1	3,899,451	3.9	1.70E-02	<i>TP73, TNFRSF14, PRDM16</i>
6q26	162,484,045	162,860,227	0.4	2.30E-02	<i>PARK2</i>
12p.13.31	6,053,015	6,703,364	0.7	3.32E-02	<i>ING4, ZNF384</i>
11p15.5	1	3,445,116	3.4	4.99E-02	<i>CARS, CDKN1C, HRAS</i>
15q15.1	38,166,323	43,737,772	5.6	1.59E-01	<i>BUB1B, CASC5, HMG2P46</i>
7q36.3	154,841,836	158,821,424	4.0	1.94E-01	<i>MNX1</i>
2q37.3	239,627,375	242,951,149	3.3	2.21E-01	<i>HDAC4</i>

**TABLE 6. Summary of frequently amplified regions.**

<b>Location</b>	<b>Start</b>	<b>End</b>	<b>Size (Mb)</b>	<b>q-value</b>	<b>Genes in the region</b>
1q42.3	232,761,812	233,339,801	0.6	7.50E-02	
6p21.1	43,690,070	44,128,387	0.4	7.50E-02	
7p11.2	54,569,040	55,378,541	0.8	7.50E-02	<i>EGFR</i>
7q21.2	91,763,193	92,799,547	1.0	7.50E-02	<i>CDK6</i>
8q24.21	128,415,541	128,832,045	0.4	7.50E-02	<i>MYC</i>
12p12.1	24,958,055	25,551,845	0.6	7.50E-02	<i>KRAS</i>
13q22.1	72,589,238	73,243,304	0.7	7.50E-02	
15q26.3	94,252,474	100,338,915	6.1	7.50E-02	<i>IGF1R</i>
17q12	34,863,650	35,320,545	0.5	7.50E-02	<i>ERBB2, CDK12</i>
18q11.2	18,375,383	18,600,004	0.2	7.50E-02	
19q12	34,820,561	35,362,277	0.5	7.50E-02	<i>CCNE1</i>
20q13.33	59,924,052	62,435,964	2.5	7.50E-02	<i>SS18L1, CDH4</i>
11p13	34,438,885	35,369,622	0.9	8.41E-02	
10q26.13	123,195,729	123,531,151	0.3	1.12E-01	<i>FGFR2</i>
5p13.1	25,319,140	41,417,176	16.1	1.56E-01	<i>IL7R, LIFR</i>
2q36.3	227,164,747	228,310,820	1.1	2.03E-01	
3q26.2	170,897,488	172,571,402	1.7	2.08E-01	

**TABLE 7. Summary of regions with homozygous deletions in more than one sample.**

<b>Chr</b>	<b>Start</b>	<b>End</b>	<b>Size (Mb)</b>	<b># samples</b>	<b>Genes in the region</b>
1	554,484	3,784,133	3.23	3	<i>TP73, PRDM16</i>
2	239,627,708	242,697,433	3.07	2	<i>HDAC4</i>
3	60,278,544	60,572,503	0.29	5	
4	26,568	4,023,209	4.00	2	<i>WHSC1, FGFR3, GAK</i>
5	81,949	2,012,324	1.93	2	<i>TERT</i>
6	1,851,860	2,109,454	0.26	2	
8	103,565	2,188,481	2.08	4	<i>CSMD1</i>
8	2,188,792	4,197,712	2.01	3	<i>CSMD1</i>
8	4,198,154	11,468,631	7.27	2	<i>CSMD1</i>
8	11,532,066	43,898,071	32.37	2	<i>PCMI, WHSC1L1, FGFR1, WRN, HOOK3, MAP2K1</i>
9	21,007,240	22,088,619	1.08	2	<i>CDKN2A, CDKN2B</i>
11	188,510	3,443,300	3.25	2	<i>CDKN1C, HRAS, CARS</i>
12	6,415,872	6,555,449	0.14	2	
14	21,443,181	22,053,063	0.61	2	
14	83,885,524	85,056,916	1.17	2	
14	98,674,664	106,356,482	7.68	2	<i>AKT1, CDC42BPB, BCL11B, HSP90AA1</i>
16	26,671	2,203,517	2.18	2	<i>TSC2</i>
16	77,251,379	77,276,096	0.02	2	<i>WWOX</i>
16	84,294,768	84,458,328	0.16	3	
16	86,129,781	88,690,776	2.56	3	<i>FANCA, CBFA2T3</i>
18	46,825,370	69,678,634	22.85	2	<i>BCL2, MALT1, DCC, KDSR, SMAD4</i>
18	69,679,190	75,288,153	5.61	3	
18	75,300,011	76,116,029	0.82	4	
20	59,334,918	62,382,907	3.05	2	<i>SS18L1</i>
21	9,928,594	19,664,061	9.74	3	
21	19,668,806	22,054,808	2.39	2	
21	42,256,842	46,519,823	4.26	2	<i>U2AF1, SIK1</i>



**TABLE 8. Strong association between mutations in *TP53* hotspots and LOH at *TP53*.**

		<i>TP53</i> mutated	<i>TP53</i> wild type	P-value by Fisher's exact test
<b>LOH at <i>TP53</i></b>	<b>Yes</b>	27	12	4.8x10 <sup>-4</sup>
	<b>No</b>	3	14	

**TABLE 9. Cox proportional hazards analysis provides no evidence that LOH proportion influences prognosis.**

<b>Variable</b>	<b>coef</b>	<b>exp(coef)</b>	<b>se(coef)</b>	<b>z</b>	<b>p</b>
LOH proportion	0.1948	1.215	1.09	0.179	0.86
Age	0.0096	1.01	0.016	0.598	0.55
Gender	-0.4245	0.654	0.448	-0.949	0.34
TNM Stage	1.0303	2.802	0.276	3.731	0.00019
Adjuvanttreatment	0.2045	1.227	0.453	0.451	0.65
Intestinal type	0.6084	1.837	0.452	1.347	0.18
Mixed/other type	0.71	2.034	0.834	0.851	0.39

### 3.9 Supplementary Table

**TABLE 10. Values of different parameters used in ASCAT analysis.**

Sample	ASCAT version	$\mu$ value	Germline BAF limits	Min Goodness of Fit	Segment length
970005	2.0	1.5	0.35, 0.65	85	100
980011	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
980021	2.0	1.5	0.35, 0.65	85	100
980029	2.0	1.5	0.35, 0.65	85	100
980097	2.0	1.5	0.35, 0.65	85	100
980156	2.0	1.5	0.35, 0.65	85	100
980369	2.0	1.5	0.35, 0.65	65	100
980390	2.0	1.5	0.35, 0.65	65	800
980401	2.0	1.5	0.35, 0.65	85	100
980417	2.0	1.5	0.35, 0.65	85	100
980418	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
980437	2.0	1.5	0.35, 0.65	85	100
980447	2.0	1.5	0.35, 0.65	65	100
990005	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
990010	2.0	1.9	0.35, 0.65	85	100
990041	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
990044	2.0	1.5	0.35, 0.65	85	100
990046	2.0	1.5	0.35, 0.65	65	100
990060	2.0	1.5	0.35, 0.65	85	100
990069	2.0	1.5	0.35, 0.65	85	100
990071	2.0	1.5	0.35, 0.65	85	100
990090	2.0	1.5	0.35, 0.65	85	100
990097	2.0	1.5	0.35, 0.65	85	100
990098	2.0	1.5	0.35, 0.65	65	800
990108	2.0	1.5	0.35, 0.65	65	100
990111	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
990119	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
990170	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
990172	2.0	1.9	0.35, 0.65	85	100
990195	2.0	1.5	0.35, 0.65	85	100
990203	2.0	1.5	0.35, 0.65	85	100
990228	2.0	1.5	0.35, 0.65	85	100
990247	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
990275	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
990300	2.0	1.5	0.35, 0.65	85	100
990339	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
990355	2.0	1.5	0.35, 0.65	85	100
990396	2.0	1.5	0.35, 0.65	85	100

---

990412	2.0	1.5	0.35, 0.65	65	100
990474	2.0	1.5	0.35, 0.65	85	100
990475	2.0	1.9	0.35, 0.65	85	100
990489	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
990515	2.0	1.5	0.35, 0.65	85	100
2000040	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
2000068	2.0	1.5	0.35, 0.65	85	100
2000085	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
2000088	2.0	1.5	0.35, 0.65	85	100
2000169	2.0	1.5	0.35, 0.65	65	100
2000175	2.0	1.5	0.35, 0.65	85	100
2000201	2.0	1.5	0.35, 0.65	85	100
2000242	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
2000286	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
2000303	2.0	1.5	0.3, 0.7	85	100
2000362	2.0	1.5	0.35, 0.65	85	100
2000403	2.0	1.5	0.35, 0.65	65	100
2000433	2.0	1.5	0.35, 0.65	85	100
2000441	2.0	1.5	0.35, 0.65	85	100
2000877	2.0	1.5	0.35, 0.65	85	100
2000892	2.0	1.5	0.35, 0.65	85	100
20020011	2.0	1.9	0.35, 0.65	85	100
20020448	2.0	1.5	0.35, 0.65	65	800
20020720	2.0	1.5	0.35, 0.65	85	100
20263644	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
32226415	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
38877042	2.0	1.5	0.35, 0.65	85	100
46404174	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1
47492137	2.0	1.9	0.35, 0.65	85	100
57689477	2.0	1.5	0.35, 0.65	85	100
57701999	2.0	1.5	0.35, 0.65	85	100
58947266	2.0	1.5	0.35, 0.65	85	100
61669256	2.0	1.5	0.35, 0.65	85	100
66811693	2.0	1.5	0.35, 0.65	85	100
73291145	2.0	1.5	0.35, 0.65	85	100
76629543	2.0	1.5	0.35, 0.65	80	100
87622942	2.0	1.5	0.35, 0.65	85	100
91228050	2.0	1.9	0.35, 0.65	85	100
96141474	2.1	1.5	0.35, 0.65	85	as determined by ASCAT 2.1

---

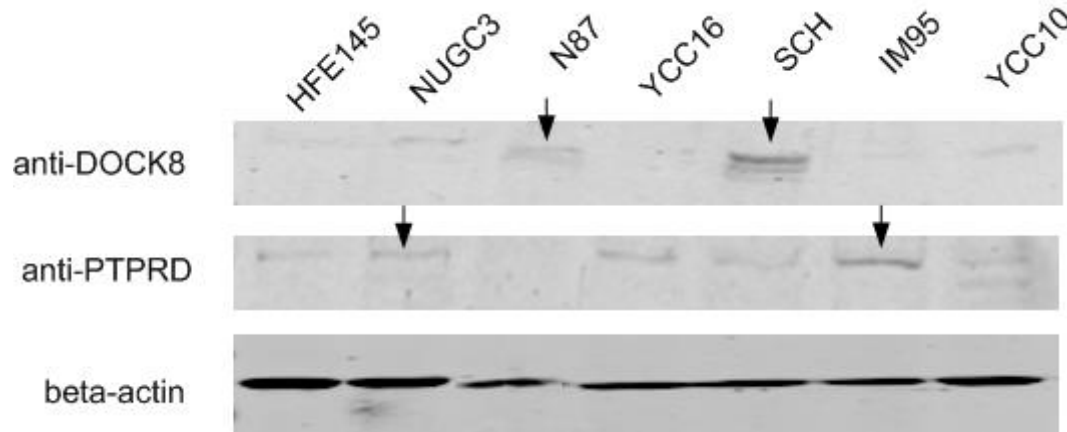
**TABLE 11. Tumors for which ASCAT was unable to estimate allele-specific copy numbers.**

<b>Sample ID</b>	<b>ASCAT fail reason</b>	<b>Laurén classification</b>	<b>Gender</b>
970003	Excessively variable LRR	Diffuse	Female
970017	Flat BAF	Diffuse	Female
980035	Flat BAF	Intestinal	Male
980251	Flat BAF	Diffuse	Female
980319	Flat BAF	Mixed	Male
980327	Flat BAF	Mixed	Female
980344	Excessively variable LRR	Diffuse	Female
980386	Flat BAF	Diffuse	Female
980436	Flat BAF	Intestinal	Female
980442	Flat BAF	Diffuse	Female
990015	Flat BAF	Intestinal	Male
990024	Flat BAF	Intestinal	Male
990068	Excessively variable LRR	Diffuse	Male
990070	Apparently low tumor content	Diffuse	Male
990089	Apparently low tumor content	Intestinal	Male
990129	Excessively variable LRR	Intestinal	Female
990136	Excessively variable LRR	Intestinal	Male
990205	Excessively variable LRR	Intestinal	Female
990413	Flat BAF	Diffuse	Male
990424	Flat BAF	Diffuse	Male
2000114	Flat BAF	Intestinal	Male
2000159	Flat BAF	Intestinal	Female
2000178	Flat BAF	Intestinal	Female
2000238	Flat BAF	Diffuse	Female
2000291	Flat BAF	Diffuse	Male
2000346	Flat BAF	Diffuse	Female
2000479	Flat BAF	Intestinal	Male
2000617	Excessively variable LRR	Intestinal	Male
2000920	Excessively variable LRR	Intestinal	Male
2001159	Flat BAF	Intestinal	Male
2001226	Excessively variable LRR	Diffuse	Male
2001229	Flat BAF	Diffuse	Female
2001241	Excessively variable LRR	Diffuse	Female
37262942	Excessively variable LRR	Diffuse	Female
43658255	Excessively variable LRR	Intestinal	Male
65256293	Flat BAF	Diffuse	Female

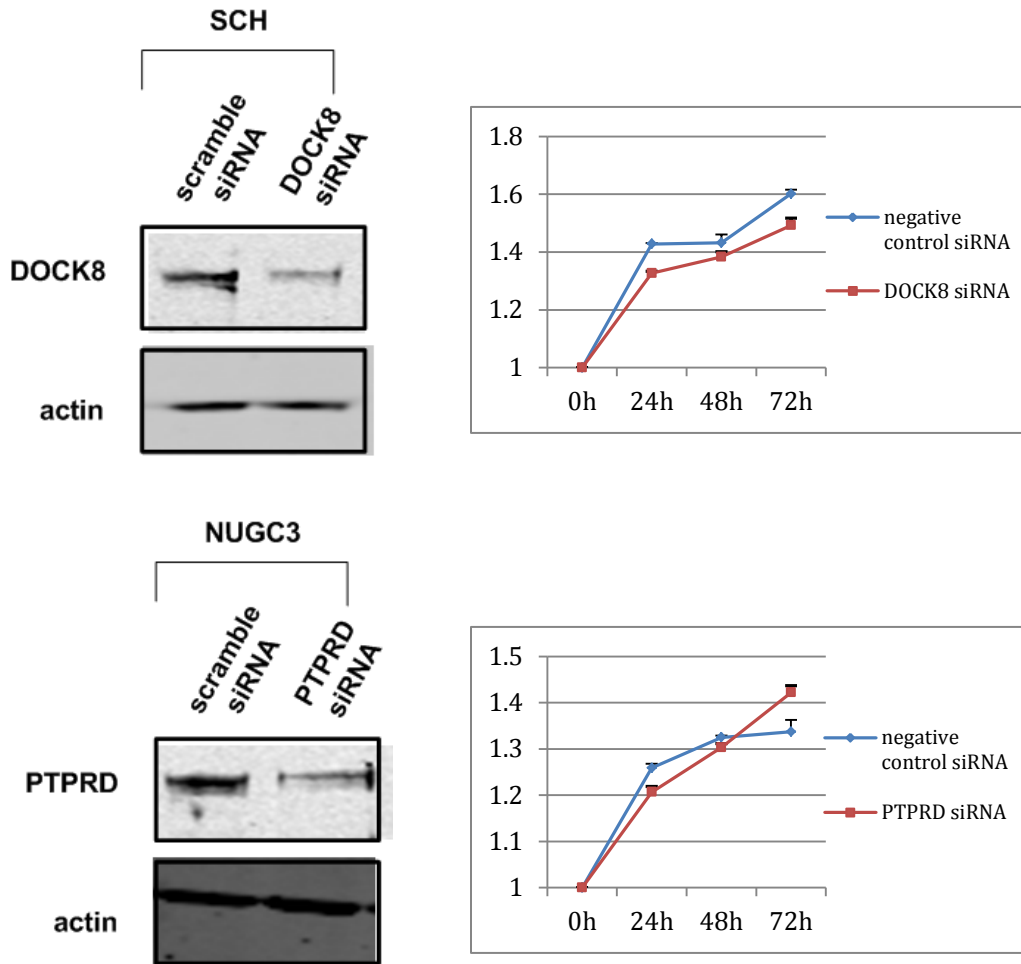
### 3.10 Supplementary Figure:

(Experimental analysis did not show significant evidence that PTPRD and DOCK8 have functional relationship to tumorigenesis)

**Figure 22. Western blot of proteins PTPRD and DOCK8 using various cell lines.**  
(Black arrows: cell lines selected for corresponding gene knock-down experiments)



**FIGURE 23. DOCK8 and PTPRD siRNA knock-down analysis show no significant effect of these two genes on cell proliferation.**



## CHAPTER 4 LOH ANALYSIS IN CANCER RESEARCH

### 4.1 Results

#### 4.1.1 LOH Analysis of Candidate TSG detected by Whole-Exome Sequencing in Gastric Cancer

Exome sequencing by next-generation sequencing approaches is a newly developed technique. It is cheaper compared to whole-genome sequencing as it targets only the coding sequences. The exome is the part of the genome that is composed by exons, the coding regions of genes that can be translated into protein and the untranslated regions flanking them (UTRs). Although exons only occupy approximately 1% of the genome, mutations on exons comprise around 85% disease-related mutations [247]. Therefore, exome sequencing is an efficient way to discover novel oncogenes and tumor suppressor genes (TSGs) with functional changes due to mutations in various diseases.

Loss of heterozygosity (LOH) is usually a second hit in the Knudson's two-hit model to lose the second normal allele after one allele of the TSG has already been lost because of mutation [4]. Thus, LOH analysis facilitates the discovery of TSGs and supports the findings of mutational analysis by exome sequencing.

We have introduced the epidemiology and etiology in Chapter 1. In our recent study, we sequenced the coding regions of ~18,000 genes in 15 gastric cancers with their matched normal samples and found 718 nonsynonymous mutations in 661 genes [78]. In addition to well-known tumor related genes such as *TP53*, *PIK3CA* and *CTNNB1*, we also revealed 26 genes that are recurrently mutated in gastric cancer (mutations found in more than 2 tumors out of the 15 discovery dataset). Among these genes, *FAT4* is especially interesting. It belongs to the E-cadherin family and may be involved in the



Wnt/planar cell polarity signaling pathway. Two non-silent mutations were found in the 15 gastric adenocarcinomas by exome sequencing and a further Sanger sequencing in the additional 95 gastric cancers discovered four more mutations. We applied the LOH analysis (as described in Chapter 2 Methods) using ASCAT [186] on these six tumors with *FAT4* mutations and found that 4 out of 6 samples also have LOH at *FAT4* (Figure 25), which supports the Knudson's model regarding the loss of function of TSGs. The combination of the exome sequencing mutational analysis and the SNP array LOH analysis reveals that *FAT4* is a candidate TSG in gastric cancer. Therefore, the *FAT4* silencing experiments were conducted on cell lines with wild-type *FAT4* and *FAT4* silencing result in a significant increase in cell proliferation. This suggests that *FAT4* functions as a tumor suppressor to suppress the tumor proliferation.

#### **4.1.2 CNA and LOH Analysis in Both SNP Array and Next-Generation**

##### **Sequence Data**

Although next-generation sequencing is now widely used to detect somatic mutations in cancers, the data can also be used for CNA and LOH analysis. The data can be used to detect: (1) gains or losses of chromosomal segments, and (2) LOH, or more generally, allelic imbalance, which is the presence of unequal numbers of maternal and paternal chromosome segments. We used a modified version of ASCAT algorithm [186], which is called RDAAC (Read Depth And Allele Count), to identify CNA and LOH from next-generation sequencing data and compared the results with those from the Affymetrix SNP 6.0 arrays. The general mechanisms of ASCAT and RDAAC are the same, except that RDAAC calculated LRR and BAF based on read depths of alleles and did not need to adjust for BAF contraction effect. To get a measure analogous to LRR obtained from

SNP arrays, RDAAC uses the  $\log_2$  ratio of read depths in the tumor sample to read depth in the matched non-malignant sample. To get a measure of the BAF, RDAAC uses the ratio of reads with the variant allele (i.e., the B allele), to total number of reads.

Results of analyses based on SNP array data and on whole-exome sequencing data are similar (Figure 26). In general, RDAAC analysis on next-generation sequencing and ASCAT analysis on SNP array data found that the tumor is hypertetraploid (average ploidy of 4.38 by RDAAC and 4.37 by ASCAT) and that the sample has ~70% tumor content (67% by RDAAC and 71% by ASCAT). We can observe a clearer separation of BAF values in the next-generation sequencing data than in the SNP array data (for example, on chromosomes 3 and 4). This may be a consequence of a nearly direct assessment of BAF in next-generation sequencing data compared to possibly non-linear measurements of allele-specific intensities in the Affymetrix SNP chip technology. Therefore, RDAAC analysis on next-generation sequencing is more efficient at detecting LOH and allelic imbalance in samples with lower tumor content. However, we also notice a higher coverage of SNPs in the Affymetrix data, which facilitates the detection of small alterations in CN and allelic balance. The larger number of SNPs in the Affymetrix data compared to the exome-data is a simple consequence of the fact that the number of SNPs present in exons is less than the number assayed by the Affymetrix chip.

## 4.2 Figures

### **FIGURE 24. The ASCAT profile of two gastric tumors assayed by Affymetrix SNP 6.0.**

LOH was found at *FAT4* in tumors that also have missense somatic mutations in *FAT4*. Panels A, B, and C refer to tumor 990515 and panels D, E, and F refer to 2000068. *FAT4* is located in the middle of the long arm of chromosome 4.

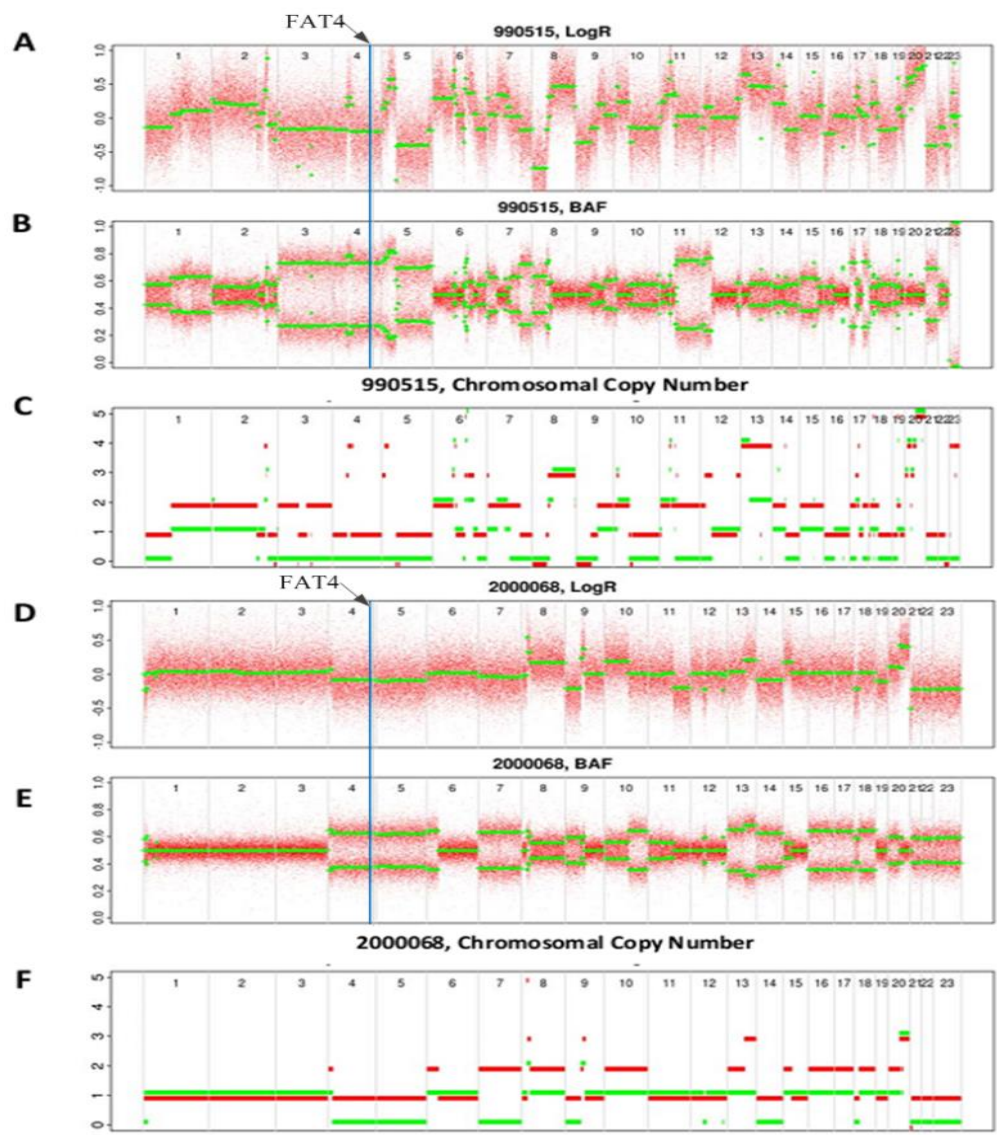
(A, D) LogR ratio (LRR). Each red dot shows, for a single SNP, the log<sub>2</sub>ratio of the total (i.e. both alleles combined) probe intensities in the tumor to the total probe intensities in the matched non-malignant sample. The overlaid green dots show the segmentation (i.e. smoothing) of these data.

(B, E) B allele frequency (BAF). Each red dot shows the proportion of non-reference alleles in the tumor sample at sites that are heterozygous in the matched non-malignant sample. As in panels A and B, the green dots show the segmentation of these data. Regions where the green dots are simultaneously displaced to values higher and lower than 0.5 are regions of LOH or allelic imbalance.

(C, F) ASCAT estimates the genomic copy number of the two parental copies of each chromosome (arbitrarily colored red and green). Note that (the “green” copy) is completely deleted (copy number 0), leading to LOH.

In both 990515 and 2000068, ASCAT estimates that one copy of chromosome 4 is completely lost (minor copy number is 0), and combined with the widely separated BAF data, we can infer LOH with copy loss at *FAT4* in both two tumors from the ASCAT profile.

\*The graph was original produced by the author of the thesis and was utilized in the supplementary of [78].



**FIGURE 25. Comparison of RDAAC analysis using next-generation sequencing data and ASCAT analysis using Affymetrix SNP 6.0 data.**

(A) Analysis on next generation whole-exome sequencing data. (B) Analysis on Affymetrix SNP 6.0 data.

The top panel shows the logR ratio (LRR), which is the  $\log_2$  of the ratio of the total signal intensity from the tumor data to the matched germ-line sample. Each red dot represents the LRR value at a site that is heterozygous in the germ-line, and the green dots show the segmentation of these data.

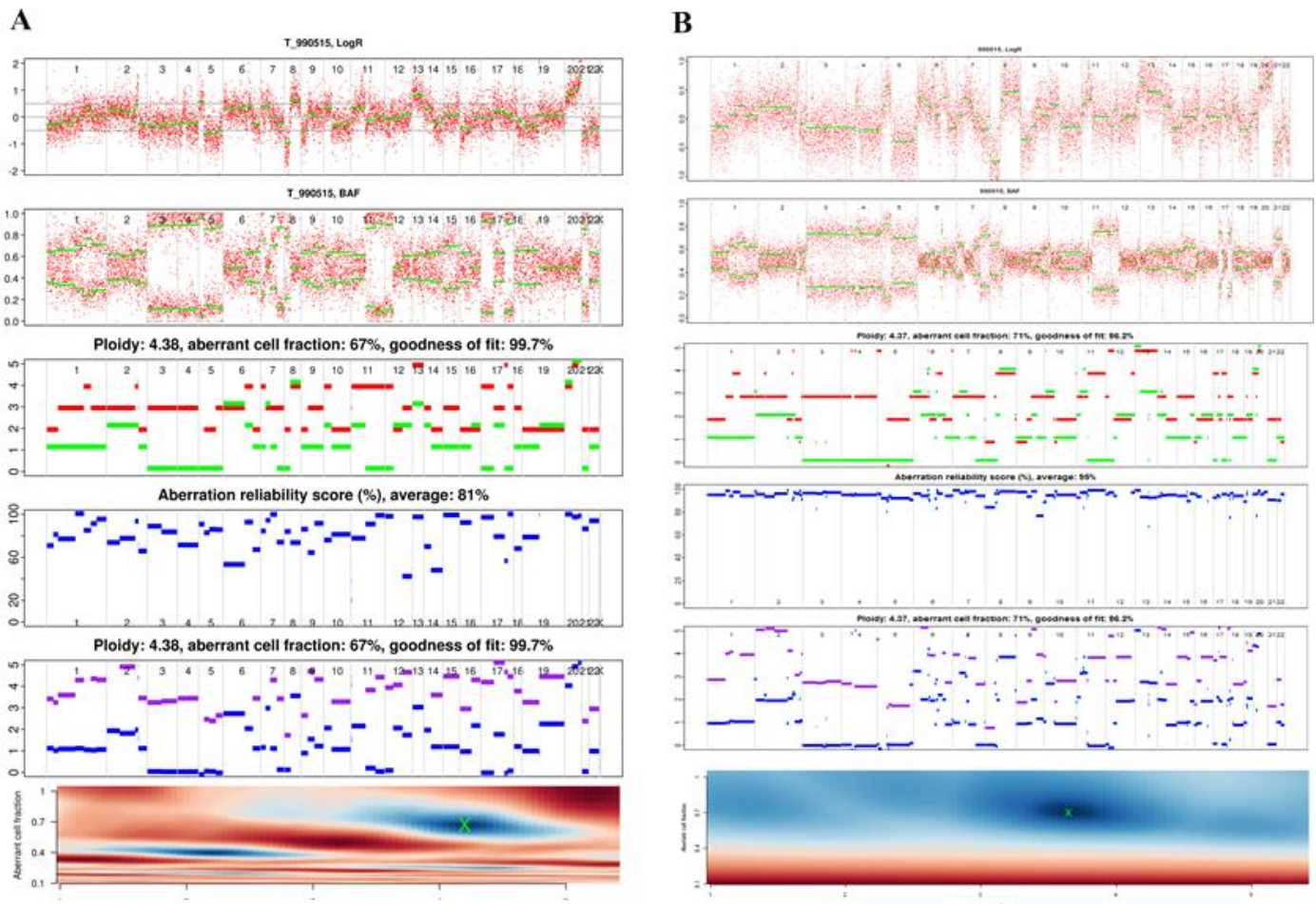
The second panel is the B allele frequency (BAF) data, which shows the proportion of non-reference alleles in the tumor sample at sites that are heterozygous in the germ-line sample. The green dots also show the segmentation of these data. Regions where the green dots are simultaneously displaced to values higher and lower than 0.5 are regions of LOH or allelic imbalance.

The third panel is the allelic-specific copy number that shows the estimated genomic copy number of the two parental copies of each chromosome (colored red: major allele copy number and green: minor allele copy number).

The fourth panel is the aberration reliability score that shows ASCAT's confidence in its estimates of chromosomal copy number at all locations that do not have one copy of each allele.

The fifth panel is the predicted chromosome-segment counts before rounding to integer values.

The bottom panel is the solution space, with the best score marked by "X".



\*(A) was produced by Dr. John Richard McPherson.

## CHAPTER 5 CONCLUDING REMARKS

Gastric cancer is one of the leading causes of cancer death worldwide. Most cases are detected as advanced stage disease, at which point treatment options are few and usually of limited benefit. However, gastric cancer has been much less intensively studied than many other common tumors, and the molecular mechanisms of gastric cancer formation and progression remain poorly understood.

CNA and LOH are common mutational events in gastric carcinogenesis. LOH is a key indicator of genomic instability and can be used to identify candidate tumor suppressor genes (TSGs). Studies of the spectrum of LOH in gastric cancer will improve our understanding of genetic alterations in gastric cancer tumors and identify possible new TSG.

In this work, we developed a modified version of the Allele Specific Copy Number Analysis of Tumors (ASCAT) algorithm that improved analysis of Affymetrix SNP6 data. Compared to other algorithms and tools (Section 2.4.1), this analysis offers better ways to:

- (1) Analyze noisy data
- (2) Avoid genotyping errors
- (3) Estimate total and allele-specific copy number in samples from aneuploid tumors
- (4) Tackle samples with low tumor content

We delineated the genome-wide landscape of CNA and LOH in gastric adenocarcinoma, including several regions, such as 9p and 17p, that frequently undergo LOH. The LOH landscape suggested the existence of novel TSGs, including *PTPRD* and *DOCK8*.

However, our functional analysis of these two genes failed to link their function to tumor proliferation. Therefore, further functional assays such as invasion and adhesion assays, are required to test their tumor suppressor functions.

Although this study utilized data generated from Affymetrix SNP 6.0 arrays, the probes of which cover the whole cancer genome with high densities, we suspect that data from Illumina Infinium whole genome genotyping (WGG) arrays might be less noisy, based on our very limited experience with this latter chip (data not shown). Thus, introducing more tumor samples and a repetition of experiments on other platforms may provide a refinement of the results in this study.

Furthermore, our experiences lead us to believe that there is still room for improvement in the development of algorithms for genome-wide assessment of CNA and LOH, especially using next-generation sequencing data, including use of whole-genome (as opposed to whole-exome) data, along with information on copy number breakpoints that whole-genome data can often provide. Beyond the basic studies of candidate TSGs, a comprehensive experimental analysis is required in subsequent studies to confirm the roles that these genes play in cancer initiation or progression.

## REFERENCES

1. Jemal, A., et al., *Global cancer statistics*. CA Cancer J Clin, 2011. **61**(2): p. 69-90.
2. *Survival rates for stomach cancer*. Available from: <http://www.cancer.org/Cancer/StomachCancer/OverviewGuide/stomach-cancer-overview-survival-rates>.
3. Volchenboum, S.L., et al., *Comparison of primary neuroblastoma tumors and derivative early-passage cell lines using genome-wide single nucleotide polymorphism array analysis*. Cancer Res, 2009. **69**(10): p. 4143-9.
4. Knudson, A.G., Jr., *Mutation and cancer: statistical study of retinoblastoma*. Proc Natl Acad Sci U S A, 1971. **68**(4): p. 820-3.
5. Law, S. and J. Wong, *Changing disease burden and management issues for esophageal cancer in the Asia-Pacific region*. J Gastroenterol Hepatol, 2002. **17**(4): p. 374-81.
6. Hamashima, C., et al., *The Japanese guidelines for gastric cancer screening*. Jpn J Clin Oncol, 2008. **38**(4): p. 259-67.
7. Layke, J.C. and P.P. Lopez, *Gastric cancer: diagnosis and treatment options*. Am Fam Physician, 2004. **69**(5): p. 1133-40.
8. *Japanese classification of gastric carcinoma: 3rd English edition*. Gastric Cancer, 2011. **14**(2): p. 101-12.
9. Wittekind, C. and B. Oberschmid, *TNM classification of malignant tumors 2010*. Pathologe, 2010. **31**(5): p. 333-338.
10. *5-year survival statistics by stage*. American Cancer Society 2012; Available from: <http://www.cancer.org/Cancer/StomachCancer/DetailedGuide/stomach-cancer-survival-rates>.
11. Lauren, P., *The Two Histological Main Types of Gastric Carcinoma: Diffuse and So-Called Intestinal-Type Carcinoma. An Attempt at a Histo-Clinical Classification*. Acta Pathol Microbiol Scand, 1965. **64**: p. 31-49.
12. Zheng, H.C., et al., *Mixed-type gastric carcinomas exhibit more aggressive features and indicate the histogenesis of carcinomas*. Virchows Arch, 2008. **452**(5): p. 525-34.
13. Carneiro, F., M. Seixas, and M. Sobrinho-Simoes, *New elements for an updated classification of the carcinomas of the stomach*. Pathol Res Pract, 1995. **191**(6): p. 571-84.



14. Vauhkonen, M., H. Vauhkonen, and P. Sipponen, *Pathology and molecular biology of gastric cancer*. Best Pract Res Clin Gastroenterol, 2006. **20**(4): p. 651-74.
15. Hamilton, S.R., et al., *Pathology and genetics of tumours of the digestive system*. 2000: IARC Press.
16. Tahara, E., *Genetic pathways of two types of gastric cancer*. IARC Sci Publ, 2004(157): p. 327-49.
17. Correa, P., *Human gastric carcinogenesis: a multistep and multifactorial process--First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention*. Cancer Res, 1992. **52**(24): p. 6735-40.
18. Ohtsu, A., et al., *Randomized phase III trial of fluorouracil alone versus fluorouracil plus cisplatin versus uracil and tegafur plus mitomycin in patients with unresectable, advanced gastric cancer: The Japan Clinical Oncology Group Study (JCOG9205)*. J Clin Oncol, 2003. **21**(1): p. 54-9.
19. Vanhoefer, U., et al., *Final results of a randomized phase III trial of sequential high-dose methotrexate, fluorouracil, and doxorubicin versus etoposide, leucovorin, and fluorouracil versus infusional fluorouracil and cisplatin in advanced gastric cancer: A trial of the European Organization for Research and Treatment of Cancer Gastrointestinal Tract Cancer Cooperative Group*. J Clin Oncol, 2000. **18**(14): p. 2648-57.
20. Lei, Z., et al., *Subtypes of Human Gastric Cancer Show Systemic Differences in Genomic and Epigenetic Characteristics and Responses to 5-FU and PI3K Inhibitors*. In review, 2012.
21. Kim, J.S., et al., *Biomarker analysis in stage III-IV (M0) gastric cancer patients who received curative surgery followed by adjuvant 5-fluorouracil and cisplatin chemotherapy: epidermal growth factor receptor (EGFR) associated with favourable survival*. Br J Cancer, 2009. **100**(5): p. 732-8.
22. Scartozzi, M., et al., *Chemotherapy for advanced gastric cancer: across the years for a standard of care*. Expert Opin Pharmacother, 2007. **8**(6): p. 797-808.
23. Deng, N., et al., *A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets*. Gut, 2012.
24. Amieva, M.R. and E.M. El-Omar, *Host-bacterial interactions in Helicobacter pylori infection*. Gastroenterology, 2008. **134**(1): p. 306-23.
25. Milne, A.N., et al., *Nature meets nurture: molecular genetics of gastric cancer*. Hum Genet, 2009. **126**(5): p. 615-28.

26. Parkin, D.M., *The global health burden of infection-associated cancers in the year 2002*. Int J Cancer, 2006. **118**(12): p. 3030-44.
27. Goldblum, J.R., et al., *Inflammation and intestinal metaplasia of the gastric cardia: the role of gastroesophageal reflux and H. pylori infection*. Gastroenterology, 1998. **114**(4): p. 633-9.
28. Ikeno, T., et al., *Helicobacter pylori-induced chronic active gastritis, intestinal metaplasia, and gastric ulcer in Mongolian gerbils*. Am J Pathol, 1999. **154**(3): p. 951-60.
29. Satoh, K., et al., *Distribution of inflammation and atrophy in the stomach of Helicobacter pylori-positive and -negative patients with chronic gastritis*. Am J Gastroenterol, 1996. **91**(5): p. 963-9.
30. Sipponen, P., et al., *Helicobacter pylori infection and chronic gastritis in gastric cancer*. J Clin Pathol, 1992. **45**(4): p. 319-23.
31. Ye, W., et al., *Helicobacter pylori infection and gastric atrophy: risk of adenocarcinoma and squamous-cell carcinoma of the esophagus and adenocarcinoma of the gastric cardia*. J Natl Cancer Inst, 2004. **96**(5): p. 388-96.
32. Basso, D., et al., *Clinical relevance of Helicobacter pylori cagA and vacA gene polymorphisms*. Gastroenterology, 2008. **135**(1): p. 91-9.
33. Polk, D.B. and R.M. Peek, Jr., *Helicobacter pylori: gastric cancer and beyond*. Nat Rev Cancer. **10**(6): p. 403-14.
34. Kim, H.J., et al., *Dietary factors and gastric cancer in Korea: a case-control study*. Int J Cancer, 2002. **97**(4): p. 531-5.
35. Kono, S. and T. Hirohata, *Nutrition and stomach cancer*. Cancer Causes Control, 1996. **7**(1): p. 41-55.
36. Ward, M.H. and L. Lopez-Carrillo, *Dietary factors and the risk of gastric cancer in Mexico City*. Am J Epidemiol, 1999. **149**(10): p. 925-32.
37. Fox, J.G., et al., *High-salt diet induces gastric epithelial hyperplasia and parietal cell loss, and enhances Helicobacter pylori colonization in C57BL/6 mice*. Cancer Res, 1999. **59**(19): p. 4823-8.
38. Hertog, M.G., et al., *Fruit and vegetable consumption and cancer mortality in the Caerphilly Study*. Cancer Epidemiol Biomarkers Prev, 1996. **5**(9): p. 673-7.
39. Sasazuki, S., S. Sasaki, and S. Tsugane, *Cigarette smoking, alcohol consumption and subsequent gastric cancer risk by subsite and histologic type*. Int J Cancer, 2002. **101**(6): p. 560-6.

40. Siman, J.H., et al., *Tobacco smoking increases the risk for gastric adenocarcinoma among Helicobacter pylori-infected individuals*. Scand J Gastroenterol, 2001. **36**(2): p. 208-13.
41. Shin, V.Y., et al., *Nicotine promotes gastric tumor growth and neovascularization by activating extracellular signal-regulated kinase and cyclooxygenase-2*. Carcinogenesis, 2004. **25**(12): p. 2487-95.
42. Sung, N.Y., et al., *Smoking, alcohol and gastric cancer risk in Korean men: the National Health Insurance Corporation Study*. Br J Cancer, 2007. **97**(5): p. 700-4.
43. Capelle, L.G., et al., *Risk and epidemiological time trends of gastric cancer in Lynch syndrome carriers in the Netherlands*. Gastroenterology, 2010. **138**(2): p. 487-92.
44. Varley, J.M., et al., *An extended Li-Fraumeni kindred with gastric carcinoma and a codon 175 mutation in TP53*. J Med Genet, 1995. **32**(12): p. 942-5.
45. Fishel, R., et al., *The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer*. Cell, 1994. **77**(1): p. 1 p following 166.
46. Papadopoulos, N., et al., *Mutation of a mutL homolog in hereditary colon cancer*. Science, 1994. **263**(5153): p. 1625-9.
47. Ou, J., et al., *Biochemical characterization of MLH3 missense mutations does not reveal an apparent role of MLH3 in Lynch syndrome*. Genes Chromosomes Cancer, 2009. **48**(4): p. 340-50.
48. Nicolaidis, N.C., et al., *Mutations of two PMS homologues in hereditary nonpolyposis colon cancer*. Nature, 1994. **371**(6492): p. 75-80.
49. Varley, J.M., *Germline TP53 mutations and Li-Fraumeni syndrome*. Hum Mutat, 2003. **21**(3): p. 313-20.
50. Guilford, P., et al., *E-cadherin germline mutations in familial gastric cancer*. Nature, 1998. **392**(6674): p. 402-5.
51. Kaurah, P. and D.G. Huntsman, *Hereditary Diffuse Gastric Cancer*. 1993.
52. Nakatsuru, S., et al., *Somatic mutation of the APC gene in gastric cancer: frequent mutations in very well differentiated adenocarcinoma and signet-ring cell carcinoma*. Hum Mol Genet, 1992. **1**(8): p. 559-63.
53. Kaurah, P., et al., *Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer*. JAMA, 2007. **297**(21): p. 2360-72.
54. Ebert, M.P., et al., *Increased beta-catenin mRNA levels and mutational alterations of the APC and beta-catenin gene are present in intestinal-type gastric cancer*. Carcinogenesis, 2002. **23**(1): p. 87-91.

55. Park, W.S., et al., *Frequent somatic mutations of the beta-catenin gene in intestinal-type gastric cancer*. *Cancer Res*, 1999. **59**(17): p. 4257-60.
56. Jakubowska, A., et al., *BRCA2 gene mutations in families with aggregations of breast and stomach cancers*. *Br J Cancer*, 2002. **87**(8): p. 888-91.
57. El-Omar, E.M., et al., *Interleukin-1 polymorphisms associated with increased risk of gastric cancer*. *Nature*, 2000. **404**(6776): p. 398-402.
58. El-Omar, E.M., et al., *Increased risk of noncardia gastric cancer associated with proinflammatory cytokine gene polymorphisms*. *Gastroenterology*, 2003. **124**(5): p. 1193-201.
59. Abnet, C.C., et al., *A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma*. *Nat Genet*, 2010. **42**(9): p. 764-7.
60. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation*. *Cell*, 2011. **144**(5): p. 646-74.
61. Boland, C.R., et al., *Infection, inflammation, and gastrointestinal cancer*. *Gut*, 2005. **54**(9): p. 1321-31.
62. Ranzani, G.N., et al., *p53 gene mutations and protein nuclear accumulation are early events in intestinal type gastric cancer but late events in diffuse type*. *Cancer Epidemiol Biomarkers Prev*, 1995. **4**(3): p. 223-31.
63. Strickler, J.G., et al., *p53 mutations and microsatellite instability in sporadic gastric cancer: when guardians fail*. *Cancer Res*, 1994. **54**(17): p. 4750-5.
64. Tamura, G., et al., *Detection of frequent p53 gene mutations in primary gastric cancer by cell sorting and polymerase chain reaction single-strand conformation polymorphism analysis*. *Cancer Res*, 1991. **51**(11): p. 3056-8.
65. Yamada, Y., et al., *p53 gene mutations in gastric cancer metastases and in gastric cancer cell lines derived from metastases*. *Cancer Res*, 1991. **51**(21): p. 5800-5.
66. Dolcet, X., et al., *NF-kB in development and progression of human cancer*. *Virchows Arch*, 2005. **446**(5): p. 475-82.
67. Karin, M. and Y. Ben-Neriah, *Phosphorylation meets ubiquitination: the control of NF-[kappa]B activity*. *Annu Rev Immunol*, 2000. **18**: p. 621-63.
68. Isomoto, H., et al., *Implication of NF-kappaB in Helicobacter pylori-associated gastritis*. *Am J Gastroenterol*, 2000. **95**(10): p. 2768-76.

69. Keates, S., et al., *Helicobacter pylori* infection activates NF-kappa B in gastric epithelial cells. *Gastroenterology*, 1997. **113**(4): p. 1099-109.
70. Zarrilli, R., V. Ricci, and M. Romano, *Molecular response of gastric epithelial cells to Helicobacter pylori-induced cell damage*. *Cell Microbiol*, 1999. **1**(2): p. 93-9.
71. Clements, W.M., et al., *beta-Catenin mutation is a frequent cause of Wnt pathway activation in gastric cancer*. *Cancer Res*, 2002. **62**(12): p. 3503-6.
72. Horii, A., et al., *The APC gene, responsible for familial adenomatous polyposis, is mutated in human gastric cancer*. *Cancer Res*, 1992. **52**(11): p. 3231-3.
73. Smith, M.G., et al., *Cellular and molecular aspects of gastric cancer*. *World J Gastroenterol*, 2006. **12**(19): p. 2979-90.
74. Taketo, M.M., *Wnt signaling and gastrointestinal tumorigenesis in mouse models*. *Oncogene*, 2006. **25**(57): p. 7522-30.
75. Park, W.S., et al., *Somatic mutations of the trefoil factor family 1 gene in gastric cancer*. *Gastroenterology*, 2000. **119**(3): p. 691-8.
76. Corso, G., et al., *Oncogenic mutations in gastric cancer with microsatellite instability*. *Eur J Cancer*, 2011. **47**(3): p. 443-51.
77. Lee, S.H., et al., *BRAF and KRAS mutations in stomach cancer*. *Oncogene*, 2003. **22**(44): p. 6942-5.
78. Zang, Z.J., et al., *Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes*. *Nat Genet*, 2012.
79. Wang, K., et al., *Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer*. *Nat Genet*, 2011. **43**(12): p. 1219-23.
80. El-Rifai, W., et al., *Consistent genetic alterations in xenografts of proximal stomach and gastro-esophageal junction adenocarcinomas*. *Cancer Res*, 1998. **58**(1): p. 34-7.
81. Kimura, Y., et al., *Genetic alterations in 102 primary gastric cancers by comparative genomic hybridization: gain of 20q and loss of 18q are associated with tumor progression*. *Mod Pathol*, 2004. **17**(11): p. 1328-37.
82. Koizumi, Y., et al., *Changes in DNA copy number in primary gastric carcinomas by comparative genomic hybridization*. *Clin Cancer Res*, 1997. **3**(7): p. 1067-76.
83. Kokkola, A., et al., *17q12-21 amplicon, a novel recurrent genetic change in intestinal type of gastric carcinoma: a comparative genomic hybridization study*. *Genes Chromosomes Cancer*, 1997. **20**(1): p. 38-43.

84. Carneiro, F. and M. Sobrinho-Simoes, *The prognostic significance of amplification and overexpression of c-met and c-erb B-2 in human gastric carcinomas*. *Cancer*, 2000. **88**(1): p. 238-40.
85. Hattori, Y., et al., *K-sam, an amplified gene in stomach cancer, is a member of the heparin-binding growth factor receptor genes*. *Proc Natl Acad Sci U S A*, 1990. **87**(15): p. 5983-7.
86. Lemoine, N.R., et al., *Amplification and overexpression of the EGF receptor and c-erbB-2 proto-oncogenes in human stomach cancer*. *Br J Cancer*, 1991. **64**(1): p. 79-83.
87. Park, J.B., et al., *Amplification, overexpression, and rearrangement of the erbB-2 protooncogene in primary human stomach carcinomas*. *Cancer Res*, 1989. **49**(23): p. 6605-9.
88. Baffa, R., et al., *Loss of FHIT expression in gastric carcinoma*. *Cancer Res*, 1998. **58**(20): p. 4708-14.
89. Li, Q.L., et al., *Causal relationship between the loss of RUNX3 expression and gastric cancer*. *Cell*, 2002. **109**(1): p. 113-24.
90. Hippo, Y., et al., *Global gene expression analysis of gastric cancer by oligonucleotide microarrays*. *Cancer Res*, 2002. **62**(1): p. 233-40.
91. Kim, H.K., et al., *Distinctions in gastric cancer gene expression signatures derived from laser capture microdissection versus histologic macrodissection*. *BMC Med Genomics*, 2011. **4**: p. 48.
92. Oue, N., et al., *Gene expression profile of gastric carcinoma: identification of genes and tags potentially involved in invasion, metastasis, and carcinogenesis by serial analysis of gene expression*. *Cancer Res*, 2004. **64**(7): p. 2397-405.
93. Chen, X., et al., *Variation in gene expression patterns in human gastric cancers*. *Mol Biol Cell*, 2003. **14**(8): p. 3208-15.
94. Boussioutas, A., et al., *Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer*. *Cancer Res*, 2003. **63**(10): p. 2569-77.
95. Wu, M.S., et al., *Genetic alterations in gastric cancer: relation to histological subtypes, tumor stage, and Helicobacter pylori infection*. *Gastroenterology*, 1997. **112**(5): p. 1457-65.
96. Yokota, J., et al., *Genetic alterations of the c-erbB-2 oncogene occur frequently in tubular adenocarcinoma of the stomach and are often accompanied by amplification of the v-erbA homologue*. *Oncogene*, 1988. **2**(3): p. 283-7.

97. Halling, K.C., et al., *Origin of microsatellite instability in gastric cancer*. Am J Pathol, 1999. **155**(1): p. 205-11.
98. Suzuki, H., et al., *Distinct methylation pattern and microsatellite instability in sporadic gastric cancer*. Int J Cancer, 1999. **83**(3): p. 309-13.
99. Keller, G., et al., *Analysis for microsatellite instability and mutations of the DNA mismatch repair gene hMLH1 in familial gastric cancer*. Int J Cancer, 1996. **68**(5): p. 571-6.
100. Thibodeau, S.N., et al., *Altered expression of hMSH2 and hMLH1 in tumors with microsatellite instability and genetic alterations in mismatch repair genes*. Cancer Res, 1996. **56**(21): p. 4836-40.
101. Falchetti, M., et al., *Gastric cancer with high-level microsatellite instability: target gene mutations, clinicopathologic features, and long-term survival*. Hum Pathol, 2008. **39**(6): p. 925-32.
102. Wu, M.S., et al., *Distinct clinicopathologic and genetic profiles in sporadic gastric cancer with different mutator phenotypes*. Genes Chromosomes Cancer, 2000. **27**(4): p. 403-11.
103. An, J.Y., et al., *Microsatellite instability in sporadic gastric cancer: its prognostic role and guidance for 5-FU based chemotherapy after R0 resection*. Int J Cancer, 2011.
104. Feinberg, A.P. and B. Vogelstein, *Hypomethylation distinguishes genes of some human cancers from their normal counterparts*. Nature, 1983. **301**(5895): p. 89-92.
105. Jones, P.A. and S.B. Baylin, *The epigenomics of cancer*. Cell, 2007. **128**(4): p. 683-92.
106. Toyota, M., et al., *Aberrant methylation in gastric cancer associated with the CpG island methylator phenotype*. Cancer Res, 1999. **59**(21): p. 5438-42.
107. Oliveira, C., et al., *Quantification of epigenetic and genetic 2nd hits in CDH1 during hereditary diffuse gastric cancer syndrome progression*. Gastroenterology, 2009. **136**(7): p. 2137-48.
108. Mitani, Y., et al., *Histone H3 acetylation is associated with reduced p21(WAF1/CIP1) expression by gastric carcinoma*. J Pathol, 2005. **205**(1): p. 65-73.
109. Yasui, W., et al., *Histone acetylation and gastrointestinal carcinogenesis*. Ann N Y Acad Sci, 2003. **983**: p. 220-31.
110. Tan, I.B., et al., *Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy*. Gastroenterology, 2011. **141**(2): p. 476-85, 485 e1-11.

111. Roukos, D.H., *Innovative genomic-based model for personalized treatment of gastric cancer: integrating current standards and new technologies*. *Expert Rev Mol Diagn*, 2008. **8**(1): p. 29-39.
112. Kuramochi, M., et al., *TSLC1 is a tumor-suppressor gene in human non-small-cell lung cancer*. *Nat Genet*, 2001. **27**(4): p. 427-30.
113. Okada, S., et al., *Loss of Heterozygosity at BRCA1 Locus Is Significantly Associated with Aggressiveness and Poor Prognosis in Breast Cancer*. *Ann Surg Oncol*, 2011.
114. Solomon, E., et al., *Chromosome 5 allele loss in human colorectal carcinomas*. *Nature*, 1987. **328**(6131): p. 616-9.
115. Fearon, E.R. and B. Vogelstein, *A genetic model for colorectal tumorigenesis*. *Cell*, 1990. **61**(5): p. 759-67.
116. Eccles, D.M., et al., *Early loss of heterozygosity on 17q in ovarian cancer. The Abe Ovarian Cancer Genetics Group*. *Oncogene*, 1992. **7**(10): p. 2069-72.
117. Ehlen, T. and L. Dubeau, *Loss of heterozygosity on chromosomal segments 3p, 6q and 11p in human ovarian carcinomas*. *Oncogene*, 1990. **5**(2): p. 219-23.
118. Lee, J.H., et al., *Frequent loss of heterozygosity on chromosomes 6q, 11, and 17 in human ovarian carcinomas*. *Cancer Res*, 1990. **50**(9): p. 2724-8.
119. Girard, L., et al., *Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering*. *Cancer Res*, 2000. **60**(17): p. 4894-906.
120. Shiseki, M., et al., *Comparative allelotype of early and advanced stage non-small cell lung carcinomas*. *Genes Chromosomes Cancer*, 1996. **17**(2): p. 71-7.
121. Tseng, R.C., et al., *Genomewide loss of heterozygosity and its clinical associations in non small cell lung cancer*. *Int J Cancer*, 2005. **117**(2): p. 241-7.
122. Ruivenkamp, C., et al., *LOH of PTPRJ occurs early in colorectal cancer and is associated with chromosomal loss of 18q12-21*. *Oncogene*, 2003. **22**(22): p. 3472-4.
123. Sato, N., et al., *Loss of heterozygosity on 10q23.3 and mutation of the tumor suppressor gene PTEN in benign endometrial cyst of the ovary: possible sequence progression from benign endometrial cyst to endometrioid carcinoma and clear cell carcinoma of the ovary*. *Cancer Res*, 2000. **60**(24): p. 7052-6.
124. Wistuba, II, et al., *Allelic losses at chromosome 8p21-23 are early and frequent events in the pathogenesis of lung cancer*. *Cancer Res*, 1999. **59**(8): p. 1973-9.



125. Campo, E., et al., *Loss of heterozygosity of p53 gene and p53 protein expression in human colorectal carcinomas*. *Cancer Res*, 1991. **51**(16): p. 4436-42.
126. Christiansen, D.H., M.K. Andersen, and J. Pedersen-Bjergaard, *Mutations with loss of heterozygosity of p53 are common in therapy-related myelodysplasia and acute myeloid leukemia after exposure to alkylating agents and significantly associated with deletion or loss of 5q, a complex karyotype, and a poor prognosis*. *J Clin Oncol*, 2001. **19**(5): p. 1405-13.
127. Huang, Y., et al., *Loss of heterozygosity involves multiple tumor suppressor genes in human esophageal cancers*. *Cancer Res*, 1992. **52**(23): p. 6525-30.
128. Bartoletti, R., et al., *Loss of P16 expression and chromosome 9p21 LOH in predicting outcome of patients affected by superficial bladder cancer*. *J Surg Res*, 2007. **143**(2): p. 422-7.
129. Brenner, A.J. and C.M. Aldaz, *Chromosome 9p allelic loss and p16/CDKN2 in breast cancer and evidence of p16 inactivation in immortal breast epithelial cells*. *Cancer Res*, 1995. **55**(13): p. 2892-5.
130. Papadimitrakopoulou, V., et al., *Frequent inactivation of p16INK4a in oral premalignant lesions*. *Oncogene*, 1997. **14**(15): p. 1799-803.
131. Reed, A.L., et al., *High frequency of p16 (CDKN2/MTS-1/INK4A) inactivation in head and neck squamous cell carcinoma*. *Cancer Res*, 1996. **56**(16): p. 3630-3.
132. Choi, S.W., et al., *Prognostic implications of microsatellite genotypes in gastric carcinoma*. *Int J Cancer*, 2000. **89**(4): p. 378-83.
133. Panani, A.D., *Cytogenetic and molecular aspects of gastric cancer: clinical implications*. *Cancer Lett*, 2008. **266**(2): p. 99-115.
134. Sano, T., et al., *Frequent loss of heterozygosity on chromosomes 1q, 5q, and 17p in human gastric carcinomas*. *Cancer Res*, 1991. **51**(11): p. 2926-31.
135. Yustein, A.S., et al., *Allelotype of gastric adenocarcinoma*. *Cancer Res*, 1999. **59**(7): p. 1437-41.
136. Bae, S.C. and J.K. Choi, *Tumor suppressor activity of RUNX3*. *Oncogene*, 2004. **23**(24): p. 4336-40.
137. Huiping, C., et al., *High frequency of LOH, MSI and abnormal expression of FHIT in gastric cancer*. *Eur J Cancer*, 2002. **38**(5): p. 728-35.
138. Mao, L., et al., *Frequent abnormalities of FHIT, a candidate tumor suppressor gene, in head and neck cancer cell lines*. *Cancer Res*, 1996. **56**(22): p. 5128-31.

139. Sozzi, G., et al., *Loss of FHIT function in lung cancer and preinvasive bronchial lesions*. Cancer Res, 1998. **58**(22): p. 5032-7.
140. Boynton, R.F., et al., *Loss of heterozygosity involving the APC and MCC genetic loci occurs in the majority of human esophageal cancers*. Proc Natl Acad Sci U S A, 1992. **89**(8): p. 3385-8.
141. Fodde, R., *The APC gene in colorectal cancer*. Eur J Cancer, 2002. **38**(7): p. 867-71.
142. Hugel, A. and N. Wernert, *Loss of heterozygosity (LOH), malignancy grade and clonality in microdissected prostate cancer*. Br J Cancer, 1999. **79**(3-4): p. 551-7.
143. Medeiros, A.C., et al., *Loss of heterozygosity affecting the APC and MCC genetic loci in patients with primary breast carcinomas*. Cancer Epidemiol Biomarkers Prev, 1994. **3**(4): p. 331-3.
144. Cairns, P., et al., *Frequent inactivation of PTEN/MMAC1 in primary prostate cancer*. Cancer Res, 1997. **57**(22): p. 4997-5000.
145. Li, J., et al., *PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer*. Science, 1997. **275**(5308): p. 1943-7.
146. Feugeas, O., et al., *Loss of heterozygosity of the RB gene is a poor prognostic factor in patients with osteosarcoma*. J Clin Oncol, 1996. **14**(2): p. 467-72.
147. Gouyer, V., et al., *Loss of heterozygosity at the RB locus correlates with loss of RB protein in primary malignant neuro-endocrine lung carcinomas*. Int J Cancer, 1994. **58**(6): p. 818-24.
148. Wang, S.I., et al., *Somatic mutations of PTEN in glioblastoma multiforme*. Cancer Res, 1997. **57**(19): p. 4183-6.
149. Xu, H.J., et al., *Loss of RB protein expression in primary bladder cancer correlates with loss of heterozygosity at the RB locus and tumor progression*. Int J Cancer, 1993. **53**(5): p. 781-4.
150. Berx, G., et al., *Mutations of the human E-cadherin (CDH1) gene*. Hum Mutat, 1998. **12**(4): p. 226-37.
151. Simpson, P.T., et al., *Molecular evolution of breast cancer*. J Pathol, 2005. **205**(2): p. 248-54.
152. Fey, M.F., et al., *Clonal allele loss in gastrointestinal cancers*. Br J Cancer, 1989. **59**(5): p. 750-4.
153. Xiao, Y.P., et al., *Loss of heterozygosity and microsatellite instabilities of fragile histidine triad gene in gastric carcinoma*. World J Gastroenterol, 2006. **12**(23): p. 3766-9.

154. Tamura, G., et al., *Allelotype of adenoma and differentiated adenocarcinoma of the stomach*. J Pathol, 1996. **180**(4): p. 371-7.
155. Ma, H., et al., *Extensive analysis of D7S486 in primary gastric cancer supports TESTIN as a candidate tumor suppressor gene*. Mol Cancer, 2010. **9**: p. 190.
156. Nishizuka, S., et al., *Loss of heterozygosity during the development and progression of differentiated adenocarcinoma of the stomach*. J Pathol, 1998. **185**(1): p. 38-43.
157. Baffa, R., et al., *Definition and refinement of chromosome 8p regions of loss of heterozygosity in gastric cancer*. Clin Cancer Res, 2000. **6**(4): p. 1372-7.
158. Choi, S.W., et al., *Fractional allelic loss in gastric carcinoma correlates with growth patterns*. Oncogene, 1998. **17**(20): p. 2655-9.
159. Sangodkar, J., et al., *Functional role of the KLF6 tumour suppressor gene in gastric cancer*. Eur J Cancer, 2009. **45**(4): p. 666-76.
160. Motomura, K., et al., *Loss of alleles at loci on chromosome 13 in human primary gastric cancers*. Genomics, 1988. **2**(2): p. 180-4.
161. Ranzani, G.N., et al., *Loss of heterozygosity and K-ras gene mutations in gastric cancer*. Hum Genet, 1993. **92**(3): p. 244-9.
162. Uchino, S., et al., *Frequent loss of heterozygosity at the DCC locus in gastric cancer*. Cancer Res, 1992. **52**(11): p. 3099-102.
163. Wang, K., et al., *PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data*. Genome Res, 2007. **17**(11): p. 1665-74.
164. Garnis, C., et al., *High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH*. Int J Cancer, 2006. **118**(6): p. 1556-64.
165. van Beers, E.H. and P.M. Nederlof, *Array-CGH and breast cancer*. Breast Cancer Research, 2006. **8**(3): p. 210.
166. Wilhelm, M., et al., *Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer*. Cancer Res, 2002. **62**(4): p. 957-60.
167. Pinkel, D. and D.G. Albertson, *Array comparative genomic hybridization and its applications in cancer*. Nat Genet, 2005. **37 Suppl**: p. S11-7.
168. McCarroll, S.A., et al., *Integrated detection and population-genetic analysis of SNPs and copy number variation*. Nat Genet, 2008. **40**(10): p. 1166-74.

169. Rigaiil, G., et al., *ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays*. *Bioinformatics*, 2008. **24**(6): p. 768-74.
170. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*. *Bioinformatics*, 2003. **19**(2): p. 185-93.
171. Li, C. and W. Hung Wong, *Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application*. *Genome Biol*, 2001. **2**(8): p. RESEARCH0032.
172. Carvalho, B., et al., *Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data*. *Biostatistics*, 2007. **8**(2): p. 485-99.
173. Bengtsson, H., P. Wirapati, and T.P. Speed, *A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6*. *Bioinformatics*, 2009. **25**(17): p. 2149-56.
174. Liu, W.M., et al., *Algorithms for large-scale genotyping microarrays*. *Bioinformatics*, 2003. **19**(18): p. 2397-403.
175. Cutler, D.J., et al., *High-throughput variation detection and genotyping using microarrays*. *Genome Res*, 2001. **11**(11): p. 1913-25.
176. *BRLMM-P: a genotype calling method for the SNP 5.0 array*, in *Affymetrix. Technical report*. 2007.
177. Korn, J.M., et al., *Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs*. *Nat Genet*, 2008. **40**(10): p. 1253-60.
178. Beroukhir, R., et al., *Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays*. *PLoS Comput Biol*, 2006. **2**(5): p. e41.
179. Nannya, Y., et al., *A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays*. *Cancer Res*, 2005. **65**(14): p. 6071-9.
180. Lin, M., et al., *dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data*. *Bioinformatics*, 2004. **20**(8): p. 1233-40.
181. Greenman, C.D., et al., *PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data*. *Biostatistics*. **11**(1): p. 164-75.
182. Colella, S., et al., *QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data*. *Nucleic Acids Res*, 2007. **35**(6): p. 2013-25.

183. Popova, T., et al., *Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays*. Genome Biol, 2009. **10**(11): p. R128.
184. Chen, H., H. Xing, and N.R. Zhang, *Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays*. PLoS Comput Biol, 2011. **7**(1): p. e1001060.
185. Rasmussen, M., et al., *Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity*. Genome Biol, 2011. **12**(10): p. R108.
186. Van Loo, P., et al., *Allele-specific copy number analysis of tumors*. Proc Natl Acad Sci U S A, 2010. **107**(39): p. 16910-5.
187. Lengauer, C., K.W. Kinzler, and B. Vogelstein, *Genetic instabilities in human cancers*. Nature, 1998. **396**(6712): p. 643-9.
188. Lieberfarb, M.E., et al., *Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP*. Cancer Res, 2003. **63**(16): p. 4781-5.
189. Zhao, X., et al., *An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays*. Cancer Res, 2004. **64**(9): p. 3060-71.
190. Primdahl, H., et al., *Allelic imbalances in human bladder cancer: genome-wide detection with high-density single-nucleotide polymorphism arrays*. J Natl Cancer Inst, 2002. **94**(3): p. 216-23.
191. Pfeifer, D., et al., *Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays*. Blood, 2007. **109**(3): p. 1202-10.
192. Staaf, J., et al., *Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays*. Genome Biol, 2008. **9**(9): p. R136.
193. Assie, G., et al., *SNP arrays in heterogeneous tissue: highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples*. Am J Hum Genet, 2008. **82**(4): p. 903-15.
194. Winchester, L., C. Yau, and J. Ragoussis, *Comparing CNV detection methods for SNP arrays*. Brief Funct Genomic Proteomic, 2009. **8**(5): p. 353-66.
195. *Copy Number and Loss of Heterozygosity Estimation Algorithms for the GeneChip Human Mapping Array Sets*, in Affymetrix. White Paper. 2007.
196. Pique-Regi, R., et al., *Sparse representation and Bayesian detection of genome copy number alterations from microarray data*. Bioinformatics, 2008. **24**(3): p. 309-18.

197. Dellinger, A.E., et al., *Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays*. *Nucleic Acids Res*, 2010. **38**(9): p. e105.
198. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data*. *Biostatistics*, 2004. **5**(4): p. 557-72.
199. Fiegler, H., et al., *Accurate and reliable high-throughput detection of copy number variation in the human genome*. *Genome Res*, 2006. **16**(12): p. 1566-74.
200. *DNA copy number and loss of heterozygosity analysis algorithms*, in *Illumina*. 2010.
201. Hupe, P., et al., *Analysis of array CGH data: from signal ratio to gain and loss of DNA regions*. *Bioinformatics*, 2004. **20**(18): p. 3413-22.
202. ; Available from: <http://www.biodiscovery.com/software/nexus-copy-number/>.
203. Eckel-Passow, J.E., et al., *Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform*. *BMC Bioinformatics*, 2011. **12**: p. 220.
204. *Affymetrix Power Tools*. Available from: <http://www.affymetrix.com/support/developer/powertools/index.affx>.
205. Bengtsson, H., et al., *Aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory*. Report 745, Department of Statistics, University of California, Berkeley, . 2008.
206. Scharpf, R.B., et al., *A multilevel model to address batch effects in copy number estimation using SNP arrays*. *Biostatistics*, 2011. **12**(1): p. 33-50.
207. Greenman, C.D., et al., *PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data*. *Biostatistics*, 2010. **11**(1): p. 164-75.
208. Olshen, A.B., et al., *Parent-specific copy number in paired tumor-normal studies using circular binary segmentation*. *Bioinformatics*, 2011. **27**(15): p. 2038-46.
209. Grigorova, M., et al., *Chromosome abnormalities in 10 lung cancer cell lines of the NCI-H series analyzed with spectral karyotyping*. *Cancer Genet Cytogenet*, 2005. **162**(1): p. 1-9.
210. *Affymetrix, Genotyping Console 4.0 User Manual*. 2008.
211. *International HapMap Project*. Available from: <http://hapmap.ncbi.nlm.nih.gov/hapmappopulations.html.en>.
212. Di, X., et al., *Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays*. *Bioinformatics*, 2005. **21**(9): p. 1958-63.

213. Kennedy, G.C., et al., *Large-scale genotyping of complex DNA*. Nat Biotechnol, 2003. **21**(10): p. 1233-7.
214. Bengtsson, H., P. Neuvial, and T.P. Speed, *TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays*. BMC Bioinformatics, 2010. **11**: p. 245.
215. Ferlay, J., et al., *Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008*. Int J Cancer, 2010. **127**(12): p. 2893-917.
216. Hartgrink, H.H., et al., *Gastric cancer*. Lancet, 2009. **374**(9688): p. 477-90.
217. Tada, M., et al., *Prognostic significance of genetic alterations detected by high-density single nucleotide polymorphism array in gastric cancer*. Cancer Sci, 2010.
218. Rhyu, M.G., et al., *Allelic deletions of MCC/APC and p53 are frequent late events in human gastric carcinogenesis*. Gastroenterology, 1994. **106**(6): p. 1584-8.
219. Baffa, R., et al., *Loss of heterozygosity for chromosome 11 in adenocarcinoma of the stomach*. Cancer Res, 1996. **56**(2): p. 268-72.
220. Vauhkonen, H., et al., *DNA copy number aberrations in intestinal-type gastric cancer revealed by array-based comparative genomic hybridization*. Cancer Genet Cytogenet, 2006. **167**(2): p. 150-4.
221. Beroukhi, R., et al., *Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma*. Proc Natl Acad Sci U S A, 2007. **104**(50): p. 20007-12.
222. Smith, M.L. and A.J. Fornace, Jr., *Genomic instability and the role of p53 mutations in cancer cells*. Curr Opin Oncol, 1995. **7**(1): p. 69-75.
223. Overholtzer, M., et al., *The presence of p53 mutations in human osteosarcomas correlates with high levels of genomic instability*. Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11547-52.
224. Wu, Y., *Genome-Wide Analysis of Loss of Heterozygosity and Discovery of Novel Tumor Suppressor Genes in Gastric Cancer*, in *Singapore-MIT Alliance*. 2013, National University of Singapore: Singapore.
225. Fang, D.C., et al., *Infrequent loss of heterozygosity of APC/MCC and DCC genes in gastric cancer showing DNA microsatellite instability*. J Clin Pathol, 1999. **52**(7): p. 504-8.
226. Vauhkonen, M., et al., *Differences in genomic instability between intestinal- and diffuse-type gastric cancer*. Gastric Cancer, 2005. **8**(4): p. 238-44.

227. Cachia, A.R., et al., *CDKN2A mutation and deletion status in thin and thick primary melanoma*. Clin Cancer Res, 2000. **6**(9): p. 3511-5.
228. Foulkes, W.D., et al., *The CDKN2A (p16) gene and human cancer*. Mol Med, 1997. **3**(1): p. 5-20.
229. Lang, J.C., et al., *Frequent mutation of p16 in squamous cell carcinoma of the head and neck*. Laryngoscope, 1998. **108**(6): p. 923-8.
230. Muscarella, P., et al., *Genetic alterations in gastrinomas and nonfunctioning pancreatic neuroendocrine tumors: an analysis of p16/MTS1 tumor suppressor gene inactivation*. Cancer Res, 1998. **58**(2): p. 237-40.
231. Wu, M.S., et al., *Intragenic homozygous deletions of MTS1 gene in gastric cancer in Taiwan*. Jpn J Cancer Res, 1996. **87**(10): p. 1052-5.
232. Zhao, G.H., et al., *Relationship between inactivation of p16 gene and gastric carcinoma*. World J Gastroenterol, 2003. **9**(5): p. 905-9.
233. Tang, S., et al., *Relationship between alterations of p16(INK4a) and p14(ARF) genes of CDKN2A locus and gastric carcinogenesis*. Chin Med J (Engl), 2003. **116**(7): p. 1083-7.
234. Wu, M.S., et al., *Overexpression of mutant p53 and c-erbB-2 proteins and mutations of the p15 and p16 genes in human gastric carcinoma: with respect to histological subtypes and stages*. J Gastroenterol Hepatol, 1998. **13**(3): p. 305-10.
235. Chung, Y.J., et al., *Microsatellite instability-associated mutations associate preferentially with the intestinal type of primary gastric carcinomas in a high-risk population*. Cancer Res, 1996. **56**(20): p. 4662-5.
236. Purdie, K.J., et al., *Allelic imbalances and microdeletions affecting the PTPRD gene in cutaneous squamous cell carcinomas detected using single nucleotide polymorphism microarray analysis*. Genes Chromosomes Cancer, 2007. **46**(7): p. 661-9.
237. Kohno, T., et al., *A catalog of genes homozygously deleted in human lung cancer and the candidacy of PTPRD as a tumor suppressor gene*. Genes Chromosomes Cancer, 2010. **49**(4): p. 342-52.
238. Solomon, D.A., et al., *Mutational inactivation of PTPRD in glioblastoma multiforme and malignant melanoma*. Cancer Res, 2008. **68**(24): p. 10300-6.
239. Veeriah, S., et al., *The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers*. Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9435-40.



240. Giefing, M., et al., *High resolution ArrayCGH and expression profiling identifies PTPRD and PCDH17/PCH68 as tumor suppressor gene candidates in laryngeal squamous cell carcinoma*. Genes Chromosomes Cancer, 2011. **50**(3): p. 154-66.
241. Nair, P., et al., *Aberrant splicing of the PTPRD gene mimics microdeletions identified at this locus in neuroblastomas*. Genes Chromosomes Cancer, 2008. **47**(3): p. 197-202.
242. Gonzalez-Quevedo, R., et al., *Receptor tyrosine phosphatase-dependent cytoskeletal remodeling by the hedgehog-responsive gene MIM/BEG4*. J Cell Biol, 2005. **168**(3): p. 453-63.
243. Sato, M., et al., *Identification of chromosome arm 9p as the most frequent target of homozygous deletions in lung cancer*. Genes Chromosomes Cancer, 2005. **44**(4): p. 405-14.
244. Takahashi, K., et al., *Homozygous deletion and reduced expression of the DOCK8 gene in human lung cancer*. Int J Oncol, 2006. **28**(2): p. 321-8.
245. Kang, J.U., et al., *Frequent silence of chromosome 9p, homozygous DOCK8, DMRT1 and DMRT3 deletion at 9p24.3 in squamous cell carcinoma of the lung*. Int J Oncol, 2010. **37**(2): p. 327-35.
246. Therneau, T.M. and P.M. Grambsch, *Modeling Survival Data: Extending the Cox Model*. 2000, New York: Springer.
247. Choi, M., et al., *Genetic diagnosis by whole exome capture and massively parallel DNA sequencing*. Proc Natl Acad Sci U S A, 2009. **106**(45): p. 19096-101.