Investigation of Role of Chromosomal Aberrations in Carcinogenesis by Undertaking Bioinformatic Approaches

LAM, Man Ting

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

in

Medical Sciences

The Chinese University of Hong Kong

September 2011

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

Abstract of thesis entitled

Investigation of Role of Chromosomal Aberrations in Carcinogenesis by Undertaking Bioinformatic Approaches

Submitted by LAM, Man Ting

for the degree of Doctor of Philosophy

at the Chinese University of Hong Kong

in September 2011

Cancer is the leading cause of death worldwide. Gene mutations, chromosomal aberrations and epigenetic changes are the major genetic defects contributing to carcinogenesis. In the present research study, the role of chromosomal aberrations in carcinogenesis was focused.

By undertaking bioinformatic approaches, the objectives of the part 1 of this research study were: 1) to extract chromosomal breakpoint data from public CGH database; 2) to determine the baseline chromosomal breakage in carcinogenesis; 3) to explore the possible mechanism underlying chromosomal breakage in cancer; and 4) to investigate whether there are gender bias in breakpoint frequency.

Rates of baseline random breakage and cancer-associated breakage in individual cancer types were successfully extracted and characterized from CGH data of a total of 14,322 cases. CNVRs and centromeres/pericentromeres were identified as the key structural elements associated with baseline random chromosomal breakages. Furthermore, comparison of the baseline breakpoint frequencies in males and females revealed that the higher baseline rate of chromosomal breakage in males was one of the possible causes contributing to the higher occurrence rate of cancers in males. It is important to note that there was no

gender bias in cancer-associated breakage events, suggesting that the gender bias in the baseline chromosomal breakage happens at or before the stage of tumour initiation, but not during tumour progression.

Among all the cancers, gastric cancer is one of the most common types worldwide, and is also one of the leading causes of cancer-related mortality. The results from the first part of the present study supports that there should be two types of chromosomal imbalances, i.e., random and non-random events. The non-random recurrent chromosomal imbalances in gastric cancer remained inconclusive. It was worthwhile to characterize the chromosomal aberrations in gastric cancer.

The objectives of the part 2 of this research study were: 1) to identify the recurrent non-random chromosomal aberrations in gastric adenoma and gastric cancer by meta-analysis of CGH studies; 2) to develop a tumour progression model from the recurrent chromosomal imbalances; and 3) to identify potential cancer driver genes in the recurrent chromosomal imbalances.

A list of non-random recurrent chromosomal imbalances was identified by meta-analysis of CGH data from 41 GA, 198 IGC and 117 DGC cases. Tumour progression models of IGC and DGC were constructed on the basis of these non-random events. A web-based genome browser, Gbrowse, was established for identification of novel cancer driver genes through the mapping between the regions of recurrent chromosomal imbalances and differential transcriptomic data. Identification of KPNA2 as a potential cancer driver gene in gastric cancer showed proof-of-principle for the presence of potential cancer-driver genes, tumour suppressor genes and therapeutic targets in these recurrent chromosomal imbalances.

In conclusion, bioinformatics is useful in characterizing the chromosomal aberrations of cancers as well as extracting the hidden information that provide novel insights into their possible roles in carcinogenesis.

摘要

癌症是全球導致死亡的首要病因， 而基因突變、染色體畸變和表觀遺傳變異則爲引發癌變的主要遺傳缺陷。在這項研究中，我們專注於研究染色體畸變在引發癌變中的角色。

本研究的第一部份，利用生物信息學的方法，預期達致以下目標：1）從公共染色體變異數據庫中提取染色體斷點數據；2）測定在癌變過程中染色體斷裂的基準線； 3）探索腫瘤中染色體斷裂的潛在機制；及 4）調查染色體斷裂頻率是否存在性別差異。

我們從公共染色體變異數據庫的 14,322 宗癌症病例中成功提取了隨機基準染色體斷裂率及癌症相關的染色體斷裂率。複製數變異區段及著絲粒 / 著絲粒周邊區經確認爲與基準隨機染色體斷裂率相關的重要染色體的結構元素。此外，透過比較男女癌症患者的基準染色體斷裂率，我們發現男性的基準染色體斷裂率遠高於女性，此發現可能是導致男性罹患較高癌症比率的其中一個原因。另一個重點發現，就是癌症相關染色體斷裂率並沒有出現性別差異。因此我們提出基準染色體斷裂率的性別差異早於腫瘤啓動之時或之前已出現，而並非在腫瘤發展時出現的推論。

在各類癌症中，胃癌是世上最普遍的癌症，也是主要的癌症殺手之一。本研究第一部分的結果指出有兩種類型的染色體失衡，即隨機性失衡和非隨機性失衡。現時學術界對於胃癌中的染色體非隨機性失衡仍沒有定論，故此我們認爲值得去研究胃癌中染色體畸變的特徵。

本研究第二部份的目標如下：1）綜合相對性基因雜交技術(CGH) 的研究數據並進行薈萃分析，以區分胃腺瘤和胃癌中非隨機性經常出現的染色體變異；2）

探究這些非隨機性出現的染色體變異構建腫瘤的發展模式並 3）從中找出潛在的癌症基因。

透過綜合 41 宗胃腺瘤、198 腸道型胃癌及 117 宗瀰漫型胃癌的 CGH 研究數據，我們進行薈萃分析，找出了一系列胃腺瘤和胃癌中非隨機性經常出現的染色體變異，並根據這些非隨機性出現的染色體變異構建構胃癌的發展模式。

接著我們建立一個網上界面 Gbrowse，以識別這些非隨機性出現的染色體變異與所表達的基因差異的數據，並從中尋找潛在的癌症基因。由此我們鑑定了 KPNA2 基因爲一個潛在的癌症基因，驗證了潛在的癌症基因、腫瘤抑制基因及癌症治療目標存在於非隨機性出現的染色體變異差區域的這項原則。

總括來說，生物信息學長於解構腫瘤染色體畸變的特徵，以及撮取其中的隱藏信息，並藉此爲染色體畸變在引發癌變中的角色給我們提供新的見解。

ACKNOWLEDGEMENTS

Ad Majorem Dei Gloriam – To the Greater Glory of God

I am heavily indebted to my supervisor, Prof. Terence C. W. Poon for his professional guidance, teaching and encouragement in the past four years. It is impossible to complete this project without his support.

I am thankful for the financial support from Li Ka Shing Foundation for supporting 'Studies in Integrative Medical Sciences' which makes this project feasible throughout my Ph. D. phase.

I would like to express my gratitude to Prof. Joseph J. Y. Sung and Prof. Francis K. L. Chan for their support in this project; Prof. H.C. Jin, former Research Assistant Professor of Institute of Digestive Disease, Department of Medicine and Therapeutics, CUHK, and his research group for their teaching and support on the cell culture and other molecular biotechnology techniques and the provision of patient samples for this study; Prof. J. Yu, Dr. Eagle Chu and other members of Institute of Digestive Disease, CUHK, for their help and support in all the laboratory work.

I would like to thank Mr. Patrick L. H. Ho for his guidance and support on the computer programming and bioinformatics analysis; former and current members of the Medical Proteomics Laboratory, Department of Paediatrics, CUHK, for their support and companionship.

I would also like to thank Prof. Ka-fai To, Prof. Nathalie Wong, Dr. Arthur K. K. Ching and Dr. Joanna H. M. Tong, Department of Anatomical & Cellular Pathology, CUHK; Prof. Richard Choy, Department of Obstetrics & Gynaecology, CUHK and Prof. Paul B. S. Lai, Department of Surgery, CUHK, for their valuable discussions and advice; and Dr. Scott Cain, GMOD Coordinator on Gbrowse, Cold Spring Harbor Laboratory, for his technical support on the construction of the web-based interface Gbrowse.

The special thank goes to Miss Jenny Tse for her help in polishing this thesis. My gratitude is also extended to all my friends for their encouragement and prayers. Last but not least, I am extremely grateful to my parents for their unlimited love, care, tolerance, emotional support and prayers.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

TABLES

FIGURES

ABBREVIATIONS

°C	Degree Celsius
µg	Microgram
µL	Microlitre
µM	Micromolar
18SrRNA	18S ribosomal RNA
ABL	V-abl Abelson murine leukemia viral oncogene homolog 1
aCGH	Array comparative genomic hybridization
APC	Adenomatous polyposis coli
BCL2	B-cell lymphoma 2
BCR	Breakpoint cluster region
BCR-ABL	Fusion gene of breakpoint cluster region and V-abl Abelson murine leukemia viral oncogene homolog 1
BDIM	Birth, death and innovation models
BL-BP	Baseline breakpoint
BP	Breakpoint
Ca-BP	Cancer-associated breakpoint
CDH1	E-cadherin
cDNA	Complementary deoxyribonucleic acid
CGH	Comparative genomic hybridization
CNV	Copy number variation
CNVR	Copy number variation region
CpG island	-C-phosphate-G- island
DAPI	4',6-diamidino-2-phenylindole
DGC	Diffuse-type gastric cancer
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide triphosphates
EBV	Epstein-Barr virus
FAB M5	French-American-British M5 subtype
FDR	False discovery rate
FISH	Fluorescence in situ hybridization
GA	Gastric adenoma
gDNA	Genomic DNA
GMOD	Generic Model Organism Database
GTP	Guanosine-5'-triphosphate
HCC	Hepatocellular carcinoma
HOXB6	Homeobox B6
HOXB7	Homeobox B7
HTLV-1	Human T-cell lymphotropic virus type 1
IARC	International Agency for Research on Cancer
ICD-10	The International Statistical Classification of Diseases and Related Health Problems, 10th Revision
ICD-O-3	International Classification of Diseases for Oncology, 3rd

	Edition
IGC	Intestinal-type gastric cancer
IM	Intestinal metaplasia
Indel	Insertion-deletion
Inv	Inversions
KPNA2	Karyopherin alpha 2 (RAG cohort 1, importin alpha 1)
KRAS	V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
M/F ratios	Male to female ratio
max-BP	Maximum breakpoint
M-FISH	Multi-fluorescence in situ hybridization
$MgCl_2$	Magnesium chloride
Min-BP	Minimum breakpoint
mL	Millilitre
ML	Malignant Lymphoma
mM	Millimolar
MMBIR	Microhomology-mediated break-induced replication
MTS	3-(4,5-dimethylthiazol-2-yl)-2-(4-sulfophenyl)-2H-tetrazolium
MUC1	Mucin 1, cell surface associated
NA	Not applicable
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NK	Natural Killer
nm	Nanometre
NOS	Not otherwise specified
PCA	Principal component analysis
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
RPMI	Roswell Park Memorial Institute
RT-PCR	Real-time polymerase chain reaction
SD	Segmental duplication
S.D.	Standard deviation
SFRPs	Secreted frizzled-related proteins
Silver (NOR) Stain	Silver Nucleolar Organizing Region Staining
siRNA	Small interfering RNA
SKY	Spectral karyotype
SMD	Stanford Microarray Database
SOM	Self-organizing maps
SOTA	Self Organizing Tree Algorithm
SV	Structural variation
TOP2A	Topoisomerase (DNA) II alpha
UBE2C	Ubiquitin-conjugating enzyme E2 C
UCSC	University of California, Santa Cruz
WHO	World Health Organization

CHAPTER 1

LITERATURE REVIEW

1.1 Introduction of carcinogenesis

Cancer, by definition, is a group of abnormal cells with malfunctioning growth control mechanism with the potential to invade other parts of the body[1]. It remains the major causes of mortality throughout the world, accounting for about 13% of all deaths in 2008. Although scientists had started investigating the origin of cancer more than a century ago, as Boveri suggested that abnormal chromosomal context could result in tumour[2], the mechanism of carcinogenesis is still not fully understood. With an advanced understanding of the chromosomal structures, three major theories on carcinogenesis were proposed[3].

1.2 Theories of carcinogenesis

The first one is the gene mutation theory. It is well-accepted that somatic gene mutations can transform a normal cell into a tumour cell and result in carcinogenesis through clonal expansion of tumour cells. There are numerous studies reporting the involvement of mutation in various genes in tumourigenesis by the activation of proto-oncogenes, alternation in the cell functions such as cell proliferation, cell cycle control and chromosomal repairing functions, etc. For example, mutation at p53, a well-known tumour-suppressor gene, was commonly found in various cancer types. Such mutation would result in defect in cell cycle control and other cellular functions such as DNA repair and recombination[4]. Another example is the mutation in V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog (KRAS). KRAS is a well-studied oncogene that has been found with mutation(s) in 17-25% of all human tumours. Mutated KRAS oncogene encodes a mutant protein that keeps binding to Guanosine-5'-triphosphate (GTP) and shows reduced GTPase activity. This affects

the subsequent signal transduction pathways, which in turn alters various cellular functions, e.g. stimulation of cell proliferation[5].

The second theory on carcinogenesis is the aneuploidy theory. The transformation of a normal cell to a tumour cell requires changes in gene expression in massive number of genes, which cannot be explained solely by gene mutation. Aneuploidy, that is numerical aberration of chromosomes, results in copy-number change in massive number of genes in tumour cells, leading to a massive change in the gene expression pattern[6]. Random aneuploidy can be induced by carcinogens. It can catalyze karyotypic evolutions by destabilizing the karyotype, disturbing the balance in the functional proteins involved in segregation, synthesis and repair of chromosomes[7].

The third theory is the epigenetic theory. The transcription of a gene can be switched on or off by methylation of specific DNA sequence adjacent to the gene, especially in the promoter region. Thus, even without chromosomal aberration, the expression of a gene can be dysregulated[8]. Besides methylation at cytosine within a CpG island, covalent modifications of histones and other chromatin components also participate in the epigenetic gene silencing. Aberrant epigenetic gene silencing contributes to carcinogenesis, as illustrated in the silencing of p16, SFRPs, GATA-4, GATA-5, and APC in colon cancer that promotes abnormal clonal expansion[9].

It is impossible to use any single theory to explain the complex process of carcinogenesis. Gene mutation, chromosomal aberrations and epigenetic gene silencing can co-occur in the same case of solid tumour. In the current study, the role of chromosomal aberrations in carcinogenesis was focused.

1.3 Introduction of chromosomal aberrations

Stability of the chromosome is essential for cell survival and maintenance of normal cellular functions. Numerical and structural variations in the chromosomes that can be lethal to a cell are called chromosomal aberrations.

1.3.1 Types of numerical aberrations

There are two types of numerical aberrations. Polyploidy refers to the total number of chromosomes being a multiple of the normal number of chromosomes. Aneuploidy refers to the total number of chromosomes higher or lower than the normal number of chromosomes but not being an exact multiple of the normal number of chromosomes[10-12].

1.3.2 Types of structural aberrations

Among structural aberrations, chromosomal changes can be further classified into different types. Some structural aberrations involve gain or loss in chromosomal materials, e.g. deletion or duplication, while some just involve changes in loci position, e.g. translocation and inversion.

Deletion

Certain parts of the chromosomes can be detached upon chromosomal breakages. A part without the centromere lags in anaphase. Subsequently, it is digested by nuclease and eventually lost. For a single breakage in a chromosome, loss of the acentric part from a terminus to the position of breakage is known as terminal deletion. In the case of two breakages in the same chromosome arm,

interstitial deletion may occur by losing the segment between the two breakage positions and reunion of the flanking regions[10-12].

A classical example of human diseases associated with chromosomal deletion is the cri-du-chat syndrome, which is caused by the partial loss of the short arm of chromosome 5[13].

1.3.2.1 Inversion

Inversion refers to a situation in which two breaks occur in a chromosome and a segment of the chromosome detaches from and rejoins with the same chromosome in a reversed orientation. This intrachromosomal aberration does not involve any gain or loss of chromosomal materials.

Inversions were classified into two types, paracentric inversion and pericentric inversion. Paracentric inversion does not involve the centromere as the two breakages occur on the same chromosomal arm, while pericentric inversion involves the centromere as the chromosomal breakages happen on different chromosomal arms. In some cases overlapping inversion occurs as secondary inversion takes place inside the chromosomal segment of the first inversion[10-12].

Since there is no gain or loss of any genetic materials, inversion is usually tolerated by an organism, though there may be changes in the phenotype due to the effect of altered gene position. In addition, inversion also contributes to the evolution of new species, e.g. divergence between the *Drosophila pseudoobscura* and *D. persimilis*[14].

1.3.2.2 Translocation

Translocation is an interchromosomal structural aberration with the transfer of a segment from one chromosome to a non-homologous chromosome. Based on the

position of the detached segment and the incorporation position in the receiving non-homologous chromosome, these aberrations could be classified into different types, as illustrated below:[10-12]

The first type is the simple translocation, where chromosomal segment from the donating chromosome attaches to the terminus of the receiving chromosome. However, this type of translocation is rare as the terminus of the receiving chromosome is protected by a telomere.

The second type is called reciprocal translocation. With a single breakages occurring in each of the two non-homologous chromosomes, the detached segment would leave the donating chromosome to attach to the receiving chromosome. Reciprocal translocation exists in either homozygous state or heterozygous state. Homozygous reciprocal refers to a mutual exchange of chromosomal segments between each of the two homologous copies of the donating chromosome with that of the receiving chromosome, while heterozygous reciprocal refers to the exchange between only one of the homologous copies.

For both simple and reciprocal translocation, the two new chromosomes would be retained by a cell if each of them possesses only one centromere. However, if they fail to be retained due to improper segregation in anaphase, acentric chromosomal segment or dicentric chromosomes will result.

The third type is the multiple translocation, where more than two pairs of non-homologous chromosomes are involved in the exchange of chromosomal segments. Multiple translocation can lead to the formation of translocation complex in form of a chromosome ring at the metaphase.

Translocation of chromosomal segments plays an important role in carcinogenesis. For example, Reciprocal translocation of chromosomal segment

between chromosome 9 and chromosome 22, which is known as Philadelphia chromosome, results in the molecular juxtaposition of two genes, BCR and ABL. The aberrant BCR-ABL gene formed on chromosome 22 is found to be associated with chronic myelogenous leukaemia and also with some acute leukaemias[15].

1.4 Cause of chromosomal aberrations

There are various causes of chromosomal aberrations, including ultra-violet radiation, gamma radiation[16], X-ray radiation[17], infection by virus, e.g. Hepatitis B virus[18], BK virus[19], Herpes simplex virus[20], oncovirus Rous sarcoma virus and simian adenovirus Sa7[21], reactive oxygen species from air pollutant[22], and carcinogenic chemicals, e.g. benzene[23]aflatoxin B1, Aroclor 1254, benzidine, benzo[a]pyrene and 20-methylcholanthrene[24]etc. Studies also suggested that aging is strongly associated with chromosomal aberrations[25, 26].

1.5 Cytogenetic techniques for assessment of chromosomal breakages and aberrations

As mentioned before, chromosomal breakage is critical for the occurrence of structural aberrations. Patients with chromosomal breakage syndromes, e.g. Fanconi anemia and Nijmegen breakage syndrome, are cancer-prone[27, 28].

Cytogenetic techniques are useful in identifying and assessing chromosomal breakage and aberrations in these patients and cancer patients. These methods include the conventional staining methods for karyotyping, fluorescence in situ hybridization (FISH), comparative genomic hybridization (CGH), CGH arrays (aCGH) and next generation sequencing (NGS).

1.5.1 Chromosome staining

In the early 1970s, cytogeneticists studied the chromosome morphology by studying the darker or lighter bands using various staining techniques (as shown in Table 1.1 [29]). The highest resolution of the karyotype obtained was up to a band level of about 1000[30], which allowed the identification of disease related to microdeletion and microduplication, e.g. cri-du-chat syndrome with 5p deletion[31].

As the karyotype obtained from chromosome banding is of low resolution (> 5 Mb[32]), the point of chromosomal breakage could only be roughly mapped to a chromosome band with a higher degree of uncertainty[33].

Table 1.1 Summary of conventional chromosome banding techniques[29].

Stain or Banding Technique	Investigator	Year
Q-banding [34]	Caspersson, Zech, Johansson	1970
G-banding (by trypsin) [35]	Seabright	1971
G-banding (by acetic-saline) [36]	Sumner, Evans, Buckland	1971
C-banding [37]	Arrighi, Hsu	1971
R-banding (by heat and Giemsa) [38]	Dutrillaux, Lejeune	1971
G-11 stain [39]	Bobrow, Madan, Pearson	1972
Antibody bands [40]	Dev *et al.*	1972
R-banding (by fluorescence) [41]	Bobrow, Madan	1973
In vitro bands (by actinomycin D) [42]	Shafer	1973
T-banding [43]	Dutrillaux	1973
Replication banding [44]	Latt	1973
Silver (NOR) Stain [45]	Howell, Denton, Diamond	1973
High resolution banding [46]	Yunis	1975
DAPI/distamycin A stain [47]	Schweizer, Ambros, Andrle	1978
Restriction endonuclease banding [48]	Sahasrabuddhe, Pathak, Hsu	1978

1.5.2 Fluorescence in situ hybridization (FISH)

The continuous advancement in molecular biology has given rise to new cytogenetic techniques. For instance, fluorescence in situ hybridization (FISH) has allowed the study of chromosomal morphology at a higher resolution (1-5 Mb[32]) and could detect symmetric translocations of terminal bands with same band pattern which could not be otherwise detected accurately by solid staining[49].

Through the in situ hybridization of a labelled probe to the fixed cell or chromosomes, target DNA segment complimentary to the probe would be visualized[50]. For example, chromosomal rearrangement were identified in chromosome 1p in cases of head and neck carcinomas by FISH and a breakpoint

was found in the pericentromeric region[51]. Breakpoints in BCL2 gene were also found in follicular lymphomas by using FISH assay[52].

SKY[53] and M-FISH[54] were developed on the basis of FISH. With the use of multiple probes labelled with different fluorophores and advance image capture device, each chromosome can be labelled with different colours, and thus allows easy identification of translocation. For example, translocation breakpoints in 3p14 in renal cell carcinoma were identified by multi-colour FISH[55].

1.5.3 Comparative Genomic Hybridization (CGH)

In order to examine tissue specimens with unknown chromosomal aberrations, a genome-wide fluorescent molecular cytogenetic technique was developed, where chromosomal gains and losses were identified from the ratio of the hybridization signals that were resulted from competitive hybridization of differentially labelled DNA from a test sample and a reference sample on metaphase chromosome spreads[56].

Since this technology allows the identification of novel chromosomal gain or loss, CGH have been widely used in the study of tumour cytogenetics. Various databases had been set up to integrate CGH data from world-wide studies of different cancer types, e.g. National Center for Biotechnology Information (NCBI) Entrez Cancer Chromosomes site (http://www.ncbi.nlm.nih.gov/cancerchromosomes) [57] and Progenetix[58] (http://www.progenetix.net). However, the use of metaphase chromosomes has been limited the resolution to 5 Mb[32].

1.5.4 Array Comparative Genomic Hybridization (aCGH)

In order to improve the resolution of CGH, metaphase chromosomes were replaced by an array of DNA probes immobilized on a glass slide. This high-throughput technique is known as array CGH. Higher resolution (>50 kb [32]) can be achieved by adjusting the sizes and the density of the probes in the array. Commercially available aCGH platform can consist up to 44,000 synthetic 60-mer oligonucleotide probes, covering the whole genome with an average spatial resolution of 30–35 kb, and have been validated for clinical use[59]. Nowadays aCGH has been being used in routine service for molecular karyotyping[60].

Submicroscopic structural variations of chromosomal segments can now be easily detected by aCGH. Unlike a reference genome, variations of copy number of DNA segments with the size of 1 kb - 50kb on certain chromosomes which occur in the form of blocks of variations are defined as copy-number variations. These DNA segments, known as copy number variants (CNV) are further classified as insertions, deletions and cyclic repeats[61, 62].

Segments of DNA with high similarity (>90%) in the sequences with the size of > 1kb that occurs in more than two copies per haploid genome and can be found at more than one site are known as segmental duplication (SD). When SDs are present in a variable copy number within one site, they can be also considered as CNVs[61].

SDs and CNVs detected by aCGH in breast cancer[63], colon cancer[64], gastric cancer[65], prostate cancer[66] and lymphoma[67] provide useful information for prognosis of cancer and possible therapeutic targets.

1.5.5 Next generation sequencing (NGS)

The advent of high-throughput next generation sequencing (NGS) technologies has made possible parallel sequencing of hundreds of samples[68]. This allows effective shotgun sequencing of the whole genome from tumour and reference DNA samples. The number of sequences from the tumour sample aligned to the reference sample revealed the relative copy number of such sequence in the tumour sample, thus this high resolution technology enables the detection of CNVs with size of 1 kb [69].

Moreover, balanced translocation of chromosomal materials cannot be detected by CGH and aCGH. By using specific primers flanking a breakpoint, the junction fragment of a translocated chromosomal segment can be amplified by PCR. Then the sequence of the junction fragment can easily be obtained by submitting the PCR products to NGS. Alignment of this sequence to the reference sequence allows mapping of the breakpoint [70].

1.6 Overview of Gastric Cancer

1.6.1 Epidemiology of gastric cancer

Gastric cancer refers to the malignancy developed in the stomach. Gastric cancer, also know as stomach cancer, ranked fourth in the cancer incidence and mortality worldwide in 2008, which generated 989,000 new cases and resulted in 738,000 deaths[71]. Analysis of geographic distribution revealed high incidence and mortality rate of gastric cancer in Eastern Asia, with estimated age-standardized incidence rates of 42.4 male and 18.3 female per 100,000 and estimated age-standardized mortality rates of 28.1 male and 13 female per 100,000. (Figure 1.1)

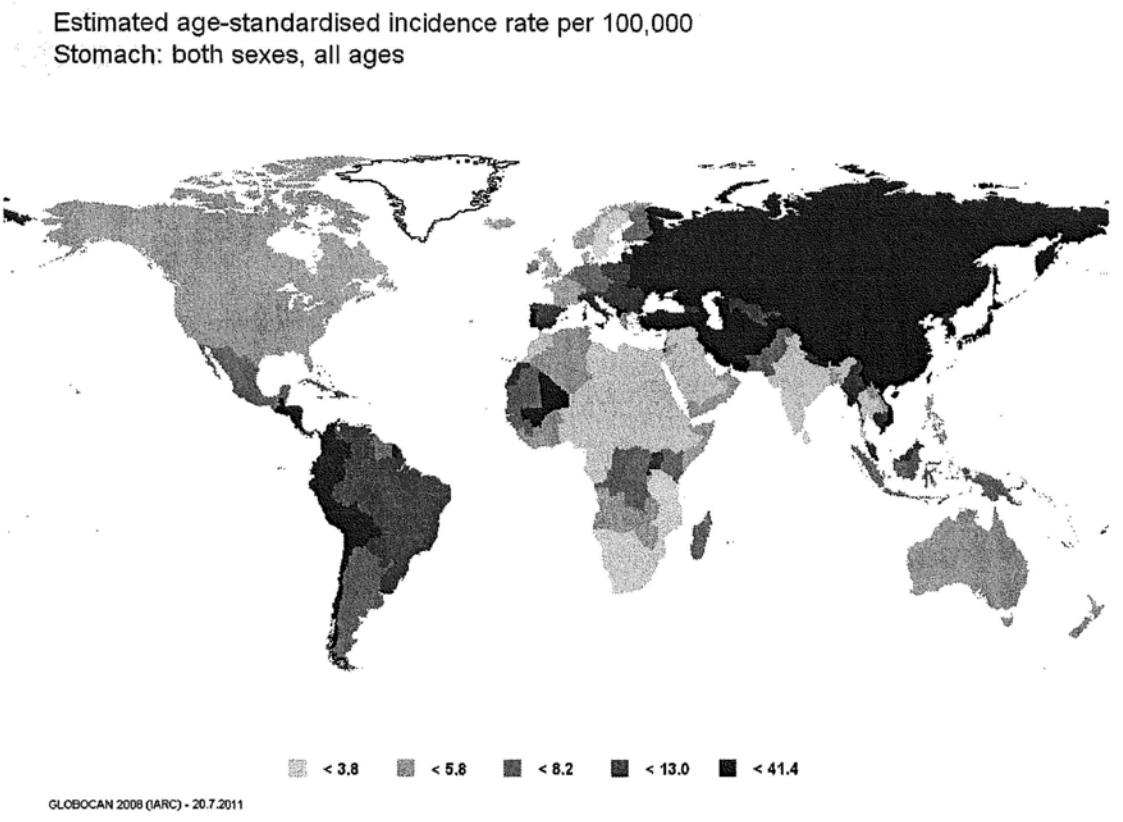


Figure 1.1 Incidence of gastric cancer world-wide with estimated age-standardized incidence rates (Adopted from GLOBOCAN 2008[72])

According to Hong Kong Cancer Registry[73], local data showed similar trend as gastric cancer is one of the leading cancers in Hong Kong. In 2008, gastric cancer ranked 6th in cancer incidence with 1058 new cases of 654 male patients and 404 female patients, and 4th in cause of cancer death with 625 cases of death with 391 male cases and 234 female cases.

The median age of cancer onset in Hong Kong is 70 for male patients and 73 for female patients. Cases of patients under the age of 30 are rare.

1.6.2 Aetiology and Risk Factors of gastric cancer

These are several risk factors for gastric cancer. Genetic and environmental factors, such as family history of gastric cancer, genetic alterations and diet are associated with the development of gastric cancer. Precursor conditions including chronic atrophic gastritis and intestinal metaplasia, pernicious anaemia, and gastric adenomatous polyps are risk factors for gastric cancer.[74]

1.6.2.1 Diet

Overconsumption of salt and nitrate compound and low intake of vegetable are risks factors for gastric cancer. High salt diet may cause gastric irritation as salt can wash out the mucous layer on the epithelium, inducing superficial atrophic gastritis. Previous study had confirmed the association of high salt diet and gastric cancer mortality.[75] The relationship between consumption of smoked or poorly preserved foods and high nitrate and nitrite content with gastric cancer has also been suggested[76]. Bacteria would convert the nitrate and nitrite to N-nitroso compounds, such as nitrosoamines, which are carcinogenic to human by facilitating the progression from chronic atrophic gastritis to gastric cancer.[77]

On the other hand, higher intake of vegetables and fruits shows protective effect against cancer. Various substances in vegetable and fruits, e.g. folate[78] and vitamin C[79], were found to be associated with a decreased risk of gastric cancer.[80-82]

1.6.2.2 Tobacco & alcohol

Smoking is another risk factor for gastric cancer. Tre'daniel *et al.*[83] found a risk of 1.5-1.6 on gastric cancer for smokers when compared to non-smokers, and Sjödahl *et al.*[84] found a risk of 1.88. Consumption of alcohol in addition to smoking can further increase the risk of gastric cancer[84, 85]. Sole consumption of alcohol without smoking also increases the risk of gastric cancer, as shown by Sung *et al.* whom compared non-drinkers with people having a daily intake of 25 gram alcohol in Korea[85].

1.6.2.3 Helicobacter pylori infection

Helicobacter pylori is a gram-negative bacterium that colonize in mucous layer overlying the human gastric epithelial cells. More than half of the adult world-wide are infected with this pathogen. The prevalence in developed country is about 50% but it rises up to 90% in developing countries.[86, 87] In 1982, *H. pylori* was first isolated[88]. There are increasing evidences of a strong association between *H. pylori* infection and gastric cancer, especially for intestinal type gastric carcinoma.

In 1994, the International Agency for Research on Cancer classified *H. pylori* as carcinogenic to humans.[86] *H. pylori* causes chronic active gastritis and are thus associated with atrophy as well as with progressive types of metaplasia.[89] A histological cascade leading to the intestinal type gastric carcinoma was suggested

by Correa *et al.* It is proposed that chronic active gastritis leads to atrophy, followed by intestinal metaplasia, dysplasia/adenoma and finally gastric cancer[90, 91].

Thus patients with chronic gastritis caused by *H. pylori* infection have increased risk of gastric cancer. A recent meta-analysis showed that eradication of *H. pylori* reduced the risk of gastric cancer.[92]

1.6.2.4 Epstein-Barr virus infection

Epstein-Barr virus (EBV) is a gamma-1 herpesvirus commonly found worldwide with a prevalence of over 90% in adults.[93] It is associated with several lymphoid disorders, e.g. Burkitt's lymphoma, etc. and epithelial tumours (nasopharyngeal carcinoma and gastric carcinoma). A recent meta-analysis[94] has showed that 9 % of the 5081 gastric cancer cases studied were EBV positive, which supported the classification of EBV-associated gastric cancer into a distinct aetiologic entity. However, it is still unclear whether EBV infection is a cause or an effect of carcinogenesis.

1.6.2.5 Other factors

In addition to *H. pylori* infection, EBV, dietary and environmental factors, multiple genetic and epigenetic alterations play important roles in both familial and sporadic gastric cancer development. These genetic alternations disrupt the cell cycle control through the activation of oncogenes and/or inactivation of tumour suppressor genes.

Hereditary diffuse gastric cancer (HDGC), which is a subgroup of gastric adenocarcinoma, is caused by a germ line mutation in CDH1 gene. CDH1 gene

encodes for E-cadherin, which is essential for cell-cell adhesion of epithelial cells.[95]

Recent study also suggests that in addition to mutation, methylation of CDH1 promoter can also result in silencing CDH1 through transcriptional down-regulation[96]. Tanaka *et al.* investigated the correlation between expression level of E-cadherin and MUC1, and suggested that an inverse correlation of these two genes could be used as a prognostic marker[97].

1.6.3 Histotypes of gastric cancer

95% of tumour cases in stomach are adenocarcinoma, the remaining cases consist of rare cancers like lymphoma, gastrointestinal stromal tumour and carcinoid tumour[98]. Primary lymphoma, which is a cancer in the immune system, is sometimes found in the lymphatic tissue in stomach. Gastrointestinal stromal tumour is a kind of connective tissue tumour developed from the interstitial cells of Cajal along the digestive track, mostly in stomach. Canceroid tumour is often a malignant type of neuroendocrine tumour that originates from the hormone-secreting cells in various organs, including stomach.

Adenocarcinoma is the most common type of tumours in stomach, as this malignant epithelial tumour develops from the glandular epithelium of the gastric mucosa. It eventually infiltrates into the muscularis mucosae, submucosa and muscularis propria.

There are a variety of classification methods for gastric adenocarcinoma. According to the world health organization, gastric adenocarcinoma is classified to four subtypes, papillary adenocarcinoma, tubular adenocarcinoma, mucinous adenocarcinoma and signet-ring cell carcinoma[99].

Besides that, Lauren's method is another worldwide accepted classification method for gastric adenocarcinoma. According to Lauren's method, gastric adenocarcinoma is divided into two subtypes, the intestinal type and diffuse type [100]. The intestinal type is characterized by gland like tubular structures composed of cohesive neoplastic cells, whereas in diffuse type there are little glandular structures, and characterized by poorly cohesive neoplastic cells diffusely infiltrating the gastric wall. It eventually thickens the stomach wall without forming a discrete mass.

1.6.4 Pathology of gastric adenocarcinoma

Not all gastric adenocarcinoma can be classified into intestinal and diffuse type, and mixed type occurs. However, the histological differences in these two subtypes reflect different epidemiologic and etiologic factors. The widely accepted pathway of gastric carcinogenesis is shown in Figure 1.2

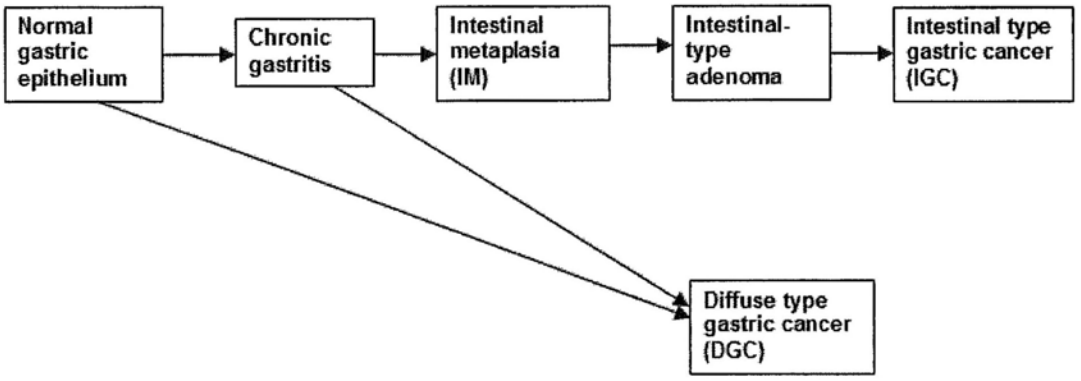


Figure 1.2 Simplified illustration of pathway of the gastric carcinogenesis adopted from Yuasa[101]

1.6.4.1 Intestinal-type gastric adenocarcinoma (IGC)

1.6.4.1.1 Chronic gastritis

As mentioned in above sections, upon inflection by *Helicobacter pylori*, inflammation of the normal gastric mucosa occurs[89]. Prolonged inflammation results in chronic gastritis and eventually transforms the normal gastric epithelium to intestinal-type epithelium, which is known as intestinal metaplasia (IM).

1.6.4.1.2 Intestinal metaplasia

IM can be histopathologically classified into three subgroups depending on the mucin content of the columnar-type and globet cells. In complete IM (type I), the epithelium lost acid secretary function, contained absorptive columnar cells and paneth cells, and showed intestinal phenotype, whereas in the incomplete IM, the paneth cells were absent in the epithelium, and the secretion of acidic sialomucins (type II) or sulphomucins (Type III) was observed [102]. Both complete and incomplete IM may co-occur, but incomplete IM is associated with higher risk of gastric cancer[101, 103, 104].

1.6.4.1.3 Dysplasia and adenoma

Gastric dysplasia refers to the non-invasive epithelial alternation that is a precursor lesion of gastric cancer. It may be present as a raised, circumscribed lesion (adenoma) or as a flat or depressed lesion.[105] Dysplasia can be classified as low-grade dysplasia and high-grade dysplasia. Tosi *et al.* [106] confirmed that type III IM lesions met the nuclear and architectural criteria for low-grade dysplasia through the comparison of the results from qualitative and quantitative

classifications. The result acts as a proof to Correa's cascade[90, 91], in which gastric carcinogenesis is originated from gastritis. Lauwers [107] concluded the importance of gastric dysplasia in prediction of gastric cancer. Further advancement of dysplasia to high grade dysplasia showed increased the risk of developing invasive gastric cancer. Lansdown *et al.*'s study showed that most patients with high-grade dysplasia developed malignancy in stomach within 15 months.[108]

1.6.4.2 Diffuse-type gastric adenocarcinoma (DGC)

1.6.4.2.1 H. pylori infection

High prevalence of *H. pylori* infection was found in resected stomach from DGC patients, indicating that *H. pylori* infection also played a role in the pathogenesis of DGC. However, the association between *H. pylori* and DGC were not through intestinal metaplasia as in the case of IGC. Genetic variation between *H. pylori* strains in DGC and IGC was found, indicating that *H. pylori* may play different roles in the pathogenesis of DGC and IGC.[109]

1.6.4.2.2 Non-metaplastics dysplasia

Ghandur-Mnaymneh *et al.* identified the association between DGC and a rare type of tubule neck dysplasia of non-metaplastics epithelium, which differed from the metaplastic dysplasia commonly found in IGC.[110]

1.6.4.2.3 Genetics alternation

Besides H. pylori infection and non-metaplastic dysplasia, genetic defects play an important role in the development of DGC. As mentioned before, defects in CDH1, which encodes E-cadherin, is one of the major cause of DGC.

Down-regulation, mutation, methylation and chromosomal loss of CDH1 were observed in DGC cases[111-114]. Germline truncation of this gene among members of Maori families in New Zealand was found to be the cause of their hereditary DGC.[115]

As E-cadherin is essential for cell-cell adhesion of epithelial cells[96], loss or mutation of this protein results in the detachment of cells. Such small cluster of loosen malignant cells can proliferate and infiltrate the gastric wall, leading to the thickening of the wall[116].

1.7 Analysis and interpretation of CGH data

1.7.1 Identification of non-random chromosomal aberrations

As mentioned in previous section, chromosomal aberrations play important roles in carcinogenesis. CGH had been adopted for the cytogenetic analysis of various cancers. However, among all the chromosomal gain and loss events captured by CGH, not all of them were related to carcinogenesis. Therefore it is necessary to identify non-random recurrent chromosomal aberration events from the CGH data in order to get a clearer picture of significant aberrations that are solely related to carcinogenesis.

Taetle *et al.* identified non-random breakpoints by comparing the breakpoint distribution in a chromosome region with the expected multinomial distribution pattern, which represents random occurrence of breakpoints at that region. Such breakpoint distribution analysis also included computation of a p-value with protection for multiple testing.[117]

Another well-established method for the identification of non-random chromosomal aberrations involved the use of random simulation[118]. In the null hypothesis, it is expected that all chromosomal aberration events occur randomly. Thus the dataset obtained from random simulation can provide a random distribution of such events as in the situation of the null hypothesis. This method requires that in the null hypothesis the prior probabilities for events to occur randomly are proportional to the size of the involved chromosomal region, and assumes that there are no preferences on chromosomal gain or loss. A score is computed by relating each CGH event to its prior probability. Such scores are calculated on 10,000 sets of simulated data, and the maximum scores that indicate the most frequently occurring events in each set of simulated data are recorded.

Calculations of this score are also carried out on the observed dataset. Once the score of an observed event exceed the 95th percentile from the 10,000 maximum scores, such event is considered as non-random. This method had been adopted in the analysis of CGH data in renal cell carcinoma[119], breast cancer[120], and colorectal cancer[121].

Our team members modified this approach by using a different method to calculate the prior probability. Instead of being a functional proportional to the size of the involved chromosomal region, the prior probability was obtained by pooling 8631 cases from 314 publications of various cancer types (from progenetix.net)[58], such that the effect of type specific tumour-related aberration events was diluted and the prior probability of such events happen solely by chance. [122] Using this modified approach, non-random CGH events in hepatocellular carcinoma were identified.

1.7.2 Information underlying the CGH dataset

In order to retrieve the hidden information in a CGH dataset, statistical methods[123] and machine learning based systems [124] were developed or adopted for the interpretation of CGH data.

Gaussian Mixture Models (GMM) [125] and principal component analysis were used to search for the underlying structure of the CGH data[126-130].

Hierarchical clustering method using Euclidean distances or Pearson correlation as distance metric were shown to be useful in forming clusters of CGH events with implication of the occurrence time points in tumour progression [128, 130] and metastasis[131].

A steady increase in publications on using machine learning methods to predict cancer risk, recurrence and outcome was observed by Cruz and Wishart[132]. Machine learning algorithms, including self-organizing maps (SOM)[133, 134], support vector machines[124, 135], K-Nearest Neighbors[124], decision tree induction[124], feed-forward neural networks[124], SOTA and hybrid method of Self Organizing Maps Artificial Neural Networks (SOMO-ANN) [125], were adopted to the classification of cancer subtypes on the basis of the CGH data.

Self-organizing tree algorithm (SOTA) is an unsupervised neural network which clusters data into a binary tree. SOTA was introduced to biological sciences for the construction of phylogenetic trees basing on protein sequences[136]. It was also adopted for the construction of tree models for oncogenesis basing on CGH data[122]..

1.7.3 Construction of tumour progression tree from non-random chromosomal aberrations

Development of cancer is a prolonged process with the accumulation of the effects of a series of genetic and chromosome changes. Application of oncogenetic tree modelling to the non-random chromosomal gains and losses obtained from CGH analysis would allow elucidation of the sequence of occurrence of chromosomal aberrations during the course of carcinogenesis.

Two approaches have been developed to construct the oncogenetic tree. They are the branching tree method[137] and the distance-based tree method[138]. These two methods have used to construct the carcinogenesis path models for colon cancer[137], hepatocellular carcinoma[122], renal cell carcinoma[119], head and neck squamous cell carcinoma[139], and nasopharyngeal carcinoma[140].

1.7.4 Mapping and visualization of CGH data along with transcriptomic data

There are two types of genome browser, desktop-based and web-based. The Integrated Genome Browser [141] and Integrative Genomics Viewer[142] are examples of the desktop-based genome browsers. Both of them allow the visualization of different types of data (e.g. annotations, cDNA array data, chromosomal aberrations, etc.) in different horizontal tracks.

There are several classic web-based genome browsers, including the UCSC Genome Browser Database[143], Ensembl[144], and Gbrowse[145].

These server-centred applications fetch data from the database and render the whole image containing all the tracks of data. Such image is then sent to the clients via embedment in a static page. Thus this requires bandwidth for the transfer of the whole static page when the client makes changes on the query[146].

JBrowse[147], which is the javascript version of Gbrowse, makes use of the AJAX technology that distribute the rendering work between server and client and thus Jbrowse allows smoother visualization. Another example of a new generation genome browser is OmicBrowse[148], which allows peer-to-peer server communications.

CHAPTER 2

RATIONALE AND OBJECTIVES

Cancer has been a major cause of death worldwide. Possible causes of carcinogenesis are gene mutations, chromosomal aberrations and epigenetic changes.

Chromosomal aberrations refer to numerical and structural variations in the chromosomes caused by radiation, viral infection and carcinogenic chemicals. Chromosomal breakages are critical for the occurrence of structural aberrations and can be detected by chromosome staining, FISH, CGH, aCGH and NGS[50, 56, 59, 68].

Despite the efforts of using all these techniques to obtain the locations and sequences of different types of chromosomal aberrations and chromosome breakpoints, no solid conclusion can be made on the mechanism of chromosomal breakage and rearrangement, and also the association of these changes with carcinogenesis. Moreover, the baseline rate of in vivo chromosomal breakage in tumour cells and its sex differences have yet to be examined and explored.

Among all the cancers, gastric cancer has been one of the most common ones worldwide, and is also one of the leading causes of cancer-related mortality. In 2008, gastric cancer was the 6th commonest cancer and the 4th leading cause of cancer deaths in Hong Kong[73]. Chromosomal aberrations were well observed in both intestinal-type and diffuse-type gastric cancers (IGC and DGC, respectively). However, the recurrent chromosomal imbalances in gastric cancer remained inconclusive. Furthermore, occurrence time points of these observed chromosomal aberrations have not yet been deciphered. Furthermore, the recurrent chromosomal aberrations in gastric adenomas, which are considered as premalignant lesions, have been inconclusive [149, 150].

In the present research study, the role of chromosomal aberrations in carcinogenesis was focused. The present study was divided into two major parts.

Regardless of the cancer types, the general mechanism for occurrence of chromosomal aberrations was investigated in the first part of the study. To achieve this, chromosomal breakpoint data were extracted from publicly available CGH data. Subsequently, their distribution patterns, their associations with genomic structural components and gender status were examined. Detailed bioinformatic characterization of these data would provide novel insights on the role of chromosomal breakage in carcinogenesis. The second part of the study was mainly focused on gastric cancer, in which the possible role of chromosomal aberrations in the carcinogenesis of gastric cancer was examined.

By undertaking bioinformatic approaches, the first part of this research study aimed to achieve the following objectives:

1) To extract chromosomal breakpoint data from public CGH database;
2) To determine the baseline chromosomal breakage in carcinogenesis;
3) To explore the possible mechanism underlying chromosomal breakage in cancer;
4) To investigate the possibility of gender bias in breakpoint frequency.

The objectives of the second part of this research study were:

1) To identify the recurrent chromosomal aberrations in gastric adenoma and gastric cancer by meta-analysis of CGH studies,
2) To develop a tumour progression model from the recurrent chromosomal imbalances;
3) To identify potential cancer driver genes in the recurrent chromosomal imbalances.

CHAPTER 3

BIOINFORMATIC STUDIES ON THE OCCURRENCE OF CHROMOSOMAL BREAKAGES IN CANCER AND ITS ASSOCIATION WITH GENDER STATUS

3.1 Introduction

CGH allows the detection of novel chromosomal gain and loss in tumour cells at a relatively high resolution, regardless of the limitation of this technique in detecting balanced translocation along the chromosomal segments. However, in carcinogenesis, the effect of alternation in gene dosage due to chromosomal duplication and deletion is more prominent than chromosomal translocation. Hence, CGH is an effective technique utilized to understand the association of chromosomal aberrations and carcinogenesis.

Since the development of CGH in 1992, this technique has been applied in the studies of all sorts of malignancies. Large amount of CGH data have been published and integrated into public databases., including the National Center for Biotechnology Information (NCBI) Entrez Cancer Chromosomes site (http://www.ncbi.nlm.nih.gov/cancerchromosomes)[57] and Progenetix[58] (http://www.progenetix.net).

Among several databases of CGH data, the Progenetix web site has collected the largest number of case-specific CGH data, in which 14,322 cases from 139 cancer types were obtained from this database for analysis in this study. From this large cohort of CGH data, it is possible to extract the breakpoint information and determine the baseline chromosomal breakage in carcinogenesis. Although recent technologies like aCGH and NGS can provide data with higher resolution, public data generated from these technologies were not used in this study because of the inadequate number of cancer cases analyzed by these methods.

The bioinformatic algorithms developed in this study were based on 2 key hypotheses. First, because of defective correction systems for spontaneous genomic errors, random chromosomal aberrations can be retained in the cancer cells

regardless of whether they can facilitate cancer growth or not. Second, those random chromosomal aberrations offering growth advantages will be enriched or selected by the cancer cells. Based on our hypotheses, two types of chromosomal aberrations should be observed in a tumour specimen. They are random chromosomal aberrations and cancer-enriched chromosomal aberrations, in which the former was expected to exist in a smaller amount than the latter. For each chromosomal locus in each cancer type, we calculated the number of cases with BPs associated with the gain of chromosomal material and that with the loss of chromosomal material. At a particular locus, the type of BP having the lower number of cases was defined as the minimum BP (min-BP) type, while that having the higher number was defined as the maximum BP (max-BP) type. We speculated that min-BP and maxi-BP data could contain the random BP and cancer enriched BP information, respectively.

Previous studies have showed that examination of the distribution pattern of genomic data could provide insights on the mechanism in the formation of the genomic changes. Frigyesi et al.'s study on the distribution of the number of chromosomal aberrations per tumour in breast, colorectal and renal cell carcinomas showed that the distribution patterns in all 3 cancer types followed power-law distributions with exponents close to unity [151]. It was concluded that the obtained distributions were the consequences of a common mechanism operating in malignant epithelial tumours. Kim et al. found significantly association between the positional distribution of segmental duplications (SDs) and regions of copy number of variations (CNV), and hypothesized that an SD-rich region would generate more CNVs than other regions, some of which, in turn, would become fixed as SDs [152]. Using similar approaches, we attempted to explore the possible mechanism for the

formations of random and cancer-enriched chromosomal aberrations by examining the frequency distribution and positional distribution patterns of the BP data.

Last but not least, according to the statistics from WHO [1], it is observed that both cancer incidence and cancer mortality were high in male than female worldwide. The existence of gender bias in baseline chromosomal breakage was investigated at the end of this part of study.

3.2 Materials and Methods

3.2.1 Extraction of Public CGH data

The CGH data were extracted from Progenetix oncogenomic online resource (http://www.progenetix.net/), which consisted of a collection of 17,023 CGH, aCGH and SKY experiments from 667 publications as of 15/11/2007[58]. Another set of 3066 cases of aCGH data from 67 publications were extracted from Progenetix database for the validation of the distribution patterns of chromosomal breakage.

Cases were classified as benign, in situ, or uncertain whether benign or malignant according to ICD-O-3 code (International Classification of Diseases for Oncology, 3rd Edition)[153] and cancer types with less than 10 cases were excluded from this study. In addition, CGH data from only 14,322 cases from 139 types of cancers obtained were included while in the validation aCGH dataset, only 2,906 cases from 37 types of cancer were included. For gender comparison, the cases with gender-specific cancers (e.g. prostate cancer, ovarian cancer, etc.) were excluded.

3.2.2 Conversion of CGH Data to Breakpoint Data

All these 14,322 cases in the CGH dataset and 2,906 case from the aCGH dataset were categorized according to their cancer types, and the CGH gain/loss data were converted to the chromosome BP and standardized to 400-band resolution using in-house Perl script (Appendix 1, 2). A locus with a copy number state different from its adjacent locus was defined to contain a BP. If a BP was associated with the gain of chromosomal materials, it was classified as a "BP with gain". A BP associated with the loss of chromosomal materials was classified as a "BP with loss". Among the cases of each cancer type, the numbers of "BPs with gain" and "BPs with loss" at each locus were summed. For each locus, these sums were compared. The

higher one was regarded as the maximum BP (max-BP) events, while the lower one was regarded as the minimum BP (min-BP) events.

It is important to note that the definition of max-BP events and min-BP events were different in different cancer types. By using this novel data extraction strategy, the total numbers of min-BP events and the max-BP events of each case in 14,322 CGH cases and 2,906 aCGH cases were obtained. As false positive results at Y-chromosome occurred at high frequency, all data of the Y-chromosome were deleted.

3.2.3 Acquisition and Annotation of Structural Variation, Segmental Duplication Data and Chromosomal Locus Lengths

The structural variations (SV) data were extracted from Database of Genomic Variants (DGV) (http://projects.tcag.ca/variation/), a collection of 31615 SVs from 28 publications as of 10/11/2008, which included 487 inversions (Inv), 11336 insertion-deletion (Indel), and 19792 copy number variations (CNV). The CNV entries were tiled to 6225 non-overlapping copy number variation loci [62, 154].

The segmental duplication (SD) data were obtained from the Segmental Duplication Database (Build 36, University of Washington, http://humanparalogy.gs.washington.edu/)[155] . There SD entries were tiled to 4,407 non-redundant intrachromosomal SD (intraSD) and 5,507 non-redundant interchromosomal SD (interSD).

The chromosomal locus positions of the SVs and SDs were annotated using in-house Perl script based on the cytoband start-stop position data available at UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/cytoBand.txt.gz (Mar. 2006 GenBank freeze assembled by NCBI (hg18, Build 36.1))). Cytoband start-stop

positions were converted to 400-band resolution by in-house Perl script and were compared with the SV position data downloaded from Database of Genomic Variants. Total SV counts and SV lengths per locus were also calculated using in-house Perl script. The lengths of each chromosomal locus were calculated from their start-stop positions on the corresponding chromosomes.

3.2.4 Acquisition of constitutional chromosomal deletion data

Swerdlow *et al.* had extracted information of 2561 cases of constitutional chromosome deletions in England, Wales and Scotland in a study of cancer risk in these patients[156]. The total number of cases of deletion at each arm of each autosome was extracted from this study.

3.2.5 Acquisition of global cancer incidence data

Global cancer incidence data were gathered from the Cancer Incidence in Five Continents, Vol. IX. (http://ci5.iarc.fr/CI5i-ix/ci5i-ix.htm)[157]. The male to female cancer incidence ratios in 1998-2002 were obtained by the ratio of the number of male and female cases worldwide at different sites according to ICD-10 three-digit rubrics.

3.2.6 Statistical Analysis

In this analysis, various statistical tests were applied. The Poisson distribution describes the probability of observing a series of random events at a fixed period of time if the events occur independently at a fixed rate. Chi-square test was used to access the goodness of fit of composite Poisson distribution to the frequency distribution of the min-BP events.

EasyFit Profession (Version 5.1, MathWave Technologies) was used to examine the positional distribution patterns of the min-BP and max-BP. Since no information on distribution bounds were available, all types of bounded, unbounded, non-negative and advanced distributions available in the software were automatically fitted to the datasets. Fittings were performed using the maximum likelihood estimates method with 100 iterations. Kolmogorov-Smirnov test was used to examine the statistical significance of a curve fitting.

Spearman's rank correlation test (SPSS 16.0, SPSS Inc.) was used to examine the association between the chromosomal structural element data, the lengths of chromosomal loci and the BP data. Bonferroni correction was applied for the adjustment for multiple comparisons. Multivariate linear regression analysis was performed (SPSS 16.0) to evaluate the relative importance of these parameters in affecting the occurrence of BPs. The rank of these parameters were subjected to multiple linear regression (forward stepwise) as these parameters were not normally distributed.

Spearman's rank correlation test was used to examine the association between the number of constitutional autosome deletions and the baseline BP and cancer-associated BP events at the centromeric / pericentromeric regions of autosomes. Multivariate linear regression analysis was also performed to evaluate the relative importance of BL-BP and Ca-BP on constitutional autosome deletions.

Paired Student t-test (SPSS 16.0) with Bonferroni correction for multiple comparisons was used to compare the baseline and cancer-associated BP events per tumour between male and female patients.

Spearman's rank correlation test (SPSS 16.0, SPSS Inc.) with Bonferroni correction for multiple comparisons was also used to examine the association between the M/F ratio of the BP events and M/F ratio of the cancer incidence.

3.3 Results

By using this novel data extraction strategy, the total numbers of min-BP events and the max-BP events of each case in 14,322 CGH cases and 2,906 aCGH cases were obtained. As false positive results at Y-chromosome occurred at high frequency, all data of the Y-chromosome were deleted.

3.3.1 Minimum Breakpoint Rates and Maximum Breakpoint Rates in 14,322 Cancer Cases

The analysis of the data extracted from the public CGH databases were based upon two basic assumptions. Firstly, random chromosomal aberration, regardless of their ability to facilitate cancer growth or not, are kept in the genome of cancer cells due to defects in the correction of spontaneous genomic errors. Secondly, those random chromosomal aberrations that bestow growth advantage will be enriched by the cancer cells.

Therefore, based on these two assumptions, two types of chromosomal aberrations can be observed in a tumour specimen, namely random chromosomal aberrations and cancer-enriched chromosomal aberrations, with the latter one present at a higher frequency among cancer cases. It is conjectured that the min-BP and max-BP data could reveal information about random and cancer-enriched chromosomal aberrations respectively.

Min-BP and max-BP data were extracted from the published comparative genomic hybridization (CGH) data from Progenetix database[58], consisted of 14,322 human cancer samples from 139 cancer types, by this novel bioinformatic algorithm under these hypotheses.

Among these 139 cancers, none of them contributed to more than 5% of the 14,322 cases (Table 3.1). The total numbers of min-BP events and max-BP events from all chromosomal loci for each of the 14,322 cases were calculated. On average, there were 1.01 (standard deviation (S.D.) = 2.1) min-BP events per cancer case and 5.11 (S.D. = 6.4) max-BP events per cancer case.

Min-BP and max-BP data were also extracted from the 2,906 cases of array comparative genomic hybridization (aCGH) data similarly. (Data not shown)

Table 3.1 The percentages of the 139 cancer types used for extraction of breakpoint data

Cancer Type	no of cases	%
Acute lymphoblastic leukemia	163	1.14
Acute monoblastic leukemia [FAB M5]	18	0.13
Acute myeloblastic leukemia	43	0.30
Acute myeloid leukemia	253	1.77
Acute myelomonocytic leukemia	34	0.24
Adenocarcinoma, endocervical type	14	0.10
Adenocarcinoma, intestinal type	168	1.17
Adenoid cystic carcinoma	55	0.38
Adrenal cortical carcinoma	27	0.19
Adult T-cell leukemia_lymphoma [HTLV-1 pos.]	82	0.57
Alveolar soft part sarcoma	14	0.10
Ameloblastoma, malignant	29	0.20
Anaplastic large cell lymphoma	89	0.62
Anaplastic oligoastrocytoma	77	0.54
Angioimmunoblastic T-cell lymphoma	61	0.43
Angiosarcoma	10	0.07
Astrocytoma	139	0.97
Basal cell carcinoma, NOS	15	0.10
B-cell chronic lymphocytic leukemia_small lymphocytic lymphoma	574	4.01
Bladder squamous cell carcinoma	41	0.29
Burkitt lymphoma, NOS	83	0.58
Carcinosarcoma, NOS	23	0.16
Central osteosarcoma	10	0.07
Cervix adenocarcinoma	84	0.59
Cervix squamous cell carcinoma	178	1.24
Cholangiocarcinoma	78	0.54
Chondrosarcoma, NOS	50	0.35
Chordoma, NOS	22	0.15
Choriocarcinoma	17	0.12
Choroid plexus papilloma, malignant	15	0.10
Chronic idiopathic myelofibrosis	26	0.18
Chronic myelogenous leukemia	20	0.14
Clear cell sarcoma of kidney	31	0.22
Colorectal adenocarcinoma	451	3.15
Comedocarcinoma, NOS	11	0.08
Desmoplastic medulloblastoma	38	0.27
Diffuse large B-cell lymphoma, NOS	713	4.98
Endocrine adenocarcinoma	29	0.20
Endometrial stromal sarcoma	12	0.08
Endometrioid carcinoma, NOS	52	0.36
Enteropathy type T-cell lymphoma	41	0.29
Ependymoma, anaplastic	29	0.20
Ependymoma, NOS	165	1.15
Epithelioid mesothelioma, malignant	49	0.34
Esophagus adenocarcinoma	59	0.41
Esophagus squamous cell carcinoma	172	1.20

Table 3.1 (Continued)

Ewing sarcoma	97	0.68
Fibrosarcoma, NOS	59	0.41
Fibrous histiocytoma, malignant	163	1.14
Follicular adenocarcinoma, NOS	27	0.19
Follicular carcinoma	20	0.14
Follicular carcinoma, oxyphilic cell	19	0.13
Follicular lymphoma	237	1.65
Gastric adenocarcinoma	344	2.40
Gastrointestinal stromal tumour, malignant	168	1.17
Glioblastoma, NOS	167	1.17
Gliosarcoma	23	0.16
Granulosa cell tumour, malignant	56	0.39
Hairy cell leukemia	36	0.25
Hemangioblastic meningioma, malignant	10	0.07
Hepatoblastoma	137	0.96
Hepatocellular carcinoma, NOS	540	3.77
Head and neck squamous cell carcinoma	354	2.47
Hodgkin lymphoma	96	0.67
Infiltrating duct carcinoma, NOS	642	4.48
Invasive lobular carcinoma	64	0.45
Islet cell carcinoma	18	0.13
Large cell carcinoma, NOS	21	0.15
Large cell medulloblastoma	21	0.15
Large cell neuroendocrine carcinoma	20	0.14
Leiomyosarcoma, NOS	144	1.01
Liposarcoma	136	0.95
Malignant lymphoma	37	0.26
Malignant lymphoma, T-cell, NOS	55	0.38
Malignant melanoma	113	0.79
Malignant myoepithelioma	18	0.13
Malignant peripheral nerve sheath tumour	70	0.49
Mantle cell lymphoma	220	1.54
Marginal zone lymphoma, NOS	53	0.37
Mediastinal large B-cell lymphoma	89	0.62
Medullary carcinoma, NOS	90	0.63
Medulloblastoma, NOS	76	0.53
Meningioma, malignant	60	0.42
Merkel cell carcinoma	48	0.34
Mesothelioma	106	0.74
Mixed germ cell tumour	11	0.08
ML, large B-cell, diffuse, immunoblastic, NOS	12	0.08
Mucinous adenocarcinoma	37	0.26
Mucinous cystadenocarcinoma, NOS	19	0.13
Multiple myeloma	253	1.77
Mycosis fungoides	149	1.04
Nephroblastoma, NOS	215	1.50
Neuroblastoma, NOS	451	3.15

Table 3.1 (Continued)

Neuroendocrine carcinoma	71	0.50
NK T-cell lymphoma	24	0.17
Nasopharyngeal carcinoma	177	1.24
Non-small cell lung carcinoma (adenocarcinoma)	93	0.65
Non-small cell lung carcinoma (squamous)	121	0.84
Oligodendroglioma	95	0.66
Osteosarcoma, NOS	133	0.93
Ovarian adenocarcinoma	91	0.64
Pancreatic adenocarcinoma	203	1.42
Papillary renal cell carcinoma	197	1.38
Papillary serous cystadenocarcinoma	123	0.86
Pheochromocytoma, malignant	14	0.10
Plasma cell leukemia	15	0.10
Plasmacytoma, extramedullary	21	0.15
Pleomorphic xanthoastrocytoma	50	0.35
Polycythemia vera	29	0.20
Precursor B-cell lymphoblastic leukemia	274	1.91
Precursor T-cell lymphoblastic leukemia	104	0.73
Primary effusion lymphoma	11	0.08
Primitive neuroectodermal tumour, NOS	16	0.11
Prolymphocytic leukemia, T-cell type	42	0.29
Prostate adenocarcinoma	628	4.38
Renal cell carcinoma	197	1.38
Retinoblastoma, NOS	135	0.94
Rhabdomyosarcoma	24	0.17
Sarcoma, NOS	25	0.17
Sarcomatoid mesothelioma	51	0.36
Serous adenocarcinoma, NOS	100	0.70
Serous surface papillary carcinoma	11	0.08
Sezary syndrome	251	1.75
Signet ring cell carcinoma	30	0.21
Small cell carcinoma, NOS	70	0.49
Small intestinal adenocarcinoma	30	0.21
Spermatocytic seminoma	12	0.08
Splenic marginal zone lymphoma, NOS	40	0.28
Squamous skin carcinoma	52	0.36
Synovial sarcoma	146	1.02
Teratoma, malignant, NOS	20	0.14
Thymoma	39	0.27
Thyroid adenocarcinoma	13	0.09
Transitional cell carcinoma, NOS	251	1.75
Uterus adenocarcinoma	19	0.13
Vulva squamous cell carcinoma	42	0.29
Yolk sac tumour	11	0.08
Other adenocarcinoma	26	0.18
Other squamous cell carcinoma	30	0.21

3.3.2 Composite Poisson Distribution in the min-BP Frequencies

In order to answer whether the min-BP frequencies reflecting random chromosomal aberrations rather than cancer-enriched chromosomal aberrations, frequency distribution patterns of the number of min-BP events from all 14,322 CGH cases and individual cancer types were examined. It was found that these distribution patterns followed the composite Poisson distribution (R^2 = 0.957, p-value from Chi-square = 1, where $p > 0.1$ indicates statistical significant model fitting) (Figure 3.1).

The composite Poisson distribution model for all the 14,322 cases consisted of four subgroups of different chromosomal breakage rate, where 72.5% of the 14,322 cases had an average number of 0.25 min-BP events per tumour; 24.5% had an average of 2.5 min-BP events per tumour; 2.2% had an average number 7.5 min-BP events per tumour; 0.8% had 15 min-BP events respectively on average.

Similar distribution patterns were observed in various individual cancer types, e.g. hepatocellular carcinomas, which were all fitted into Poisson distribution models without or with 2 to 3 subgroups (Figure 3.2).

In the probability theory, the Poisson distribution is used to describe the frequency distribution pattern of the number of discrete events occurred independently and randomly at a known average rate in a fixed period of time. The well fitting of the frequency distribution pattern of the min-BP events to the composite Poisson distribution means that (i) there are different subgroups of cancer tissues with different occurrence rates of the min-BP, and (ii) in each subgroup, each min-BP events occurred independent and randomly at a known average rate in a cancer tissue.

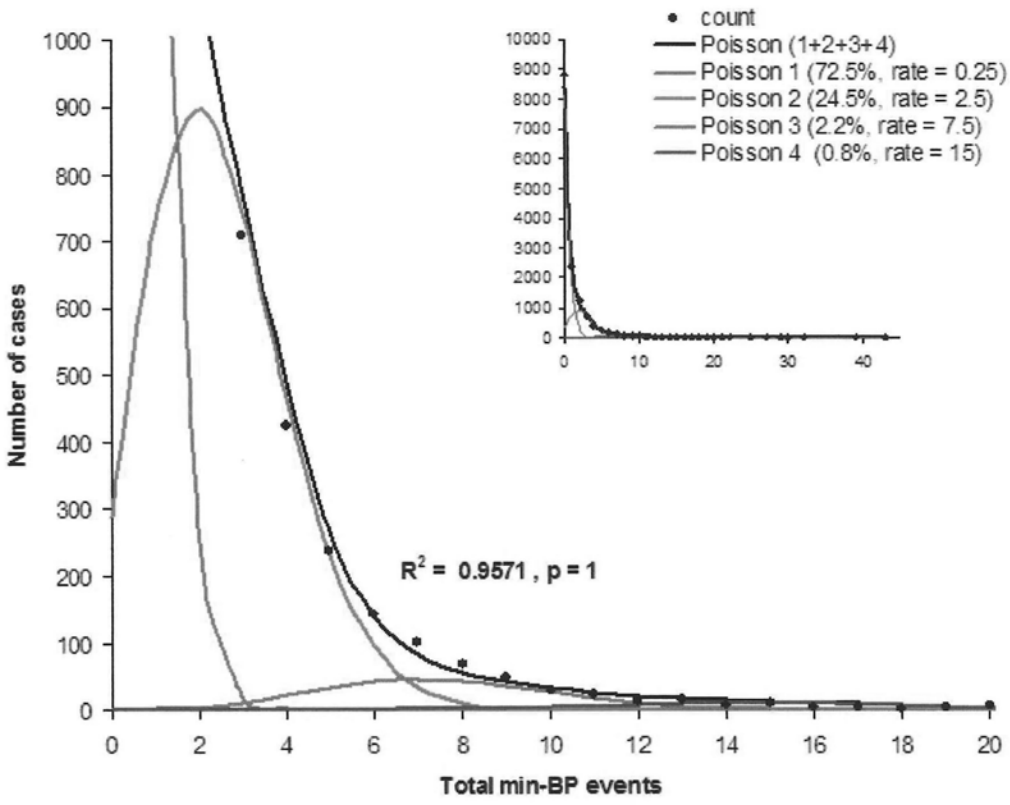


Figure 3.1 The frequency distribution pattern of the number of total min-BP events in all 14,322 cancer cases. The data followed the composite Poisson distribution model (Poisson 1+2+3+4), which was composed of 4 major subgroups (Poisson 1, 2, 3 & 4) with different rates of chromosomal breakage. Their contributions to the composite model and rates of min-BP occurrence (i.e. min-BP events per tumour) were provided in the brackets. The major graph contains only a zoom-in region, while the graph at full scale was provided as the embedment.

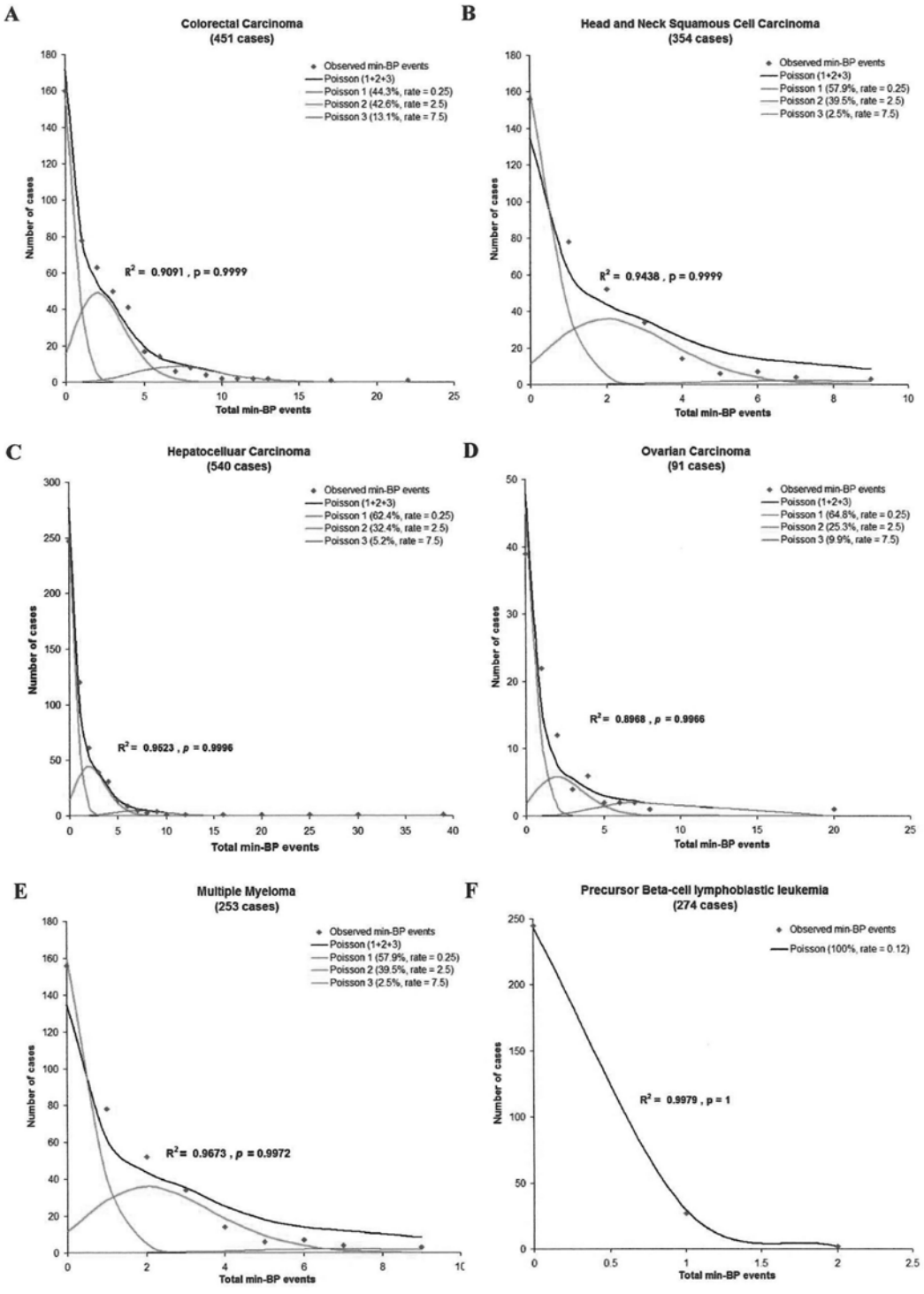


Figure 3.2 The frequency distribution patterns of the number of total min-BP events in 6 representative cancer types including colorectal carcinoma (A), head and neck squamous cell carcinoma (B), hepatocellular carcinoma (C), ovarian carcinoma (D), multiple myeloma (E) and precursor B-cell lymphoblastic leukaemia (F). All data sets followed the composite Poisson distribution model, which was composed of 1 to 3 with different rates of chromosomal breakage. Their contributions to the composite model and rates of min-BP occurrence (i.e. min-BP events per tumour) were provided in the brackets.

3.3.3 The Occurrence of min-BPs, but Not max-BPs, in Human Genome Following the Asymptotic Power Law Distribution

The compatibility of the frequency distribution pattern of the min-BP events among the cancer patients with the composite Poisson distribution suggests that the min-BP events happened randomly, either associated with gain or loss with chromosomal materials.

However, from the view of positional distribution, it is still unclear that the min-BP events randomly happened within the genome, or that certain chromosomal loci are more prone to break. Analyses of positional distribution patterns of the min-BP and max-BP in the human genome would further assist in the distinguishing between random and cancer-enriched BPs. Frigyesi *et al.*'s study on the distribution of the number of chromosomal aberrations per tumour in breast, colorectal and renal cell carcinomas showed that the distribution patterns in all 3 cancer types followed power-law distributions [151].

A power-law distribution suggests a preferential attachment mechanism for the formation of an associated network or system from a set of data [158]. In other words, it is a "rich-get-richer" phenomenon. We hence examined whether the positional distribution patterns of the min-BPs and/or max-BPs followed the power-law distribution.

To examine the positional distribution of the min-BP and max-BP in the human genome, we first calculated the number of the min-BP events and the number of the max-BP events observed in 14,322 cancer cases for individual chromosomal loci at 400-band resolution. Our curve fitting analysis confirmed that the occurrence of the min-BP events among various chromosomal loci fitted the generalized Pareto distribution without significant evidence of the lack of fit when assessed by the

Kolmogorov-Smirnov test (Figure 3.3a, Kolmogorov-Smirnov test, p = 0.338, where p > 0.1 indicates that the model fitting is significantly valid). However, the distribution patterns of the max-BP events did not fit the generalized Pareto distribution (Figure 3.4a, Kolmogorov-Smirnov test: p = 0.052). Moreover, similar curve fitting results were obtained when focusing only on the distribution patterns in the euchromatic loci (p = 0.146 for min-BP, Figure 3.3b; p = 0.070 for max-BP, Figure 3.4b).

The min-BP events from the 2,906 aCGH dataset also significantly fitted to the generalized Pareto distribution (Appendix 3a & 3b, all loci, p = 0.265; euchromatic loci only, p = 0.113), while the max-BP events did not fit the generalized Pareto distribution (Appendix 3c & 3d, all loci, p = 0.074; euchromatic loci only, p = 0.066).

3.3.4 The Maximum Extreme Value Distribution of max-BPs, but min-BPs, in the Human Genome

The maximum extreme value distribution describes the properties of random draws from the maximum tails of distributions[159, 160]. Since the early 1980s, the maximum extreme value distribution has been used to model evolutionary genetics [161, 162]. It has been proposed that mutations occur randomly in a genome, and the beneficial mutations that increase the fitness of the organism in the environment will be retained during evolution. From a mathematical standpoint, beneficial mutations represent extreme draws from the random mutations [163].

In other words, the distribution of fitness follows a maximum extreme value distribution. If the max-BPs were those events enriched during carcinogenesis, it is

expected that they follow the maximum extreme value distribution (Gumbel, Type 1) as in the case of beneficial mutation enrichment during evolution.

When factoring in all chromosomal loci, the occurrence of the max-BP events did not significantly fit the maximum extreme value distribution (Figure 4a, Kolmogorov-Smirnov test, $p = 5.33\times10^{-4}$). However, when focused on euchromatic loci, the occurrence of max-BP events significantly followed the maximum extreme value distribution ($p = 0.805$, Figure 4b). The lack of fit in the former case could be attributed to the presence of other mechanisms controlling the breakages at the pericentromeric regions in the cancers. For example, the presence of hypomethylation in pericentromeric regions in cancers would lead to genomic instability [164-166]. In contrast to the max-BPs, the occurrence of the min-BP events in the human genome did not follow the maximum extreme value distribution (Figure 3.3, all loci, $p = 9.80\times10^{-4}$; euchromatic loci only, $p = 0.012$).

Similar distribution patterns were also observed in the max-BP events from the 2,906 aCGH dataset. The max-BP events in the validation dataset also fitted the maximum extreme value distribution (Appendix 3c & 3d, all loci, $p = 0.222$; euchromatic loci only, $p = 0.611$) while the occurrence of the min-BP events did not follow the maximum extreme value distribution (appendix 3a & 3b, all loci, $p = 0.016$; euchromatic loci only, $p = 0.081$).

To prove that the significant fitting of the max-BP events to the maximum extreme value distribution was not caused by data extraction bias, it was further examined whether min-BP or max-BP events followed the minimum extreme value distribution (Gumbel, Type 1). Both min-BP events (all loci: $p = 3.01\times10^{-5}$; only euchromatic loci: $p = 8.08\times10^{-12}$) and max-BP events (all loci: $p = 1.05\times10^{-17}$; only euchromatic loci: $p = 8.93\times10^{-8}$) did not fit the minimum extreme value distribution.

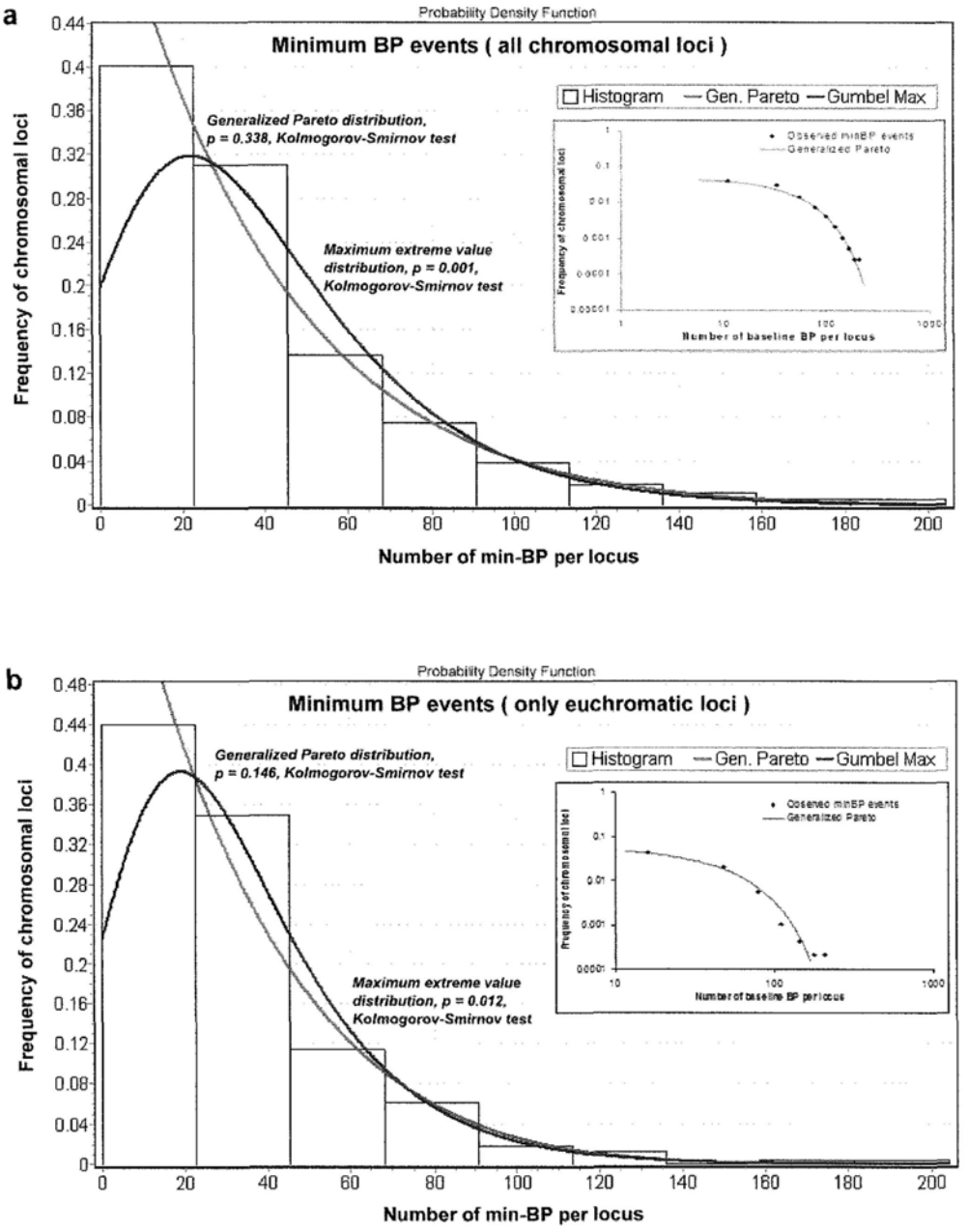


Figure 3.3 Distribution patterns of the minimum breakpoint (min-BP) events observed among all individual chromosomal loci (A) and only those within the euchromatic regions (B). The best fit curves and their associated p-values for goodness of fit were provided. A p-value > 0.1 means that the model fitting is statistically significant. The numbers of min-BP events among all individual chromosomal loci and only those among the loci within the euchromatic regions both fitted the generalized Pareto distribution (Gen. Pareto) without significant evidence of lack of fit when assessed by the Kolmogorov-Smirnov test.

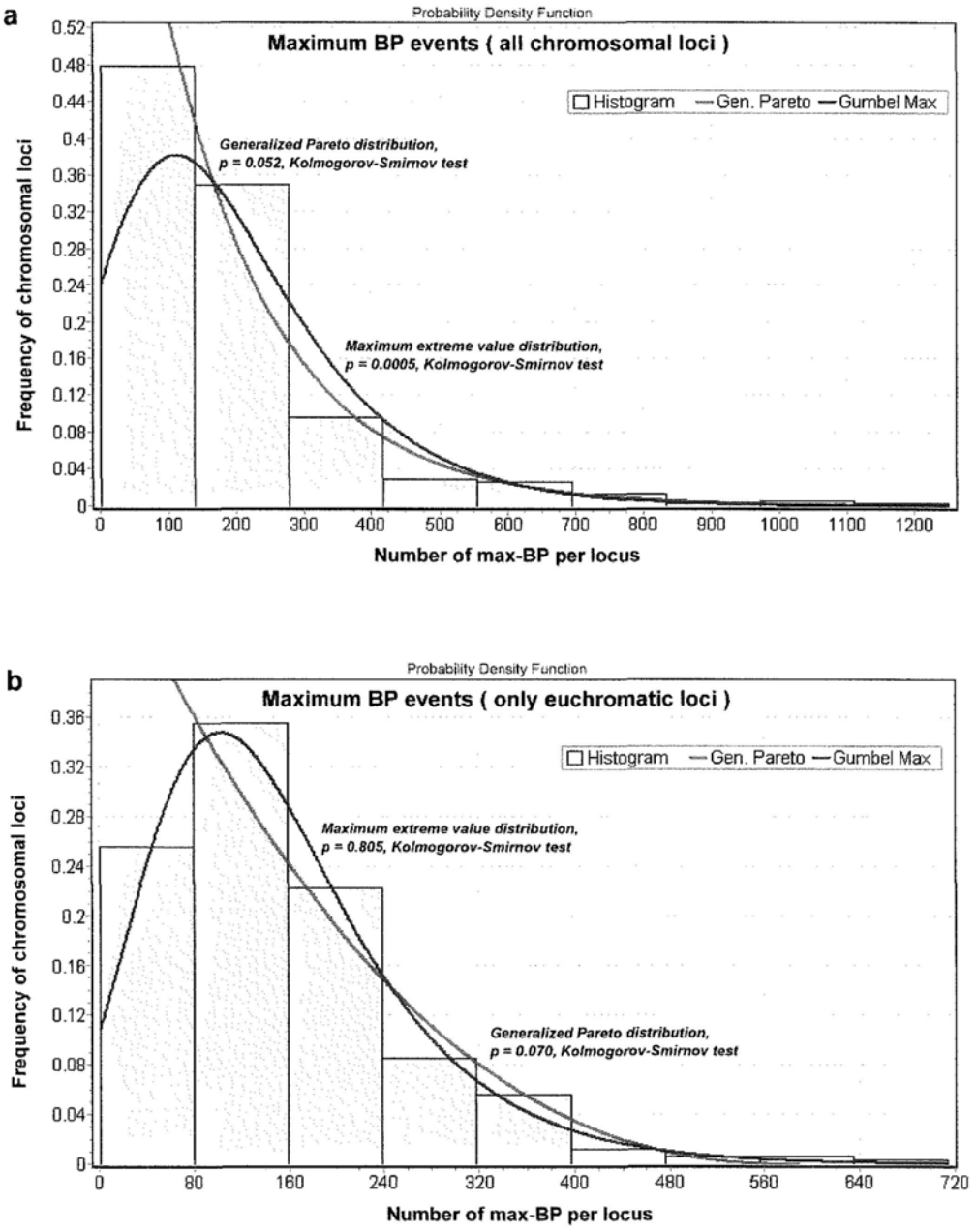


Figure 3.4 Distribution patterns of the maximum breakpoint (max-BP) events observed among all individual chromosomal loci (A) and those only within the euchromatic regions (B). The best fit curves and their associated p-values for goodness of fit were provided. A p-value > 0.1 means that the model fitting is statistically significant. When focusing only on the euchromatic regions (B), the numbers of max-BP events among individual chromosomal loci fitted the maximum extreme value distribution (Gumbel Max) without significant evidence of lack of fit when assessed by the Kolmogorov-Smirnov test.

3.3.5 Strong Association between min-BPs and Copy Number Variation Regions

Identification of structural components in the genome possibly involved in the generation of BPs in cancers and investigation of such relationship would add information to the current understanding of chromosomal BP formation in cancers. It will also facilitate the mathematical modelling of the occurrence of the min-BP and max-BP events. The approach developed by Kim *et al.* [152] regarding the exploration of the role of CNV and SD in chromosomal rearrangement were adopted in this study to investigate the formation of baseline BPs (BL-BPs). The statistically significant correlation between chromosomal structures reflects their functional association and co-localization.

Additionally, Spearman's rank correlation test and multivariate linear regression analysis were performed to explore the pairwise correlation between the frequency of BP events, the length of each chromosomal locus and the frequency of different types of chromosome structural elements (including centromeric or pericentromeric regions, Invs, Indels, copy number variation regions (CNVRs), interSDs and intraSDs) in the 387 chromosomal loci. Similar to the case of the positional distribution analysis, we first calculated the number of the min-BP events and the number of the max-BP events observed in 14,322 cancer cases for individual chromosomal loci at 400-band resolution.

When data from all chromosomal loci were analyzed by Spearman's rank correlation test, the number of min-BP events in each locus was found to be positively associated with centromeric or pericentromeric regions (i.e., p11/p11.1 and q11/q11.1) ($\rho = 0.368$, adjusted $p = 1.70\times10^{-12}$), the number of copy number variation regions (CNVRs) ($\rho = 0.245$, adjusted $p = 2.45\times10^{-5}$), the number and the length of interchromosomal segmental duplications (interSD) (number of interSD: ρ

= 0.216, adjusted p = 4.45×10^{-4}; length of interSD: ρ = 0.160, adjusted p = 0.038 7.34×10^{-4}) and the length of the chromosomal locus (ρ = 0.246, adjusted p = 2.25×10^{-5}).

Similarly, the number of max-BP events was significantly associated with centromeric or pericentromeric regions (i.e., p11/p11.1 and q11/q11.1) (ρ = 0.406, adjusted p = 2.13×10^{-15}), the number of copy number variation regions (CNVRs) (ρ = 0.196, adjusted p = 2.58×10^{-3}), the number of interSD (ρ = 0.159, adjusted p = 0.041) and the length of the chromosomal locus (ρ = 0.203, adjusted p = 1.37×10^{-3}) but not with the length of interSD. (Table 3.2)

Table 3.2 Pairwise correlation among the number of min-BP and max-BP events per locus, centromeric or pericentromeric regions, the number and length of different types of chromosome structural elements (including Invs, Indels, CNVRs and SDs) per locus and the chromosomal locus length by Spearman's rank correlation test. P values were adjusted with Bonferroni correction for the multiple comparisons in all 387 loci.

		Correlations [a]											
Spearman's rho		Centromeric / pericentromeric region	Number of Inv	Total length of Inv	Number of Indel	Total length of Indel	Number of CNVR	Total length of CNVR	Number of interSD	Total length of interSD	Number of intraSD	Total length of intraSD	Chromosomal locus length
min-BP	Correlation Coefficient	0.368**	0.098	0.047	0.09	0.079	0.245**	0.136	0.216**	0.160*	0.037	0.088	0.246**
	Sig. (2-tailed)	1.70×10^{-12}	1.000	1.000	1.000	1.000	2.45×10^{-5}	0.173	4.45×10^{-4}	0.038	1.000	1.000	2.25×10^{-5}
	N	387	387	387	387	387	387	387	387	387	387	387	387
max-BP	Correlation Coefficient	0.406**	0.055	-0.013	0.037	0.026	0.196**	0.087	0.159*	0.13	-0.017	0.035	0.203**
	Sig. (2-tailed)	2.13×10^{-15}	1.000	1.000	1.000	1.000	2.58×10^{-3}	1.000	0.041	0.249	1.000	1.000	1.37×10^{-3}
	N	387	387	387	387	387	387	387	387	387	387	387	387

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).
[a] The data from the Y-chromosome were excluded from the analyses because of its high false positive rate in chromosomal imbalances.

The use of multivariate linear regression analysis was to avoid identifying false significant correlations caused by multiple comparison problem and those caused by associations among the structural elements. The centromeric or pericentromeric regions, ranks of chromosomal structural element data, lengths of chromosomal loci and BP data were subjected to multiple linear regression (forward stepwise), in an effort to examine their independent effects on the chromosomal breakages in the cancers.

When data from all chromosomal loci were analyzed, the min-BP events was positively correlated with the centromeric/pericentromeric regions, the number of CNVR and the number of interSD and negatively correlated with the number of indel. For the max-BP, similar correlation patterns were found, except that the total length of CNVR was found to be an additional positive correlation factor, while the number of interSD was not significantly associated. (Table 3.3)

Table 3.3 Summary of the chromosomal structural elements that were significantly associated with the minimum breakpoint (min-BP) and maximum breakpoint (max-BP) events in the genome.

	Significantly associated structural elements Standardized coefficient, Beta (p-value)[a]				
	Centromeric / pericentromeric region	**Number of CNVR**	**Number of Indel**	**Number of interSD**	**Total length of CNVR**
Min-BP from all loci	0.636 (1.59×10^{-36})	0.884 (4.24×10^{-22})	-0.527 (2.45×10^{-8})	0.233 (1.30×10^{-4})	N.S.[b]
Max-BP from all loci	0.648 (2.17×10^{-37})	0.970 (5.64×10^{-26})	-0.717 (2.33×10^{-10})	N.S.[b]	0.268 (1.38×10^{-3})

[a] Multivariate linear regression analyses (stepwise forward) on the data from 387 chromosomal loci were performed to obtain the standardized coefficients and p-values for the associations between the rank of the number of breakpoint events and the ranks of the assessed parameters, including centromeric/pericentromeric region, number and total length of copy number of variation region (CNVR), number and total length of insert-deletion (Indel), number and total length of inversion (Inv), number and total length of interchromosomal segmental duplication, number and total length of intrachromosomal segmental duplication, and the length of chromosomal locus. The data from the Y-chromosome were excluded from the analyses because of its high false positive rate in chromosomal imbalances.

[b] N.S., not significant.

3.3.6 Stochastic Breakage model for the Formation of Chromosomal Aberrations

On one hand, the min-BPs was observed to happen randomly at the CNVR-rich regions and pericentromeric regions, and thus, could be considered as baseline-BP events. The maximum extreme value distribution, on the other hand, suggested the max-BPs as those BP events enriched in cancers through a process similar to adaptive evolution. Recently, Camps et al. documented a considerable number of cancer-associated BPs co-localized with CNVs in colon cancer [167]. Thus, we proposed a stochastic breakage model to explain the formation of chromosomal aberrations at the CNVR-rich chromosomal regions in tumour cells, in which recurrent events are those enriched through adaptive selection. In the model, we would like to propose two hypotheses: First, chromosomes break randomly in the tumour cells. Second, at the CNVR-rich regions those BPs associated with chromosomal aberrations that increase the fitness of the tumour cells are preferentially retained through clonal selection, while the BPs at the perichromatic regions were enriched through clonal selection or other mechanisms.

3.3.7 Baseline BP rates and Cancer-Associated BP rates in Cancers

Despite the risk of contamination of the min-BP events by a certain amount of rare non-random events benefiting tumour growth, the random distribution pattern of the min-BP events in the human genome strongly did support the hypothesis that the majority of the min-BP events were not cancer-associated. Furthermore, after pooling all the min-BP data of 14,322 tumours from 139 cancer types together, those non-random cancer-associated min-BPs should have been significantly diluted.

The average min-BP rates (i.e., average number of events per tumour) should provide a good estimation of the occurrence of baseline random BP events in human

cancers. If a BP occurs randomly, it can be associated with either the gain or loss of its adjacent chromosomal material. As the min-BPs only reflect BPs associated with either the gain or loss of chromosomal material but not both, doubling the min-BP rates should approximate the baseline random BP (BL-BP) rates in cancer tissues. Once the BL-BP rate is defined, the subtraction of the min-BP rates from the max-BP will provide a reasonable estimation of cancer-associated BP (Ca-BP) rates, i.e.,

$$\text{BL-BP} = 2 \times \text{min-BP} \quad (1)$$

$$\text{Ca-BP} = \text{max-BP} - \text{min-BP} \quad (2)$$

From the CGH data of 14,322 cases from 139 types of cancers, BL-BP and Ca-BP events per chromosomal locus were summed and are summarized in Table 3.4.

Table 3.4 The summary of the percentages of 14,322 cases having baseline breakpoint, cancer enriched breakpoint, chromosomal gain and chromosomal loss in each chromosomal locus.

Chromosomal Locus	Baseline breakpoint [a]		Cancer enriched breakpoint [a]	
	BL-BP, % (rank)		Ca-BP, % (rank)	
1p36.3	0.14	(310)	0.26	(340)
1p36.2	0.11	(328)	0.2	(353)
1p36.1	0.6	(118)	1.35	(90)
1p35	0.56	(130)	0.72	(214)
1p34.3	0.17	(298)	0.53	(271)
1p34.2	0.1	(337)	0.23	(342)
1p34.1	0.28	(245)	0.81	(189)
1p33	0.59	(123)	0.8	(189)
1p32	1.42	(21)	1.51	(76)
1p31	2.28	(3)	2.51	(26)
1p22	0.99	(53)	1.29	(93)
1p21	0.77	(81)	1.5	(77)
1p13	0.74	(85)	1.41	(81)
1p12	0.39	(194)	0.82	(177)
1p11	1.35	(25)	4.4	(8)
1q11	0.78	(78)	8.33	(1)
1q12	0.15	(303)	1.82	(46)
1q21	0.4	(185)	3.84	(11)
1q22	0.17	(298)	1.28	(88)
1q23	0.2	(279)	1.67	(88)
1q24	0.21	(271)	1.02	(54)
1q25	0.18	(284)	1.38	(125)
1q31	0.66	(106)	2.34	(77)
1q32	0.81	(75)	2.08	(28)
1q41	0.46	(160)	1.00	(34)
1q42	0.31	(233)	0.64	(127)
1q43	0.13	(320)	0.54	(210)
1q44	0.08	(341)	0.14	(242)
2p25	0.21	(271)	0.21	(340)
2p24	0.38	(198)	1.10	(325)
2p23	0.49	(153)	1.01	(105)
2p22	0.28	(245)	0.84	(123)
2p21	0.45	(168)	0.86	(157)

Table 3.4 (Continued)				
2p16	0.31	(233)	1.09	(154)
2p15	0.13	(320)	0.54	(106)
2p14	0.18	(284)	0.67	(237)
2p13	0.22	(266)	0.87	(198)
2p12	0.21	(271)	0.54	(148)
2p11.2	0.03	(363)	0.31	(235)
2p11.1	0.43	(175)	1.47	(291)
2q11.1	1.01	(51)	1.15	(72)
2q11.2	0	(371)	0.18	(99)
2q12	0.13	(320)	0.32	(319)
2q13	0.14	(310)	0.29	(288)
2q14.1	0.28	(245)	0.47	(293)
2q14.2	0	(371)	0.05	(256)
2q14.3	0.11	(328)	0.28	(329)
2q21	0.56	(130)	0.86	(296)
2q22	0.8	(77)	1.01	(150)
2q23	0.4	(185)	0.51	(120)
2q24	0.73	(89)	0.75	(241)
2q31	0.52	(143)	0.84	(174)
2q32	0.77	(81)	1.1	(150)
2q33	0.71	(94)	0.82	(103)
2q34	0.45	(168)	0.72	(153)
2q35	0.27	(254)	0.57	(180)
2q36	0.29	(238)	0.7	(216)
2q37	0.21	(271)	0.4	(184)
3p26	0.11	(328)	0.19	(256)
3p25	0.36	(204)	0.5	(302)
3p24	0.52	(143)	0.69	(235)
3p23	0.15	(303)	0.36	(184)
3p22	0.47	(155)	0.65	(259)
3p21	1.28	(29)	1.64	(190)
3p14	0.88	(65)	1.53	(54)
3p13	0.4	(185)	0.9	(65)
3p12	1.05	(44)	1.23	(134)
3p11	0.87	(66)	4.11	(88)
3q11.1	0.64	(108)	4.35	(8)
3q11.2	0.07	(346)	0.42	(7)
3q12	0.22	(266)	0.53	(244)
3q13.1	0.45	(168)	1.2	(221)

Table 3.4 (Continued)				
3q13.2	0.03	(363)	0.27	(87)
3q13.3	0.18	(284)	0.54	(272)
3q21	0.34	(222)	1.54	(215)
3q22	0.18	(284)	0.74	(60)
3q23	0.13	(320)	0.8	(163)
3q24	0.38	(198)	1.51	(60)
3q25	0.28	(245)	1.17	(149)
3q26.1	0.31	(233)	1.49	(62)
3q26.2	0.07	(346)	0.3	(86)
3q26.3	0.21	(271)	1.03	(63)
3q27	0.25	(260)	0.94	(259)
3q28	0.17	(298)	0.53	(102)
3q29	0.06	(355)	0.24	(115)
4p16	0.29	(238)	0.48	(210)
4p15.3	0.43	(175)	0.93	(266)
4p15.2	0.11	(328)	0.29	(222)
4p15.1	0.35	(210)	0.83	(116)
4p14	0.35	(210)	0.74	(254)
4p13	0.18	(284)	0.8	(135)
4p12	0.4	(185)	0.59	(154)
4p11	0.91	(64)	1.47	(142)
4q11	1.79	(12)	2.14	(186)
4q12	0.61	(115)	0.87	(142)
4q13	0.77	(81)	0.96	(64)
4q21	0.74	(85)	0.9	(30)
4q22	0.6	(118)	0.59	(127)
4q23	0.14	(310)	0.43	(109)
4q24	0.25	(260)	0.59	(119)
4q25	0.17	(298)	0.48	(180)
4q26	0.43	(175)	0.59	(214)
4q27	0.24	(264)	0.5	(180)
4q28	0.52	(143)	0.82	(209)
4q31.1	0.32	(228)	0.59	(180)
4q31.2	0.15	(303)	0.2	(200)
4q31.3	0.25	(260)	0.5	(132)
4q32	0.35	(210)	0.85	(179)
4q33	0.18	(284)	0.46	(251)
4q34	0.18	(284)	0.49	(198)
4q35	0.08	(341)	0.2	(128)

Table 3.4 (Continued)				
5p15.3	0.08	(341)	0.23	(204)
5p15.2	0.04	(359)	0.49	(200)
5p15.1	0.42	(183)	0.62	(247)
5p14	0.68	(97)	1.25	(241)
5p13	0.45	(168)	0.91	(200)
5p12	0.28	(245)	1.05	(170)
5p11	0.68	(97)	5.13	(76)
5q11.1	1.45	(19)	2.45	(115)
5q11.2	0.52	(143)	0.87	(93)
5q12	0.47	(155)	0.94	(3)
5q13	0.64	(108)	0.96	(21)
5q14	0.85	(71)	1.35	(120)
5q15	0.4	(185)	0.6	(108)
5q21	0.53	(138)	1.33	(105)
5q22	0.36	(204)	0.61	(65)
5q23	1.28	(29)	1.78	(167)
5q31	0.98	(55)	1.03	(67)
5q32	0.28	(245)	0.61	(162)
5q33	0.43	(175)	0.63	(39)
5q34	0.39	(194)	0.55	(93)
5q35	0.18	(284)	0.27	(160)
6p25	0.13	(320)	0.22	(155)
6p24	0.15	(303)	0.37	(167)
6p23	0.34	(222)	0.5	(216)
6p22	0.73	(89)	1.02	(223)
6p21.3	0.6	(118)	1.35	(194)
6p21.2	0.08	(341)	0.36	(179)
6p21.1	0.5	(148)	1.56	(95)
6p12	0.38	(198)	1.03	(64)
6p11.2	0.1	(337)	0.36	(191)
6p11.1	1.2	(35)	3.1	(54)
6q11	1.82	(10)	3.67	(91)
6q12	1.09	(41)	0.83	(189)
6q13	0.35	(210)	0.89	(14)
6q14	0.56	(130)	0.92	(8)
6q15	0.38	(198)	0.8	(111)
6q16	0.92	(61)	1.28	(104)
6q21	1.05	(44)	1.54	(97)
6q22	1.21	(33)	1.75	(37)

Table 3.4 (Continued)				
6q23	0.87	(66)	1.34	(60)
6q24	0.92	(61)	1.05	(82)
6q25	0.53	(138)	1.04	(85)
6q26	0.2	(279)	0.54	(158)
6q27	0.1	(337)	0.19	(208)
7p22	0.21	(271)	0.36	(179)
7p21	0.64	(108)	1.18	(69)
7p15	0.32	(228)	1.05	(81)
7p14	0.4	(185)	0.63	(140)
7p13	0.2	(279)	0.69	(129)
7p12	0.28	(245)	0.66	(130)
7p11.2	0.22	(266)	0.74	(118)
7p11.1	1.05	(44)	2.37	(21)
7q11.1	1.21	(33)	1.97	(28)
7q11.2	0.59	(123)	1.37	(57)
7q21	0.57	(128)	2.06	(26)
7q22	0.53	(138)	1.55	(48)
7q31	1.24	(32)	1.9	(31)
7q32	0.46	(160)	1.05	(75)
7q33	0.17	(298)	0.34	(172)
7q34	0.14	(310)	0.5	(152)
7q35	0.46	(160)	0.57	(137)
7q36	0.13	(320)	0.35	(166)
8p23	0.2	(279)	0.82	(97)
8p22	0.35	(210)	1.79	(32)
8p21	0.74	(85)	2.11	(25)
8p12	1.47	(18)	1.98	(25)
8p11.2	0.63	(113)	1.08	(68)
8p11.1	1.35	(25)	5.11	(3)
8q11.1	0.87	(66)	6.65	(1)
8q11.2	0.31	(233)	0.92	(78)
8q12	0.32	(228)	1.3	(49)
8q13	0.5	(148)	1.14	(62)
8q21.1	0.34	(222)	1.61	(37)
8q21.2	0.07	(346)	0.31	(162)
8q21.3	0.38	(198)	0.73	(101)
8q22	0.59	(123)	2.19	(21)
8q23	0.75	(84)	2.15	(21)
8q24.1	0.61	(115)	2.26	(20)

Table 3.4 (Continued)				
8q24.2	0.14	(310)	0.61	(118)
8q24.3	0	(371)	0.21	(171)
9p24	0.21	(271)	0.4	(144)
9p23	0.61	(115)	0.89	(77)
9p22	0.27	(254)	0.79	(89)
9p21	1.09	(41)	2.29	(19)
9p13	0.66	(106)	1.08	(59)
9p12	0.4	(185)	0.83	(80)
9p11	1.75	(13)	3.48	(7)
9q11	1.31	(28)	1.96	(19)
9q12	0.59	(123)	0.35	(142)
9q13	0.43	(175)	0.76	(88)
9q21	1.02	(48)	0.96	(66)
9q22	0.87	(66)	1.14	(55)
9q31	0.63	(113)	0.73	(88)
9q32	0.39	(194)	0.51	(123)
9q33	0.78	(78)	0.79	(82)
9q34	0.85	(71)	1.16	(52)
10p15	0.14	(310)	0.35	(135)
10p14	0.27	(254)	0.31	(140)
10p13	0.34	(222)	0.64	(98)
10p12	0.4	(185)	0.82	(75)
10p11.2	0.32	(228)	0.56	(108)
10p11.1	1.2	(35)	1.54	(35)
10q11.1	0.85	(71)	1.51	(36)
10q11.2	0.38	(198)	0.41	(124)
10q21	1.17	(38)	0.71	(87)
10q22	0.98	(55)	0.92	(63)
10q23	0.73	(89)	0.96	(62)
10q24	0.64	(108)	1.01	(58)
10q25	0.52	(143)	1.26	(42)
10q26	0.35	(210)	0.43	(115)
11p15	0.71	(94)	0.87	(66)
11p14	0.98	(55)	0.73	(78)
11p13	0.42	(183)	0.52	(107)
11p12	0.43	(175)	0.47	(110)
11p11.2	0.47	(155)	0.59	(95)
11p11.1	1.09	(41)	1.38	(38)
11q11	1.45	(19)	1.41	(37)

Table 3.4 (Continued)				
11q12	0.49	(153)	1.75	(22)
11q13	1.82	(10)	4.08	(4)
11q14	1.72	(14)	2.44	(15)
11q21	0.68	(97)	0.82	(63)
11q22	1.54	(17)	1.67	(25)
11q23	1.68	(15)	1.94	(17)
11q24	0.5	(148)	0.73	(70)
11q25	0.15	(303)	0.1	(138)
12p13	0.56	(130)	0.91	(53)
12p12	0.95	(60)	1.61	(26)
12p11.2	0.18	(284)	0.57	(85)
12p11.1	1.42	(21)	3	(9)
12q11	0.67	(103)	1.64	(24)
12q12	0.29	(238)	1.17	(38)
12q13	0.45	(168)	1.74	(20)
12q14	0.5	(148)	1.58	(24)
12q15	0.46	(160)	1.68	(21)
12q21	1.26	(31)	1.74	(20)
12q22	0.57	(128)	1.08	(37)
12q23	1.02	(48)	1.21	(32)
12q24.1	0.98	(55)	1.18	(32)
12q24.2	0.06	(355)	0.29	(103)
12q24.3	0.11	(328)	0.17	(115)
13p13	0	(371)	0.02	(131)
13p12	0.01	(368)	0.04	(125)
13p11.2	0	(371)	0.02	(130)
13p11.1	0.01	(368)	0.34	(97)
13q11	2	(7)	4.84	(2)
13q12	0.54	(136)	2.41	(13)
13q13	0.39	(194)	1.24	(27)
13q14	1.1	(39)	3.27	(6)
13q21	2.85	(1)	2.88	(7)
13q22	1.33	(27)	1.52	(19)
13q31	1.83	(9)	1.75	(14)
13q32	1.02	(48)	1.03	(30)
13q33	0.22	(266)	0.63	(62)
13q34	0.06	(355)	0.26	(96)
14p13	0	(371)	0.01	(120)
14p12	0	(371)	0.03	(115)

Table 3.4 (Continued)

14p11.2	0	(371)	0.03	(115)
14p11.1	0	(371)	0.35	(87)
14q11.1	1.97	(8)	2.61	(9)
14q11.2	0.34	(222)	0.64	(58)
14q12	0.29	(238)	0.68	(53)
14q13	0.46	(160)	0.66	(53)
14q21	0.68	(97)	1.24	(22)
14q22	0.36	(204)	0.97	(30)
14q23	0.22	(266)	0.66	(51)
14q24	0.98	(55)	1.29	(19)
14q31	0.74	(85)	0.88	(34)
14q32	0.47	(155)	0.77	(42)
15p13	0	(371)	0.01	(107)
15p12	0	(371)	0.03	(104)
15p11.2	0	(371)	0.01	(106)
15p11.1	0.07	(346)	0.31	(80)
15q11.1	2.22	(4)	2.49	(9)
15q11.2	0.25	(260)	0.55	(56)
15q12	0.08	(341)	0.23	(85)
15q13	0.11	(328)	0.29	(79)
15q14	0.18	(284)	0.55	(56)
15q15	0.59	(123)	0.59	(54)
15q21	0.92	(61)	1.12	(23)
15q22	0.6	(118)	1.06	(23)
15q23	0.4	(185)	0.55	(53)
15q24	0.73	(89)	0.8	(34)
15q25	0.5	(148)	0.64	(46)
15q26	0.24	(264)	0.38	(64)
16p13.3	0.03	(363)	0.18	(79)
16p13.2	0.1	(337)	0.27	(72)
16p13.1	0.45	(168)	0.73	(40)
16p12	0.68	(97)	0.53	(54)
16p11.2	0.35	(210)	0.52	(54)
16p11.1	2.11	(5)	2.68	(8)
16q11.1	1.01	(51)	3.39	(5)
16q11.2	0.04	(359)	0.26	(67)
16q12.1	0.15	(303)	0.87	(29)
16q12.2	0.03	(363)	0.31	(64)
16q13	0.29	(238)	0.65	(41)

Table 3.4 (Continued)

16q21	0.31	(233)	0.8	(31)
16q22	0.43	(175)	0.9	(26)
16q23	0.46	(160)	0.79	(31)
16q24	0.27	(254)	0.42	(52)
17p13	0.53	(138)	0.98	(23)
17p12	0.64	(108)	1.38	(15)
17p11.2	0.54	(136)	1.16	(18)
17p11.1	2.02	(6)	6.17	(1)
17q11.1	0.78	(78)	3.99	(2)
17q11.2	0.27	(254)	0.76	(27)
17q12	0.67	(103)	0.91	(20)
17q21	0.99	(53)	2.76	(4)
17q22	0.34	(222)	1.57	(10)
17q23	0.29	(238)	0.97	(17)
17q24	0.47	(155)	1.24	(13)
17q25	0.21	(271)	0.71	(24)
18p11.3	0.14	(310)	0.29	(47)
18p11.2	0.35	(210)	0.66	(25)
18p11.1	1.38	(24)	1.97	(4)
18q11.1	1.55	(16)	3.55	(2)
18q11.2	0.36	(204)	0.71	(22)
18q12	1.19	(37)	1.65	(6)
18q21	1.1	(39)	1.93	(3)
18q22	0.71	(94)	1.49	(6)
18q23	0.14	(310)	0.26	(40)
19p13.3	0.04	(359)	0.2	(43)
19p13.2	0.18	(284)	0.36	(35)
19p13.1	0.35	(210)	0.64	(19)
19p12	0.07	(346)	0.13	(46)
19p11	0.81	(75)	1.28	(6)
19q11	0.67	(103)	1.85	(3)
19q12	0.14	(310)	0.45	(27)
19q13.1	0.56	(130)	0.62	(17)
19q13.2	0.35	(210)	0.49	(25)
19q13.3	0.18	(284)	0.34	(30)
19q13.4	0.13	(320)	0.16	(36)
20p13	0.18	(284)	0.22	(34)
20p12	0.43	(175)	0.62	(17)
20p11.2	0.35	(210)	0.54	(20)

Table 3.4 (Continued)				
20p11.1	1.05	(44)	1.02	(8)
20q11.1	1.4	(23)	4.47	(1)
20q11.2	0.36	(204)	1.05	(6)
20q12	0.32	(228)	1.15	(5)
20q13.1	0.36	(204)	0.9	(5)
20q13.2	0.15	(303)	0.54	(15)
20q13.3	0.07	(346)	0.23	(25)
21p13	0	(371)	0.12	(30)
21p12	0	(371)	0.04	(32)
21p11.2	0.01	(368)	0.15	(26)
21p11.1	0.03	(363)	0.25	(24)
21q11.1	0.85	(71)	1.75	(2)
21q11.2	0.14	(310)	0.43	(16)
21q21	0.2	(279)	0.78	(8)
21q22	0.53	(138)	0.74	(8)
22p13	0	(371)	0.06	(25)
22p12	0	(371)	0.02	(26)
22p11.2	0	(371)	0.03	(25)
22p11.1	0.06	(355)	0.15	(21)
22q11.1	2.61	(2)	3.2	(1)
22q11.2	0.13	(320)	0.57	(9)
22q12	0.45	(168)	0.84	(4)
22q13	0.73	(89)	0.89	(3)
Xp22.3	0.04	(359)	0.1	(19)
Xp22.2	0.46	(160)	0.15	(17)
Xp22.1	0.28	(245)	0.39	(11)
Xp21	0.56	(130)	0.82	(3)
Xp11.4	0.11	(328)	0.54	(7)
Xp11.3	0.07	(346)	0.13	(14)
Xp11.2	0.11	(328)	0.43	(8)
Xp11.1	0.68	(97)	1.59	(1)
Xq11	0.87	(66)	1.26	(1)
Xq12	0.18	(284)	0.36	(7)
Xq13	0.46	(160)	0.55	(4)
Xq21	0.6	(118)	0.8	(1)
Xq22	0.28	(245)	0.5	(3)
Xq23	0.11	(328)	0.34	(4)
Xq24	0.07	(346)	0.34	(4)
Xq25	0.29	(238)	0.72	(1)

Table 3.4 (Continued)

Xq26	0.35	(210)	0.61	(1)
Xq27	0.27	(254)	0.38	(1)
Xq28	0.07	(346)	0.17	(1)

[a] As false positive results at Y-chromosome occurred at high frequency, all data of the Y-chromosome were deleted.

3.3.8 Strong Association between BL-BPs and Constitutional Chromosome Deletions

Although the BL-BPs were calculated from chromosomal aberrations in cancer patients, it is worthwhile to investigate the association between the BL-BP and constitutional chromosomal aberrations.

Constitutional chromosomal abnormality refers to the chromosomal aberrations in the fertilized egg and thus such abnormality present in all cells of that person. Data of the constitutional deletions of whole chromosome arm in autosomes were extracted from the study by Swerdlow et al.[156] Such data were analyzed with the number of BL-BP and Ca-BP events from the centromeric/pericentromeric regions of autosomes using Spearman's rank correlation test. Both BL-BP (ρ = 0.589, adjusted p =7.730×10-5) and Ca-BP (ρ = 0.376, adjusted p = 0.035) showed significant association with the constitutional autosomal chromosome deletion.

Furthermore, the rank of BL-BP events and Ca-BP events from the centromeric/pericentromeric regions of autosomes and the rank of constitutional autosome deletion data were subjected to multiple linear regression (forward stepwise). Only the rank of BL-BP was found to the only correlation factor with the rank of constitutional autosome deletion data (Beta = 0.589, adjusted p =7.730×10-5).

3.3.9 Male Baseline Chromosomal breakage Rate was about Twice that of Females

In order to probe into the gender differences in BL-BP rates, gender-specific cancers were excluded, resulting in 3134 male cases from 47 cancer types and 1642 female cases from 39 cancer types for analysis. The average BL-BP events per

tumour were 1.49 (S.D. = 2.93) for the male cases and 0.82 (S.D. = 1.95) for the female cases.

There were 34 cancer groups that had BP data in both the male and the female groups. A paired comparison of the male and female values in these 34 cancer groups showed that the number of BL-BP events per tumour was significantly higher in males ($p = 0.022$, paired Student's t-test with Bonferroni correction), with the male-to-female (M/F) ratio of the average BL-BP events per tumour being 1.81.

The M/F ratio of the average BL-BP rate from all chromosomal loci was 1.74. When the centromeric or pericentromeric loci and the euchromatic loci were examined separately, the average M/F ratios were 1.36 and 2.00, respectively.

This gave rise to the hypothesis that, chromosomes, especially euchromatic chromosomes regions, in males are about twice as susceptible to breakage as those in females. For this, the observation that the M/F ratio of mutation rates for insertions and deletions in rodents is also about 2 is consistent with the result of our study and further confirms our hypothesis[168].

3.3.10 Male Cancer-associated Breakage Rate was Close to that of Females

In stark contrast to the previous observations, the average Ca-BP events per tumour were similar between the male and female patients (4.63, S.D. = 5.65 for male; 4.33, S.D. = 5.42 for female). A paired comparison of the male and female values in the 34 cancer types showed a lack of significant difference between the number of Ca-BP events per tumour of males and females ($p = 1.0$, paired Student's t-test with Bonferroni correction). The male-to-female (M/F) ratio of the average Ca-BP events per tumour was 1.06. Similar M/F ratios were obtained when centromeric or pericentromeric loci (1.04) and the euchromatic loci (1.07) were examined separately. Thus, our results suggested that, during carcinogenesis, similar numbers of chromosomal aberrations are acquired in tumours in the male and female patients.

3.3.11 Association between Sex Differences in Chromosomal Stability and Cancer Incidence

According to the IARC Scientific Publication (No. 160) of Cancer Incidence in Five Continents (Volume XI), the average M/F ratio of all cancer incidents in 1998-2002 was 1.5, which indicates a general bias toward males[157]. To examine whether the gender bias in the BL-BPs could be one of the causes of the gender bias in cancer incidence, the M/F ratios of cancer incidence across 29 non-gender specific cancer types were extracted from the IARC Scientific Publication, The pairwise correlation between the M/F ratios of BP events per tumour and the M/F ratios of cancer incidence was then examined by Spearman's rank correlation test (Table 3.5).

Among the 29 non-gender specific cancer types, the M/F ratios of the BL-BP events per tumour showed a strong positive correlation ($\rho = 0.462$, $p = 0.020$, Spearman's rank correlation with Bonferroni correction) with the M/F ratios of cancer incidence. The degree of correlation further peaked ($\rho = 0.542$, $p = 0.015$, after Bonferroni correction) when only euchromatic loci were included in the analysis. However, there was no significant correlation between the M/F ratios of Ca-BP events per tumour and the M/F ratios of cancer incidence ($\rho = 0.214$, $p = 0.264$, after Bonferroni correction). One of the typical examples is hepatocellular carcinoma (HCC), which is preferentially developed in males. The M/F ratios of cancer incidence, BL-BP events per tumour at euchromatic loci and Ca-BP per tumour were 3.02, 5.50 and 0.94, respectively.

Table 3.5 Summary of male-to-female (M/F) ratios of cancer incidence, baseline breakpoint (BL-BP) events per tumor, and cancer-associated breakpoint (Ca-BP) events per tumor in 29 cancer groups.

Cancer Type	M/F ratio of cancer incidence	M/F ratio of BL-BP events per tumor[a]			M/F ratio of Ca-BP events per tumor[a]		
		All loci	Euchromatic loci	Centromeric/ Pericentromeric loci	All loci	Euchromatic loci	Centromeric/ Pericentromeric loci
Hepatocellular carcinoma	3.02	1.98	5.5	0.81	1.02	0.94	1.08
Transitional cell carcinoma	2.96	4.26	3.26	17.27	1.07	0.92	1.48
Esophagus squamous cell carcinoma	2.86	2.42	7.48	0.73	1.13	1.55	0.63
Head and neck squamous cell carcinoma	2.71	2.18	7.12	1.19	0.83	0.84	0.81
Non-small cell lung carcinoma (squamous)	2.62	6.76	13.55	3.36	1.08	0.99	1.23
Renal cell carcinoma & Papillary renal cell carcinoma	1.68	1.59	2.13	1.09	1.29	1.28	1.33
Desmoplastic medulloblastoma & Medulloblastoma	1.68	N.A.[b]	N.A.	N.A.	0.9	1.22	0.67
Gastric adenocarcinoma	1.62	2.35	4.05	1.83	1.52	1.74	1.06
Hepatoblastoma	1.49	2.86	2.57	3.43	1.77	1.73	1.85
T-cell and NK-cell lymphoma	1.44	1.06	1	1.16	0.76	0.72	0.89
Small cell carcinoma, & Neuroendocrine carcinoma	1.33	0.96	0.96	0.96	0.87	0.89	0.82
Astrocytoma, Glioblastoma, Pleomorphic xanthoastrocytoma	1.33	0.89	0.39	1.09	1.21	1.08	1.5
Endocrine adenocarcinoma	1.32	0.54	0.36	1.07	0.73	0.79	0.64
Osteosarcoma	1.26	0.91	0.75	1.18	0.9	0.85	1.04
Anaplastic oligoastrocytoma & Oligodendroglioma	1.24	0.66	N.A.	0.57	0.92	0.88	0.95
Gastrointestinal stromal tumor, malignant	1.23	1.07	N.A.	1.07	0.92	0.5	0.99
Hodgkin's lymphoma	1.23	5.6	5.42	N.A.	1.01	0.97	1.44
Acute myeloid leukemia	1.21	N.A.	N.A.	N.A.	1.64	1.74	1.44
Leiomyosarcoma	1.19	0.79	0	1.19	0.9	0.24	1.66
Non-small cell lung carcinoma (adenocarcinoma)	1.19	0.86	0.74	1.66	0.81	0.77	0.89
B-cell lymphoma	1.19	2.33	2.83	1.42	0.89	0.91	0.82
Cholangiocarcinoma	1.17	1.21	0.88	3.18	0.77	0.78	0.76
Malignant melanoma	1.16	0.42	0.6	0.29	0.92	0.93	0.89
Ependymoma	1.14	N.A.	N.A.	N.A.	1.46	1.59	1.42
Small intestinal adenocarcinoma	1.13	0.5	1	0.25	0.48	0.48	0.48
Liposarcoma	1.13	1.22	0.76	N.A.	0.78	0.81	0.6
Colorectal adenocarcinoma	1.06	1.32	1.46	0.94	0.81	0.82	0.76
Pancreatic adenocarcinoma	0.99	1.94	2.42	1.43	1.19	1.56	0.89
Nephroblastoma	0.97	N.A.	N.A.	N.A.	1.23	1.12	1.3
Spearman's rank correlation with M/F ratios of cancer incidence, ρ (1 tailed p-value after Bonferroni correction)	--	*0.462 (0.020)*	*0.542 (0.015)*	*0.198 (0.732)*	*0.214 (0.264)*	*0.217 (0.515)*	*0.194 (0.626)*

[a] As false positive results on the Y-chromosome occurred at high frequencies, all data for the Y-chromosome were excluded from analysis. [b] N.A., not applicable.

3.4 Discussion

3.4.1 The Mechanisms Underlying Chromosomal Breakage and the Production of Recurrent Chromosomal Aberrations in Cancer

Frigyesi *et al.*'s study on the distribution of the number of chromosomal aberrations per tumour showed that the distribution patterns of breast, colorectal and renal cell carcinomas followed power-law distributions with exponents close to unity. It led to arrive at the conclusion that the obtained distributions were the consequences of a common mechanism operating in malignant epithelial tumours[151].

As chromosomal aberrations conferring growth advantages were selectively retained and enriched in cancers, it is important to differentiate baseline chromosomal changes from those chromosomal changes enriched during cancer development and to examine their distribution patterns separately, which were achieved by the analysis of BPs instead of chromosomal aberrations in this study.

The min-BP and max-BP events in the human genome contained the random baseline breakage and non-random cancer-associated breakage information, respectively, and followed the different distribution patterns in the cancer cell genome. The consistence of distribution patterns on min-BP and max-BP events in both 14,322 cases of CGH dataset and 2,906 case of aCGH dataset showed a strong proof that the random baseline BP events and the cancer associated BP events followed an asymptotic power law distribution and a maximum extreme value distribution, respectively. The significant fitting of the max-BP events to the maximum extreme value distribution was not caused by data extraction bias as both min-BP or max-BP events did not followed the minimum extreme value distribution.

Such distribution patterns suggested a stochastic breakage model for the formation of chromosomal aberrations in tumour cells, in which recurrent events were enriched through adaptive selection.

Furthermore, CNVRs, interSDs, Indels and centromeres/pericentromeres were identified as the key structural elements associated with baseline random chromosomal breakage. The co-occurrence of CNVRs and interSDs had been reported in previous studies[152, 169], and in this study, significant correlation between CNVRs and interSDs had also been identified (Spearman's rank correlation test, $\rho = 0.631$, adjusted $p = 5.22\times10\text{-}43$). Therefore, it is not surprised that interSD was also identified as one of the key elements associated with the baseline random chromosomal breakage.

In this study, the chromosomal locus length was found to be significantly correlated with the number of min-BP and max-BP per locus by Spearman's rank correlation test, which was due to the correlation between the CNVRs and the chromosomal locus length ($\rho = 0.888$, adjusted $p = 1.11\times10\text{-}130$). When the data were subjected to multivariate linear regression analysis, the chromosomal locus length was excluded from the model. This indicated that while looking at the chromosomal locus length as an independent factor, it only showed a non-significant effect on the chromosomal breakages in the cancers.

Therefore, CNVR rich regions and centromeres/pericentromeres are more prone to break, whereas indel rich regions are more resistant to break. Proposed mechanisms for chromosomal breakage and the production of recurrent chromosomal aberrations were summarized in Figure 3.5.

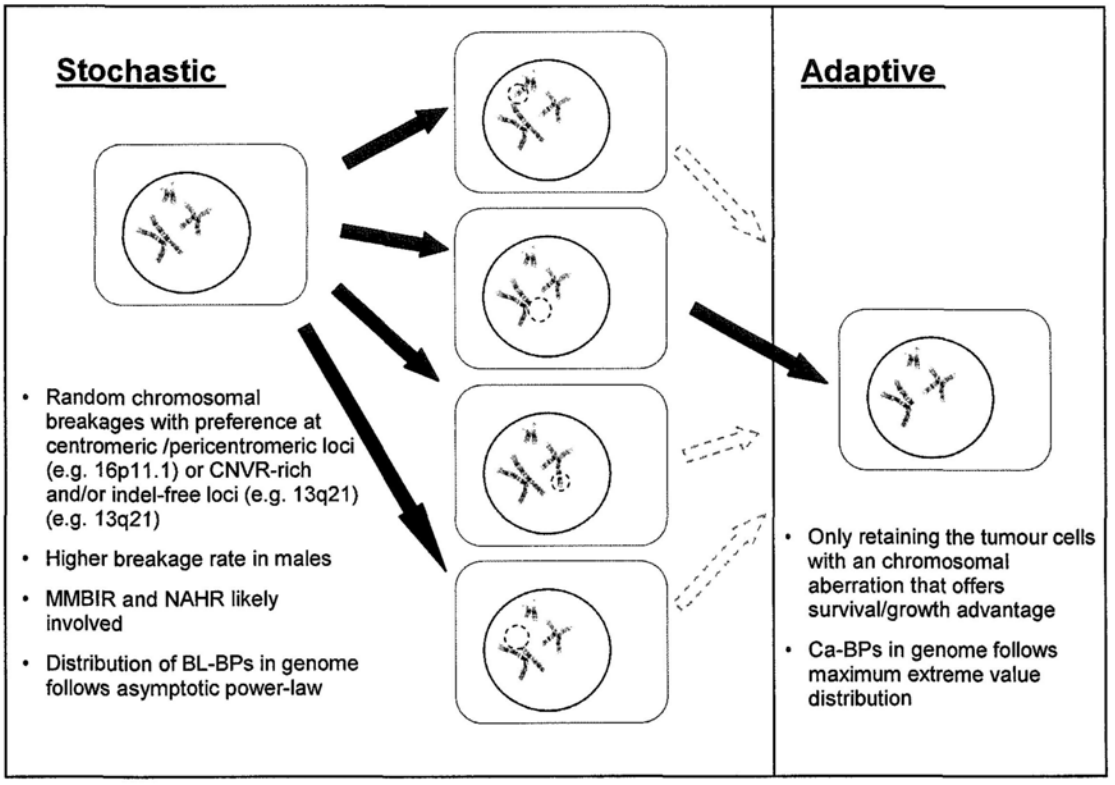


Figure 3.5 A stochastic breakage model for the formation of chromosomal aberrations in tumour cells.

When a dataset follows a power-law distribution, it suggests a preferential attachment mechanism for the formation of its associated network or system[158]. In other words, it is a "rich-get-richer" phenomenon. Besides the min-BP events, we also found that distribution of CNVRs in the human genome followed the asymptotic power law distribution (data not shown). For the CNVRs, the asymptotic power law distribution may suggest a preferential attachment mechanism for distribution of chromosomal breakage sites in the human genome. This also implies that there is an imbalance in the distribution of breakage sites in the genome, leading to some chromosomal loci being more prone to breakage.

It is worth noting that segmental duplications (SDs) also follow a power-law distribution in the genome[152]. This inferred that SD formation preferentially occurs in regions with many previously existing SDs. Kim et al. hypothesized that an SD-rich region would generate more CNVs than other regions, some of which, in turn, would become fixed as SDs[152]. In this study, it would imply the involvement of specific chromosomal element(s) in chromosomal breakages, and suggested that the breakage sites preferentially occur in regions with many previously existing breakage sites were very likely to be related to CNVRs.

The distribution pattern of protein structural domains also followed the asymptotic power law[170]. One of the possible explanations for the emergence of the asymptotic power law and the related skewed distribution of the domains and family sizes in the protein universe is an evolutionary BDIM[170, 171]. In the BDIM for protein structural domains, birth and death refer to the duplication and deletion of an existing protein domain, respectively, while innovation refers to the spontaneous creation or "invention" of a new protein domain[170, 171]. This BDIM suggests that genome evolution might largely be a stochastic process[170, 171]. Such a BDIM

might also explain the asymptotic power law distribution of the baseline BPs. During genome evolution, a stochastic process in which existing breakage sites, such as CNVRs, were duplicated, deleted or invented might have governed the formation of breakage sites in the chromosomes.

The predominant association of CNVRs with the baseline BPs at euchromatic loci is consistent with the observation that a considerable number of cancer-associated BPs co-localize with CNVs in colon cancer[167]. Recently, a model of microhomology-mediated break-induced replication (MMBIR) has been proposed for the origin of CNVs in order to explain the observation that de novo CNVs arise by rearrangements at sites that lack extensive homology[172]. Similar to the case of CNVs, many junctions of translocation endpoints have microhomology in leukemia cells[173, 174]. MMBIR, in fact, can be classified as a specific type of BDIM. The results of the current study strongly support that MMBIR at CNVRs is one of the possible important mechanisms controlling chromosomal stability in both normal cells and cancers, leading to the asymptotic power-law distribution of baseline BPs.

Although the BL-BPs were calculated from chromosomal aberrations in cancer patients, it is possible to apply the basebline BP data to non-tumour related chromosomal aberrations. The strong association between the BL-BPs and the constitutional autosome deletions suggested that the BL-BPs identified were reflecting the random baseline chromosomal breakage.

3.4.2 Baseline Chromosomal Stability and Gender Bias in Tumour Development

In the present study, BL-BPs and Ca-BPs were successfully extracted and characterized from CGH data of a total of 14,322 cases. To the best of my

knowledge, this is the first study to report in vivo baseline chromosomal breakage data and the presence of a gender bias in the baseline chromosomal breakage rate in human cancers. The male-to-female (M/F) ratio of the average BL-BP rate from all chromosomal loci was 1.74. This is in line with the observation that the M/F mutation rate ratio for insertions and deletions in rodents is also about 2[168].

The presence of gender bias in the chromosomal breakage baselines may have already existed in specific normal organs in males. However, it is also possible this gender difference in chromosomal stability is acquired during carcinogenesis. Interestingly, previous studies on the telomere length support the former possibility. Large amount of in vivo and in vitro studies showed that shortening of telomere length induced chromosome instability, and enhanced tumor initiation[175]. A recent 10-years follow-up study of 787 participants has confirmed that shorter leukocyte telomere length is associated with higher risk of cancer development [176].

In consistence to our result, gender bias is also present in the telomere length. The telomere length decreases with age. At birth, sex difference in length is not obvious. The shortening rate in males is higher than that in females, leading to shorter telomeres in males in general[177, 178]. The shorter telomeres in males may be the underlying cause of the higher baseline chromosome breakage rates in the tumor tissues from male cancer patients.

In general, there are more males than females suffering from cancer[157]. The estrogen-related gender difference in the regulation of inflammation is one of the possible mechanisms leading to the higher susceptibility of hepatocellular carcinoma in males[179]. This mechanism may also extend to other cancers, such as colorectal cancer[180].

3.5 Conclusion

Rates of baseline breakpoint in tumour and cancer-associated breakpoint were defined based on the analysis of the breakpoint data extracted from public CGH database. CNVRs and centromeres/pericentromeres were identified as the key structural elements associated with baseline random chromosomal breakage. This strongly supported MMBIR at CNVRs as one of the possible important mechanisms controlling chromosomal stability in both normal cells and cancers, leading to the asymptotic power-law distribution of baseline BPs.

On the other hand, the maximum extreme value distribution of the max-BPs suggested that the BPs benefiting tumour survival and/or growth were enriched in cancers through a process similar to adaptive evolution.

Furthermore, comparison of the baseline breakpoint frequency in male and female revealed that the higher baseline rate of chromosomal breakage in males is one of the other possible causes contributing to the increased occurrence of cancers in males. It is important to note that there was no gender bias in Ca-BP events, suggesting that the gender bias in the baseline chromosomal breakage happens at the stage of tumour initiation, not tumour progression.

CHAPTER 4

BIOINFORMATIC ANALYSIS OF CYTOGENETIC DATA AND IDENTIFICATION OF POTENTIAL CANCER DRIVER GENES IN GASTRIC CANCER

4.1 Introduction

Gastric cancer is one of the most common cancers worldwide, and is also one of the leading causes of cancer-related mortality. In 2008, it ranked fourth in the cancer incidence and mortality worldwide [71],and it was the 6th commonest cancer and the 4th leading cause of cancer deaths in Hong Kong [73]. The 5-year survival rate of patients with advanced gastric cancer is low (10-30%)]. Chemotherapy is the main treatment option for these patients, but with limited efficacy. New anticancer drugs are heavily demanded[181]. To develop effective anti-cancer drugs, it is important to identify genes that are directly driving the neoplastic process (i.e., cancer driver genes), especially those with abnormal activation at the early stage of cancer development.

Copy number gain and loss is one of the common mechanisms for activating a cancer driver gene and for inactivating a tumour suppressor gene, respectively. Comparative genomic hybridization (CGH) is an important tool for the profiling of chromosomal imbalances. Though previous studies identified chromosomal imbalances in gastric cancer, it is difficult to draw a solid conclusion on the key recurrent events from these studies [182]. This should have been caused by the limitations present in each study, such as inadequate number of samples [183]. Similarly, the recurrent chromosomal aberrations in gastric adenomas, which are considered as premalignant lesions, have been inconclusive [149, 150].

Specific chromosomal aberrations were shown to be associated with histological features and metastasis in gastric cancer [184]. However, occurrence time points of these aberrations were unknown. Moreover, earlier a gene being activated during cancer development, more likely it plays a pivotal role in the neoplastic process and acts as a cancer driver gene. Therefore, there is a need to

decipher the occurrence time points for individual the recurrent chromosomal aberrations in gastric cancer in order to understand their roles in oncogenesis.

Furthermore, the results from the first part of the present research project supports that there should be 2 types of chromosomal imbalances, i.e., random and non-random events. The non-random events are those retained by a tumour for offering growth advantage. However, for gastric cancer, no attempts were made to differentiate the non-random CGH events from the random events. Identification of significant events in adenoma and carcinoma through statistical means may help to explain how cancer develops in stomach. Recently our team developed novel statistics approaches for identification of non-random CGH events and for deciphering the time sequence for the occurrence of chromosomal imbalances during the oncogenesis of hepatocellular carcinoma [122]. Gains of 1q21-23 and 8q22-24 were identified as genomic events associated with the early development of HCC; and gain of 3q22-24 as a late genomic event significantly associated with tumour recurrence and poor overall patient survival.

In this study, we attempted to improve our previous statistics approach for identification of non-random CGH events by incorporating a function for controlling the false discovery rate (FDR). We subsequently use it to identify recurrent chromosomal imbalances in gastric adenoma and carcinoma through meta-analysis of CGH data previously published by various research teams. The increase in the number of cases analyzed and implementation of FDR estimation would greatly improve the accuracy of the identification of statistically significant chromosomal imbalances. The identified recurrent chromosomal imbalances were subsequently used to construct tumour progression models.

Expression levels of many of the genes with copy number change do not seem to be altered. Gene expression microarray, which profiles genome-wide transcription pattern in a tissue, is another important technology in cancer research. It has successfully identified gene expression signatures for molecular classification, prognosis as well as for therapeutic drug development[185].

However, gene expression profile is highly dynamic. Microarray-based gene expression profile only provides a snapshot of complex genetic interactions. It is difficult to differentiate the cancer driver genes from the down-stream effector genes.

We hypothesize that (i) chromosomal regions with copy number gain in gastric adenoma and in early gastric cancer contain cancer driver genes playing pivotal roles in the neoplastic process; and (ii) chromosomal regions with copy number gain in early gastric cancer, but not gastric adenoma, contain cancer driver genes that transform benign tumour cells into malignant tumour cells and/or promote cancer progression. In the last part of the present study, we attempted to identify potential cancer driver genes through mapping and visualization of the recurrent chromosomal imbalances and differential transcriptomic data, and then verified them by a series of gene expression and functional experiments.

4.2 Materials and Methods

4.2.1 Acquisition of public CGH data

CGH data of gastric adenoma cases and gastric adenocarcinoma cases was extracted from Progenetix oncogenomic online resource [186] and from publications obtained through literature search.

After discarding cases of low resolution (i.e. data showing only gain or loss of the whole chromosome arm), 41 gastric adenoma (GA) cases, 198 cases of intestinal type gastric cancer (IGC) cases and 117 cases of diffuse type gastric cancer (DGC) cases were included in this study. The number of research studies involved and the corresponding publications are summarized in Table 4.1.

Table 4.1 List of gastric adenoma and carcinoma cases from various publications

	No. of cases	No. of Studies	Investigaors
Gastric Adenoma (GA)	41	4	Buffart *et al.*[150]; Kokkola *et al.*[149]; van Dekken *et al.*[187]; Weiss *et al.*[188]
Intestinal type gastric cancer (IGC)	198	11	Chan *et al.*[189]; van Grieken *et al.*[183]; Kokkola *et al.*[190]; Larramendy *et al.*[191]; Noguchi *et al.*[192]; Oga *et al.*[193]; El-Rifai *et al.*[194]; Sud *et al.*[195]; Tay *et al.*[196]; Varis *et al.*[197]; Wu *et al.*[184];
Diffuse type gastric cancer (DGC)	117	7	Chan *et al.*[189]; Kokkola *et al.*[190]; Larramendy *et al.*[191]; Peng *et al.*[198]; El-Rifai *et al.*[194]; Sud *et al.*[195]; Wu *et al.*[184]

A Perl script was developed for reformatting of the public CGH data. DNA copy number gain and loss at the same chromosomal region were treated as independent CGH events. Data on the chromosome Y had been excluded due to high false-positive rate.

4.2.2 Implementation of an in-house statistics tool for identification of non-random CGH events with consideration of false discovery rate

As mentioned, CGH has been a useful tool for identification of the chromosomal aberrations in cancer cases. However, not all chromosomal aberrations captured were tumour-related. A statistical tool had been developed by our team members[122] for the identification of non-random events in the CGH data.

The prior probability was obtained by pooling 8631 cases from 314 publications of various cancer types (from progenetix.net)[58], such that the effect of type specific tumour-related aberration events was diluted and the prior probability of such events happen solely by chance.

Random simulation datasets were produced basing on the prior probability and processed as the observed dataset. A chromosomal aberration event was identified as non-random when the observed frequency obtained from bootstrapping was significantly higher than the corresponding frequency from random simulations.

In this study, modification on this existing tool was carried out to implement a statistical function for the estimation of false discovery rate. Those CGH events found significant from the random simulations were considered as false positive events and thus the false discovery rate (FDR) could be calculated. The tool was implemented with the Matlab 6.0 (The MathWorks, Inc.) (Scripts shown in Appendix 4).

4.2.3 Identification of non-random CGH events in GA, IGC and DGC

CGH data from 41 GA cases, 198 IGC cases and 117 DGC cases were analyzed separately by using our in-house statistics tool for identification of non-random events. In each group CGH events were considered as statistically significant (i.e. non-random) when a p-value was < 0.05 and the average FDR was 0%.

4.2.4 Development of tumour progression model

Self-organizing tree algorithm (SOTA) is an unsupervised neural network which clusters data into a binary tree. SOTA had been introduced to biological sciences for the construction of phylogenetic trees of organisms basing on protein sequences[136]. It had also been adopted for the construction of tree models for oncogenesis basing on chromosomal aberrations[122].

In the present research study, SOTA was applied to analysing GA, IGC and DGC cases separately. Minimal overlapping regions of the non-random CGH events were subjected to SOTA (MultiExperiment Viewer (MeV v4.0), part of the TM4 Microarray Software Suite[199, 200] to construct oncogenetic trees. Default parameters were used in the application of SOTA as previously described by our team members [122]. Cases were clustered into several groupings. Subsequently, CGH data of each SOTA group were subjected to our in-house statistics tool again for identification of non-random events. In each SOTA group CGH events were considered as statistically significant when a p-value was < 0.05 and the average FDR was 2%. Higher FDR were adopted to compensate to the lower statistical power due to the reduction in sample size.

Frequencies of the minimal overlapping regions of significant CGH events in each SOTA group were subjected to hierarchical cluster analysis (MeV v4.0) to organize the non-random chromosomal aberration events, and the tumour progression model was constructed, as previous described by our team members [122].

4.2.5 Identification of target genes through mapping of cytogenetic and transcriptomic data

A web-based interface was written to set up a Perl-based open source genome browser: Generic Genome Browser (Gbrowse)[201] for the mapping and visualization of the recurrent CGH events obtained in the present study and the meta-analysis results of gene expression microarray data provided by our team members. Gastric cancer microarray data sets from five independent research centres in Australia [202], Brazil [203], Hong Kong [204], Japan [205] and Korea [206] were downloaded from the NCBI Gene Expression Omnibus and the Stanford Microarray Database (SMD/GSE2701, GSE2444, GSE2669, GSE2685 and GSE3438, and subjected to meta-analysis by using our in-house novel algorithm (unpublished data; manuscript in preparation). Together with the position information of the genes and the cytobands from NCBI[207], the differentially expressed genes and the chromosomal aberration events were mapped according to their positions. Genes significantly over-expressed in regions of non-random chromosomal gain and genes significantly under-expressed in regions of non-random chromosomal loss were identified as potential cancer driver genes and tumour suppressor genes, respectively.

4.2.6 Preparation of RNA from Cell lines

7 gastric cancer cells lines (AGS, Kato III, MKN28, MKN45, NCI-N87, SNU1 and SNU16) were cultured in RPMI 1640 medium (Invitrogen, Carlsbad, CA, USA) supplemented with 10% fetal bovine serum (Invitrogen, Carlsbad, CA, USA). The cells are incubated at 37 °C in a 5% carbon dioxide-containing atmosphere with 95% humidity. After 72 hours, the cells were harvested for RNA extraction using TRIzol (Invitrogen, Carlsbad, CA, USA) according to manufacturer's instructions. The quality and quantity of total RNA were spectrophotometerically assessed by Nanodrop 1000 (Nanodrop, Wilmington, DE, USA).

4.2.7 RNA extracted from clinical specimens

RNA extracted from 25 pairs of tumour and adjacent non-tumourous tissue from gastric cancer patients and from 10 healthy individuals were provided by Institute of Digestive Disease, CUHK. The quality and quantity of total RNA were also spectrophotometerically assessed by Nanodrop 1000 (Nanodrop, Wilmington, DE, USA).

4.2.8 Assessment of expression level of target genes in various GC cell line and clinical specimens

Reverse transcription were performed with 1μg of total RNA, using High Capacity cDNA Reverse Transcription kit (Applied Biosystems, Foster City, CA, USA). Expression level of target genes in various gastric cancer cell lines were assessed by conventional polymerase chain reaction using GoTaq polymerase (Promega, Madison, WI, USA). Each reaction was carried out in a 12.5 μL reaction volume which contained 2.5 μL of 5X Green GoTaq® Flexi Buffer, 1 μL of 25mM magnesium chloride (MgCl2), 0.25 μL of 10mM deoxynucleotide triphosphates (dNTPs), 0.25 μL of each of 10μM forward and reverse primers (Table 4.2), 0.0625 μL of Go-Taq® DNA Polymerase at the concentration of 5U/μL, 6.1875 μL of nuclease-free water and 2 μL of cDNA. The thermal profile for PCR was 95°C for 2 minutes and 40 cycles of 95°C for 30 seconds; 55-60°C for 30 seconds; 72°C for 30 seconds; and finally 72°C for 5 minutes for finally extension. PCR products were electrophoresed in 2% agarose gel and stained with GelRed (Biotium, Hayward, CA, USA).

Table 4.2 List of primer sequences used in polymerase chain reaction

Gene name	Forward primer	Reverse primer
18SrRNA	AAACGGCTACCACATCCAAG	GAAATTGCTATCTGCCAGTT
HOXB6	GAGACAGAAGAGCAGAAGTG	CGTCAGGTAGCGATTGTAGT
HOXB7	ATCTACCCCTGGATGCGAAG	ATCTGTCTTTCCGTGAGGCA
KPNA2	TAGAGGTCAATGTGGAGCTG	TAGAGGTCAATGTGGAGCTG
TOP2A	TATTTTGCTCCGCCCAGACA	CCCTTTGTTTGTTGTCCGCA
UBE2C	CCTGTATGATGTCAGGACCA	AAAGACGACACAAGGACAGG

Besides conventional PCR, quantitative real-time PCR (RT-PCR) were also performed in triplicate using Power SYBR® Green PCR Master Mix Kit in ABI 7500 Real-Time PCR System or ABI 7900 HT Real-Time PCR System (Applied Biosystems, Foster City, CA, USA).

Reactions were carried out in 20 μL scale in 96 well plate (catalogue number: 4346906, Applied Biosystems, Foster City, CA, USA) or in 10 μL scale in 384 well plate (catalogue number: 4309849, Applied Biosystems, Foster City, CA, USA). In each 20 μL reaction, the reaction mix contained 10 μL 2 × Power SYBR® Green PCR Master Mix, 1 μL of each of 10uM forward and reverse primers, 6 μL of nuclease-free water and 2 μL of cDNA. The amounts of reagents were halved in a 10 μL reaction. The thermal profile for RT-PCR was pre-incubation at 95°C for 10 minutes, and 40 cycles of 95°C for 15 seconds and 60°C for 1 minute.

The RT-PCR data were collected and processed in SDS software (version 2.4). Data among three replicates were averaged and were normalized to the internal control, ribosomal 18S RNA, by the $2^{-\Delta\Delta Ct}$ method[208].

4.2.9 Examination of the effect of in vitro siRNA knockdown of KPNA2 on gastric cancer cell growth

Cells were seeded at a concentration of 1×105 per well in 12 well plate and transfected with KPNA2 siRNA or control siRNA (SI02780631; AllStars Negative control siRNA, FlexiTube, Qiagen, Hilden, Germany) using Lipofectamine™ 2000 reagent (Invitrogen, Carlsbad, CA, USA) as described by the manufacturer. The transfected cells were seeded to a 96 well plate (25-50 μL/mL of cells) 24 hours after transfection and incubated at 37 °C in 5% carbon dioxde for further 48 hours.

Cell viability was assessed by using 3-(4,5-dimethylthiazol-2-yl)-2-(4-sulfophenyl)-2H-tetrazolium (MTS) assay (Promega, Madison, USA). After incubating the cells with CellTiter 96® AQueous One Solution Cell Proliferation reagent at 37 °C for 1 hour, the absorbance at 490nm (ASYS UVM 340 Microplate Reader, Biochrom, Cambridge, UK) was measured .

The transfected cells were then harvested. Total RNA was extracted, and reverse transcribed to cDNA. Quantitative real time PCR was carried out to determine the percentage of knockdown.

4.2.10 Statistical analysis

Kruskal-Wallis test was carried out for the comparison of the number of chromosomal aberrations from three SOTA groups in both IGC and DGC as these chromosomal aberrations were not normally distributed.

Gene expression data from quantitative RT-PCR were calculated by the $2^{-\Delta\Delta Ct}$ method[208]. Wilcoxon matched pairs tests, quadratic curve fitting, scatter plot and box-plot were carried out using R (R Foundation for Statistical Computing) [209] for the comparison of the tumour and non-tumour pairs. A p-value <0.05 indicated a significant difference between two groups.

4.3 Results

4.3.1 Identification of significant non-random CGH events by using a modified statistics tool with consideration of FDR

CGH data of 41 GA cases, 198 IGC and 117 DGC cases were extracted from 4, 11 and 7 research studies, respective, and re-formatted as chromosomal gain events and chromosomal loss events at 320-band resolution. Chromosomal gain events and chromosomal loss events were treated as independent events. The statistical tool that had been developed by Poon *et al.* for identification of non-random CGH events [122] was successfully modified to implement a function for calculation of the FDR. The plot of FDRs against the corresponding p values for GA, IGC and GDC cases were shown at figure 4.1.

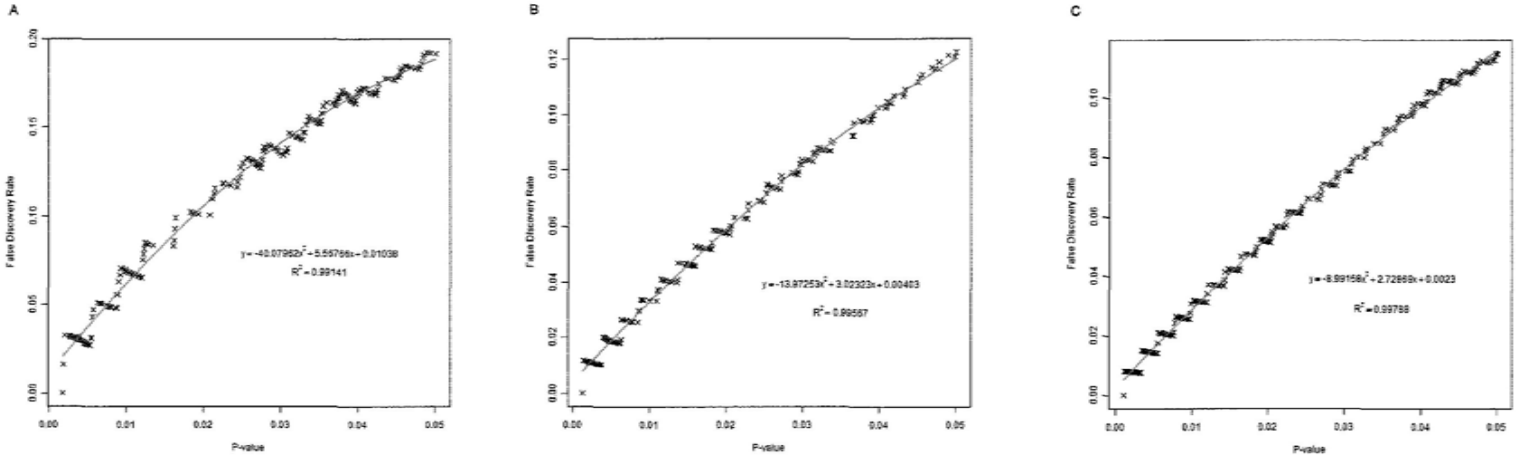


Figure 4.1 Plot of FDRs against the corresponding p values for GA(A), IGC(B) and DGC (C) cases. The R-squared values, i.e. the coefficients of multiple determination, for all three curves were close to 1, indicating a good fit.

With the new FDR function, significant (i.e. non-random) recurrent chromosomal imbalances were identified in the GA, IGC and DGC cases with p-value < 0.05 and cut-off of FDR at 0%.

Consecutive significant loci of gain or consecutive significant loci of loss were considered as a recurrent event of gain or loss, respectively. The recurrent CGH events in GA, IGC and DGC are summarized in Table 4.4. In GA, 8 recurrent regions of gain and 7 recurrent regions of loss were identified; in IGC, 18 recurrent region of gain and 10 recurrent regions of loss were identified; in DGC, 23 recurrent regions of gain and 9 recurrent regions of loss were identified. It is important to note that all 8 recurrent regions of gain in GA were overlapped with the recurrent regions of gain in IGC as well as in DGC.

However, the recurrent regions of loss in GA, only 2 of the 7 regions (i.e., 5q12-23 and 9p21-24) were overlapped with those in IGC, and only 1 (i.e., 9p21-24) was overlapped with those in DGC. When comparing between IGC and DGC, 14 recurrent regions of gain (1q11-42, 2p24-25, 2p13-14, 3q25-26, 5p11-14, 6p21, 6p23-25, 7, 8q, 11q12-13, 12q24, 16p13, 17q, 20) were overlapped, but only 3 recurrent regions of loss (9pter-12, 16q22-23, 18q12-ter) were overlapped in IGC and DGC.

Table 4.3 Summary of commonly recurrent events of chromosomal gain and loss identified in GA, IGC and DGC by meta-analysis. The values inside brackets indicate the frequency of occurrences.

Group	Gain events	Frequency	Loss events	Frequency
GA (41 cases)	6p21-11	(11.38%)	5q12-23	(24.04%)
	7q11	(7.32%)	5q33-34	(12.20%)
	8q22-ter	(7.32%)	6p12-11	(13.42%)
	9q33-ter	(9.76%)	6q11-16	(15.85%)
	17q21	(9.76%)	9p24-21	(11.59%)
	17q24-ter	(9.76%)	13q21	(19.51%)
	20q11	(7.32%)	13q31	(21.95%)
	20q13	(12.2%)		
	[8 regions]		[7 regions]	
IGC (198 cases)	1q	(15.31%)	3p23	(9.09%)
	2p	(7.49%)	3p25-26	(8.76%)
	2q11-13	(7.07%)	4	(16.28%)
	2q37	(10.10%)	5q11-31	(15.21%)
	3q25-ter	(11.42%)	9pter-12	(13.38%)
	5p	(9.19%)	15q11-21	(8.25%)
	6p	(10.03%)	16p	(12.46%)
	7	(14.60%)	16q11-23	(8.08%)
	8q	(25.25%)	18q12-ter	(17.17%)
	9q31-34	(11.87%)	21q11	(8.59%)
	11q12-13	(18.19%)		
	12p	(7.58%)		
	12q13-15	(6.40%)		
	12q24	(9.09%)		
	16p13	(10.61)		
	17q	(28.07%)		
	19	(12.63%)		
	20	(34.93%)		
	[18 regions]		[7 regions]	
IGC (117 cases)	1p36	(17.95%)	9p	(10.79%)
	1p34	(11.97%)	9q11-12	(9.40%)
	1q11-42	(11.03%)	9q34	(14.53%)
	2p24-25	(8.12%)	10p15	(8.55%)
	2p13-14	(5.98%)	11p	(8.03%)
	3q25-26	(10.69%)	16q22-ter	(9.12%)
	3q28-ter	(10.26%)	17p12-13	(19.23%)
	5p11-14	(4.07%)	18q12-ter	(14.11%)
	6p25-23	(8.55%)	19p	(14.53%)
	6p21	(11.11%)		
	7	(14.64%)		
	8p23	(15.38%)		
	8p11-12	(14.11%)		
	8q	(19.53%)		
	11q12-13	(10.26%)		
	12q23-ter	(6.41%)		
	15q24	(8.55%)		
	16p12-13	(11.12%)		
	17	(18.29%)		
	18p11	(9.40%)		
	20	(22.37%)		
	22q	(13.39%)		
	X	(18.07%)		
	[23 regions]		[9 regions]	

4.3.2 Construction of tumour progression model

4.3.2.1 Evolution tree for classifying IGC and DGC cases

By submitting the significant chromosomal CGH events to SOTA, an evolution tree was constructed for classifying the 198 IGC cases, and one was constructed for classifying 117 DGC cases. In both evolution trees, the cases were classified into 4 groups at 3 different evolution levels. (Figure 4.2) For each evolution tree, the first group diverge from the root was assigned as SOTA 1, representing tumour cases at the early stage, and the next diverged group was assigned as SOTA 2, representing tumour cases at the intermediate stage. The remaining two groups at the bottom taxonomy level were combined and assigned as SOTA 3, representing tumour cases at the advance stage.

For IGC, 117, 45 and 36 cases were clustered in SOTA 1, 2 and 3 respectively. On average, there were 5.91 (S.D. = 3.97), 6.93 (S.D. = 3.03) and 12.86 (S.D. = 4.18) chromosomal aberrations per case in SOTA 1, 2 and 3 respectively.

For DGC, 75, 25 and 17 cases were clustered in SOTA 1, 2 and 3 respectively. On average, there were 5.40 (S.D. = 5.14), 8.52 (S.D. = 4.24) and 9.71 (S.D. = 3.39) chromosomal aberrations per case in SOTA 1, 2 and 3 respectively.

In both evolution trees, the number of chromosomal aberrations significantly increased from three SOTA groups differed significantly increased from SOTA 1 to SOTA 2 and further to SOTA 3 (Kruskal-Wallis test, $P < 0.001$ for both IGC and DGC cases).

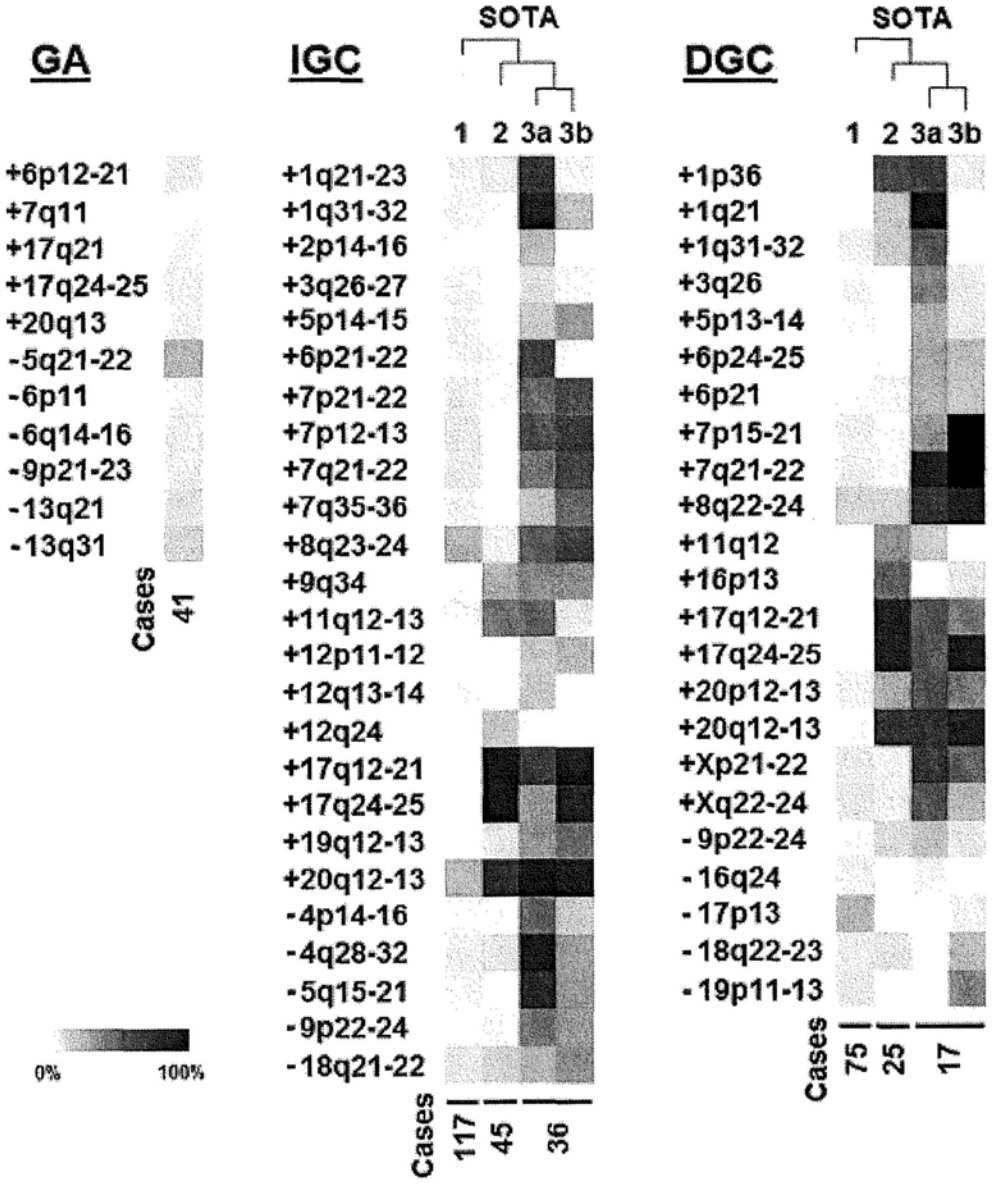


Figure 4.2 Two evolution trees for classifying the IGC and DGC cases into classes at different evolution level. The trees were constructed by using SOTA. For each trees, the class closest to the root was assigned as SOTA 1, and the 1 in the middle as SOTA 2. There were 2 classes at the bottom taxonomy level, which were grouped as SOTA 3, were 2 subgroups A and B in the SOTA 3. The grey colour intensity indicates the percentage of cases having the corresponding chromosomal aberration in each class. For the ease of comparison, the recurrent chromosomal aberrations in GA were also shown.

4.3.2.2 Significant chromosomal CGH events in SOTA groups

Significant non-random CGH events were identified from each SOTA group at $p < 0.05$ and FDR = 2%. Those remained significant in all SOTA groups were considered as the earliest events in the carcinogenesis and were chosen to represent early stage of carcinogenesis. Those events found to be significant in both SOTA 2 and SOTA 3 were chosen as representatives of the intermediate stage. Lastly, those events that were significant only in SOTA 3 represented the advanced stage (Table 4.4).

Table 4.4 and 4.5 summarize those events representing different SOTA groups in IGC and DGC. Frequency patterns of minimal overlapping regions of these representative chromosomal imbalances in different SOTA groups are shown in Figure 4.3 and 4.4.

Table 4.4 Summary of the representative chromosomal imbalances in 3 SOTA groups of IGC. Those representative events of SOTA 1 were also identified as non-random events in SOTA 2 and SOTA 3, and those of SOTA 2 were also identified as non-random event in SOTA

Group	SOTA 1 (117 cases)	SOTA 2 (45 cases)	SOTA 3 (36 cases)
Representative chromosomal imbalances	+1q21-23 (15%, p<0.0001),	+9q34 (31%, p<0.0001),	+1q31-32 (61%, p<0.0001),
	+8q23-24 (32%, p<0.0001),	+11q12-13 (49%, p<0.0001),	+2p14-16 (19%, p<0.0045),
	+20q12-13 (32%, p<0.0001)	+17q12-21 (93%, p<0.0001),	+5p14-15 (31%, p<0.0001),
		+17q24-25 (87%, p<0.0001),	+6p21-22 (39%, p<0.0001),
		+19q12-13 (16%, p<0.0002),	+7p21-22 (56%, p<0.0001),
		-4q28-32 (18%, p<0.0004)	+7p12-13 (64%, p<0.0001),
			+7q21-22 (53%, p<0.0001),
			+7q35-36 (39%, p<0.0001),
			+12p11-12 (28%, p<0.0001),
			-4p14-16 (42%, p<0.0001),
			-5q14-21 (58%, p<0.0001),
			-9p22-24 (42%, p<0.0001)

Table 4.5 Summary of the representative chromosomal imbalances in 3 SOTA groups of DGC. Those representative events listed in SOTA 1 were also identified as non-random events in SOTA 2 and SOTA 3, and those listed in SOTA 2 were also identified as non-random event in SOTA 3.

Group	SOTA 1 (75 cases)	SOTA 2 (25 cases)	SOTA 3 (17 cases)
Representative chromosomal aberration events	+1q31-32 (15%, p<0.0022), +8q22-24 (24%, p<0.0001), +20p12-13 (11%, p<0.0001), +20q12-13 (7%, p<0.0022)	+1q21 (28%, p<0.0001), +17q12-21 (76%, p<0.0001), +17q24-25 (72%, p<0.0001)	+5p13-14 (24%, p<0.0007), +6p24-25 (29%, p<0.0007), +6p21 (29%, p<0.0004), +7p15-21 (65%, p<0.0001), +7q21-22 (88%, p<0.0001), +Xp21-22 (59%, p<0.0001), +Xq22-24 (41%, p<0.0001)

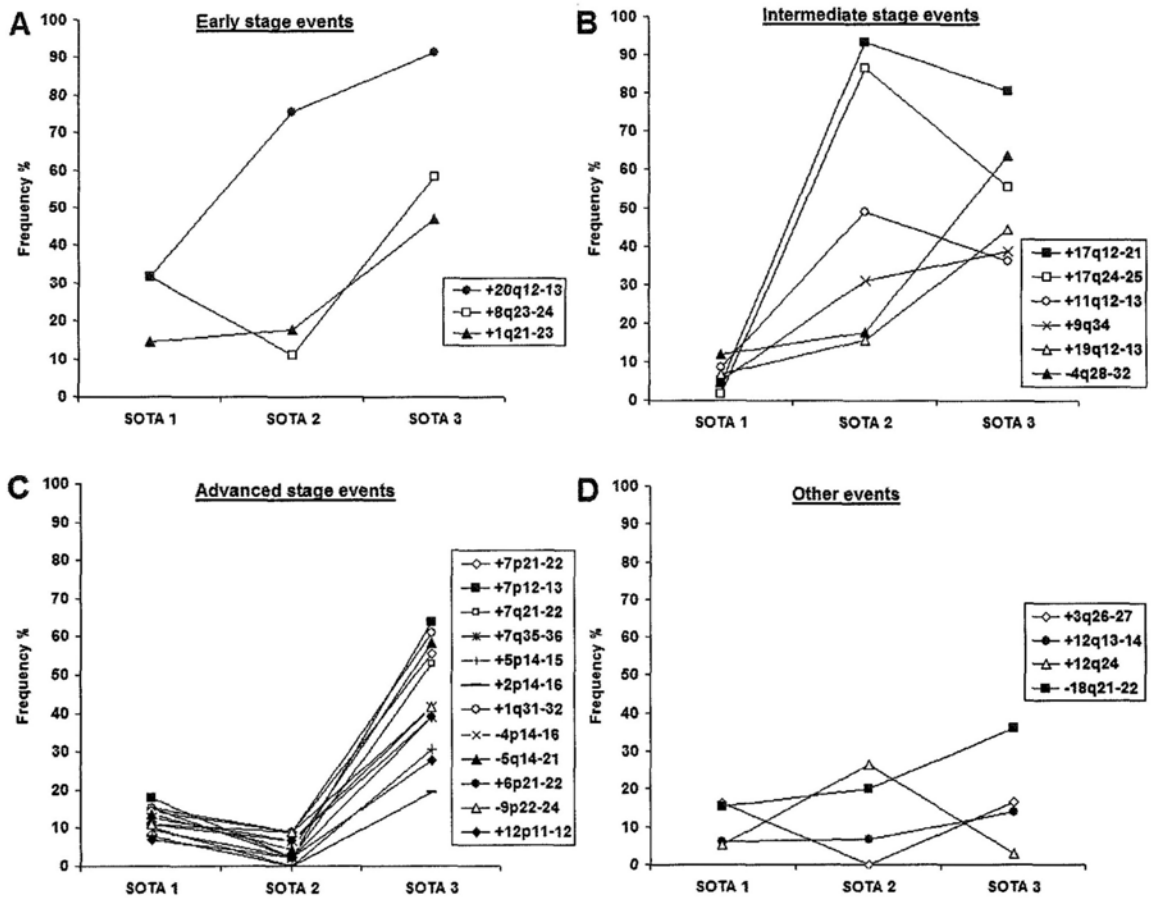


Figure 4.3 Frequency patterns of minimal overlapping regions of representative chromosomal imbalances in the SOTA groups 1, 2 and 3 of IGC, which potentially represent the early (A), intermediate (B), advanced (C) events, respectively. Frequency patterns of other non-random CGH events (D) that were not considered as representative events are also provided here.

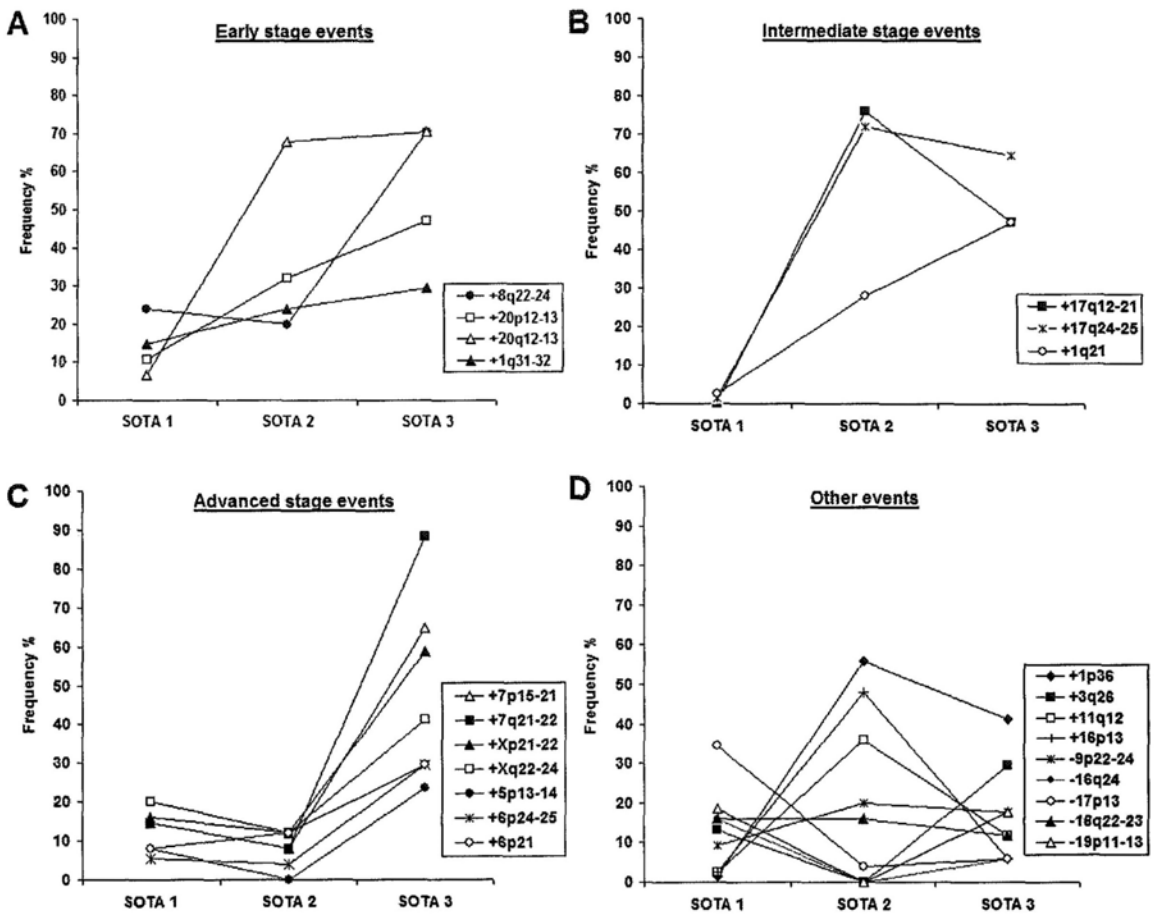


Figure 4.4 Frequency patterns of minimal overlapping regions of representative chromosomal imbalances in the SOTA groups 1, 2 and 3 of DGC, which represent the early (A), intermediate (B), advanced (C) events, respectively. Frequency patterns of other CGH events (D) that were not considered as representative events are also provided here.

4.3.2.3 Evolutionary chromosomal changes in the carcinogenesis of gastric cancer

Figure 4.5 shows the cluster trees of the representative CGH events (minimal overlapping regions) in the individual SOTA groups of IGC and DGC.

For IGC (Figure 4.5A), in SOTA 1, gain of 20q12-13 appears as an earlier branch, suggesting its importance in the earliest stage of carcinogenesis. In SOTA 2, gain of 17q12-21 and 17q24-25 are clustered together in the earliest branch, suggesting their prior roles in the intermediate stage of tumour progression. In SOTA 3, it also appears that there are also 2 major tumour progression paths in the late stage of tumour progression. One involves gains of 7p21-22 and 7p12-13, and one involves gain of 1q31-32 and loss of 4p14-16.

For DGC (Figure 4.5B), in SOTA 1, gain of 8q22-24 appears as an earlier branch and suggests its importance in the earliest stage of carcinogenesis. In SOTA 2, gains of 17q12-21 and 17q24-25 are clustered together in the earliest branch, suggesting their prior roles in the intermediate stage of tumour progression. In SOTA 3, similarly gains of 7p15-21 and 7q21-22 are clustered together in the earliest branch suggesting their important roles in the advanced stage of tumour progression. Later branches in the cluster tree suggest the presence of different pathways of tumour progression in the advanced stage.

After combining the clustering information, cytogenetic pathways for tumour progression of IGC and DGC were proposed and are illustrated in Figure 4.6.

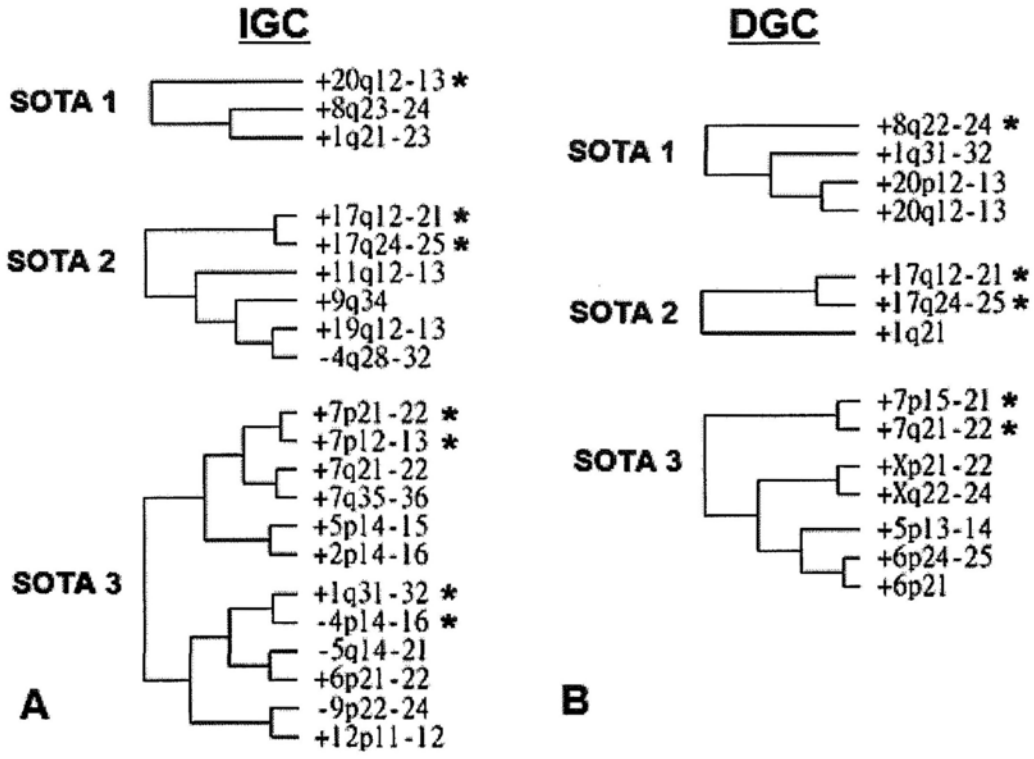
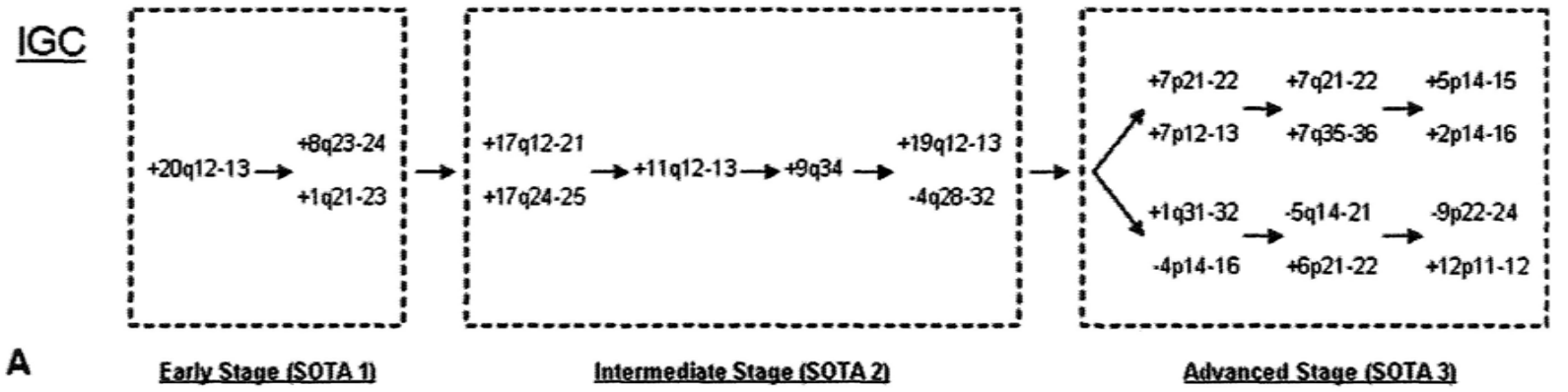


Figure 4.5 Hierarchical clustering of the representative CGH events of SOTA groups 1, 2 and 3 in IGC (A) and DGC (B). * The predicted early events that demonstrated the highest occurrence frequency in each branch.



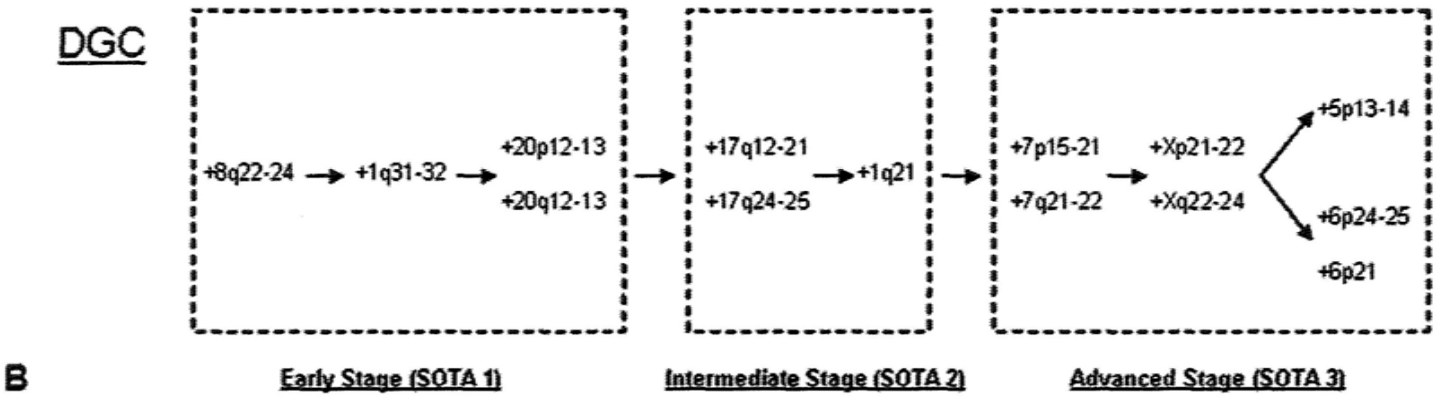


Figure 4.6 Proposed evolutionary changes in the carcinogenetic pathways of IGC (A) and DGC (B). The minimal overlapping regions of the recurrent chromosomal imbalances are presented.

4.3.3 Potential cancer driver genes identified through mapping of cytogenetic and transcriptomic data

A web-based interface (Gbrowse) was successfully established for the mapping of the recurrent cytogenetic (CGH) obtained in the present study and the differential transcriptomic (microarray) data provided by our team members (unpublished data). The differential transcriptomic data were obtained by meta-analysis of the gastric cancer microarray data sets from 5 independent research centres in Australia [202], Brazil [203], Hong Kong [204], Japan [205] and Korea [206]. Using the user friendly Gbrowse interface, information of chromosomal locus or gene of interest can be accessed easily by text search or by clicking at the corresponding location in the ideogram. (Figure 4.7)

Using the well-known oncogene cyclin E1 (CCNE1) as an example, Figure 4.8 illustrated the co-visualization of the gene expression data from the meta-analysis and the cytogenetic data. Another example is topoisomerase II alpha gene (TOP2A), which is a well-known over-expressed gene in gastric cancer[210], and is shown in Figure 4.9.

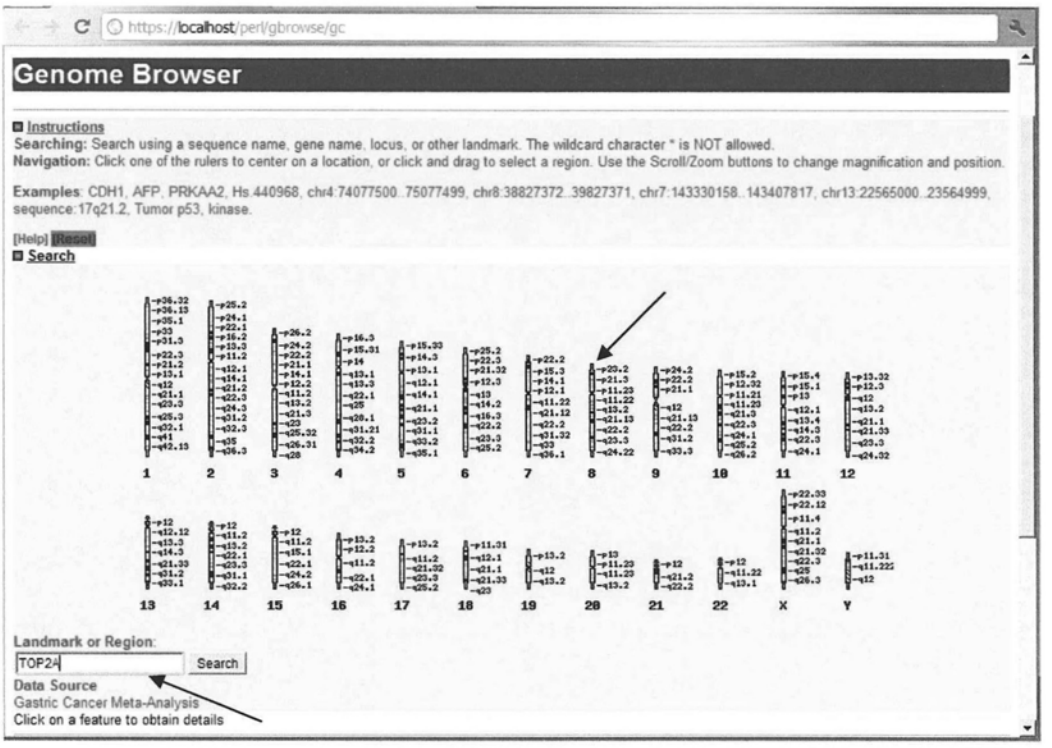


Figure 4.7 Screenshot of a web-based interface in which the Gbrowse was implemented for combining and presenting the recurrent CGH data and the differential transcriptomic data obtained by meta-analysis. Arrows indicate that query could be performed either by clicking at the chromosomal locus at the ideogram or by text query.

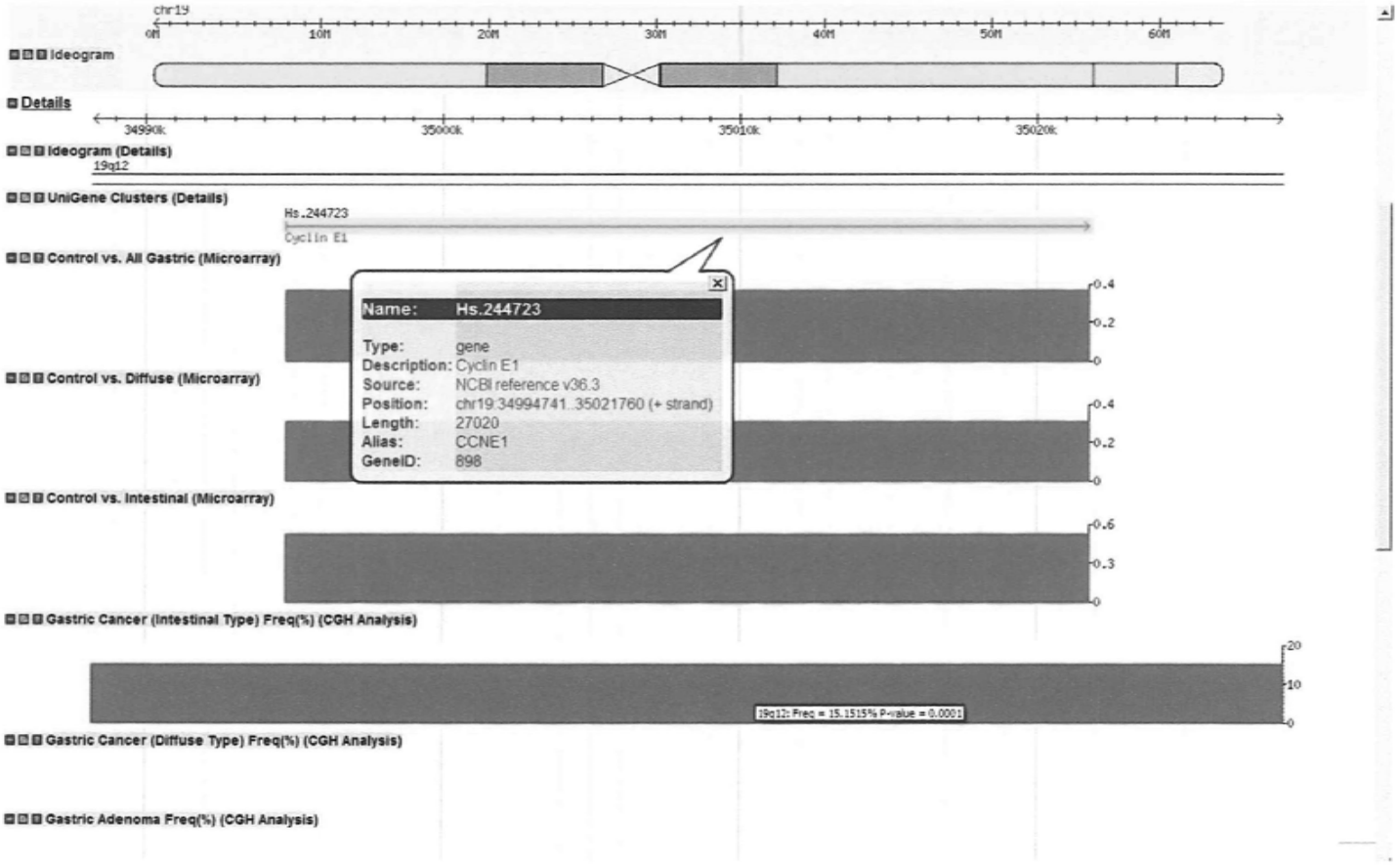


Figure 4.8 Visualization of the scores from the comparison of gene-expression between different cases and frequency of chromosomal aberrations in different tracks by using our web-based interface. Annotation of the gene were available by clicking at the bar representing the gene at the Unigene clusters track. Exact value of the frequency of chromosomal aberration events and the p-value were shown in the tooltip upon mouse over at the track displaying the result of the CGH analysis. The region containing cyclin E1 is shown as a positive control.

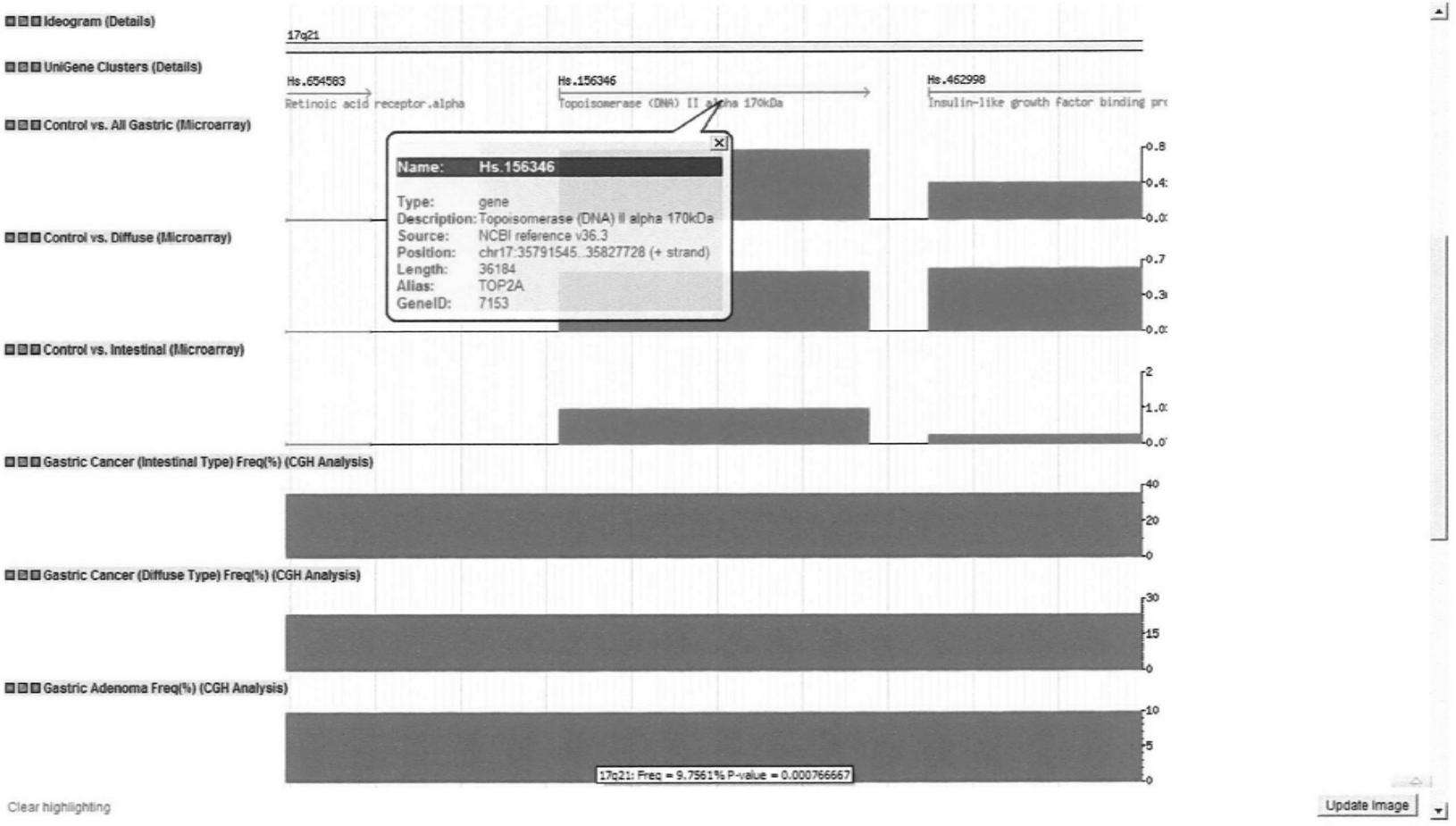


Figure 4.9 Co-visualization of chromosomal gain at 17q21.2 and over-expression of TOP2A in both IGC and DGC cases by using our web-based interface.

Successful identification of CNNE1 and TOP2A supported that our web-based interface was useful in the identification of potential cancer driver genes, which were over-expressed and located in a locus of chromosomal gain. Similarly, potential tumour suppressor genes could be identified by looking for those which were under-expressed and located in a locus of chromosomal loss.

Our CGH meta-analysis results showed that gains of 20q13, 17q21 and 17q24-25 were observed as recurrent events in GA (Table 4.4), and gain of 20q12-13 and gains of 17q12-21 and 17q24-25 as key recurrent events at the early stage and intermediate stage of gastric cancer development in both IGC and DGC, respectively (Figure 4.6).

Using our web-based interface, we attempted to identify novel potential cancer driver genes in gastric cancer within the chromosomal loci 20q13, 17q21 17q24-25. In 20q13, UBE2C was found to be significantly over-expressed in both IGC and DGC; in 17q21, besides TOP2A, HOXB6 and HOXB7 were identified; in 17q24-25, KPNA2 were identified. (Screenshots in Appendix 6)

4.3.4 Assessment of expression levels of UBE2C, HOXB6, HOXB7 and KNPA2 in gastric cancer cell lines and clinical tissue specimens

Expression levels of potential cancer driver genes in seven gastric cancer cell lines (AGS, Kato III, MKN28, MKN45, NCI-N87, SNU1 and SNU16) were assessed by semi-quantitative PCR and agarose gel electrophoresis. (Figure 4.10) Over-expression of HOXB6, HOXB7 and KPNA2 were observed in the gastric cancer cell lines when compared to a normal gastric cancer tissue. However, it was difficult to determine whether.

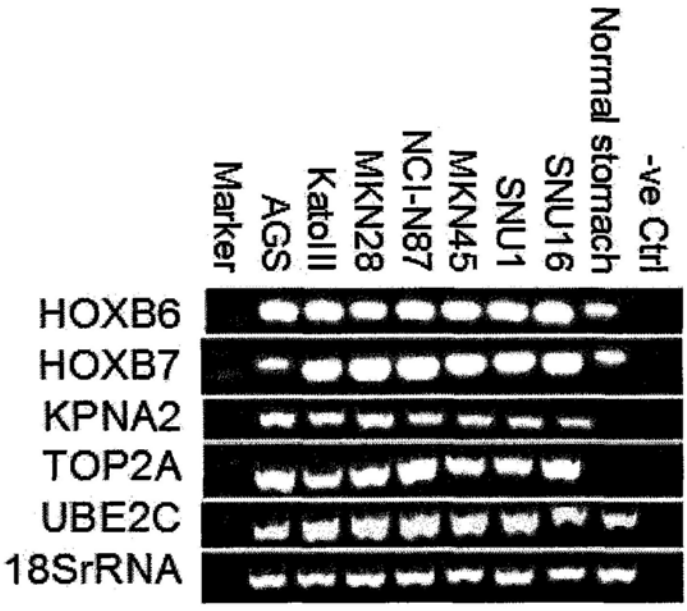


Figure 4.10 Expression of HOXB6, HOXB7, KPNA2 and UBE2C in various gastric cancer cell line and normal stomach. 18SrRNA was used as the internal control while TOP2A was used as a positive control.

The expression of these target genes in cell lines were also determined by quantitative real-time RT-PCR. Using 18SrRNA as internal control, the relative expressions of these genes are shown in Figure 4.11. Upon comparison with the normal, over-expression of KPNA2 was observed in 6 out of 7 gastric cancer cell lines. For HOXB6 and HOXB7, 2 and 3 out of the 7 cell lines showed over-expression, respectively. Over-expression of UBE2C was only found in AGS cells.

Because KPNA2 was the only gene showing over-expression in most of the cell lines, we further examined its expression levels in gastric cancer tumour tissues and the corresponding adjacent non-tumourous gastric tissues collected from 25 patients. 8 pairs of samples showing a CT value > 40 were considered as contaminated, and such pairs of samples were excluded, leaving 17 pairs of tissues for statistical analysis. Significant over-expression of KPNA2 in the gastric cancer tissues was confirmed (1-sided Wilcoxon match pairs test: $p = 0.001$, Figure 4.12). The expression of KPNA2 in the tumour tissues was increased by 1.5 folds.

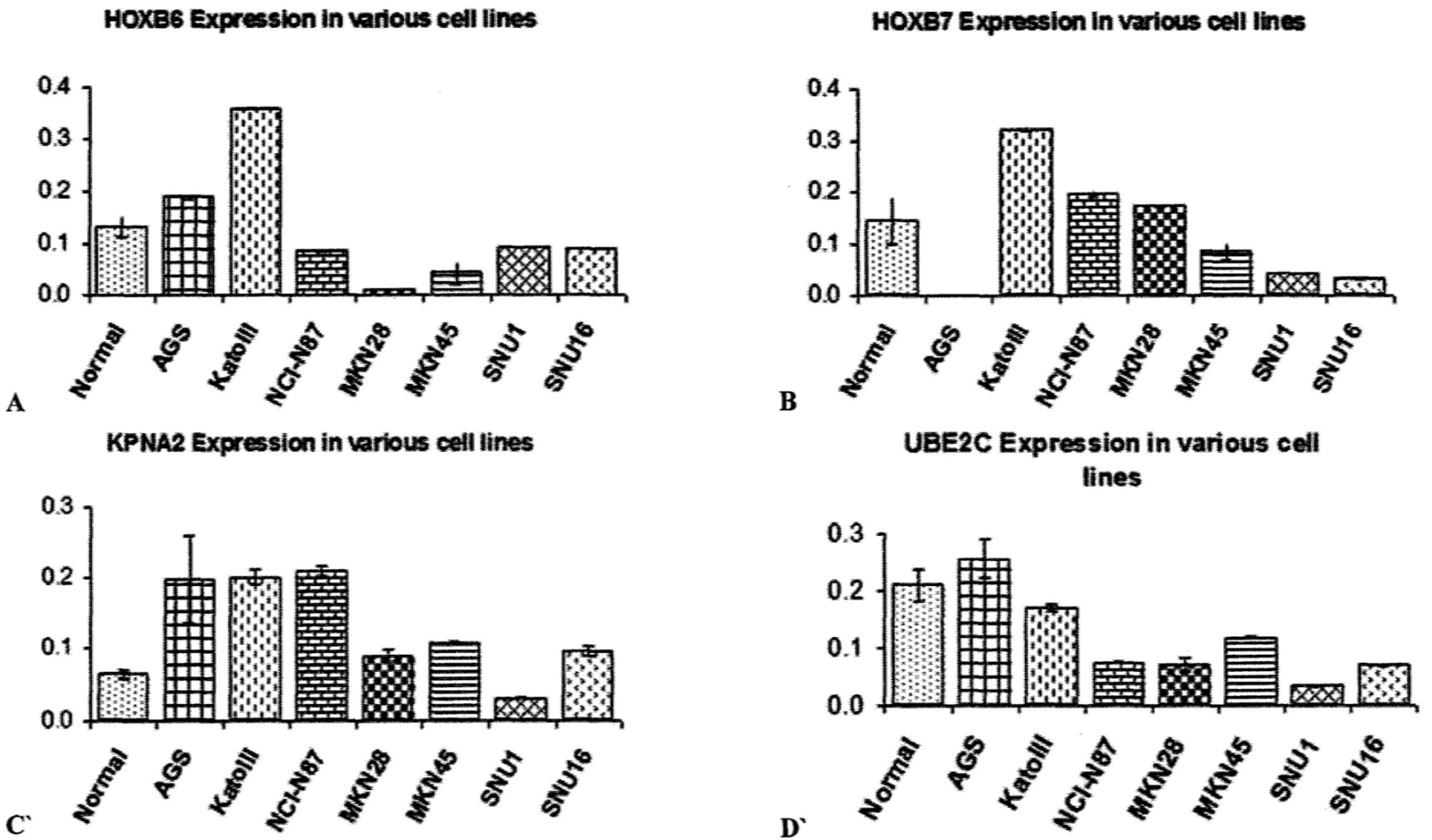


Figure 4.11 Relative expression of HOXB6 (A), HOXB7 (B), KPNA2 (C) and UBE2C (D) in various gastric cancer cell line and normal stomach examined by RT-PCR . 18SrRNA was used as the internal control

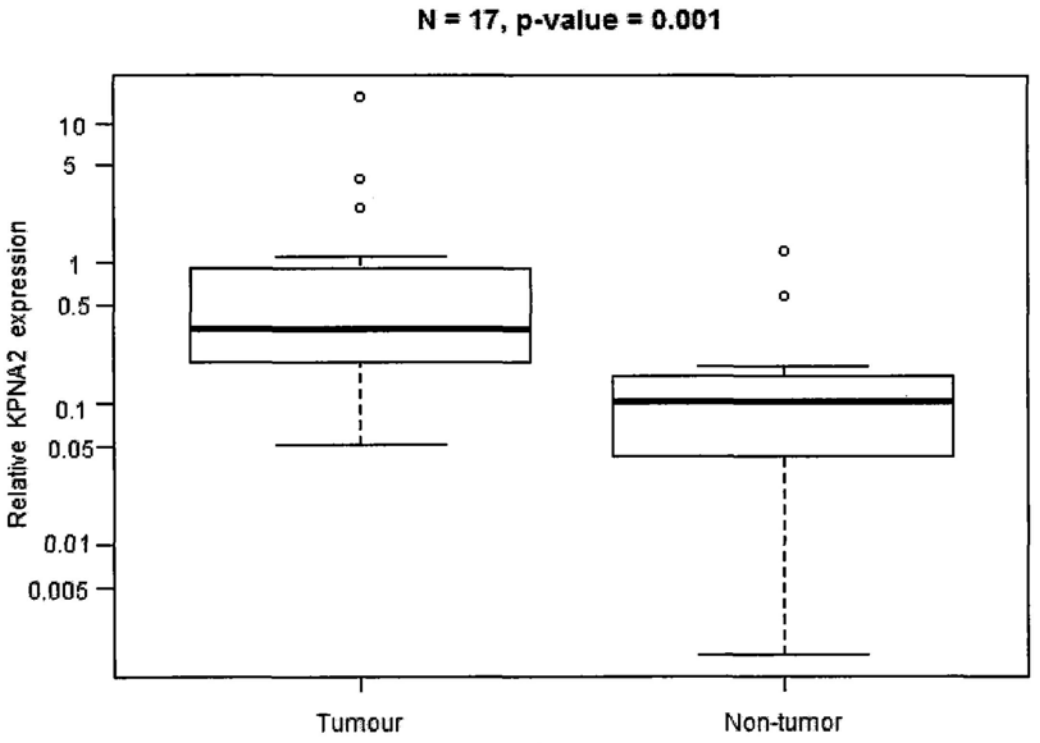


Figure 4.12 The box-plot of relative gene expression levels of KPNA2 in 17 pairs of gastric cancer tissues and adjacent non-tumourous gastric tissue. The expression levels were examined by RT-PCR. 18SrRNA gene was used as the internal control for normalization. P-value was calculated using 1-sided Wilcoxon match pairs test since it was hypothesized that there would be an over-expression of KPNA2 in the tumourous tissue basing on the result from the cell lines. The error bars indicates 25th and 75th percentiles.

4.3.5 Effect of knockdown of KPNA2 on the growth of gastric cancer cell lines

In gastric cancer tissues, over-expression of KPNA2 genes and recurrent chromosomal gains of 17q12-21, in which KPNA2 is located, support KPNA2 is potential cancer driver gene in carcinogenesis. To further examine this possibility, the effect of KPNA2 on the growth of gastric cancer was examined *in vitro* by silencing its expression with siRNA. Two gastric cancer lines, KatoIII and SNU16, which over-expressed KPNA2, were examined in this experiment. Quantitative RT-PCR confirmed that over 70% of mRNA levels of KPNA2 was knocked down in both cell lines after treating with KPNA2 siRNA. Down-regulation of KPNA2 led to a reduction of cell growth in both cell lines by 30% (Figure 4.13).

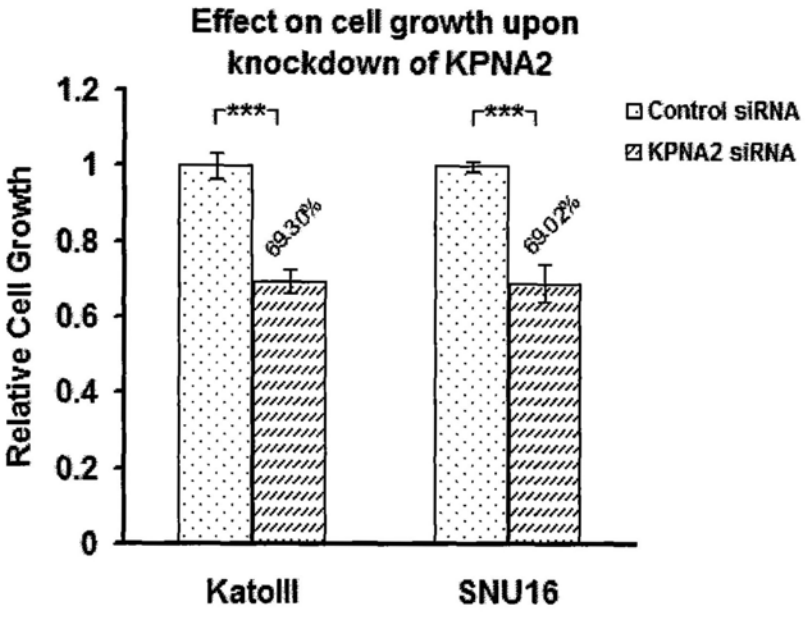


Figure 4.13 Relative cell growth of 2 gastric cancer cell lines, KatoIII and SNU16, upon the knockdown of KPNA2. The cell growth was assessed by MTS assay. P-value was calculated using Mann-Whitney 1 tail test. P-value < 0.005 as shown by ***, indicating significant decrease in cell growth.

4.4 Discussions

4.4.1 Identification of recurrent chromosomal imbalances in gastric adenoma and cancer by meta-analysis

The results from the Part 1 of the present study supports that there should be two types of chromosomal imbalances, i.e., random and non-random events. The non-random events are those retained by a tumour for offering growth advantage. However, in most of the CGH studies, including those studying gastric cancer, no attempts were made to differentiate the non-random events from the random events.

It was not feasible for the identification of non-random chromosomal aberrations events in gastric cancer through the analysis of the small number of cases in each individual studies. Therefore, by the meta-analysis of public CGH data from various studies, the increase in the sample size would improve the statistical power for the identification of recurrent chromosomal imbalances.

The statistics tool for identification of non-random CGH events, which was originally developed by Poon *et al.* [122], was successfully modified to include a function for controlling FDR. With the p-value and FDR information, the identified non-random CGH events will be less likely false discovery events, and hence be more reliable. By using this new tool, meta-analysis of the CGH data from 11 international studies on IGC, 7 international studies on DGC, and 4 international studies on GA were performed. As far as we know, this is the first study provides statistical evidence for the meta-analysis of non-random recurrent chromosomal imbalances identified in GA, IGC and DGC.

4.4.2 Construction of tumour progression model for gastric carcinogenesis from the recurrent chromosomal imbalances in gastric adenoma and cancer

The identified non-random recurrent CGH events of IGC and DGC were subjected to SOTA, and both IGC and DGC cases were classified into 3 groups, representing the early, intermediate and advance stages of carcinogenesis. Again, this is the first study that tumour progression models for IGC and DGC were constructed for deciphering the probable time points when various chromosomal imbalances are involved in oncogenesis.

Among the non-random CGH events identified for GA, IGC and DGC (Appendix 5), 4 key gain events in the regions of 8q22-24, 17q12-21, 17q24-25 and 20q11-13 are highlighted, as they occurred coincidently in GA, and in early/intermediate stages of IGC and DGC. Cancer driver genes may locate in these regions.

Another highlight is the gain of 1q21-23 which was only found significant in early stage of IGC and intermediate stage of DGC but not in GA. This region may harbour genes that facilitate tumour progression and promote the conversion from benign cells to malignant cells. Previous study on HCC [122] also identified significant chromosomal gain at 1q21-23 and 8q22-24 tumour initiation state and such region harbours 56% of the possible cancer driver genes identified in HCC[211].

Besides these 5 highlighted chromosomal imbalances, it is also worth noting that gain of chromosome 7 was found with highest occurrence frequency in the advanced stage of both IGC (+7p21-22, 56%; +7p12-13, 64%; +7q21-22, 53%; +7q35-36, 39%) and DGC (+7p15-21, 65%; +7q21-22, 88%).

4.4.3 Potential cancer driver genes identified in the recurrent chromosomal aberrations

For proof-of-principle, attentions were paid to the chromosomal regions mentioned above. With the help of Gbrowse that mapped and visualized the meta-analysis results of the cytogenetic and transcriptomic data, various well-known proto-oncogenes, oncogenes or genes reported with increased expressions in the gastric cancer tissues were identified in the regions of recurrent chromosomal gains. These genes included MYC at 8q24 and TOP2A at 17q21. Besides the well-known potential cancer driver genes, the results of the present study suggest that KPNA2, which is located at 17q24, is a potential cancer driver gene.

MYC is a well-studied oncogene for gastric cancer. MYC gene is commonly amplified in gastric cancer[212] and was shown to be over-expressed in IM, and associated with H. pylori infection[213]. Amplification and over-expression of TOP2A had been reported in gastric cancer [210, 214]

4.4.4 Over-expression of KPNA2 in gastric cancer

Among the over-expressed genes mapped to 20q13, 17q21 17q24-25 (i.e., UBE2C, HOXB6, HOXB7, KPNA2), quantitative RT-PCR analysis on 7 gastric cancer cell lines only confirmed that KPNA2 was commonly over-expressed. Therefore, in the subsequent experiments, we focused on KPNA2. In our pilot study, we only examined and compared the mRNA levels of KPNA2 in 17 pairs of tumour tissues and adjacent non-tumourous gastric tissues. Although our result confirmed the significant over-expression of KPNA2 in gastric cancer; the small size prohibited us to perform any subgroup analysis. In the future, another study with larger sample size should be carried out to examine the expression patterns of KPNA2 in gastric

tumours at different TMN stages. If amplification and over-expression of KPNA2 are key events in the intermediate stage of tumourigenesis as suggested in our tumour progression models for IGC and DGC, KPNA2 should be significantly over-expressed in both early and advanced cases. The associations between expression of levels of KPNA and clinical features should be also examined. Last but not least, the prognostic value of KPNA2 in gastric cancer should be investigated.

4.4.5 Knockdown of KPNA2 leading to reduction in cell proliferation of KatoIII and SNU16

Transfection of target gene specific siRNAs to KatoIII and SNU16 resulted in the reduced KPNA2 expression, and in turn reduced the cell growth. Although this result supports that KPNA2 is a potential cancer driver gene, more experiments should be performed to investigate the roles of KPNA2 in gastric cancer development. For in vitro experiments, it is important to examine whether the growth reduction was caused by decreased cell proliferation or by increased cell apoptosis. Furthermore, the effect of siRNA knockdown of KPNA2 on colony formation and on cell migration should also be investigated. KPNA2 has been found to be involved in the regulation of two cell cycle checkpoint mediators, NBS1 and CHK2.[215, 216]. It is important to investigate the molecular mechanism that confers the cancer driving effect of KPNA2. Last but not least, the in vivo cancer promoting effect of KPNA2 should be investigated by using transgenic mice in which KPNA2 is over-expressed in the gastric epithelium.

4.4.6 Long term significance of the our results

The identification of KPNA2 as a potential cancer driver gene is just a proof-of-principle study to illustrate the presence of valuable pathological information in the recurrent chromosomal imbalances and in the tumour progression

models identified/developed by us. After publication of our results, our database of the well-organized CGH data of gastric cancer will be made available to the biomedical research community. Other research teams can extract useful information from our database to help them identifying novel cancer driver genes, tumour suppressor genes and therapeutic drug targets in the recurrent chromosomal imbalances by undertaking a systematic screening approach. The results of these studies should improve our understanding on the molecular pathogenesis of gastric cancer. Ultimately, the overall survival of gastric cancer patients will be improved by using the novel therapeutic strategies.

4.5 Conclusion

A list of non-random recurrent chromosomal imbalances was identified by meta-analysis of GA, IGC and DGC cases. Tumour progression model of gastric cancer were constructed based on these non-random events. Key chromosomal imbalances that were identified in GA and in early/intermediate stage of IGC and DGC contain potential cancer driving genes that are involved in tumour initiation, while those imbalances that were only observed in early stage of GC but not in GA contain cancer driver genes that transform benign tumour cells into malignant tumour cells and/or promote cancer progression.

A web-based genome browser, Gbrowse, was established and facilitated the easy identification of novel cancer driver genes through the mapping between the regions of recurrent chromosomal imbalances and differential transcriptomic data. Subsequent expression and functional studies of KPNA2 had shown proof-of-principle for the presence of potential cancer-driver genes, tumour suppressor genes and therapeutic targets in these recurrent chromosomal imbalances. Further identification of their downstream pathways should allow us to choose the best combination of existing drugs for achieving maximum treatment efficiency, as well as to provide an important foundation for future drug development.

In conclusion, *tumour* progression models were constructed for IGC and DGC, and revealed chromosomal imbalances associated with different developmental stages. Systematic genome-wide analysis at these key chromosomal imbalances should allow us to discover a set of potential cancer driver genes, tumour suppressor genes and therapeutic targets in gastric cancer. One the one hand, the results from the present study provide an important foundation for future studies on the pathogenesis

of gastric cancer. On the other hand, the approach developed in the present study can be applied to the studies of other cancer types.

REFERENCES

1. World Health Organization, "Cancer Fact Sheet N° 297," **2011**(February 2011).
2. K. L. Manchester, "Theodor Boveri and the origin of malignant tumours," *Trends Cell Biol.* **5**, 384 (1995).
3. P. U. Devi, "Basics of carcinogenesis," *Health Administrator* **17**, 16 (2005).
4. L. H. Hartwell, M. B. Kastan, "Cell cycle control and cancer," *Science* **266**, 1821 (1994).
5. O. Kranenburg, "The KRAS oncogene: past, present, and future," *Biochim. Biophys. Acta* **1756**, 81 (2005).
6. D. Rasnick, P. H. Duesberg, "How aneuploidy affects metabolic control and causes cancer." *Biochem. J.* **340**, 621 (1999).
7. J. M. Nicholson, P. Duesberg, "On the karyotypic origin and evolution of cancer cells," *Cancer Genet. Cytogenet.* **194**, 96 (2009).
8. R. Holliday, "A new theory of carcinogenesis," *Br. J. Cancer* **40**, 513 (1979).
9. P. A. Jones, S. B. Baylin, "The epigenomics of cancer," *Cell* **128**, 683 (2007).
10. S. S. Rajan, *Cytogenetics* (Anmol Publications Pvt. Ltd, , 2002).
11. B. A. Pierce, *Genetics :a conceptual approach* (W.H. Freeman and Co., New York, 2003), pp. 709, [80].
12. K. B. Ahluwalia, *Genetics* (New Age International (P) Ltd, New Delh, India, ed. 2nd Edition, 2009).
13. J. LEJEUNE *et al.*, "3 Cases of Partial Deletion of the Short Arm of a 5 Chromosome," *C. R. Hebd. Seances. Acad. Sci.* **257**, 3098 (1963).
14. C. A. Machado, T. S. Haselkorn, M. A. F. Noor, "Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in Drosophila pseudoobscura and D. persimilis," *Genetics* **175**, 1289 (2007).
15. R. Kurzrock, H. M. Kantarjian, B. J. Druker, M. Talpaz, "Philadelphia chromosome–positive leukemias: from basic mechanisms to molecular therapeutics," *Ann. Intern. Med.* **138**, 819 (2003).
16. H. G. Griggs, M. A. Bender, "Photoreactivation of ultraviolet-induced chromosomal aberrations," *Science* **179**, 86 (1973).
17. R. Saraswathy, A. T. Natarajan, "Frequencies of X-ray induced chromosome aberrations in lymphocytes of xeroderma pigmentosum and Fanconi anemia patients estimated by Giemsa and fluorescence in situ hybridization staining techniques," *Genetics and Molecular Biology* **23**, 893 (2000).
18. O. Hino, T. B. Shows, C. E. Rogler, "Hepatitis B virus integration site in hepatocellular carcinoma at chromosome 17; 18 translocation," *Proceedings of the National Academy of Sciences* **83**, 8338 (1986).
19. C. Trabanelli *et al.*, "Chromosomal Aberrations Induced by BK Virus T Antigen in Human Fibroblasts* 1," *Virology* **243**, 492 (1998).
20. C. Chenet-Monte, F. Mohammad, C. M. Celluzzi, P. A. Schaffer, F. E. Farber, "Herpes simplex virus gene products involved in the induction of chromosomal aberrations," *Virus Res.* **6**, 245 (1986).
21. S. Nabirochkin *et al.*, "Oncoviral DNAs induce transposition of endogenous mobile elements in the genome of Drosophila melanogaster," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **403**, 127 (1998).
22. E. Nagy, S. Adachi, T. Takamura-Enya, M. Zeisig, L. Möller, "DNA adduct formation and oxidative stress from the carcinogenic urban air pollutant

3-nitrobenzanthrone and its isomer 2-nitrobenzanthrone, in vitro and in vivo," *Mutagenesis* **22**, 135 (2007).
23. B. Holecková, E. Piesova, K. Sivikova, J. Dianovskỳ, "Chromosomal aberrations in humans induced by benzene." *Annals of Agricultural and Environmental Medicine: AAEM* **11**, 175 (2004).
24. K. Al-Sabti, "Carcinogenic-mutagenic chemicals induced chromosomal aberrations in the kidney cells of three cyprinids," *Comparative Biochemistry and Physiology Part C: Comparative Pharmacology* **82**, 489 (1985).
25. P. Erceg, D. P. Milosevic, N. Despotovic, M. Davidovic, "Chromosomal changes in ageing," *Journal of Genetics* **86**, 277 (2007).
26. N. Parveen, G. Shadab, "Chromosome and Ageing," *World Journal of Zoology* **5**, 298 (2010).
27. M. A. Nowak *et al.*, "The role of chromosomal instability in tumor initiation," *Proc. Natl. Acad. Sci. U. S. A.* **99**, 16226 (2002).
28. W. Bodmer, "Genetic instability is not a requirement for tumor development," *Cancer Res.* **68**, 3558 (2008).
29. S. N. Choudhary, "Seminar report on chromosomal banding," **2011**(2002).
30. J. J. Yunis, O. Sanchez, "The G-banded prophase chromosomes of man." *Humangenetik* **27**, 167 (1975).
31. R. Berger, G. Touati, J. Derre, M. A. Ortiz, J. Martinetti, ""Cri du chat" syndrome with maternal insertional translocation." *Clin. Genet.* **5**, 428 (1974).
32. S. Heim, F. Mitelman, *Cancer cytogenetics* (Blackwell Pub, , ed. 3rd Edition, 2009).
33. J. R. Savage, "Application of chromosome banding techniques to the study of primary chromosome structural changes." *J. Med. Genet.* **14**, 362 (1977).
34. T. Caspersson, L. Zech, C. Johansson, "Differential binding of alkylating fluorochromes in human chromosomes* 1," *Exp. Cell Res.* **60**, 315 (1970).
35. M. Seabright, "A rapid banding technique for human chromosomes." *Lancet* **2**, 971 (1971).
36. A. Sumner, H. Evans, R. Buckland, "New technique for distinguishing between human chromosomes," *Nature* **232**, 31 (1971).
37. F. E. Arrighi, T. C. Hsu, "Localization of heterochromatin in human chromosomes," *Cytogenetic and Genome Research* **10**, 81 (1971).
38. B. Dutrillaux, J. Lejeune, "[A new technic of analysis of the human karyotype]," *C. R. Acad. Sci. Hebd. Seances. Acad. Sci. D.* **272**, 2638 (1971).
39. M. Bobrow, K. MADAN, P. L. Pearson, "Staining of some specific regions of human chromosomes, particularly the secondary constriction of No. 9," *Nature* **238**, 122 (1972).
40. V. G. Dev *et al.*, "Consistent pattern of binding of anti-adenosine antibodies to human metaphase chromosomes." *Exp. Cell Res.* **74**, 288 (1972).
41. M. Bobrow, K. Madan, "The effects of various banding procedures on human chromosomes, studied with acridine orange," *Cytogenetic and Genome Research* **12**, 145 (1973).
42. D. A. Shafer, "Banding human chromosomes in culture with actinomycin D." *Lancet* **1**, 828 (1973).
43. B. Dutrillaux, "Nouveau système de marquage chromosomique: Les bandes T," *Chromosoma* **41**, 395 (1973).
44. S. A. Latt, "Microfluorometric detection of deoxyribonucleic acid replication in human metaphase chromosomes," *Proceedings of the National Academy of Sciences* **70**, 3395 (1973).

45. W. M. Howell, T. E. Denton, J. R. Diamond, "Differential staining of the satellite regions of human acrocentric chromosomes," *Cellular and Molecular Life Sciences* **31**, 260 (1975).
46. J. J. Yunis, O. Sanchez, "The G-banded prophase chromosomes of man," *Hum. Genet.* **27**, 167 (1975).
47. D. Schweizer, P. Ambros, M. Andrle, "Modification of DAPI banding on human chromosomes by prestaining with a DNA-binding oligopeptide antibiotic, distamycin A," *Exp. Cell Res.* **111**, 327 (1978).
48. C. G. Sahasrabuddhe, S. Pathak, T. C. Hsu, "Responses of mammalian metaphase chromosomes to endonuclease digestion," *Chromosoma* **69**, 331 (1978).
49. B. J. Trask, "Fluorescence in situ hybridization: applications in cytogenetics and gene mapping," *Trends in Genetics* **7**, 149 (1991).
50. A. C. Van Prooijen-Knegt *et al.*, "In situ hybridization of DNA sequences in human metaphase chromosomes visualized by an indirect fluorescent immunocytochemical procedure* 1," *Exp. Cell Res.* **141**, 397 (1982).
51. Y. Jin, C. Jin, J. Wennerberg, F. Mertens, M. Höglund, "Cytogenetic and fluorescence in situ hybridization characterization of chromosome 1 rearrangements in head and neck carcinomas delineate a target region for deletions within 1p11–1p13," *Cancer Res.* **58**, 5859 (1998).
52. J. W. Vaandrager *et al.*, "Interphase FISH detection of BCL2 rearrangement in follicular lymphoma using breakpoint - flanking probes," *Genes, Chromosomes and Cancer* **27**, 85 (2000).
53. E. Schröck *et al.*, "Multicolor spectral karyotyping of human chromosomes," *Science* **273**, 494 (1996).
54. M. R. Speicher, S. G. Ballard, D. C. Ward, "Karyotyping human chromosomes by combinatorial multi-fluor FISH," *Nat. Genet.* **12**, 368 (1996).
55. C. M. Wilke *et al.*, "Multicolor FISH mapping of YAC clones in 3p14 and identification of a YAC spanning both FRA3B and the t (3; 8) associated with hereditary renal cell carcinoma," (1994).
56. A. Kallioniemi *et al.*, "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors," *Science* **258**, 818 (1992).
57. T. Knutsen *et al.*, "The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence," *Genes Chromosomes Cancer* **44**, 52 (2005).
58. M. Baudis, M. L. Cleary, "Progenetix. net: an online repository for molecular cytogenetic aberration data," *Bioinformatics* **17**, 1228 (2001).
59. Y. S. Fan *et al.*, "Detection of pathogenic gene copy number variations in patients with mental retardation by genomewide oligonucleotide array comparative genomic hybridization," *Hum. Mutat.* **28**, 1124 (2007).
60. J. R. Vermeesch *et al.*, "Guidelines for molecular karyotyping in constitutional genetic diagnosis," *European Journal of Human Genetics* **15**, 1105 (2007).
61. L. Feuk, A. R. Carson, S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics* **7**, 85 (2006).
62. A. J. Iafrate *et al.*, "Detection of large-scale variation in the human genome," *Nat. Genet.* **36**, 949 (2004).
63. G. Callagy *et al.*, "Identification and validation of prognostic markers in breast cancer with the complementary use of array - CGH and tissue microarrays," *J. Pathol.* **205**, 388 (2005).

64. K. Nakao *et al.*, "High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization," *Carcinogenesis* **25**, 1345 (2004).
65. M. M. Weiss *et al.*, "Genomic alterations in primary gastric adenocarcinomas correlate with clinicopathological characteristics and survival," *Cellular Oncology* **26**, 307 (2004).
66. P. L. Paris *et al.*, "Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors," *Hum. Mol. Genet.* **13**, 1303 (2004).
67. J. A. Martinez-Climent *et al.*, "Transformation of follicular lymphoma to diffuse large cell lymphoma is associated with a heterogeneous set of DNA copy number and gene expression alterations," *Blood* **101**, 3109 (2003).
68. J. M. Rothberg, J. H. Leamon, "The development and impact of 454 sequencing," *Nat. Biotechnol.* **26**, 1117 (2008).
69. D. Y. Chiang *et al.*, "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nature Methods* **6**, 99 (2008).
70. W. Chen *et al.*, "Mapping translocation breakpoints by next-generation sequencing," *Genome Res.* **18**, 1143 (2008).
71. J. Ferlay *et al.*, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *International Journal of Cancer*(2010).
72. J. Ferlay *et al.*, "GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 [Internet]." *Lyon, France: International Agency for Research on Cancer; 2010*(20/7/2011).
73. Hong Kong Hospital Authority, "Hong Kong Cancer Registry," **2011**(2008).
74. S. Quade, "Screening for gastric cancer," **2011**(11/2/2003).
75. S. Kono, M. Ikeda, S. Tokudome, M. Kuratsune, "A case-control study of gastric cancer and diet in northern Kyushu, Japan," *Jpn. J. Cancer Res.* **79**, 1067 (1988).
76. J. V. Joossens *et al.*, "Dietary salt, nitrate and stomach cancer mortality in 24 countries. European Cancer Prevention (ECP) and the INTERSALT Cooperative Research Group," *Int. J. Epidemiol.* **25**, 494 (1996).
77. M. J. Hill, "Bacterial N-nitrosation and gastric carcinogenesis in humans," *Ital. J. Gastroenterol.* **23**, 17 (1991).
78. H. Shen *et al.*, "Polymorphisms of 5, 10 - methylenetetrahydrofolate reductase and risk of gastric cancer in a Chinese population: A case - control study," *International Journal of Cancer* **95**, 332 (2001).
79. W. You *et al.*, "Gastric dysplasia and gastric cancer: Helicobacter pylori, serum vitamin C, and other risk factors," *J. Natl. Cancer Inst.* **92**, 1607 (2000).
80. G. Block, B. Patterson, A. Subar, "Fruit, vegetables, and cancer prevention: a review of the epidemiological evidence." *Nutr. Cancer* **18**, 1 (1992).
81. K. A. Steinmetz, J. D. Potter, "Vegetables, Fruit, and Cancer Prevention:: A Review," *J. Am. Diet. Assoc.* **96**, 1027 (1996).
82. Z. F. Zhang *et al.*, "Adenocarcinomas of the esophagus and gastric cardia: the role of diet." *Nutr. Cancer* **27**, 298 (1997).
83. J. Trédaniel, P. Boffetta, E. Buiatti, R. Saracci, A. Hirsch, "Tobacco smoking and gastric cancer: Review and meta - analysis," *International Journal of Cancer* **72**, 565 (1997).
84. K. Sjödahl *et al.*, "Smoking and alcohol drinking in relation to risk of gastric cancer: A population - based, prospective cohort study," *International Journal of Cancer* **120**, 128 (2007).
85. N. Y. Sung *et al.*, "Smoking, alcohol and gastric cancer risk in Korean men: the National Health Insurance Corporation Study," *Br. J. Cancer* **97**, 700 (2007).

86. Y. T. H. P. van Duynhoven, R. De Jonge, "Transmission of Helicobacter pylori: for food?" *Bull. World Health Organ.* **79**, 455 (2001).
87. R. H. Hunt *et al.*, "Helicobacter pylori in developing countries," (2010).
88. J. Robin Warren, B. Marshall, "UNIDENTIFIED CURVED BACILLI ON GASTRIC EPITHELIUM IN ACTIVE CHRONIC GASTRITIS," *Lancet* **321**, 1273 (1983).
89. E. J. Kuipers, "Review article: Relationship between Helicobacter pylori, atrophic gastritis and gastric cancer," *Aliment. Pharmacol. Ther.* **12 Suppl 1**, 25 (1998).
90. P. Correa, W. Haenszel, C. Cuello, S. Tannenbaum, M. Archer, "A model for gastric cancer epidemiology," *Lancet* **2**, 58 (1975).
91. P. Correa, "A human model of gastric carcinogenesis," *Cancer Res.* **48**, 3554 (1988).
92. L. Fuccio *et al.*, "Meta-analysis: can Helicobacter pylori eradication treatment reduce the risk for gastric cancer?" *Ann. Intern. Med.* **151**, 121 (2009).
93. P. G. Murray, L. S. Young, "Themed issue: the biology and pathology of the Epstein-Barr virus," *Molecular Pathology* **53**, 219 (2000).
94. M. C. Camargo *et al.*, "Determinants of Epstein-Barr virus-positive gastric cancer: an international pooled analysis," *Br. J. Cancer*(2011).
95. M. Scartozzi *et al.*, "Molecular biology of sporadic gastric cancer: prognostic indicators and novel therapeutic approaches," *Cancer Treat. Rev.* **30**, 451 (2004).
96. B. Humar *et al.*, "Association of CDH1 haplotypes with susceptibility to sporadic diffuse gastric cancer," *Oncogene* **21**, 8192 (2002).
97. M. Tanaka, Y. Kitajima, S. Sato, K. Miyazaki, "Combined evaluation of mucin antigen and E-cadherin expression may help select patients with gastric cancer suitable for minimally invasive therapy," *Br. J. Surg.* **90**, 95 (2003).
98. W. El-Rifai, S. M. Powell, "Molecular biology of gastric cancer," *Semin. Radiat. Oncol.* **12**, 128 (2002).
99. L. A. Aaltonen, S. R. Hamilton, International Agency for Research on Cancer, World Health Organization, *Pathology and genetics of tumours of the digestive system* (IARC Press, Lyon, 2000), pp. 314.
100. P. LAUREN, "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.* **64**, 31 (1965).
101. Y. Yuasa, "Control of gut differentiation and intestinal-type gastric carcinogenesis," *Nature Reviews Cancer* **3**, 592 (2003).
102. M. Voutilainen, M. Färkkilä, M. Juhola, J. P. Mecklin, P. Sipponen, "Complete and incomplete intestinal metaplasia at the oesophagogastric junction: prevalences and associations with endoscopic erosive oesophagitis and gastritis," *Gut* **45**, 644 (1999).
103. W. C. You *et al.*, "Evolution of precancerous lesions in a rural Chinese population at high risk of gastric cancer," *International Journal of Cancer* **83**, 615 (1999).
104. M. I. Filipe *et al.*, "Incomplete sulphomucin-secreting intestinal metaplasia for gastric cancer. Preliminary data from a prospective study from three centres." *Gut* **26**, 1319 (1985).
105. N. S. Goldstein, K. J. Lewin, "Gastric epithelial dysplasia and adenoma: historical review and histological criteria for grading," *Hum. Pathol.* **28**, 127 (1997).
106. P. Tosi *et al.*, "Gastric intestinal metaplasia type III cases are classified as low - grade dysplasia on the basis of morphometry," *J. Pathol.* **169**, 73 (1993).

107. G. Y. Lauwers, R. H. Riddell, "Gastric epithelial dysplasia," *Gut* **45**, 784 (1999).
108. M. Lansdown, P. Quirke, M. F. Dixon, A. T. Axon, D. Johnston, "High grade dysplasia of the gastric mucosa: a marker for gastric carcinoma." *Gut* **31**, 977 (1990).
109. Y. Handa *et al.*, "Association of Helicobacter pylori and diffuse type gastric cancer," *J. Gastroenterol.* **31 Suppl 9**, 29 (1996).
110. L. Ghandur-Mnaymneh, J. Paz, E. Roldan, J. Cassady, "Dysplasia of nonmetaplastic gastric mucosa: a proposal for its classification and its possible relationship to diffuse-type gastric carcinoma," *Am. J. Surg. Pathol.* **12**, 96 (1988).
111. T. Oda *et al.*, "E-cadherin gene mutations in human gastric carcinoma cell lines." *Proceedings of the National Academy of Sciences* **91**, 1858 (1994).
112. K. Yoshiura *et al.*, "Silencing of the E-cadherin invasion-suppressor gene by CpG methylation in human carcinomas," *Proceedings of the National Academy of Sciences* **92**, 7416 (1995).
113. S. Hirohashi, "Inactivation of the E-cadherin-mediated cell adhesion system in human cancers," *Am. J. Pathol.* **153**, 333 (1998).
114. K. F. Becker *et al.*, "E-cadherin gene mutations provide clues to diffuse type gastric carcinomas," *Cancer Res.* **54**, 3845 (1994).
115. P. Guilford *et al.*, "E-cadherin germline mutations in familial gastric cancer," *Nature* **392**, 402 (1998).
116. H. T. Lynch, W. Grady, G. Suriano, D. Huntsman, "Gastric cancer: new genetic developments," *J. Surg. Oncol.* **90**, 114 (2005).
117. R. Taetle *et al.*, "Chromosome abnormalities in ovarian adenocarcinoma: I. Nonrandom chromosome abnormalities from 244 cases," *Genes, Chromosomes and Cancer* **25**, 290 (1999).
118. G. M. Brodeur, A. A. Tsiatis, D. L. Williams, F. W. Luthardt, A. A. Green, "Statistical analysis of cytogenetic abnormalities in human cancer cells," *Cancer Genet. Cytogenet.* **7**, 137 (1982).
119. F. Jiang *et al.*, "Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data," *Cancer Res.* **60**, 6503 (2000).
120. T. Kainu *et al.*, "Somatic deletions in hereditary breast cancers implicate 13q21 as a putative novel breast cancer susceptibility locus," *Proceedings of the National Academy of Sciences* **97**, 9603 (2000).
121. X. Li *et al.*, "– 8p12–23 and 20q Are Predictors of Subtypes and Metastatic Pathways in Colorectal Cancer: Construction of Tree Models Using Comparative Genomic Hybridization Data," *OMICS: A Journal of Integrative Biology* **15**, 37 (2011).
122. T. C. Poon *et al.*, "A tumor progression model for hepatocellular carcinoma: bioinformatic analysis of genomic data," *Gastroenterology* **131**, 1262 (2006).
123. F. Picard, S. Robin, M. Lavielle, C. Vaisse, J. J. Daudin, "A statistical approach for array CGH data analysis," *BMC Bioinformatics* **6**, 27 (2005).
124. Machine learning models for lung cancer classification using array comparative genomic hybridization. (American Medical Informatics Association, , 2002).
125. F. Menolascina *et al.*, "Hybrid Intelligent Data Mining Techniques and Array CGH in Breast Cancer Profiling," *Proceedings of 2006 Workshop on Intelligent Computing & Bioinformatics of CAS*, pp. 93-99 (2006).
126. M. Höglund *et al.*, "Multivariate analyses of genomic imbalances in solid tumors reveal distinct and converging pathways of karyotypic evolution," *Genes, Chromosomes and Cancer* **31**, 156 (2001).

127. M. Höglund *et al.*, "Identification of cytogenetic subgroups and karyotypic pathways in transitional cell carcinoma," *Cancer Res.* **61**, 8241 (2001).
128. M. Höglund, D. Gisselsson, T. Säll, F. Mitelman, "Coping with complexity:: multivariate analysis of tumor karyotypes," *Cancer Genet. Cytogenet.* **135**, 103 (2002).
129. M. Höglund, D. Gisselsson, G. B. Hansen, T. Säll, F. Mitelman, "Multivariate analysis of chromosomal imbalances in breast cancer delineates cytogenetic pathways and reveals complex relationships among imbalances," *Cancer Res.* **62**, 2675 (2002).
130. M. Höglund *et al.*, "Dissecting karyotypic patterns in malignant melanomas: temporal clustering of losses and gains in melanoma karyotypic evolution," *International Journal of Cancer* **108**, 57 (2004).
131. M. R. Teixeira *et al.*, "Assessment of clonal relationships in ipsilateral and bilateral multiple breast carcinomas by comparative genomic hybridisation and hierarchical clustering analysis," *Br. J. Cancer* **91**, 775 (2004).
132. J. A. Cruz, D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informatics* **2**, 59 (2006).
133. T. Mattfeldt, H. Wolter, R. Kemmerling, H. W. Gottfried, H. A. Kestler, "Cluster analysis of comparative genomic hybridization (CGH) data using self-organizing maps: application to prostate carcinomas," *Analytical Cellular Pathology* **23**, 29 (2001).
134. T. Mattfeldt, D. Trijic, H. W. Gottfried, H. A. Kestler, "Incidental carcinoma of the prostate: clinicopathological, stereological and immunohistochemical findings studied with logistic regression and self - organizing feature maps," *BJU Int.* **93**, 284 (2004).
135. J. Liu, S. Ranka, T. Kahveci, "Classification and feature selection algorithms for multi-class CGH data," *Bioinformatics* **24**, i86 (2008).
136. L. G. D. L. Fraga, J. M. Carazo, H. C. Wang, J. Dopazo, Y. P. Zhu, "Self - organizing tree - growing network for the classification of protein sequences," *Protein Science* **7**, 2613 (1998).
137. R. Desper *et al.*, "Distance-based reconstruction of tree models for oncogenesis," *Journal of Computational Biology* **7**, 789 (2000).
138. R. Simon *et al.*, "Chromosome abnormalities in ovarian adenocarcinoma: III. Using breakpoint data to infer and test mathematical models for oncogenesis," *Genes, Chromosomes and Cancer* **28**, 106 (2000).
139. Q. Huang *et al.*, "Genetic differences detected by comparative genomic hybridization in head and neck squamous cell carcinomas from different tumor sites: construction of oncogenetic trees for tumor progression," *Genes, Chromosomes and Cancer* **34**, 224 (2002).
140. Z. Huang *et al.*, "Construction of tree models for pathogenesis of nasopharyngeal carcinoma," *Genes, Chromosomes and Cancer* **40**, 307 (2004).
141. J. W. Nicol, G. A. Helt, S. G. Blanchard, A. Raja, A. E. Loraine, "The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets," *Bioinformatics* **25**, 2730 (2009).
142. J. T. Robinson *et al.*, "Integrative genomics viewer," *Nat. Biotechnol.* **29**, 24 (2011).
143. D. Karolchik *et al.*, "The UCSC genome browser database," *Nucleic Acids Res.* **31**, 51 (2003).
144. J. Stalker *et al.*, "The Ensembl Web site: mechanics of a genome browser," *Genome Res.* **14**, 951 (2004).

145. L. D. Stein *et al.*, "The generic genome browser: a building block for a model organism system database," *Genome Res.* **12**, 1599 (2002).
146. M. Mader, R. Simon, S. Steinbiss, S. Kurtz, "FISH Oracle: a web server for flexible visualization of DNA copy number data in a genomic context," *Journal of Clinical Bioinformatics* **1**, 20 (2011).
147. M. E. Skinner, A. V. Uzilov, L. D. Stein, C. J. Mungall, I. H. Holmes, "JBrowse: A next-generation genome browser," *Genome Res.* **19**, 1630 (2009).
148. T. Toyoda, Y. Mochizuki, K. Player, N. K. Heida, "OmicBrowse: a browser of multidimensional omics annotations," *Bioinformatics* **23**, 524 (2007).
149. A. Kokkola *et al.*, "Presence of high-level DNA copy number gains in gastric carcinoma and severely dysplastic adenomas but not in moderately dysplastic adenomas," *Cancer Genet. Cytogenet.* **107**, 32 (1998).
150. T. E. Buffart *et al.*, "DNA copy number profiles of gastric cancer precursor lesions," *BMC Genomics* **8**, 345 (2007).
151. A. Frigyesi, D. Gisselsson, F. Mitelman, M. Höglund, "Power law distribution of chromosome aberrations in cancer," *Cancer Res.* **63**, 7094 (2003).
152. P. M. Kim *et al.*, "Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history," *Genome Res.* **18**, 1865 (2008).
153. A. G. Fritz, *International classification of diseases for oncology: ICD-O* (World Health Organization, , 2000).
154. J. Zhang, L. Feuk, G. E. Duggan, R. Khaja, S. W. Scherer, "Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome," *Cytogenet. Genome Res.* **115**, 205 (2006).
155. J. A. Bailey *et al.*, "Recent segmental duplications in the human genome," *Science* **297**, 1003 (2002).
156. A. J. Swerdlow, M. J. Schoemaker, C. D. Higgins, A. F. Wright, P. A. Jacobs, "Cancer risk in patients with constitutional chromosome deletions: a nationwide British cohort study," *Br. J. Cancer* **98**, 1929 (2008).
157. M. P. Curado, International Agency for Research on Cancer, *Cancer incidence in five continents* (International Agency for Research on Cancer, , 2008).
158. A. L. Barabási, R. Albert, "Emergence of scaling in random networks," *Science* **286**, 509 (1999).
159. E. J. Gumbel, *Statistics of extremes* (Columbia University Press, New York, ed. 1st Ed., 1958).
160. M. R. Leadbetter, G. Lindgren, H. Rootzen, "Extremes and related properties of random sequences and processes." (1983).
161. J. H. Gillespie, "A simple stochastic gene substitution model," *Theor. Popul. Biol.* **23**, 202 (1983).
162. J. H. Gillespie, "Molecular evolution over the mutational landscape," *Evolution* **38**, 1116 (1984).
163. H. A. Orr, "The genetic theory of adaptation: a brief history," *Nature Reviews Genetics* **6**, 119 (2005).
164. K. D. Robertson, "DNA methylation and human disease," *Nature Reviews Genetics* **6**, 597 (2005).
165. A. Narayan *et al.*, "Hypomethylation of pericentromeric DNA in breast adenocarcinomas," *International Journal of Cancer* **77**, 833 (1998).
166. N. Wong *et al.*, "Hypomethylation of chromosome 1 heterochromatin DNA correlates with q-arm copy gain in human hepatocellular carcinoma," *Am. J. Pathol.* **159**, 465 (2001).

167. J. Camps *et al.*, "Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome," *Cancer Res.* **68**, 1284 (2008).
168. K. D. Makova, S. Yang, F. Chiaromonte, "Insertions and deletions are male biased too: a whole-genome analysis in rodents," *Genome Res.* **14**, 567 (2004).
169. D. F. Conrad *et al.*, "Origins and functional impact of copy number variation in the human genome," *Nature* **464**, 704 (2009).
170. E. V. Koonin, Y. I. Wolf, G. P. Karev, "The structure of the protein universe and genome evolution," *Nature* **420**, 218 (2002).
171. G. P. Karev, F. S. Berezovskaya, E. V. Koonin, "Modeling genome evolution with a diffusion approximation of a birth-and-death process," *Bioinformatics* **21**, iii12 (2005).
172. P. J. Hastings, G. Ira, J. R. Lupski, "A microhomology-mediated break-induced replication model for the origin of human copy number variation," *PLoS Genetics* **5**, e1000327 (2009).
173. Y. Zhang *et al.*, "Genomic DNA breakpoints in AML1/RUNX1 and ETO cluster with topoisomerase II DNA cleavage and DNase I hypersensitive sites in t (8; 21) leukemia," *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3070 (2002).
174. Y. Zhang *et al.*, "Characterization of genomic breakpoints in MLL and CBP in leukemia patients with t (11; 16)," *Genes, Chromosomes and Cancer* **41**, 257 (2004).
175. M. W. Djojosubroto, Y. S. Choi, H. W. Lee, K. L. Rudolph, "Telomeres and telomerase in aging, regeneration and cancer," *Mol. Cells* **15**, 164 (2003).
176. P. Willeit *et al.*, "Telomere length and risk of incident cancer and cancer mortality," *JAMA: The Journal of the American Medical Association* **304**, 69 (2010).
177. S. Mayer *et al.*, "Sex-specific telomere length profiles and age-dependent erosion dynamics of individual chromosome arms in humans," *Cytogenetic and Genome Research* **112**, 194 (2006).
178. P. Möller *et al.*, "Sex-related differences in length and erosion dynamics of human telomeres favor females," *Aging (Albany NY)* **1**, 733 (2009).
179. W. E. Naugler *et al.*, "Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production," *Science* **317**, 121 (2007).
180. S. Rakoff-Nahoum, R. Medzhitov, "Regulation of spontaneous intestinal tumorigenesis through the adaptor protein MyD88," *Science* **317**, 124 (2007).
181. A. Ohtsu, N. Fuse, T. Yoshino, M. Tahara, T. Doi, "Future perspectives of chemotherapy for advanced gastric cancer," *Gastric Cancer* **12**, 60 (2009).
182. A. D. Panani, "Cytogenetic and molecular aspects of gastric cancer: clinical implications," *Cancer Lett.* **266**, 99 (2008).
183. N. C. van Grieken *et al.*, "Helicobacter pylori-related and -non-related gastric cancers do not differ with respect to chromosomal aberrations," *J. Pathol.* **192**, 301 (2000).
184. C. W. Wu *et al.*, "Clinical implications of chromosomal abnormalities in gastric adenocarcinomas," *Genes Chromosomes Cancer* **35**, 219 (2002).
185. M. Nannini *et al.*, "Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives," *Cancer Treat. Rev.* **35**, 201 (2009).
186. M. Baudis, "cancer genome data @ progenetix.org," **9/8/2007**.
187. H. van Dekken *et al.*, "Molecular cytogenetic evaluation of gastric cardia adenocarcinoma and precursor lesions," *Am. J. Pathol.* **158**, 1961 (2001).
188. M. M. Weiss *et al.*, "Genome wide array comparative genomic hybridisation analysis of premalignant lesions of the stomach," *Mol. Pathol.* **56**, 293 (2003).

189. W. Y. Chan *et al.*, "Recurrent genomic aberrations in gastric carcinomas associated with Helicobacter pylori and Epstein-Barr virus," *Diagn. Mol. Pathol.* **11**, 127 (2002).
190. A. Kokkola *et al.*, "17q12-21 Amplicon, a Novel Recurrent Genetic Change in Intestinal Type of Gastric Carcinoma: a Comparative Genomic Hybridization Study," *Genes Chromosomes Cancer* **20**, 38 (1997).
191. M. L. Larramendy *et al.*, "Comparative genomic hybridization reveals differences in DNA copy number changes between sporadic gastric carcinomas and gastric carcinomas from patients with hereditary nonpolyposis colorectal cancer," *Cancer Genet. Cytogenet.* **106**, 62 (1998).
192. T. Noguchi, H. C. Wirtz, S. Michaelis, H. E. Gabbert, W. Mueller, "Chromosomal imbalances in gastric cancer. Correlation with histologic subtypes and tumor progression," *Am. J. Clin. Pathol.* **115**, 828 (2001).
193. A. Oga, G. Kong, Y. Ishii, H. Izumi, K. Chang Young Park Sasaki, "Preferential loss of 5q14-21 in intestinal-type gastric cancer with DNA aneuploidy." *Communications in Clinical Cytometry* **46**, 57 (2001).
194. W. El-Rifai *et al.*, "Genetic differences between adenocarcinomas arising in Barrett's esophagus and gastric mucosa," *Gastroenterology* **121**, 592 (2001).
195. R. Sud, D. Wells, I. C. Talbot, J. D. Delhanty, "Genetic alterations in gastric cancers from British patients," *Cancer Genet. Cytogenet.* **126**, 111 (2001).
196. S. T. Tay *et al.*, "A combined comparative genomic hybridization and expression microarray analysis of gastric cancer reveals novel molecular subtypes," *Cancer Res.* **63**, 3309 (2003).
197. A. Varis *et al.*, "DNA copy number changes in young gastric cancer patients with special reference to chromosome 19," *Br. J. Cancer* **88**, 1914 (2003).
198. D. F. Peng, H. Sugihara, K. Mukaisho, Y. Tsubosa, T. Hattori, "Alterations of chromosomal copy number during progression of diffuse-type gastric carcinomas: metaphase- and array-based comparative genomic hybridization analyses of multiple samples from individual tumours," *J. Pathol.* **201**, 439 (2003).
199. A. I. Saeed *et al.*, "TM4: a free, open-source system for microarray data management and analysis," *BioTechniques* **34**, 374 (2003).
200. A. I. Saeed *et al.*, "TM4 microarray software suite," *Methods Enzymol.* **411**, 134 (2006).
201. L. D. Stein *et al.*, "The Generic Genome Browser: A building block for a model organism system database." *Genome Res.* **12**, 1599 (2002).
202. A. Boussioutas *et al.*, "Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer," *Cancer Res.* **63**, 2569 (2003).
203. L. I. Gomes *et al.*, "Expression profile of malignant and nonmalignant lesions of esophagus and stomach: differential activity of functional modules related to inflammation and lipid metabolism," *Cancer Res.* **65**, 7127 (2005).
204. S. Y. Leung *et al.*, "Phospholipase A2 group IIA expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis," *Proceedings of the National Academy of Sciences* **99**, 16203 (2002).
205. Y. Hippo *et al.*, "Global gene expression analysis of gastric cancer by oligonucleotide microarrays," *Cancer Res.* **62**, 233 (2002).
206. S. Y. Kim *et al.*, "Meta-and gene set analysis of stomach cancer gene expression data." *Mol. Cells* **24**, 200 (2007).
207. National Center for Biotechnology Information, "Map Viewer," **2008**(2008).

208. K. J. Livak, T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative PCR and the 2-[Delta][Delta] CT method," *Methods* **25**, 402 (2001).
209. R Development Core Team, *R: A language and environment for statistical computing.*(R Foundation for Statistical Computing, Vienna, Austria, 2011).
210. S. Y. Kanta *et al.*, "Topoisomerase II [alpha] gene amplification in gastric carcinomas: correlation with the HER2 gene. An immunohistochemical, immunoblotting, and multicolor fluorescence in situ hybridization study," *Hum. Pathol.* **37**, 1333 (2006).
211. H. G. Woo *et al.*, "Identification of potential driver genes in human liver carcinoma by genomewide screening," *Cancer Res.* **69**, 4059 (2009).
212. M. Shibuya, J. Yokota, Y. Ueyama, "Amplification and expression of a cellular oncogene (c-myc) in human gastric adenocarcinoma cells." *Mol. Cell. Biol.* **5**, 414 (1985).
213. G. Nardone *et al.*, "Effect of Helicobacter pyloriinfection and its eradication on cell proliferation, DNA status, and oncogene expression in patients with chronic gastritis," *Gut* **44**, 789 (1999).
214. A. Varis *et al.*, "Targets of gene amplification and overexpression at 17q in gastric cancer," *Cancer Res.* **62**, 2625 (2002).
215. S. Tseng, C. Chang, K. Wu, S. Teng, "Importin KPNA2 is required for proper nuclear localization and multiple functions of NBS1," *J. Biol. Chem.* **280**, 39594 (2005).
216. L. Zannini *et al.*, "Karyopherin-alpha2 protein interacts with Chk2 and contributes to its nuclear import." *J. Biol. Chem.* **278**, 42346 (2003).

PUBLICATIONS

1. **Lam MT**, Ho LHP, Wong N, Choy KW, Lai PBS, Johnson PJ, Sung JJY, Poon TCW. Bioinformatic characterization of chromosomal breakages in 14,322 Cancer Cases and Association with Sex Differences in Cancer Incidence. (Submitted)

2. **Lam MT**, Ho LHP, Wong N, Lai PBS, Johnson PJ, Sung JJY, Poon TCW. Tumor progression models for intestinal and diffuse types of gastric cancer: meta-analysis of chromosomal imbalances. (Submitted)

RAW DATA

The raw data collected for this research project is available on request.

APPENDIX

Appendix 1 Perl script for classifying CGH data from Progenetix database according to ICD-O-3 code and other criteria

```
#!/usr/bin/perl
#
# filter.pl
# Little utility to format the CGH data downloaded from Progenetix
# 1) to extract the CGH data according to the ICD-O-3 code
# 2) to filter sex-related cancers
# 3) to include only male or female data
#
# Developed by LAM Man Ting
# Department of Medicine and Therapeutics
# The Chinese University of Hong Kong
#
# For the Partial Fulfillment of the Requirements for
# the Degree of Doctor of Philosophy in Medical Sciences
#

if ($#ARGV < 0) {
   print "Usage: perl filter.pl filter(0/1) flag(0/1) sex(m/f/a)\n"; exit;}

my $filter = $ARGV[0];
#filter = 0 for all types except adenocarcinoma and squamous cell carcinoma,
# 1 for adenocarcinoma and squamous cell carcinoma cases only

my $flag = $ARGV[1]; #flag=1 to exclude sex-related cancers
my $sex = $ARGV[2]; #sex=m or f to filter by patients' gender, a to include all cases

my $filename;
if ($filter ==0){ system("md type"); $filename ="icdo3list.txt"; }
#list of icd-o-3 cacner codes
else {$filename ="icdo3list2.txt";}
#list of icd-o-3 for Adenocarcinoma and squamous cell carcinoma to group
# adenocarcinoma and squamous cell carcinoma from different tumour site

open (ICDO, "<$filename") or die "Cannot open file: $filename\n";
my @type = <ICDO>;
foreach (@type) { chomp;}
close ICDO;

my %hash = (); #database
open ICDODB, "icdo.txt" or die "Cannot open icdo.txt : $!";
my @icdo = <ICDODB>;
close ICDODB;
foreach (@icdo) { chomp;my ($key, $value) = split(/\t/,$_);$hash{ $key } = $value;}

my @lines;
open (FILE, "<862_bands_ninf_no_aberrationmatrix.txt") or die "Cannot open file:
862_bands_ninf_no_aberrationmatrix.txt\n";
```

```
# file downloaded from progenetix database

my @tmp = <FILE>;
my $headline=$tmp[0];

for my $k (1..$#tmp) {
chomp;
my $tag =1;
my $current = $tmp[$k];
if ($sex eq "a"){if ($current!~m/\tmale\t/){if $current!~m/\tfemale\t/){next;} } }
if ($sex eq "m"){if ($current!~m/\tmale\t/){next;} }
if ($sex eq "f"){
                    if ($current=~m/cervix ca./ or $current=~m/uterus ca./ or
            $current=~m/vulva ca./ ){$tag=0;}
if ($tag){if ($current!~m/\tfemale\t/){next;} }
}
# Removal of gender specific cancer

if ($flag){
if ($current=~m/cervix ca./ or $current=~m/breast ca./ or $current=~m/ovarial ca./ or
$current=~m/prostate adenoca./ or $current=~m/uterus ca./ or $current=~m/vulva
ca./ or $current=~m/Germ cell neoplasms/){next;}
if ($current=~m|8384/3| or $current=~m|8380/3| or $current=~m|9100/3| or
$current=~m|8500/3| or $current=~m|8501/3| or $current=~m|8620/3| or
$current=~m|9063/3| or $current=~m|9071/3| or $current=~m|8930/3| or
$current=~m|8931/3|){next;}}

push(@lines,$current);
}
close FILE;

chomp $headline;
my @header = split (/\t/, $headline);
my $titleline=extract(@header);

for my $r(0..$#type){

$current = $type[$r];
my $dir = "\"$hash{$current}\"";
my $dir2 = $hash{$current};
$dir=~s|\/|_|;  $dir2=~s|\/|_|;
system("md type\\$dir");
open (OUT, ">type\\$dir2\\$dir2.txt") or die "Cannot open type\\$dir2\\$dir2.txt\n";
print OUT "$titleline\n";
print "$r\r";

for my $i(0..$#lines){my @cell = split (/\t/, $lines[$i]);my $qtype = $cell[867];
#grouping of related cancer into a single group with large number of cancer cases

if ($qtype =~m|9826/3|){$qtype =a;}if ($qtype =~m|9835/3|){$qtype =a;}
```

```
if ($qtype =~m|9874/3|){$qtype =b;}if ($qtype =~m|9872/3|){$qtype =b;}if ($qtype
=~m|9873/3|){$qtype =b;}
if ($qtype =~m|9895/3|){$qtype =c;}if ($qtype =~m|9861/3|){$qtype =c;}
if ($qtype =~m|9867/3|){$qtype =d;}if ($qtype =~m|9871/3|){$qtype =d;}
if ($qtype =~m|9718/3|){$qtype =e;}if ($qtype =~m|9714/3|){$qtype =e;}
if ($qtype =~m|8930/3|){$qtype =f;}if ($qtype =~m|8931/3|){$qtype =f;}
if ($qtype =~m|9695/3|){$qtype =g;}if ($qtype =~m|9691/3|){$qtype =g;}
if ($qtype =~m|9698/3|){$qtype =g;}if ($qtype =~m|9690/3|){$qtype =g;}
if ($qtype =~m|8970/3|){$qtype =h;}if ($qtype =~m|8171/3|){$qtype =h;}
if ($qtype =~m|9702/3|){$qtype =i;}if ($qtype =~m|9709/3|){$qtype =i;}
if ($qtype =~m|8761/3|){$qtype =j;}if ($qtype =~m|8720/3|){$qtype =j;}
if ($qtype =~m|9053/3|){$qtype =k;}if ($qtype =~m|9050/3|){$qtype =k;}
if ($qtype =~m|9663/3|){$qtype =965;}
if ($qtype =~m|8312/3|){$qtype =l;}if ($qtype =~m|8317/3|){$qtype =l;}if ($qtype
=~m|8319/3|){$qtype =l;}

if ($qtype =~m/^$current/){ my $out = extract(@cell); print OUT "$out\n";}
}
}

sub extract{
my @input=@_;
my @output;
for my $m(0..869){push(@output, $input[$m]);}
push(@output, $input[872]);
my $outline=join("\t", @output);
return $outline;
}
```

Appendix 2 Perl script for classifying CGH data from Progenetix database according to ICD-O-3 code and other criteria

```
#!/usr/bin/perl
#
# split_bp.pl
#!/usr/bin/perl
#
# Little utility to format the CGH data downloaded from Progenetix
# 1) extract the CGH data according to the chromosome
# 2) convert these data to breakpoints
#
# Developed by LAM Man Ting
# Department of Medicine and Therapeutics
# The Chinese University of Hong Kong
#
# For the Partial Fulfillment of the Requirements for
# the Degree of Doctor of Philosophy in Medical Sciences
#

#use strict;
#use warnings;

if ($#ARGV < 0) { print "Usage: perl split_bp.pl file\n"; exit;}
$file = $ARGV[0];
open (FILE, "<$file") or die "Cannot open flip table file: $file\n";
@lines = <FILE>;
$headline = shift @lines;
chomp $headline;
@header = split (/\t/, $headline);

for ($i = 1; $i <= $#header; $i++) { #chr 1234556....xy
      if ($header[$i] =~ /^cY/) {$loc{$i} = "cy";}
# split to 1p and 1q to prevent selecting c10-c19
elsif ($header[$i] =~ /^c1p/){$loc{$i} = "c1";}
elsif ($header[$i] =~ /^c1q/){$loc{$i} = "c1";}
elsif ($header[$i] =~ /^c10/){ $loc{$i} = "c10";}
elsif ($header[$i] =~ /^c11/){ $loc{$i} = "c11";}
elsif ($header[$i] =~ /^c12/){ $loc{$i} = "c12";}
elsif ($header[$i] =~ /^c13/){ $loc{$i} = "c13";}
elsif ($header[$i] =~ /^c14/){ $loc{$i} = "c14";}
elsif ($header[$i] =~ /^c15/){ $loc{$i} = "c15";}
elsif ($header[$i] =~ /^c16/){ $loc{$i} = "c16";}
elsif ($header[$i] =~ /^c17/){ $loc{$i} = "c17";}
elsif ($header[$i] =~ /^c18/){ $loc{$i} = "c18";}
elsif ($header[$i] =~ /^c19/){ $loc{$i} = "c19";}
# split to 2p and 2q to prevent selecting c20,c21,c22
          elsif ($header[$i] =~ /^c2p/){$loc{$i} = "c2";}
elsif ($header[$i] =~ /^c2q/){ $loc{$i} = "c2";}
elsif ($header[$i] =~ /^c20/){ $loc{$i} = "c20";}
```

```
elsif ($header[$i] =~ /^c21/){ $loc{$i} = "c21";}
elsif ($header[$i] =~ /^c22/){ $loc{$i} = "c22";}
elsif ($header[$i] =~ /^c3/){ $loc{$i} = "c3";}
elsif ($header[$i] =~ /^c4/){ $loc{$i} = "c4";}
elsif ($header[$i] =~ /^c5/){ $loc{$i} = "c5";}
elsif ($header[$i] =~ /^c6/){ $loc{$i} = "c6";}
elsif ($header[$i] =~ /^c7/){ $loc{$i} = "c7";}
elsif ($header[$i] =~ /^c8/){ $loc{$i} = "c8";}
elsif ($header[$i] =~ /^c9/){ $loc{$i} = "c9";}
elsif ($header[$i] =~ /^cX/){ $loc{$i} = "cx";}
   }

open GCHR1, ">gchr1.txt" or die "unable to open chr1.txt $!";
open GCHR2, ">gchr2.txt" or die "unable to open chr2.txt $!";
open GCHR3, ">gchr3.txt" or die "unable to open chr3.txt $!";
open GCHR4, ">gchr4.txt" or die "unable to open chr4.txt $!";
open GCHR5, ">gchr5.txt" or die "unable to open chr5.txt $!";
open GCHR6, ">gchr6.txt" or die "unable to open chr6.txt $!";
open GCHR7, ">gchr7.txt" or die "unable to open chr7.txt $!";
open GCHR8, ">gchr8.txt" or die "unable to open chr8.txt $!";
open GCHR9, ">gchr9.txt" or die "unable to open chr9.txt $!";
open GCHR10, ">gchr10.txt" or die "unable to open chr10.txt $!";
open GCHR11, ">gchr11.txt" or die "unable to open chr11.txt $!";
open GCHR12, ">gchr12.txt" or die "unable to open chr12.txt $!";
open GCHR13, ">gchr13.txt" or die "unable to open chr13.txt $!";
open GCHR14, ">gchr14.txt" or die "unable to open chr14.txt $!";
open GCHR15, ">gchr15.txt" or die "unable to open chr15.txt $!";
open GCHR16, ">gchr16.txt" or die "unable to open chr16.txt $!";
open GCHR17, ">gchr17.txt" or die "unable to open chr17.txt $!";
open GCHR18, ">gchr18.txt" or die "unable to open chr18.txt $!";
open GCHR19, ">gchr19.txt" or die "unable to open chr19.txt $!";
open GCHR20, ">gchr20.txt" or die "unable to open chr20.txt $!";
open GCHR21, ">gchr21.txt" or die "unable to open chr21.txt $!";
open GCHR22, ">gchr22.txt" or die "unable to open chr22.txt $!";
open GCHRX, ">gchrx.txt" or die "unable to open chrx.txt $!";
open GCHRY, ">gchry.txt" or die "unable to open chry.txt $!";

open LCHR1, ">lchr1.txt" or die "unable to open chr1.txt $!";
open LCHR2, ">lchr2.txt" or die "unable to open chr2.txt $!";
open LCHR3, ">lchr3.txt" or die "unable to open chr3.txt $!";
open LCHR4, ">lchr4.txt" or die "unable to open chr4.txt $!";
open LCHR5, ">lchr5.txt" or die "unable to open chr5.txt $!";
open LCHR6, ">lchr6.txt" or die "unable to open chr6.txt $!";
open LCHR7, ">lchr7.txt" or die "unable to open chr7.txt $!";
open LCHR8, ">lchr8.txt" or die "unable to open chr8.txt $!";
open LCHR9, ">lchr9.txt" or die "unable to open chr9.txt $!";
open LCHR10, ">lchr10.txt" or die "unable to open chr10.txt $!";
open LCHR11, ">lchr11.txt" or die "unable to open chr11.txt $!";
open LCHR12, ">lchr12.txt" or die "unable to open chr12.txt $!";
open LCHR13, ">lchr13.txt" or die "unable to open chr13.txt $!";
```

```
open LCHR14, ">lchr14.txt" or die "unable to open chr14.txt $!";
open LCHR15, ">lchr15.txt" or die "unable to open chr15.txt $!";
open LCHR16, ">lchr16.txt" or die "unable to open chr16.txt $!";
open LCHR17, ">lchr17.txt" or die "unable to open chr17.txt $!";
open LCHR18, ">lchr18.txt" or die "unable to open chr18.txt $!";
open LCHR19, ">lchr19.txt" or die "unable to open chr19.txt $!";
open LCHR20, ">lchr20.txt" or die "unable to open chr20.txt $!";
open LCHR21, ">lchr21.txt" or die "unable to open chr21.txt $!";
open LCHR22, ">lchr22.txt" or die "unable to open chr22.txt $!";
open LCHRX, ">lchrx.txt" or die "unable to open chrx.txt $!";
open LCHRY, ">lchry.txt" or die "unable to open chry.txt $!";
####################
# for printing the header for debug
 for ($i = 1; $i <= $#header; $i++) {
      if ($loc{$i} eq "c1" && $header[$i] ne "") {
      print GCHR1 "$header[$i]\t";print LCHR1 "$header[$i]\t";}
      elsif ($loc{$i} eq "c2" && $header[$i] ne "") {
      print GCHR2 "$header[$i]\t";print LCHR2 "$header[$i]\t";}
      elsif ($loc{$i} eq "c3" && $header[$i] ne "") {
      print GCHR3 "$header[$i]\t";print LCHR3 "$header[$i]\t";}
      elsif ($loc{$i} eq "c4" && $header[$i] ne "") {
      print GCHR4 "$header[$i]\t";print LCHR4 "$header[$i]\t";}
      elsif ($loc{$i} eq "c5" && $header[$i] ne "") {
      print GCHR5 "$header[$i]\t";print LCHR5 "$header[$i]\t";}
      elsif ($loc{$i} eq "c6" && $header[$i] ne "") {
      print GCHR6 "$header[$i]\t";print LCHR6 "$header[$i]\t";}
      elsif ($loc{$i} eq "c7" && $header[$i] ne "") {
      print GCHR7 "$header[$i]\t";print LCHR7 "$header[$i]\t";}
      elsif ($loc{$i} eq "c8" && $header[$i] ne "") {
      print GCHR8 "$header[$i]\t";print LCHR8 "$header[$i]\t";}
      elsif ($loc{$i} eq "c9" && $header[$i] ne "") {
      print GCHR9 "$header[$i]\t";print LCHR9 "$header[$i]\t";}
      elsif ($loc{$i} eq "c10" && $header[$i] ne "") {
      print GCHR10 "$header[$i]\t";print LCHR10 "$header[$i]\t";}
      elsif ($loc{$i} eq "c11" && $header[$i] ne "") {
      print GCHR11 "$header[$i]\t";print LCHR11 "$header[$i]\t";}
      elsif ($loc{$i} eq "c12" && $header[$i] ne "") {
      print GCHR12 "$header[$i]\t";print LCHR12 "$header[$i]\t";}
      elsif ($loc{$i} eq "c13" && $header[$i] ne "") {
      print GCHR13 "$header[$i]\t";print LCHR13 "$header[$i]\t";}
      elsif ($loc{$i} eq "c14" && $header[$i] ne "") {
      print GCHR14 "$header[$i]\t";print LCHR14 "$header[$i]\t";}
      elsif ($loc{$i} eq "c15" && $header[$i] ne "") {
      print GCHR15 "$header[$i]\t";print LCHR15 "$header[$i]\t";}
      elsif ($loc{$i} eq "c16" && $header[$i] ne "") {
      print GCHR16 "$header[$i]\t";print LCHR16 "$header[$i]\t";}
      elsif ($loc{$i} eq "c17" && $header[$i] ne "") {
      print GCHR17 "$header[$i]\t";print LCHR17 "$header[$i]\t";}
      elsif ($loc{$i} eq "c18" && $header[$i] ne "") {
      print GCHR18 "$header[$i]\t";print LCHR18 "$header[$i]\t";}
```

```
        elsif ($loc{$i} eq "c19" && $header[$i] ne "") {
        print GCHR19 "$header[$i]\t";print LCHR19 "$header[$i]\t";}
        elsif ($loc{$i} eq "c20" && $header[$i] ne "") {
        print GCHR20 "$header[$i]\t";print LCHR20 "$header[$i]\t";}
        elsif ($loc{$i} eq "c21" && $header[$i] ne "") {
        print GCHR21 "$header[$i]\t";print LCHR21 "$header[$i]\t";}
        elsif ($loc{$i} eq "c22" && $header[$i] ne "") {
        print GCHR22 "$header[$i]\t";print LCHR22 "$header[$i]\t";}
        elsif ($loc{$i} eq "cx" && $header[$i] ne "") {
        print GCHRX "$header[$i]\t";print LCHRX "$header[$i]\t";}
        elsif ($loc{$i} eq "cy" && $header[$i] ne "") {
        print GCHRY "$header[$i]\t";print LCHRY "$header[$i]\t";}
}
print GCHR1 "\n"; print LCHR1 "\n"; print GCHR2 "\n"; print LCHR2 "\n";
print GCHR3 "\n"; print LCHR3 "\n"; print GCHR4 "\n"; print LCHR4 "\n";
print GCHR5 "\n"; print LCHR5 "\n"; print GCHR6 "\n"; print LCHR6 "\n";
print GCHR7 "\n"; print LCHR7 "\n"; print GCHR8 "\n"; print LCHR8 "\n";
print GCHR9 "\n"; print LCHR9 "\n"; print GCHR10 "\n"; print LCHR10 "\n";
print GCHR11 "\n"; print LCHR11 "\n"; print GCHR12 "\n"; print LCHR12 "\n";
print GCHR13 "\n"; print LCHR13 "\n"; print GCHR14 "\n"; print LCHR14 "\n";
print GCHR15 "\n"; print LCHR15 "\n"; print GCHR16 "\n"; print LCHR16 "\n";
print GCHR17 "\n"; print LCHR17 "\n"; print GCHR18 "\n"; print LCHR18 "\n";
print GCHR19 "\n"; print LCHR19 "\n"; print GCHR20 "\n"; print LCHR20 "\n";
print GCHR21 "\n"; print LCHR21 "\n"; print GCHR22 "\n"; print LCHR22 "\n";
print GCHRX "\n"; print LCHRX "\n"; print GCHRY "\n"; print LCHRY "\n";
######################

# Conversion of CGH data to breakpoint
foreach $line (@lines) {
   chomp $line;
   @data = split (/\t/, $line);
      for ($i = 1; $i <= $#header; $i++) {
         if ($loc{$i} eq "cy" && $data[$i] ne "") {
         if ($data[$i] == 1){print GCHRY "$data[$i]\t";print LCHRY "0\t";}
       elsif ($data[$i] == -1){print GCHRY "0\t";print LCHRY "1\t";}
          else{print GCHRY "$data[$i]\t";print LCHRY "$data[$i]\t";}
           }
          elsif ($loc{$i} eq "c1" && $data[$i] ne "") {
           if ($data[$i] == 1){print GCHR1 "$data[$i]\t";print LCHR1 "0\t";}
           elsif ($data[$i] == -1){print GCHR1 "0\t";print LCHR1 "1\t";}
           else{print GCHR1 "$data[$i]\t";print LCHR1 "$data[$i]\t";}
           }
elsif ($loc{$i} eq "c2" && $data[$i] ne "") {
          if ($data[$i] == 1){print GCHR2 "$data[$i]\t"; print LCHR2 "0\t";}
           elsif ($data[$i] == -1){print GCHR2 "0\t";print LCHR2 "1\t";}
           else{print GCHR2 "$data[$i]\t"; print LCHR2 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c3" && $data[$i] ne "") {
           if ($data[$i] == 1){print GCHR3 "$data[$i]\t" ;print LCHR3 "0\t";}
           elsif ($data[$i] == -1){print GCHR3 "0\t";print LCHR3 "1\t";}
```

```
            else{print GCHR3 "$data[$i]\t";print LCHR3 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c4" && $data[$i] ne "") {
            if ($data[$i] == 1){print GCHR4 "$data[$i]\t";print LCHR4 "0\t";}
            elsif ($data[$i] == -1){print GCHR4 "0\t";print LCHR4 "1\t";}
            else{print GCHR4 "$data[$i]\t";print LCHR4 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c5" && $data[$i] ne "") {
            if ($data[$i] == 1){print GCHR5 "$data[$i]\t";print LCHR5 "0\t";}
            elsif ($data[$i] == -1){print GCHR5 "0\t";print LCHR5 "1\t";}
            else{print GCHR5 "$data[$i]\t";print LCHR5 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c6" && $data[$i] ne "") {
            if ($data[$i] == 1){print GCHR6 "$data[$i]\t";print LCHR6 "0\t";}
            elsif ($data[$i] == -1){print GCHR6 "0\t";print LCHR6 "1\t";}
            else{print GCHR6 "$data[$i]\t";print LCHR6 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c7" && $data[$i] ne "") {
            if ($data[$i] == 1){print GCHR7 "$data[$i]\ "; print LCHR7 "0\t";}
            elsif ($data[$i] == -1){print GCHR7 "0\t"; print LCHR7 "1\t";}
            else{print GCHR7 "$data[$i]\t"; print LCHR7 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c8" && $data[$i] ne "") {
            if ($data[$i] == 1){print GCHR8 "$data[$i]\t"; print LCHR8 "0";}
            elsif ($data[$i] == -1){print GCHR8 "0\t"; print LCHR8 "1\t";}
            else{print GCHR8 "$data[$i]\t"; print LCHR8 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c9" && $data[$i] ne "") {
            if ($data[$i] == 1){print GCHR9 "$data[$i]\t"; print LCHR9 "0\t";}
            elsif ($data[$i] == -1){print GCHR9 "0\t"; print LCHR9 "1\t";}
            else{print GCHR9 "$data[$i]\t"; print LCHR9 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c10" && $data[$i] ne "") {
            if ($data[$i] == 1){print GCHR10 "$data[$i]\t";print LCHR10 "0\t";}
            elsif ($data[$i] == -1){print GCHR10 "0\t";print LCHR10 "1\t";}
            else{print GCHR10 "$data[$i]\t";print LCHR10 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c11" && $data[$i] ne "") {
            if ($data[$i] == 1){print GCHR11 "$data[$i]\t";print LCHR11 "0\t";}
            elsif ($data[$i] == -1){print GCHR11 "0\t";print LCHR11 "1\t";}
            else{print GCHR11 "$data[$i]\t";print LCHR11 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c12" && $data[$i] ne "") {
            if ($data[$i] == 1){print GCHR12 "$data[$i]\t";print LCHR12 "0\t";}
            elsif ($data[$i] == -1){print GCHR12 "0\t";print LCHR12 "1\t";}
            else{print GCHR12 "$data[$i]\t";print LCHR12 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c13" && $data[$i] ne "") {
          if ($data[$i] == 1){print GCHR13 "$data[$i]\t";print LCHR13 "0\t";}
          elsif ($data[$i] == -1){print GCHR13 "0\t";print LCHR13 "1\t";}
```

```
        else{print GCHR13 "$data[$i]\t";print LCHR13 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c14" && $data[$i] ne "") {
        if ($data[$i] == 1){print GCHR14 "$data[$i]\t";print LCHR14 "0\t";}
        elsif ($data[$i] == -1){print GCHR14 "0\t";print LCHR14 "1\t";}
        else{print GCHR14 "$data[$i]\t";print LCHR14 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c15" && $data[$i] ne "") {
        if ($data[$i] == 1){print GCHR15 "$data[$i]\t";print LCHR15 "0\t";}
        elsif ($data[$i] == -1){print GCHR15 "0\t";print LCHR15 "1\t";}
        else{print GCHR15 "$data[$i]\t";print LCHR15 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c16" && $data[$i] ne "") {
        if ($data[$i] == 1){print GCHR16 "$data[$i]\t";print LCHR16 "0\t";}
        elsif ($data[$i] == -1){print GCHR16 "0\t";print LCHR16 "1\t";}
        else{print GCHR16 "$data[$i]\t";print LCHR16 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c17" && $data[$i] ne "") {
        if ($data[$i] == 1){print GCHR17 "$data[$i]\t";print LCHR17 "0\t";}
        elsif ($data[$i] == -1){print GCHR17 "0\t";print LCHR17 "1\t";}
       else{print GCHR17 "$data[$i]\t";print LCHR17 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c18" && $data[$i] ne "") {
        if ($data[$i] == 1){print GCHR18 "$data[$i]\t";print LCHR18 "0\t";}
        elsif ($data[$i] == -1){print GCHR18 "0\t";print LCHR18 "1\t";}
        else{print GCHR18 "$data[$i]\t";print LCHR18 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c19" && $data[$i] ne "") {
         if ($data[$i] == 1){print GCHR19 "$data[$i]\t";print LCHR19 "0\t";}
         elsif ($data[$i] == -1){print GCHR19 "0\t";print LCHR19 "1\t";}
         else{print GCHR19 "$data[$i]\t";print LCHR19 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c20" && $data[$i] ne "") {
        if ($data[$i] == 1){print GCHR20 "$data[$i]\t";print LCHR20 "0\t";}
        elsif ($data[$i] == -1){print GCHR20 "0\t";print LCHR20 "1\t";}
        else{print GCHR20 "$data[$i]\t";print LCHR20 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c21" && $data[$i] ne "") {
         if ($data[$i] == 1){print GCHR21 "$data[$i]\t";print LCHR21 "0\t";}
         elsif ($data[$i] == -1){print GCHR21 "0\t";print LCHR21 "1\t";}
         else{print GCHR21 "$data[$i]\t";print LCHR21 "$data[$i]\t";}
}
elsif ($loc{$i} eq "c22" && $data[$i] ne "") {
         if ($data[$i] == 1){print GCHR22 "$data[$i]\t";print LCHR22 "0\t";}
         elsif ($data[$i] == -1){print GCHR22 "0\t";print LCHR22 "1\t";}
         else{print GCHR22 "$data[$i]\t";print LCHR22 "$data[$i]\t";}
}
elsif ($loc{$i} eq "cx" && $data[$i] ne "") {
         if ($data[$i] == 1){print GCHRX "$data[$i]\t";print LCHRX "0\t";}
         elsif ($data[$i] == -1){print GCHRX "0\t";print LCHRX "1\t";}
```

```
            else{print GCHRX "$data[$i]\t";print LCHRX "$data[$i]\t";}
            }
        }
print GCHR1 "\n"; print LCHR1 "\n"; print GCHR2 "\n"; print LCHR2 "\n";
print GCHR3 "\n"; print LCHR3 "\n"; print GCHR4 "\n"; print LCHR4 "\n";
print GCHR5 "\n"; print LCHR5 "\n"; print GCHR6 "\n"; print LCHR6 "\n";
print GCHR7 "\n"; print LCHR7 "\n"; print GCHR8 "\n"; print LCHR8 "\n";
print GCHR9 "\n"; print LCHR9 "\n"; print GCHR10 "\n"; print LCHR10 "\n";
print GCHR11 "\n"; print LCHR11 "\n"; print GCHR12 "\n"; print LCHR12 "\n";
print GCHR13 "\n"; print LCHR13 "\n"; print GCHR14 "\n"; print LCHR14 "\n";
print GCHR15 "\n"; print LCHR15 "\n"; print GCHR16 "\n"; print LCHR16 "\n";
print GCHR17 "\n"; print LCHR17 "\n"; print GCHR18 "\n"; print LCHR18 "\n";
print GCHR19 "\n"; print LCHR19 "\n"; print GCHR20 "\n"; print LCHR20 "\n";
print GCHR21 "\n"; print LCHR21 "\n"; print GCHR22 "\n"; print LCHR22 "\n";
print GCHRX "\n"; print LCHRX "\n"; print GCHRY "\n"; print LCHRY "\n";
}
close GCHR1;close LCHR1; close GCHR2;close LCHR2;
close GCHR3;close LCHR3; close GCHR4;close LCHR4;
close GCHR5;close LCHR5; close GCHR6;close LCHR6;
close GCHR7;close LCHR7; close GCHR8;close LCHR8;
close GCHR9;close LCHR9; close GCHR10; close LCHR10;
close GCHR11; close LCHR11; close GCHR12; close LCHR12;
close GCHR13; close LCHR13; close GCHR14; close LCHR14;
close GCHR15; close LCHR15; close GCHR16; close LCHR16;
close GCHR17; close LCHR17; close GCHR18; close LCHR18;
close GCHR19; close LCHR19; close GCHR20; close LCHR20;
close GCHR21; close LCHR21; close GCHR22; close LCHR22;
close GCHRX; close LCHRX; close GCHRY; close LCHRY;

my @input =
("gchr1.txt","gchr2.txt","gchr3.txt","gchr4.txt","gchr5.txt","gchr6.txt","gchr7.txt","g
chr8.txt","gchr9.txt","gchr10.txt","gchr11.txt","gchr12.txt","gchr13.txt","gchr14.txt",
"gchr15.txt","gchr16.txt","gchr17.txt","gchr18.txt","gchr19.txt","gchr20.txt","gchr21
.txt","gchr22.txt","gchrx.txt","gchry.txt","lchr1.txt","lchr2.txt","lchr3.txt","lchr4.txt",
"lchr5.txt","lchr6.txt","lchr7.txt","lchr8.txt","lchr9.txt","lchr10.txt","lchr11.txt","lchr
12.txt","lchr13.txt","lchr14.txt","lchr15.txt","lchr16.txt","lchr17.txt","lchr18.txt","lch
r19.txt","lchr20.txt","lchr21.txt","lchr22.txt","lchrx.txt","lchry.txt");
my @output =
("bp_gchr1.txt","bp_gchr2.txt","bp_gchr3.txt","bp_gchr4.txt","bp_gchr5.txt","bp_gc
hr6.txt","bp_gchr7.txt","bp_gchr8.txt","bp_gchr9.txt","bp_gchr10.txt","bp_gchr11.tx
t","bp_gchr12.txt","bp_gchr13.txt","bp_gchr14.txt","bp_gchr15.txt","bp_gchr16.txt"
,"bp_gchr17.txt","bp_gchr18.txt","bp_gchr19.txt","bp_gchr20.txt","bp_gchr21.txt","
bp_gchr22.txt","bp_gchrx.txt","bp_gchry.txt","bp_lchr1.txt","bp_lchr2.txt","bp_lchr
3.txt","bp_lchr4.txt","bp_lchr5.txt","bp_lchr6.txt","bp_lchr7.txt","bp_lchr8.txt","bp_
lchr9.txt","bp_lchr10.txt","bp_lchr11.txt","bp_lchr12.txt","bp_lchr13.txt","bp_lchr1
4.txt","bp_lchr15.txt","bp_lchr16.txt","bp_lchr17.txt","bp_lchr18.txt","bp_lchr19.txt
","bp_lchr20.txt","bp_lchr21.txt","bp_lchr22.txt","bp_lchrx.txt","bp_lchry.txt");
################
my @locus=(); #for storing the chr index in the first column of bpcomp.txt
```

```
my @locusloc=(); #for storing the locus index (name of that 400 locus) in the 2nd
column of bpcomp.txt
my @locuscount=(); #for storing the count index (no of 862 locus in that 400 locus)
in the 3rd column of bpcomp.txt
my $previous=0;

open (BPCOMP, "<bpcomp.txt") or die "Cannot open file: bpcomp.txt\n";
my @bpcomp = <BPCOMP>;
foreach (@bpcomp) {chomp;my ($p1, $p2,$p3)=
split("\t",$_);push(@locusloc,$p1);push(@locus,$p2);push(@locuscount,$p3);}
close BPCOMP;

################
for my $i(0..$#input){
open (FILE, "<$input[$i]") or die "Cannot open flip table file: $input[$i]\n";
open (OUT, ">$output[$i]") or die "Cannot open flip table file: $output[$i]\n";
my @lines = <FILE>;
foreach (@lines) { chomp;}
close FILE;

my $header = shift @lines;
@title = split (/\t/, $header);
my $final=$#title-1;

my @titlebar=();
for my $t(0..$#locusloc){
my $cnt;
if($i<24){ $cnt = $i;}
else {$cnt=$i-24;} # For the $i>24 the loss ones
if ($locusloc[$t] == $cnt){push (@titlebar,$locus[$t]);}
}

my $head=join("\t", @titlebar);
print OUT "$head\n";
my @bpdata=();

for my $m(0..$#lines){
my @p;
@data=split (/\t/, $lines[$m]);
@data=reverse(@data);#flip the table so that it begins with p arm
for my $c(0..$final){
$d=$c+1;
$p[$c]=$data[$c]-$data[$d]; #Breakpoint calculation
}

my $bp=join("\t", @p)."\t0";
if ($bp=~m/-1\t0/){$bp =~ s/-1\t0/0\t1/g;} #
if ($bp=~m/-1\t1/){$bp =~ s/-1\t1/0\t2/g;}
        # single locus abberation -> two breakpoints
@bpdata=split ("\t", $bp);
```

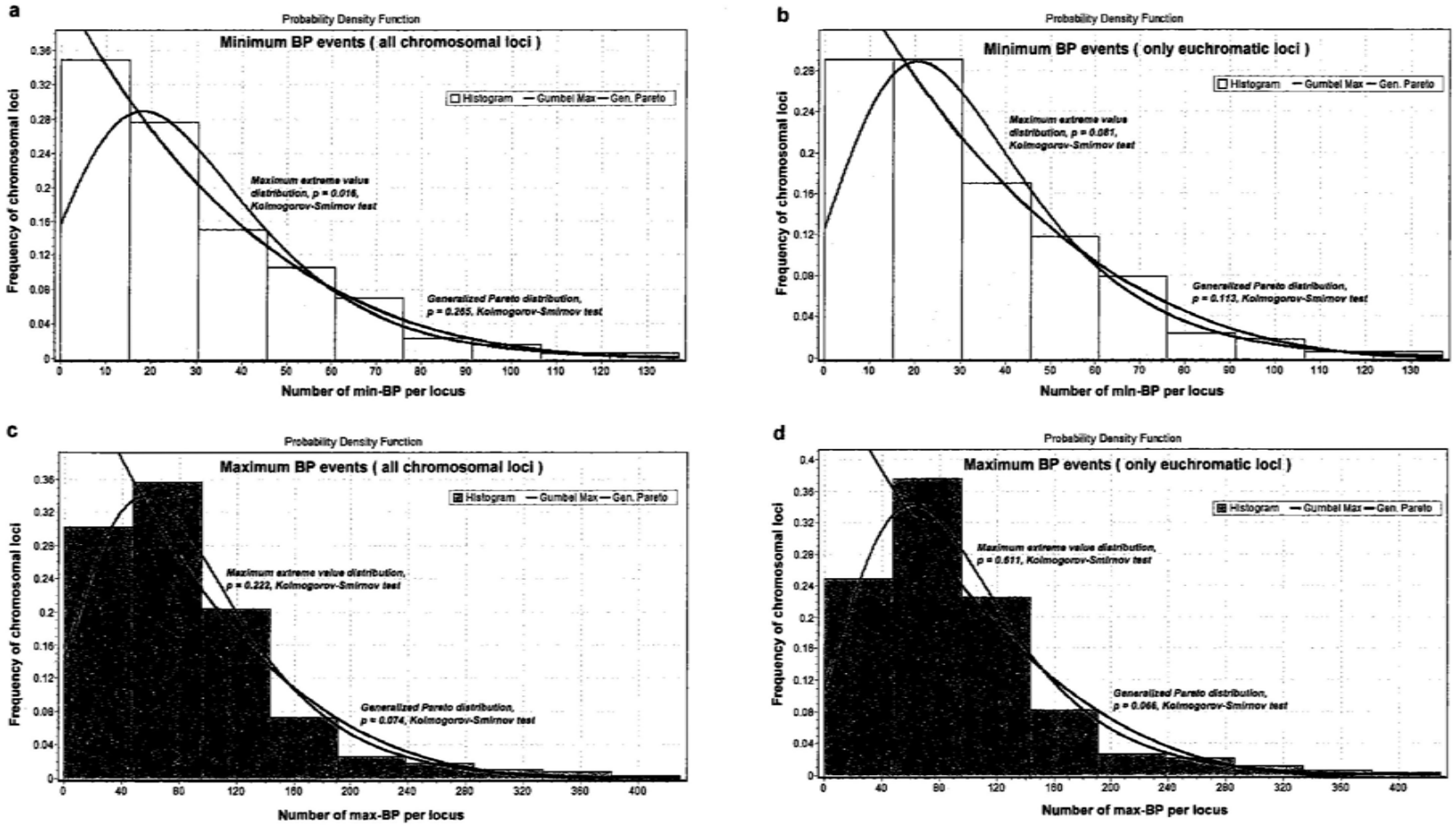
```
 #split the flipped breakpoints again for grouping into 400-band resolution
my @count=();

for my $t(0..$#locusloc){
my $cnt;
if($i<24){ $cnt = $i;}
else {$cnt=$i-24;} # For the $i>24 the loss ones
if ($locusloc[$t] == $cnt){push (@count,$locuscount[$t]);}
}

my @bpcount=(); #Array for storing the grouped 400-band breakpoints
for my $pp(0..$#count){

if ($pp == 0){$previous = 0};
#Reset the starting pointer for looping through the array(list) of 862-band
breakpoints
for my $l($previous..$count[$pp]){
                        $bpcount[$pp] = $bpcount[$pp]+$bpdata[$l];}
$previous = $count[$pp]+1; # Update the pointer

#Print the 400-band breakpoints out
if($pp == $#count){
my $bpall=join("\t", @bpcount);
print OUT "$bpall\n";
}
}
}
close OUT;
}
# Tidy up and group gain and loss to separate folders
system("md lchr"); system("move lchr*.txt lchr");
system("md gchr"); system("move gchr*.txt gchr");
system("md bpl"); system("move bp_l*.txt bpl");
system("md bpg"); system("move bp_g*.txt bpg;
```



Appendix 3 The distributions of minimum breakpoint (min-BP) events (a,b) and maximum breakpoint (max-BP) events (c,d) among all chromosomal loci (a,c) or only the euchromatic loci (b,d) of 2,906 aCGH cases. The best-fit curves and their associated *p*-values for goodness of fit are provided. A *p*-value > 0.1 indicates that the model fitting is statistically significant.

Appendix 4 Matlab scripts for identification of significant chromosome aberration events with the estimation of false discovery rate

```
% filename: batch.m
% A "marco" script for running the tool to
% identify significant chromosome aberration events
% with false discovery rate estimation
%
% Developed by LAM, Man Ting
% Department of Medicine and Therapeutics
% The Chinese University of Hong Kong
%
% For the Partial Fulfillment of the Requirements for
% the Degree of Doctor of Philosophy in Medical Sciences
%
clear all
% directory storing the CGH data
cd ('D:\CGH\matlab\GA')

% file named "GA.txt" contains the data from gastric adenoma
load GA.txt

% repeat 5 times
for L = 1:5
clear b
clear c
clear p
clear f

tic
[b,r,c,p,f] = cghrandom08(GA); % refering t the name of the input file
folder=['1koutput' num2str(L)];
mkdir(folder);
bootfreq= ['1koutput' num2str(L) '/bootfreq.txt'];
randfreq= ['1koutput' num2str(L) '/randfreq.txt'];
compare= ['1koutput' num2str(L) '/compare.txt'];
pval= ['1koutput' num2str(L) '/pval.txt'];
fdr= ['1koutput' num2str(L) '/fdr.txt'];
timespent= ['1koutput' num2str(L) '/timespent.txt'];

dlmwrite(bootfreq, b, 'delimiter', '\t','precision', 8)
dlmwrite(randfreq, r, 'delimiter', '\t','precision', 8)
dlmwrite(compare, c, 'delimiter', '\t','precision', 8)
dlmwrite(pval, p, 'delimiter', '\t','precision', 8)
dlmwrite(fdr, f, 'delimiter', '\t','precision', 8)
t=toc
dlmwrite(timespent, t)
end
delete('randout.txt');
delete('sortrandout.txt');
```

```
% filename: cghrandom08.m
%
% Little utility to identify significant chromosome aberration events
% with false discovery rate estimation
% Based on the script by Professor Terence Poon in [122]
% Implement of false discovery rate estimation by LAM, Man Ting
%
% Department of Medicine and Therapeutics
% The Chinese University of Hong Kong
%
% For the Partial Fulfillment of the Requirements for
% the Degree of Doctor of Philosophy in Medical Sciences
%

 [bootfreq,randfreq,compare, pvalue, FDR] = cghrandom08(indata)

CGHdata = indata;
[r,P]=size(indata);
pCGH = indata(:,P);
CGHdata(:,P) = [];
%indata =[];
[r,c] = size(CGHdata);

CNAsum = sum(CGHdata);
clear CGHdata;
%mkdir('1koutput')
%pCGH = CNAfreq./totalCNA;

nSimulation = 10000;
%mkdir('Rand_matrix') may write the random stimulations out for reference
RandomCount = 0;
RandomSimulation = 0;
randCNAfreq=zeros(nSimulation,r);
for s = 1:nSimulation;
    [tempCGHdata, RandomCounter] = random(CNAsum, pCGH);
    RandomCount = RandomCount + RandomCounter;
    randCNAfreq(s,:) = (sum(tempCGHdata,2))';
    RandomSimulation = RandomSimulation + 1;
    s
    clear tempCGHdata;
end


clear simuData;
randout=randCNAfreq';     %add here. This is the mean from each random
stimulation
dlmwrite('randout.txt', randout, 'delimiter', '\t','precision', 6) %%write the random
matrix out for reference
sortrandout = (sort(randCNAfreq))'; % add here, output the sorted mean from
random stimulation
```

```
dlmwrite('sortrandout.txt', sortrandout, 'delimiter', '\t','precision', 6) %%write the
random matrix out for reference
clear randout;
clear sortrandout;

randresult(1,:) = mean(randCNAfreq);
randresult(2,:) = std(randCNAfreq);
randdistribution = prctile(randCNAfreq,[1 2.5 5 10 25 50 75 90 95 95.01 95.02
95.03 95.04 95.05 95.06 95.07 95.08 95.09 95.1 95.11 95.12 95.13 95.14 95.15
95.16 95.17 95.18 95.19 95.2 95.21 95.22 95.23 95.24 95.25 95.26 95.27 95.28
95.29 95.3 95.31 95.32 95.33 95.34 95.35 95.36 95.37 95.38 95.39 95.4 95.41 95.42
95.43 95.44 95.45 95.46 95.47 95.48 95.49 95.5 95.51 95.52 95.53 95.54 95.55
95.56 95.57 95.58 95.59 95.6 95.61 95.62 95.63 95.64 95.65 95.66 95.67 95.68
95.69 95.7 95.71 95.72 95.73 95.74 95.75 95.76 95.77 95.78 95.79 95.8 95.81 95.82
95.83 95.84 95.85 95.86 95.87 95.88 95.89 95.9 95.91 95.92 95.93 95.94 95.95
95.96 95.97 95.98 95.99 96 96.01 96.02 96.03 96.04 96.05 96.06 96.07 96.08 96.09
96.1 96.11 96.12 96.13 96.14 96.15 96.16 96.17 96.18 96.19 96.2 96.21 96.22 96.23
96.24 96.25 96.26 96.27 96.28 96.29 96.3 96.31 96.32 96.33 96.34 96.35 96.36
96.37 96.38 96.39 96.4 96.41 96.42 96.43 96.44 96.45 96.46 96.47 96.48 96.49 96.5
96.51 96.52 96.53 96.54 96.55 96.56 96.57 96.58 96.59 96.6 96.61 96.62 96.63
96.64 96.65 96.66 96.67 96.68 96.69 96.7 96.71 96.72 96.73 96.74 96.75 96.76
96.77 96.78 96.79 96.8 96.81 96.82 96.83 96.84 96.85 96.86 96.87 96.88 96.89 96.9
96.91 96.92 96.93 96.94 96.95 96.96 96.97 96.98 96.99 97 97.01 97.02 97.03 97.04
97.05 97.06 97.07 97.08 97.09 97.1 97.11 97.12 97.13 97.14 97.15 97.16 97.17
97.18 97.19 97.2 97.21 97.22 97.23 97.24 97.25 97.26 97.27 97.28 97.29 97.3 97.31
97.32 97.33 97.34 97.35 97.36 97.37 97.38 97.39 97.4 97.41 97.42 97.43 97.44
97.45 97.46 97.47 97.48 97.49 97.5 97.51 97.52 97.53 97.54 97.55 97.56 97.57
97.58 97.59 97.6 97.61 97.62 97.63 97.64 97.65 97.66 97.67 97.68 97.69 97.7 97.71
97.72 97.73 97.74 97.75 97.76 97.77 97.78 97.79 97.8 97.81 97.82 97.83 97.84
97.85 97.86 97.87 97.88 97.89 97.9 97.91 97.92 97.93 97.94 97.95 97.96 97.97
97.98 97.99 98 98.01 98.02 98.03 98.04 98.05 98.06 98.07 98.08 98.09 98.1 98.11
98.12 98.13 98.14 98.15 98.16 98.17 98.18 98.19 98.2 98.21 98.22 98.23 98.24
98.25 98.26 98.27 98.28 98.29 98.3 98.31 98.32 98.33 98.34 98.35 98.36 98.37
98.38 98.39 98.4 98.41 98.42 98.43 98.44 98.45 98.46 98.47 98.48 98.49 98.5 98.51
98.52 98.53 98.54 98.55 98.56 98.57 98.58 98.59 98.6 98.61 98.62 98.63 98.64
98.65 98.66 98.67 98.68 98.69 98.7 98.71 98.72 98.73 98.74 98.75 98.76 98.77
98.78 98.79 98.8 98.81 98.82 98.83 98.84 98.85 98.86 98.87 98.88 98.89 98.9 98.91
98.92 98.93 98.94 98.95 98.96 98.97 98.98 98.99 99 99.01 99.02 99.03 99.04 99.05
99.06 99.07 99.08 99.09 99.1 99.11 99.12 99.13 99.14 99.15 99.16 99.17 99.18
99.19 99.2 99.21 99.22 99.23 99.24 99.25 99.26 99.27 99.28 99.29 99.3 99.31 99.32
99.33 99.34 99.35 99.36 99.37 99.38 99.39 99.4 99.41 99.42 99.43 99.44 99.45
99.46 99.47 99.48 99.49 99.5 99.51 99.52 99.53 99.54 99.55 99.56 99.57 99.58
99.59 99.6 99.61 99.62 99.63 99.64 99.65 99.66 99.67 99.68 99.69 99.7 99.71 99.72
99.73 99.74 99.75 99.76 99.77 99.78 99.79 99.8 99.81 99.82 99.83 99.84 99.85
99.86 99.87 99.88 99.89 99.9 99.91 99.92 99.93 99.94 99.95 99.96 99.97 99.98
99.99 100]);
randresult = [randdistribution; randresult];
clear randdistribution;
clear randCNAfreq
randfreq = randresult';
```

```
clear randresult;

% Obtain the descriptive statistics of the bootstrapped values

nboot = 10000;
bootmatrix = zeros(c,c);
bootsam = [];
for b = 1:nboot
for n = 1:c
% store the random sample number is each row
bootmatrix(n,:) = randperm(c);
    end
    % use the values in the first column as a bootstrap sample set
tempboot = bootmatrix(:,1);
    % store the bootstrap sample set into each row of the bootsam matrix
bootsam(b,:)=tempboot';
end

Bootstrapping = 0;
dataholder = [];
bootCNAfreq = [];
CGHdata = indata;
for n = 1:nboot
    %bootCNAfreq = [];
    for m = 1:c
        samplevalue = bootsam(n,m);
        dataholder(:,m) = CGHdata(:,samplevalue);
    end
        bCNAfreq = (sum(dataholder,2))';
        bootCNAfreq(n,:) = bCNAfreq;
        Bootstrapping = Bootstrapping + 1;
end
clear Bootstrapping;
clear dataholder;
bootresult(1,:) = mean(bootCNAfreq);
bootresult(2,:) = std(bootCNAfreq);
bootdistribution = prctile(bootCNAfreq,[1 2.5 5 10 25 50 75 90 95 97.5 99]);
bootresult = [bootdistribution; bootresult];
bootfreq = bootresult';
clear bootresult;
clear bootdistribution;

% Find out the p-value!!!
% The automatic comparison of the bootstrap mean (observed mean)
% to the different quartiles of the distribution obtained from the random stimulations

% round(the bootstrap mean ( observed mean ))
bootmean = round(mean(bootCNAfreq)');
clear bootCNAfreq;
```

```
compare = zeros(r,509);

for n = 1:r
    for m = 1:509
%randfreq 90, p = 0.1
        if bootmean(n)> randfreq(n,m)
            compare(n,m)=1;
        else
            break;
        end
    end
end
obs_sig=sum(compare);

pval = zeros(1,509);
load 'sortrandout.txt';
for n = 1:r
    for m = 1:nSimulation
        m
        if ((bootmean(n)<sortrandout(n,m)) && (m ~= 1))
            pval(n)=1-((m-1)/nSimulation);
            break;
        elseif ((bootmean(n)<sortrandout(n,m)) && (m == 1))
pval(n) = 1;
            break;
        elseif (bootmean(n)== sortrandout(n,m))
            pval(n)=1-((m-1)/nSimulation);
            break;
        elseif ((bootmean(n)> sortrandout(n,m)) && (m ~= nSimulation))
continue;
        elseif bootmean(n)> sortrandout(n,m)
pval(n)=1-((m-1)/nSimulation);
            break;
        end
    end
end
clear sortrandout;
clear bootmean;

pvalue=pval';
clear pval;
load 'header.txt';
compare = [header';compare];

%
% Estimation of the false discovery rate
%

sumrand = zeros(nSimulation,509);
load 'randout.txt';
```

```
for i=1:nSimulation
      randmean = randout(:,i);
      comparerand=zeros(r,509);
      i
      for n = 1:r
            for m = 1:509
% randfreq 90, p = 0.1
                  if randmean(n)> randfreq(n,m)

% Compare the rand mean to the random percentiles from 90-100
comparerand(n,m)=1;

                  else
       % If randmean value not greater than the value at this smaller percentile
break;
                  end% it must not be greater than those with higher percentile
            end
      end
      sumrand(i,:)=sum(comparerand);
end
clear comparerand;
clear randout;
mean_fpos=round(mean(sumrand));
rand_sig=[sumrand; mean(sumrand)];
clear sumrand;
FDRout = zeros(1,509);
for i=1:509
      FDRout(i) = mean_fpos(i)/(obs_sig(i) + mean_fpos(i));
end

fdrresult(1,:) = mean(FDRwSD);
fdrresult(2,:) = std(FDRwSD);
fdrdistribution = prctile(FDRwSD,[1 2.5 5 10 25 50 75 90 95 97.5 99]);
fdrresult = [fdrdistribution; fdrresult];
FDRMSD = fdrresult';
clear fdrresult;
clear fdrdistribution;
end

%
% Output the results
%

temp=([obs_sig;mean_fpos;FDRout])';
FDR = [header,temp];
clear header;
clear temp;
clear obs_sig;
clear mean_fpos;
clear FDRout;
```

```
%
% random generator
%

function [randData, RandomCounter] = random(CNAsum, pCGH)

cases = size(CNAsum,2);
r = size(pCGH,1);
%cases = c;
accumProb = pCGH;
clear pCGH;
for L = 1:r
      if L > 1
      accumProb(L) = accumProb(L) + accumProb(L-1);
      end
end

%accumProb = accumProb

loci = r;

data = zeros(loci, cases);

RandomCounter = 0;
for Case = 1:cases

      numCNAs = CNAsum(Case);
      CNA = [];
      check = 0;

      for L = 1:numCNAs
            while check < numCNAs
                  valueR = rand;
                  locus = find(accumProb > valueR);
                  if isempty(locus)
                        CNA = loci;
                  else
                        CNA = locus(1);
                  end

% Check of the locus is assigned to be gain and loss at the same time
                  if CNA > (r/2)
                        if data((CNA - r/2),Case)>0
                              break;
                        end
                  else
                        if data((CNA + r/2),Case)>0
                              break;
                        end
                  end
```

```
                data(CNA, Case) = 1;
            check = sum(data(:,Case));
            RandomCounter = RandomCounter + 1;
            end
        end
end
clear CNAsum;
randData = data;
clear data;
clear accumProb;
```

Appendix 5 List of per locus frequency and probability of significant chromosomal aberrations in GA, IGC and DGC cases

	GA (p-value cutoff at 5% FDR=0.0058)		IGC (p-value cutoff at 5% FDR=0.0185)		DGC (p-value cutoff at 5% FDR=0.0159)	
Locus	average p-value	Frequency in 41 cases	average p-value	Frequency in 198 cases	average p-value	Frequency in 117 cases
+1P36	0.37665	4.88%	0.10765	12.63%	0.00010	17.95%
+1P35	0.72132	2.44%	0.06403	12.63%	0.05687	11.11%
+1P34	0.68997	2.44%	0.05103	12.12%	0.01450	11.97%
+1P33	0.67245	2.44%	0.27155	9.60%	0.17932	8.55%
+1P32	1.00000	0.00%	0.32363	9.09%	0.15228	8.55%
+1P31	1.00000	0.00%	0.99317	5.05%	0.97097	3.42%
+1P22	1.00000	0.00%	0.99937	3.54%	0.98858	2.56%
+1P21	1.00000	0.00%	0.99987	3.03%	0.99983	0.85%
+1P13	1.00000	0.00%	0.99852	3.03%	0.99885	0.85%
+1P12	1.00000	0.00%	0.99333	3.03%	0.99710	0.85%
+1P11	1.00000	0.00%	0.98523	3.03%	0.99558	0.85%
+1Q11	1.00000	0.00%	0.00010	9.60%	0.00010	7.69%
+1Q12	1.00000	0.00%	0.00010	10.61%	0.00010	11.11%
+1Q21	1.00000	0.00%	0.00010	16.67%	0.00010	14.53%
+1Q22	1.00000	0.00%	0.00010	18.69%	0.00010	11.11%
+1Q23	1.00000	0.00%	0.00010	20.20%	0.00010	11.11%
+1Q24	1.00000	0.00%	0.00010	17.68%	0.00010	10.26%
+1Q25	1.00000	0.00%	0.00010	17.17%	0.00010	7.69%
+1Q31	1.00000	0.00%	0.00010	17.68%	0.00010	16.24%
+1Q32	0.32202	2.44%	0.00010	15.66%	0.00010	16.24%
+1Q41	0.34320	2.44%	0.00010	15.66%	0.00030	8.55%
+1Q42	0.36365	2.44%	0.00010	13.13%	0.00935	6.84%
+1Q43	0.35632	2.44%	0.00010	13.13%	0.02498	5.98%
+1Q44	0.35992	2.44%	0.00010	13.13%	0.02583	5.98%
+2P25	1.00000	0.00%	0.00390	7.58%	0.00137	8.55%
+2P24	1.00000	0.00%	0.00462	7.58%	0.00348	7.69%
+2P23	1.00000	0.00%	0.00928	7.07%	0.03132	5.98%
+2P22	1.00000	0.00%	0.00120	7.58%	0.01842	5.98%
+2P21	1.00000	0.00%	0.00125	7.58%	0.05422	5.13%
+2P16	1.00000	0.00%	0.00033	8.08%	0.01627	5.98%
+2P15	1.00000	0.00%	0.00038	8.08%	0.01600	5.98%
+2P14	1.00000	0.00%	0.00033	8.08%	0.01322	5.98%
+2P13	1.00000	0.00%	0.00525	6.57%	0.01253	5.98%
+2P12	1.00000	0.00%	0.00065	7.07%	0.02737	5.13%
+2P11	1.00000	0.00%	0.00055	7.07%	0.02518	5.13%
+2Q11	1.00000	0.00%	0.00437	7.07%	0.28885	3.42%
+2Q12	1.00000	0.00%	0.00753	7.07%	0.16107	4.27%
+2Q13	1.00000	0.00%	0.01138	7.07%	0.19373	4.27%
+2Q14	1.00000	0.00%	0.02365	7.07%	0.04800	5.98%
+2Q21	1.00000	0.00%	0.04883	7.07%	0.07960	5.98%
+2Q22	1.00000	0.00%	0.10335	7.07%	0.05573	6.84%

+2Q23	1.00000	0.00%	0.10193	7.07%	0.05975	6.84%
+2Q24	1.00000	0.00%	0.08203	7.58%	0.07273	6.84%
+2Q31	1.00000	0.00%	0.08382	7.58%	0.07148	6.84%
+2Q32	1.00000	0.00%	0.07858	7.58%	0.07110	6.84%
+2Q33	1.00000	0.00%	0.04578	8.08%	0.07257	6.84%
+2Q34	1.00000	0.00%	0.05813	8.08%	0.03897	7.69%
+2Q35	1.00000	0.00%	0.18455	7.07%	0.09155	6.84%
+2Q36	1.00000	0.00%	0.11970	7.58%	0.09150	6.84%
+2Q37	1.00000	0.00%	0.00373	10.10%	0.16632	5.98%
+3P26	1.00000	0.00%	1.00000	2.53%	0.89048	5.98%
+3P25	1.00000	0.00%	1.00000	2.53%	0.89770	5.98%
+3P24	1.00000	0.00%	0.99998	3.54%	0.89003	5.98%
+3P23	1.00000	0.00%	1.00000	3.54%	0.94460	5.13%
+3P22	1.00000	0.00%	1.00000	3.54%	0.94488	5.13%
+3P21	1.00000	0.00%	0.99960	5.05%	0.99378	3.42%
+3P14	1.00000	0.00%	0.99998	3.54%	0.99368	3.42%
+3P13	1.00000	0.00%	1.00000	4.04%	0.99238	3.42%
+3P12	1.00000	0.00%	0.99997	4.04%	0.99952	1.71%
+3P11	1.00000	0.00%	0.99978	3.54%	0.99863	1.71%
+3Q11	1.00000	0.00%	0.21978	7.58%	0.90363	2.56%
+3Q12	1.00000	0.00%	0.22370	7.58%	0.90678	2.56%
+3Q13	1.00000	0.00%	0.16250	8.08%	0.79962	3.42%
+3Q21	1.00000	0.00%	0.05378	9.09%	0.28083	5.98%
+3Q22	1.00000	0.00%	0.03932	9.09%	0.24645	5.98%
+3Q23	1.00000	0.00%	0.01937	9.60%	0.24522	5.98%
+3Q24	1.00000	0.00%	0.02180	9.60%	0.07107	7.69%
+3Q25	1.00000	0.00%	0.00412	10.61%	0.01093	9.40%
+3Q26	0.53250	2.44%	0.00063	11.62%	0.00032	11.97%
+3Q27	0.54200	2.44%	0.00030	12.12%	0.03047	8.55%
+3Q28	0.53272	2.44%	0.00053	11.62%	0.00378	10.26%
+3Q29	0.53457	2.44%	0.00148	11.11%	0.00365	10.26%
+4P16	1.00000	0.00%	0.99990	2.02%	1.00000	0.00%
+4P15	1.00000	0.00%	1.00000	1.52%	0.99975	0.85%
+4P14	1.00000	0.00%	1.00000	1.01%	1.00000	0.00%
+4P13	1.00000	0.00%	1.00000	1.01%	0.99960	0.85%
+4P12	1.00000	0.00%	1.00000	1.01%	0.99928	0.85%
+4P11	1.00000	0.00%	1.00000	1.01%	0.99932	0.85%
+4Q11	1.00000	0.00%	1.00000	1.52%	0.95715	3.42%
+4Q12	1.00000	0.00%	1.00000	2.02%	0.96020	3.42%
+4Q13	0.73895	2.44%	1.00000	2.02%	0.97422	3.42%
+4Q21	0.75395	2.44%	1.00000	2.02%	0.89622	5.13%
+4Q22	1.00000	0.00%	1.00000	1.52%	0.96122	4.27%
+4Q23	1.00000	0.00%	1.00000	1.52%	0.98765	3.42%
+4Q24	1.00000	0.00%	1.00000	1.52%	0.96812	4.27%
+4Q25	1.00000	0.00%	1.00000	1.52%	0.96628	4.27%
+4Q26	1.00000	0.00%	1.00000	1.52%	0.92680	5.13%
+4Q27	1.00000	0.00%	1.00000	1.52%	0.91845	5.13%
+4Q28	1.00000	0.00%	1.00000	1.52%	0.96258	4.27%

+4Q31	1.00000	0.00%	1.00000	1.01%	0.73332	6.84%
+4Q32	1.00000	0.00%	1.00000	1.52%	0.93865	5.13%
+4Q33	1.00000	0.00%	1.00000	1.52%	0.86402	5.98%
+4Q34	1.00000	0.00%	1.00000	1.52%	0.93288	5.13%
+4Q35	1.00000	0.00%	1.00000	1.01%	0.97108	4.27%
+5P15	0.03805	4.88%	0.00010	10.61%	0.01750	5.13%
+5P14	0.04137	4.88%	0.00010	10.10%	0.00052	7.69%
+5P13	0.22463	2.44%	0.00010	9.09%	0.00022	6.84%
+5P12	0.20772	2.44%	0.00010	9.09%	0.00505	5.13%
+5P11	1.00000	0.00%	0.00010	7.07%	0.00307	5.13%
+5Q11	1.00000	0.00%	0.98962	2.02%	0.80182	2.56%
+5Q12	1.00000	0.00%	0.99780	2.02%	0.88083	2.56%
+5Q13	1.00000	0.00%	0.99932	2.02%	0.64435	4.27%
+5Q14	1.00000	0.00%	0.99977	2.02%	0.59083	5.13%
+5Q15	1.00000	0.00%	0.99983	2.02%	0.78715	4.27%
+5Q21	1.00000	0.00%	0.99988	2.02%	0.50963	5.98%
+5Q22	1.00000	0.00%	0.99998	2.02%	0.95617	2.56%
+5Q23	1.00000	0.00%	0.99963	2.02%	0.95135	2.56%
+5Q31	1.00000	0.00%	0.97967	3.03%	0.96617	1.71%
+5Q32	1.00000	0.00%	0.98973	2.53%	0.95807	1.71%
+5Q33	1.00000	0.00%	0.99642	2.02%	0.72002	3.42%
+5Q34	1.00000	0.00%	0.96433	3.03%	0.32845	5.13%
+5Q35	1.00000	0.00%	0.88870	3.54%	0.82375	2.56%
+6P25	1.00000	0.00%	0.00025	9.09%	0.00108	8.55%
+6P24	1.00000	0.00%	0.00040	9.09%	0.00085	8.55%
+6P23	0.39603	2.44%	0.00158	8.59%	0.00175	8.55%
+6P22	0.07268	4.88%	0.00010	10.10%	0.03150	5.98%
+6P21	0.00010	14.63%	0.00010	14.14%	0.00010	11.11%
+6P12	0.00030	9.76%	0.00010	9.60%	0.25663	3.42%
+6P11	0.00042	9.76%	0.00010	9.60%	0.10103	4.27%
+6Q11	1.00000	0.00%	0.96965	4.04%	0.98393	1.71%
+6Q12	0.27027	4.88%	0.99120	4.04%	0.99347	1.71%
+6Q13	0.29617	4.88%	0.99532	4.04%	0.98087	2.56%
+6Q14	0.33332	4.88%	0.99553	4.55%	0.98787	2.56%
+6Q15	0.36272	4.88%	0.99815	4.55%	0.99275	2.56%
+6Q16	0.76105	2.44%	0.99815	5.05%	0.98405	3.42%
+6Q21	0.76890	2.44%	0.99205	6.06%	0.83585	5.98%
+6Q22	0.75693	2.44%	0.97365	6.57%	0.67838	6.84%
+6Q23	0.73102	2.44%	0.94798	6.57%	0.43985	7.69%
+6Q24	0.72322	2.44%	0.96862	6.06%	0.30372	8.55%
+6Q25	1.00000	0.00%	0.93613	6.57%	0.41545	7.69%
+6Q26	1.00000	0.00%	0.95273	6.06%	0.52142	6.84%
+6Q27	1.00000	0.00%	0.94818	6.06%	0.79675	5.13%
+7P22	0.03247	4.88%	0.00010	18.69%	0.00010	14.53%
+7P21	0.04015	4.88%	0.00010	18.69%	0.00010	17.95%
+7P15	0.03707	4.88%	0.00010	19.19%	0.00010	17.09%
+7P14	0.03370	4.88%	0.00010	20.20%	0.00010	16.24%
+7P13	0.03238	4.88%	0.00010	21.72%	0.00010	13.68%

+7P12	0.02945	4.88%	0.00010	20.71%	0.00010	11.11%
+7P11	0.02022	4.88%	0.00010	19.70%	0.00010	9.40%
+7Q11	0.00105	7.32%	0.00010	14.14%	0.00010	17.09%
+7Q21	0.02395	4.88%	0.00010	17.17%	0.00010	17.09%
+7Q22	0.03277	4.88%	0.00010	14.14%	0.00010	18.80%
+7Q31	1.00000	0.00%	0.00010	12.12%	0.00010	16.24%
+7Q32	0.33758	2.44%	0.00010	12.12%	0.00010	14.53%
+7Q33	0.33923	2.44%	0.00010	13.13%	0.00010	13.68%
+7Q34	0.35597	2.44%	0.00010	13.13%	0.00010	11.97%
+7Q35	0.35920	2.44%	0.00010	13.13%	0.00010	11.97%
+7Q36	0.05948	4.88%	0.00010	13.13%	0.00010	12.82%
+8P23	0.81423	2.44%	0.92448	9.09%	0.01377	15.38%
+8P22	0.49417	4.88%	0.79357	10.61%	0.06437	13.68%
+8P21	0.47455	4.88%	0.74032	10.61%	0.02548	14.53%
+8P12	0.74655	2.44%	0.83827	8.08%	0.00463	14.53%
+8P11	0.67045	2.44%	0.74193	7.07%	0.00140	13.68%
+8Q11	0.04327	4.88%	0.00010	19.70%	0.00010	17.09%
+8Q12	0.04323	4.88%	0.00010	21.72%	0.00010	17.09%
+8Q13	0.04927	4.88%	0.00010	22.73%	0.00010	17.09%
+8Q21	0.03932	4.88%	0.00010	26.77%	0.00010	20.51%
+8Q22	0.00342	7.32%	0.00010	27.78%	0.00010	21.37%
+8Q23	0.00412	7.32%	0.00010	28.28%	0.00010	21.37%
+8Q24	0.00372	7.32%	0.00010	29.80%	0.00010	22.22%
+9P24	1.00000	0.00%	1.00000	2.02%	0.99802	2.56%
+9P23	0.81037	2.44%	1.00000	2.02%	0.99885	2.56%
+9P22	1.00000	0.00%	1.00000	2.02%	0.99852	2.56%
+9P21	1.00000	0.00%	1.00000	2.02%	0.99997	0.85%
+9P13	1.00000	0.00%	1.00000	2.02%	1.00000	0.00%
+9P12	1.00000	0.00%	0.99998	2.02%	1.00000	0.00%
+9P11	1.00000	0.00%	0.99995	2.02%	1.00000	0.00%
+9Q11	1.00000	0.00%	0.29353	6.06%	0.81630	2.56%
+9Q12	1.00000	0.00%	0.22168	6.57%	0.65152	3.42%
+9Q13	0.52748	2.44%	0.32153	6.57%	0.72210	3.42%
+9Q21	0.54860	2.44%	0.13548	8.08%	0.76982	3.42%
+9Q22	1.00000	0.00%	0.10362	8.08%	0.74607	3.42%
+9Q31	1.00000	0.00%	0.01308	9.60%	0.72848	3.42%
+9Q32	1.00000	0.00%	0.01470	9.60%	0.72405	3.42%
+9Q33	0.00497	9.76%	0.00165	11.11%	0.73997	3.42%
+9Q34	0.00438	9.76%	0.00010	17.17%	0.01937	8.55%
+10P15	1.00000	0.00%	0.15778	9.09%	0.86742	3.42%
+10P14	1.00000	0.00%	0.17480	9.09%	0.87680	3.42%
+10P13	1.00000	0.00%	0.25420	8.59%	0.87825	3.42%
+10P12	1.00000	0.00%	0.12612	9.09%	0.85022	3.42%
+10P11	1.00000	0.00%	0.29815	7.58%	0.81823	3.42%
+10Q11	1.00000	0.00%	0.91020	5.05%	0.58020	5.13%
+10Q21	1.00000	0.00%	0.93112	5.56%	0.69538	5.13%
+10Q22	1.00000	0.00%	0.85758	6.57%	0.73053	5.13%
+10Q23	1.00000	0.00%	0.94483	6.06%	0.98590	2.56%

+10Q24	1.00000	0.00%	0.96118	6.06%	0.98957	2.56%
+10Q25	1.00000	0.00%	0.83537	7.58%	0.96577	3.42%
+10Q26	1.00000	0.00%	0.91390	6.57%	0.95805	3.42%
+11P15	1.00000	0.00%	0.92997	4.55%	0.12295	7.69%
+11P14	1.00000	0.00%	0.94648	4.55%	0.14268	7.69%
+11P13	1.00000	0.00%	0.75618	5.56%	0.32383	5.98%
+11P12	1.00000	0.00%	0.56905	6.06%	0.78110	3.42%
+11P11	0.52948	2.44%	0.24022	7.07%	0.96137	1.71%
+11Q11	0.47438	2.44%	0.17953	6.57%	0.80025	2.56%
+11Q12	0.47907	2.44%	0.00010	15.66%	0.00103	10.26%
+11Q13	0.03958	7.32%	0.00010	20.71%	0.00523	10.26%
+11Q14	1.00000	0.00%	0.66760	8.08%	0.89113	4.27%
+11Q21	1.00000	0.00%	0.96047	6.57%	0.93548	4.27%
+11Q22	1.00000	0.00%	0.99050	6.57%	0.93045	5.13%
+11Q23	0.19387	7.32%	0.98010	7.58%	0.90030	5.98%
+11Q24	1.00000	0.00%	0.97883	7.07%	0.99075	3.42%
+11Q25	1.00000	0.00%	0.99163	6.06%	0.99615	2.56%
+12P13	1.00000	0.00%	0.01135	6.57%	0.77633	1.71%
+12P12	0.34240	2.44%	0.00010	8.59%	0.93965	0.85%
+12P11	0.28258	2.44%	0.00025	7.58%	0.89465	0.85%
+12Q11	1.00000	0.00%	0.20950	2.53%	0.15712	2.56%
+12Q12	1.00000	0.00%	0.05165	3.54%	0.18422	2.56%
+12Q13	0.20023	2.44%	0.00010	6.57%	0.18832	2.56%
+12Q14	0.22480	2.44%	0.00017	6.57%	0.23778	2.56%
+12Q15	1.00000	0.00%	0.00150	6.06%	0.56582	1.71%
+12Q21	1.00000	0.00%	0.28512	3.54%	0.02898	5.13%
+12Q22	1.00000	0.00%	0.58352	2.53%	0.07392	4.27%
+12Q23	1.00000	0.00%	0.48885	3.03%	0.01217	5.98%
+12Q24	1.00000	0.00%	0.00010	9.09%	0.00412	6.84%
+13P13	1.00000	0.00%	0.78255	3.54%	1.00000	0.00%
+13P12	1.00000	0.00%	0.75448	3.54%	1.00000	0.00%
+13P11	1.00000	0.00%	0.75788	3.54%	1.00000	0.00%
+13Q11	1.00000	0.00%	0.41510	9.60%	0.99743	1.71%
+13Q12	0.76005	2.44%	0.16325	12.63%	0.98497	3.42%
+13Q13	0.80115	2.44%	0.05747	15.15%	0.99335	3.42%
+13Q14	0.86492	2.44%	0.30925	15.15%	0.98888	5.13%
+13Q21	0.66192	4.88%	0.85058	13.64%	0.99812	5.13%
+13Q22	0.62832	4.88%	0.90422	12.12%	0.97085	6.84%
+13Q31	0.87748	2.44%	0.70980	13.13%	0.97953	5.98%
+13Q32	0.83210	2.44%	0.48950	12.63%	0.87662	6.84%
+13Q33	1.00000	0.00%	0.66903	10.61%	0.80282	6.84%
+13Q34	1.00000	0.00%	0.50692	11.11%	0.76907	6.84%
+14P13	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
+14P12	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
+14P11	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
+14Q11	1.00000	0.00%	0.99970	1.52%	1.00000	0.00%
+14Q12	1.00000	0.00%	0.99692	2.53%	1.00000	0.00%
+14Q13	1.00000	0.00%	0.99485	3.03%	1.00000	0.00%

+14Q21	1.00000	0.00%	0.99938	2.53%	1.00000	0.00%
+14Q22	1.00000	0.00%	0.99990	1.52%	1.00000	0.00%
+14Q23	1.00000	0.00%	0.99998	1.52%	1.00000	0.00%
+14Q24	1.00000	0.00%	1.00000	1.01%	0.99937	0.85%
+14Q31	1.00000	0.00%	1.00000	1.52%	0.99950	0.85%
+14Q32	1.00000	0.00%	1.00000	1.01%	0.99172	1.71%
+15P13	1.00000	0.00%	1.00000	0.00%	0.05347	3.42%
+15P12	1.00000	0.00%	1.00000	0.00%	0.05388	3.42%
+15P11	1.00000	0.00%	1.00000	0.00%	0.05775	3.42%
+15Q11	1.00000	0.00%	0.99180	2.02%	0.12953	5.98%
+15Q12	1.00000	0.00%	0.99477	2.02%	0.15850	5.98%
+15Q13	1.00000	0.00%	0.99658	2.02%	0.18827	5.98%
+15Q14	1.00000	0.00%	0.99685	2.02%	0.20743	5.98%
+15Q15	1.00000	0.00%	0.99220	2.53%	0.21870	5.98%
+15Q21	1.00000	0.00%	0.99065	2.53%	0.21600	5.98%
+15Q22	1.00000	0.00%	0.72327	4.55%	0.03795	7.69%
+15Q23	1.00000	0.00%	0.13730	7.07%	0.03160	7.69%
+15Q24	1.00000	0.00%	0.03408	8.59%	0.01523	8.55%
+15Q25	1.00000	0.00%	0.42983	5.56%	0.03088	7.69%
+15Q26	1.00000	0.00%	0.87582	3.54%	0.02490	7.69%
+16P13	0.55325	2.44%	0.00647	10.61%	0.00050	11.97%
+16P12	0.54198	2.44%	0.01902	9.60%	0.00338	10.26%
+16P11	0.52143	2.44%	0.02380	9.09%	0.02052	8.55%
+16Q11	0.66837	2.44%	0.99867	3.54%	0.99967	0.85%
+16Q12	0.70208	2.44%	0.99992	3.03%	0.99985	0.85%
+16Q13	0.71920	2.44%	0.99990	3.03%	1.00000	0.00%
+16Q21	0.72930	2.44%	0.99993	3.03%	1.00000	0.00%
+16Q22	0.38330	4.88%	0.99998	3.03%	1.00000	0.00%
+16Q23	0.37532	4.88%	0.99690	5.05%	0.99990	0.85%
+16Q24	0.02857	9.76%	0.99523	5.05%	0.99980	0.85%
+17P13	1.00000	0.00%	0.44840	11.11%	0.00168	16.24%
+17P12	1.00000	0.00%	0.52075	10.61%	0.00870	14.53%
+17P11	1.00000	0.00%	0.14075	11.62%	0.00290	14.53%
+17Q11	0.06862	4.88%	0.00010	23.74%	0.00010	20.51%
+17Q12	0.07283	4.88%	0.00010	34.85%	0.00010	21.37%
+17Q21	0.00077	9.76%	0.00010	34.85%	0.00010	23.08%
+17Q22	0.00737	7.32%	0.00010	24.75%	0.00010	11.97%
+17Q23	0.00752	7.32%	0.00010	23.23%	0.00010	17.09%
+17Q24	0.00062	9.76%	0.00010	29.29%	0.00010	23.08%
+17Q25	0.00085	9.76%	0.00010	25.76%	0.00010	20.51%
+18P11	1.00000	0.00%	0.25593	6.57%	0.00610	9.40%
+18Q11	1.00000	0.00%	0.96208	4.55%	0.95887	2.56%
+18Q12	1.00000	0.00%	0.99833	4.04%	0.80780	5.13%
+18Q21	1.00000	0.00%	0.99995	3.03%	0.76705	5.98%
+18Q22	1.00000	0.00%	1.00000	2.53%	0.80175	5.98%
+18Q23	1.00000	0.00%	0.99998	2.53%	0.93000	4.27%
+19P13	0.44807	2.44%	0.00012	10.61%	0.55278	3.42%
+19P12	1.00000	0.00%	0.00012	11.11%	0.28427	4.27%

+19P11	1.00000	0.00%	0.00010	11.11%	0.26498	4.27%
+19Q11	1.00000	0.00%	0.00010	13.13%	0.06560	5.98%
+19Q12	1.00000	0.00%	0.00010	15.15%	0.06488	5.98%
+19Q13	0.45397	2.44%	0.00010	14.65%	0.09463	5.98%
+20P13	0.25287	2.44%	0.00010	22.73%	0.00010	20.51%
+20P12	0.25938	2.44%	0.00010	24.24%	0.00010	19.66%
+20P11	0.23250	2.44%	0.00010	21.21%	0.00010	17.95%
+20Q11	0.00142	7.32%	0.00010	41.92%	0.00010	23.08%
+20Q12	0.02652	4.88%	0.00010	49.49%	0.00010	26.50%
+20Q13	0.00010	12.20%	0.00010	50.00%	0.00010	26.50%
+21P13	1.00000	0.00%	0.98655	0.51%	0.83562	0.85%
+21P12	1.00000	0.00%	0.98573	0.51%	0.83322	0.85%
+21P11	1.00000	0.00%	0.98155	0.51%	0.81218	0.85%
+21Q11	1.00000	0.00%	0.99983	0.51%	0.97307	0.85%
+21Q21	1.00000	0.00%	0.99928	1.01%	0.98282	0.85%
+21Q22	1.00000	0.00%	0.99990	0.51%	0.89362	1.71%
+22P13	1.00000	0.00%	0.49650	4.04%	0.22227	4.27%
+22P12	1.00000	0.00%	0.49027	4.04%	0.22467	4.27%
+22P11	1.00000	0.00%	0.49103	4.04%	0.22237	4.27%
+22Q11	1.00000	0.00%	0.53808	7.58%	0.00492	11.97%
+22Q12	1.00000	0.00%	0.17355	10.10%	0.00185	13.68%
+22Q13	1.00000	0.00%	0.12413	11.11%	0.00088	14.53%
+XP22	1.00000	0.00%	0.83442	4.55%	0.00010	18.80%
+XP21	1.00000	0.00%	0.77938	5.05%	0.00010	18.80%
+XP11	1.00000	0.00%	0.83193	4.55%	0.00010	19.66%
+XQ11	1.00000	0.00%	0.14828	7.07%	0.00010	17.95%
+XQ12	0.50775	2.44%	0.16180	7.07%	0.00010	17.95%
+XQ13	0.51873	2.44%	0.18222	7.07%	0.00010	17.95%
+XQ21	0.54853	2.44%	0.26557	7.07%	0.00010	17.95%
+XQ22	0.54033	2.44%	0.16022	7.58%	0.00010	20.51%
+XQ23	0.53328	2.44%	0.14468	7.58%	0.00010	20.51%
+XQ24	0.53552	2.44%	0.09225	8.08%	0.00010	20.51%
+XQ25	0.55347	2.44%	0.04565	9.09%	0.00010	18.80%
+XQ26	0.55445	2.44%	0.07985	8.59%	0.00010	17.09%
+XQ27	0.53245	2.44%	0.02802	9.09%	0.00010	13.68%
+XQ28	0.52780	2.44%	0.04385	8.59%	0.00015	12.82%
-1P36	0.57040	2.44%	0.87823	4.55%	1.00000	0.00%
-1P35	0.58712	2.44%	0.90830	4.55%	1.00000	0.00%
-1P34	0.59687	2.44%	0.86000	5.05%	1.00000	0.00%
-1P33	0.56933	2.44%	0.89298	4.55%	1.00000	0.00%
-1P32	0.18442	4.88%	0.92527	4.04%	1.00000	0.00%
-1P31	0.03837	7.32%	0.11645	8.08%	1.00000	0.00%
-1P22	0.02950	7.32%	0.06318	8.08%	0.99322	0.85%
-1P21	0.02738	7.32%	0.05245	8.08%	1.00000	0.00%
-1P13	0.13665	4.88%	0.13180	7.07%	1.00000	0.00%
-1P12	1.00000	0.00%	0.39785	5.56%	1.00000	0.00%
-1P11	1.00000	0.00%	0.27985	5.56%	1.00000	0.00%
-1Q11	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%

-1Q12	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%
-1Q21	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%
-1Q22	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%
-1Q23	0.93160	2.44%	1.00000	1.52%	1.00000	0.00%
-1Q24	0.92882	2.44%	1.00000	1.52%	1.00000	0.00%
-1Q25	0.92740	2.44%	1.00000	1.52%	1.00000	0.00%
-1Q31	0.92650	2.44%	1.00000	0.51%	1.00000	0.00%
-1Q32	0.91787	2.44%	1.00000	0.51%	1.00000	0.85%
-1Q41	0.65622	4.88%	1.00000	2.53%	1.00000	0.85%
-1Q42	0.63697	4.88%	1.00000	2.53%	1.00000	0.85%
-1Q43	0.89135	2.44%	1.00000	2.02%	1.00000	0.85%
-1Q44	0.89005	2.44%	1.00000	2.02%	1.00000	0.85%
-2P25	0.24757	4.88%	1.00000	1.01%	1.00000	0.00%
-2P24	0.27080	4.88%	1.00000	1.01%	1.00000	0.00%
-2P23	0.28043	4.88%	1.00000	1.01%	1.00000	0.00%
-2P22	0.66190	2.44%	1.00000	1.01%	1.00000	0.00%
-2P21	0.64605	2.44%	0.99998	1.52%	0.99875	0.85%
-2P16	0.63227	2.44%	1.00000	1.52%	0.99892	0.85%
-2P15	0.61320	2.44%	1.00000	1.52%	0.99865	0.85%
-2P14	0.60210	2.44%	0.99997	1.52%	0.99788	0.85%
-2P13	0.58782	2.44%	0.99993	1.52%	1.00000	0.00%
-2P12	0.55705	2.44%	0.99975	1.52%	1.00000	0.00%
-2P11	1.00000	0.00%	0.99953	1.52%	1.00000	0.00%
-2Q11	1.00000	0.00%	0.86380	4.04%	1.00000	0.00%
-2Q12	1.00000	0.00%	0.88370	4.04%	1.00000	0.00%
-2Q13	0.53003	2.44%	0.89915	4.04%	1.00000	0.00%
-2Q14	0.52975	2.44%	0.89568	4.04%	1.00000	0.00%
-2Q21	0.17282	4.88%	0.92060	4.04%	0.99467	0.85%
-2Q22	0.17858	4.88%	0.86077	4.55%	0.97073	1.71%
-2Q23	0.18212	4.88%	0.77905	5.05%	0.77675	3.42%
-2Q24	0.19773	4.88%	0.83127	5.05%	0.92178	2.56%
-2Q31	0.04742	7.32%	0.84987	5.05%	0.93095	2.56%
-2Q32	0.19907	4.88%	0.95602	4.04%	0.98073	1.71%
-2Q33	0.55433	2.44%	0.96450	3.54%	0.97337	1.71%
-2Q34	0.53028	2.44%	0.94228	3.54%	0.99353	0.85%
-2Q35	0.50777	2.44%	0.85080	4.04%	1.00000	0.00%
-2Q36	0.51053	2.44%	0.85485	4.04%	0.99247	0.85%
-2Q37	0.50838	2.44%	0.85347	4.04%	0.95337	1.71%
-3P26	0.46935	2.44%	0.00537	9.09%	0.23087	5.13%
-3P25	0.47900	2.44%	0.00665	9.09%	0.06047	6.84%
-3P24	0.48897	2.44%	0.01897	8.59%	0.14160	5.98%
-3P23	0.47983	2.44%	0.01472	8.59%	0.13098	5.98%
-3P22	0.48338	2.44%	0.02937	8.08%	0.24632	5.13%
-3P21	0.50605	2.44%	0.02463	8.59%	0.16068	5.98%
-3P14	0.44203	2.44%	0.05068	7.07%	0.53962	3.42%
-3P13	0.43418	2.44%	0.13895	6.06%	0.89257	1.71%
-3P12	1.00000	0.00%	0.06270	7.07%	0.75923	2.56%
-3P11	1.00000	0.00%	0.06020	6.57%	0.87593	1.71%

-3Q11	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%
-3Q12	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%
-3Q13	0.72128	2.44%	1.00000	1.52%	0.99980	0.85%
-3Q21	0.76175	2.44%	1.00000	1.52%	0.99997	0.85%
-3Q22	0.77063	2.44%	1.00000	1.52%	0.99995	0.85%
-3Q23	0.78430	2.44%	1.00000	1.52%	1.00000	0.00%
-3Q24	0.81057	2.44%	1.00000	2.02%	1.00000	0.00%
-3Q25	1.00000	0.00%	1.00000	2.02%	1.00000	0.00%
-3Q26	1.00000	0.00%	1.00000	2.02%	1.00000	0.85%
-3Q27	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%
-3Q28	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%
-3Q29	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%
-4P16	0.51150	2.44%	0.00010	13.13%	0.95338	1.71%
-4P15	0.14393	4.88%	0.00010	14.65%	0.84065	2.56%
-4P14	0.14432	4.88%	0.00010	14.14%	0.84258	2.56%
-4P13	0.50333	2.44%	0.00010	13.64%	0.94975	1.71%
-4P12	0.51117	2.44%	0.00015	12.63%	0.95422	1.71%
-4P11	0.48797	2.44%	0.00037	11.11%	0.93717	1.71%
-4Q11	0.52855	2.44%	0.00043	11.62%	0.96002	1.71%
-4Q12	0.58377	2.44%	0.00025	13.13%	0.98000	1.71%
-4Q13	0.58978	2.44%	0.00010	15.66%	0.84015	3.42%
-4Q21	0.60638	2.44%	0.00010	16.16%	0.72175	4.27%
-4Q22	0.59647	2.44%	0.00010	16.67%	0.36407	5.98%
-4Q23	0.59222	2.44%	0.00010	17.68%	0.35400	5.98%
-4Q24	0.59047	2.44%	0.00010	18.18%	0.21058	6.84%
-4Q25	0.59550	2.44%	0.00010	18.69%	0.22783	6.84%
-4Q26	0.60047	2.44%	0.00010	19.19%	0.38293	5.98%
-4Q27	0.59750	2.44%	0.00010	19.70%	0.70040	4.27%
-4Q28	0.21570	4.88%	0.00010	20.20%	0.70473	4.27%
-4Q31	0.56525	2.44%	0.00010	19.70%	0.44215	5.13%
-4Q32	0.53952	2.44%	0.00010	19.70%	0.22423	5.98%
-4Q33	0.51563	2.44%	0.00010	17.68%	0.32555	5.13%
-4Q34	0.51378	2.44%	0.00010	17.68%	0.49957	4.27%
-4Q35	0.50787	2.44%	0.00010	17.17%	0.67867	3.42%
-5P15	0.46428	4.88%	1.00000	2.02%	1.00000	0.00%
-5P14	0.81632	2.44%	1.00000	2.02%	1.00000	0.00%
-5P13	0.80248	2.44%	1.00000	2.53%	1.00000	0.00%
-5P12	1.00000	0.00%	1.00000	2.53%	1.00000	0.00%
-5P11	0.76145	2.44%	1.00000	2.02%	1.00000	0.00%
-5Q11	0.59600	2.44%	0.00298	12.12%	0.99763	0.85%
-5Q12	0.00128	12.20%	0.00087	13.13%	0.99815	0.85%
-5Q13	0.00012	14.63%	0.00023	14.14%	0.95078	2.56%
-5Q14	0.00010	26.83%	0.00010	17.68%	0.90187	3.42%
-5Q15	0.00010	26.83%	0.00010	19.19%	0.95693	2.56%
-5Q21	0.00010	29.27%	0.00010	18.18%	0.67490	5.13%
-5Q22	0.00010	29.27%	0.00012	15.15%	0.62485	5.13%
-5Q23	0.00010	29.27%	0.00010	15.66%	0.46652	5.98%
-5Q31	0.01092	9.76%	0.01548	11.62%	0.61292	5.13%
-5Q32	0.00823	9.76%	0.08387	9.60%	0.55523	5.13%

-5Q33	0.00143	12.20%	0.05047	10.10%	0.55073	5.13%
-5Q34	0.00125	12.20%	0.05205	10.10%	0.55007	5.13%
-5Q35	0.05348	7.32%	0.05120	10.10%	0.55540	5.13%
-6P25	1.00000	0.00%	1.00000	1.01%	0.98847	3.42%
-6P24	0.77488	2.44%	1.00000	1.01%	0.98938	3.42%
-6P23	1.00000	0.00%	1.00000	1.01%	0.99718	2.56%
-6P22	0.78585	2.44%	1.00000	1.01%	0.99965	1.71%
-6P21	0.17378	7.32%	1.00000	1.52%	0.99973	1.71%
-6P12	0.00538	12.20%	1.00000	1.52%	0.99865	1.71%
-6P11	0.00047	14.63%	1.00000	1.52%	0.99790	1.71%
-6Q11	0.00010	14.63%	0.75027	5.05%	0.76123	3.42%
-6Q12	0.00010	14.63%	0.43523	7.07%	0.83703	3.42%
-6Q13	0.00013	14.63%	0.45095	7.07%	0.84438	3.42%
-6Q14	0.00010	17.07%	0.35660	7.58%	0.85150	3.42%
-6Q15	0.00010	17.07%	0.45798	7.07%	0.70432	4.27%
-6Q16	0.00010	17.07%	0.47393	7.07%	0.23837	6.84%
-6Q21	0.01082	9.76%	0.63765	6.57%	0.27022	6.84%
-6Q22	0.01097	9.76%	0.53413	7.07%	0.42107	5.98%
-6Q23	0.23752	4.88%	0.90202	5.05%	0.74792	4.27%
-6Q24	0.21875	4.88%	0.86680	5.05%	0.37115	5.98%
-6Q25	1.00000	0.00%	0.90912	4.55%	0.20357	6.84%
-6Q26	1.00000	0.00%	0.93978	4.04%	0.47377	5.13%
-6Q27	1.00000	0.00%	0.93147	4.04%	0.62770	4.27%
-7P22	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-7P21	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-7P15	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-7P14	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-7P13	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-7P12	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-7P11	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-7Q11	0.88165	2.44%	1.00000	2.53%	1.00000	0.00%
-7Q21	1.00000	0.00%	1.00000	2.53%	1.00000	0.00%
-7Q22	1.00000	0.00%	1.00000	2.53%	1.00000	0.00%
-7Q31	1.00000	0.00%	1.00000	2.53%	1.00000	0.00%
-7Q32	1.00000	0.00%	1.00000	2.53%	0.92750	7.69%
-7Q33	1.00000	0.00%	1.00000	3.54%	0.92380	7.69%
-7Q34	1.00000	0.00%	1.00000	3.54%	0.92368	7.69%
-7Q35	1.00000	0.00%	1.00000	3.54%	0.91747	7.69%
-7Q36	1.00000	0.00%	1.00000	3.54%	0.91698	7.69%
-8P23	0.58627	2.44%	0.62825	6.06%	0.99733	0.85%
-8P22	0.59933	2.44%	0.66757	6.06%	0.99797	0.85%
-8P21	0.59972	2.44%	0.67622	6.06%	0.99810	0.85%
-8P12	1.00000	0.00%	0.95327	4.55%	0.99835	0.85%
-8P11	1.00000	0.00%	0.96147	4.55%	0.99872	0.85%
-8Q11	1.00000	0.00%	1.00000	1.01%	1.00000	0.00%
-8Q12	1.00000	0.00%	1.00000	1.01%	1.00000	0.00%
-8Q13	1.00000	0.00%	1.00000	1.01%	1.00000	0.00%
-8Q21	1.00000	0.00%	1.00000	1.01%	1.00000	0.00%

-8Q22	0.74183	4.88%	1.00000	1.01%	1.00000	0.00%
-8Q23	0.76940	4.88%	1.00000	1.01%	1.00000	0.00%
-8Q24	1.00000	0.00%	1.00000	1.01%	1.00000	0.00%
-9P24	0.00485	9.76%	0.00010	15.15%	0.00027	11.97%
-9P23	0.00080	12.20%	0.00010	15.66%	0.00017	12.82%
-9P22	0.00052	12.20%	0.00010	15.15%	0.00025	11.97%
-9P21	0.00035	12.20%	0.00010	14.14%	0.00012	11.97%
-9P13	0.02853	7.32%	0.00350	10.10%	0.00180	10.26%
-9P12	0.50775	2.44%	0.00312	10.10%	0.00567	9.40%
-9P11	0.48920	2.44%	0.04005	8.08%	0.00443	9.40%
-9Q11	1.00000	0.00%	0.99713	2.02%	0.00813	9.40%
-9Q12	1.00000	0.00%	0.99815	2.02%	0.01197	9.40%
-9Q13	1.00000	0.00%	0.99930	2.02%	0.02037	9.40%
-9Q21	0.60573	2.44%	0.99952	2.02%	0.01607	10.26%
-9Q22	0.66283	2.44%	0.99998	2.02%	0.04115	10.26%
-9Q31	0.67393	2.44%	0.99995	2.02%	0.05248	10.26%
-9Q32	0.69042	2.44%	0.99990	2.53%	0.06835	10.26%
-9Q33	1.00000	0.00%	0.99983	3.03%	0.04138	11.11%
-9Q34	1.00000	0.00%	0.99992	3.54%	0.00743	14.53%
-10P15	1.00000	0.00%	0.83665	3.54%	0.00660	8.55%
-10P14	1.00000	0.00%	0.83208	3.54%	0.01970	7.69%
-10P13	1.00000	0.00%	0.83387	3.54%	0.01992	7.69%
-10P12	1.00000	0.00%	0.83763	3.54%	0.05132	6.84%
-10P11	1.00000	0.00%	0.75585	3.54%	0.03150	6.84%
-10Q11	1.00000	0.00%	0.71482	3.03%	1.00000	0.00%
-10Q21	1.00000	0.00%	0.82750	3.03%	1.00000	0.00%
-10Q22	0.43042	2.44%	0.61923	4.04%	0.89157	1.71%
-10Q23	0.40757	2.44%	0.69308	3.54%	0.87002	1.71%
-10Q24	0.42042	2.44%	0.91597	2.53%	0.87535	1.71%
-10Q25	0.44125	2.44%	0.94245	2.53%	0.73498	2.56%
-10Q26	0.43037	2.44%	0.85748	3.03%	0.30600	4.27%
-11P15	1.00000	0.00%	0.96503	2.02%	0.00240	8.55%
-11P14	1.00000	0.00%	0.65590	3.54%	0.00142	8.55%
-11P13	1.00000	0.00%	0.64733	3.54%	0.00510	7.69%
-11P12	1.00000	0.00%	0.65683	3.54%	0.00543	7.69%
-11P11	1.00000	0.00%	0.95163	2.02%	0.00498	7.69%
-11Q11	1.00000	0.00%	0.92242	3.03%	0.98552	0.85%
-11Q12	1.00000	0.00%	0.99233	3.03%	0.99700	0.85%
-11Q13	1.00000	0.00%	0.99773	4.04%	0.78478	5.13%
-11Q14	1.00000	0.00%	0.58352	6.06%	0.97677	1.71%
-11Q21	0.51833	2.44%	0.38877	6.06%	0.95422	1.71%
-11Q22	0.51043	2.44%	0.24147	6.57%	0.94987	1.71%
-11Q23	0.49773	2.44%	0.29758	6.06%	0.07505	6.84%
-11Q24	0.45970	2.44%	0.31310	5.56%	0.22152	5.13%
-11Q25	0.45878	2.44%	0.29010	5.56%	0.09778	5.98%
-12P13	1.00000	0.00%	1.00000	1.52%	1.00000	0.00%
-12P12	1.00000	0.00%	1.00000	2.02%	1.00000	0.00%
-12P11	1.00000	0.00%	1.00000	2.02%	1.00000	0.00%

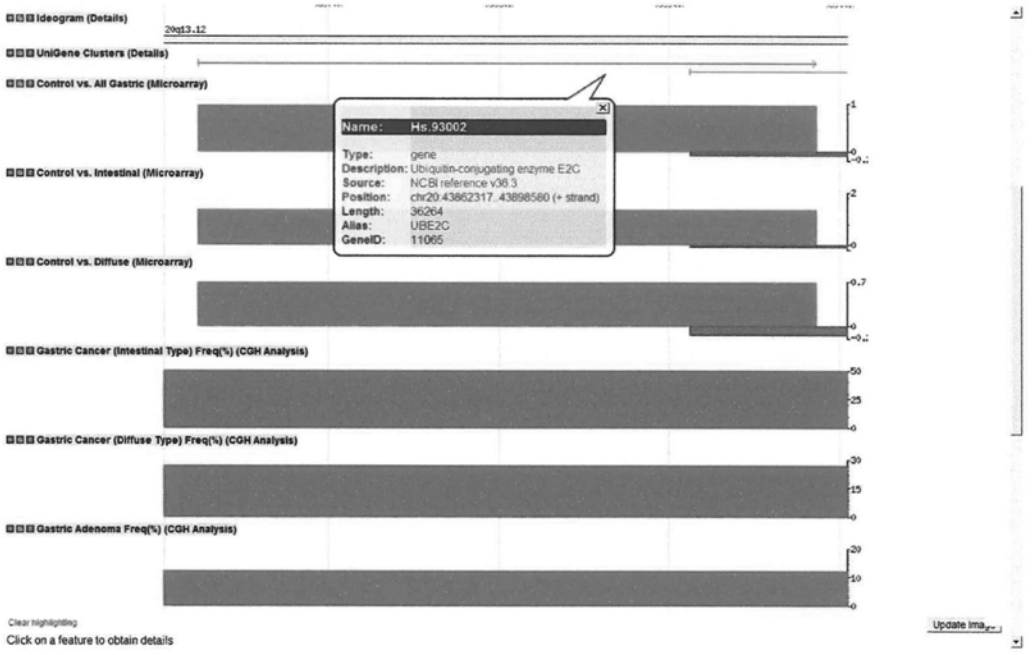
-12Q11	1.00000	0.00%	0.88735	7.07%	1.00000	0.00%
-12Q12	1.00000	0.00%	0.96680	7.07%	1.00000	0.00%
-12Q13	1.00000	0.00%	0.98437	7.07%	1.00000	0.00%
-12Q14	1.00000	0.00%	0.96925	7.58%	1.00000	0.00%
-12Q15	1.00000	0.00%	0.85747	9.09%	1.00000	0.00%
-12Q21	0.78923	2.44%	0.65557	10.61%	0.99980	1.71%
-12Q22	1.00000	0.00%	0.63098	10.10%	1.00000	0.00%
-12Q23	1.00000	0.00%	0.88085	8.59%	0.99938	1.71%
-12Q24	0.77730	2.44%	0.93058	8.08%	0.84950	5.98%
-13P13	1.00000	0.00%	0.87498	1.52%	0.87963	0.85%
-13P12	1.00000	0.00%	0.87627	1.52%	0.88437	0.85%
-13P11	1.00000	0.00%	0.86678	1.52%	0.87588	0.85%
-13Q11	1.00000	0.00%	0.81960	3.54%	0.91678	1.71%
-13Q12	0.50533	2.44%	0.90263	3.54%	0.94795	1.71%
-13Q13	1.00000	0.00%	0.92012	3.54%	0.95582	1.71%
-13Q14	0.03828	7.32%	0.88795	4.04%	0.96823	1.71%
-13Q21	0.00010	19.51%	0.11412	9.09%	0.14770	7.69%
-13Q22	0.01253	9.76%	0.16615	9.09%	0.30768	6.84%
-13Q31	0.00010	21.95%	0.70092	6.57%	0.48840	5.98%
-13Q32	0.25447	4.88%	0.92007	5.05%	0.78455	4.27%
-13Q33	0.24542	4.88%	0.91230	5.05%	0.88377	3.42%
-13Q34	0.62747	2.44%	0.94998	4.55%	0.98865	1.71%
-14P13	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-14P12	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-14P11	1.00000	0.00%	1.00000	0.00%	1.00000	0.00%
-14Q11	1.00000	0.00%	0.65210	4.04%	1.00000	0.00%
-14Q12	1.00000	0.00%	0.64027	4.55%	0.98550	0.85%
-14Q13	0.50105	2.44%	0.46552	5.56%	0.99052	0.85%
-14Q21	0.53178	2.44%	0.45202	6.06%	0.99370	0.85%
-14Q22	0.52967	2.44%	0.33217	6.57%	0.96078	1.71%
-14Q23	0.53105	2.44%	0.57397	5.56%	0.96075	1.71%
-14Q24	0.56043	2.44%	0.68557	5.56%	0.44572	5.13%
-14Q31	0.56432	2.44%	0.69997	5.56%	0.09202	7.69%
-14Q32	0.59423	2.44%	0.92453	4.55%	0.12412	7.69%
-15P13	1.00000	0.00%	0.98127	0.51%	1.00000	0.00%
-15P12	1.00000	0.00%	0.98210	0.51%	1.00000	0.00%
-15P11	1.00000	0.00%	0.98293	0.51%	1.00000	0.00%
-15Q11	0.38983	2.44%	0.00893	7.58%	0.96480	0.85%
-15Q12	0.40772	2.44%	0.00663	8.08%	0.97032	0.85%
-15Q13	0.41797	2.44%	0.00798	8.08%	0.97415	0.85%
-15Q14	0.42737	2.44%	0.01033	8.08%	0.97627	0.85%
-15Q15	0.44563	2.44%	0.00850	8.59%	0.98270	0.85%
-15Q21	0.50007	2.44%	0.01673	9.09%	0.84127	2.56%
-15Q22	0.55663	2.44%	0.06172	9.09%	0.44295	5.13%
-15Q23	0.57540	2.44%	0.13312	8.59%	0.47918	5.13%
-15Q24	0.59488	2.44%	0.26497	8.08%	0.06298	8.55%
-15Q25	1.00000	0.00%	0.37278	7.58%	0.13105	7.69%
-15Q26	1.00000	0.00%	0.22418	8.08%	0.10805	7.69%

-16P13	1.00000	0.00%	0.00677	12.63%	0.79773	4.27%
-16P12	1.00000	0.00%	0.00648	12.63%	0.90022	3.42%
-16P11	1.00000	0.00%	0.00825	12.12%	0.88583	3.42%
-16Q11	1.00000	0.00%	0.00322	8.08%	0.41783	3.42%
-16Q12	1.00000	0.00%	0.00292	8.08%	0.41583	3.42%
-16Q13	0.39183	2.44%	0.00327	8.08%	0.41147	3.42%
-16Q21	0.08563	4.88%	0.00432	8.08%	0.42568	3.42%
-16Q22	0.09037	4.88%	0.00488	8.08%	0.00685	7.69%
-16Q23	0.43735	2.44%	0.01042	8.08%	0.00400	8.55%
-16Q24	0.42928	2.44%	0.01967	7.58%	0.00010	11.11%
-17P13	1.00000	0.00%	0.05298	13.13%	0.00010	23.93%
-17P12	0.75308	2.44%	0.04165	13.64%	0.00450	14.53%
-17P11	1.00000	0.00%	0.11265	12.63%	0.98212	3.42%
-17Q11	1.00000	0.00%	0.99872	7.58%	1.00000	1.71%
-17Q12	1.00000	0.00%	0.99970	7.58%	1.00000	1.71%
-17Q21	1.00000	0.00%	0.99993	7.58%	1.00000	2.56%
-17Q22	1.00000	0.00%	0.99998	7.58%	0.99990	3.42%
-17Q23	1.00000	0.00%	1.00000	7.58%	1.00000	2.56%
-17Q24	1.00000	0.00%	1.00000	8.08%	0.99840	5.98%
-17Q25	1.00000	0.00%	1.00000	8.08%	0.98460	7.69%
-18P11	1.00000	0.00%	0.99973	3.03%	0.42038	6.84%
-18Q11	0.64948	2.44%	0.12562	10.10%	0.22788	7.69%
-18Q12	0.29720	4.88%	0.00013	16.16%	0.00155	13.68%
-18Q21	0.30538	4.88%	0.00010	18.18%	0.00205	13.68%
-18Q22	0.08567	7.32%	0.00010	17.68%	0.00015	15.38%
-18Q23	0.07742	7.32%	0.00010	16.67%	0.00078	13.68%
-19P13	1.00000	0.00%	0.84778	7.07%	0.00088	14.53%
-19P12	1.00000	0.00%	0.84665	7.07%	0.00117	14.53%
-19P11	1.00000	0.00%	0.83342	7.07%	0.00077	14.53%
-19Q11	1.00000	0.00%	0.81895	7.07%	0.11565	9.40%
-19Q12	0.68973	2.44%	0.83635	7.07%	0.12718	9.40%
-19Q13	0.69928	2.44%	0.85122	7.07%	0.13907	9.40%
-20P13	1.00000	0.00%	1.00000	1.52%	0.95030	4.27%
-20P12	0.74748	2.44%	1.00000	1.52%	0.98095	3.42%
-20P11	0.74532	2.44%	1.00000	1.52%	0.99373	2.56%
-20Q11	1.00000	0.00%	1.00000	4.04%	0.99988	2.56%
-20Q12	1.00000	0.00%	1.00000	4.04%	0.99998	2.56%
-20Q13	0.61275	4.88%	1.00000	4.04%	0.99972	3.42%
-21P13	1.00000	0.00%	0.05827	6.06%	1.00000	0.00%
-21P12	1.00000	0.00%	0.05660	6.06%	1.00000	0.00%
-21P11	1.00000	0.00%	0.05785	6.06%	1.00000	0.00%
-21Q11	0.12702	4.88%	0.01760	8.59%	0.80058	2.56%
-21Q21	0.15527	4.88%	0.04660	8.59%	0.71513	3.42%
-21Q22	0.56055	2.44%	0.11490	8.59%	0.45090	5.13%
-22P13	1.00000	0.00%	0.91952	2.53%	0.97195	0.85%
-22P12	1.00000	0.00%	0.91737	2.53%	0.97322	0.85%
-22P11	1.00000	0.00%	0.92063	2.53%	0.97163	0.85%
-22Q11	1.00000	0.00%	0.01988	10.10%	0.97642	1.71%

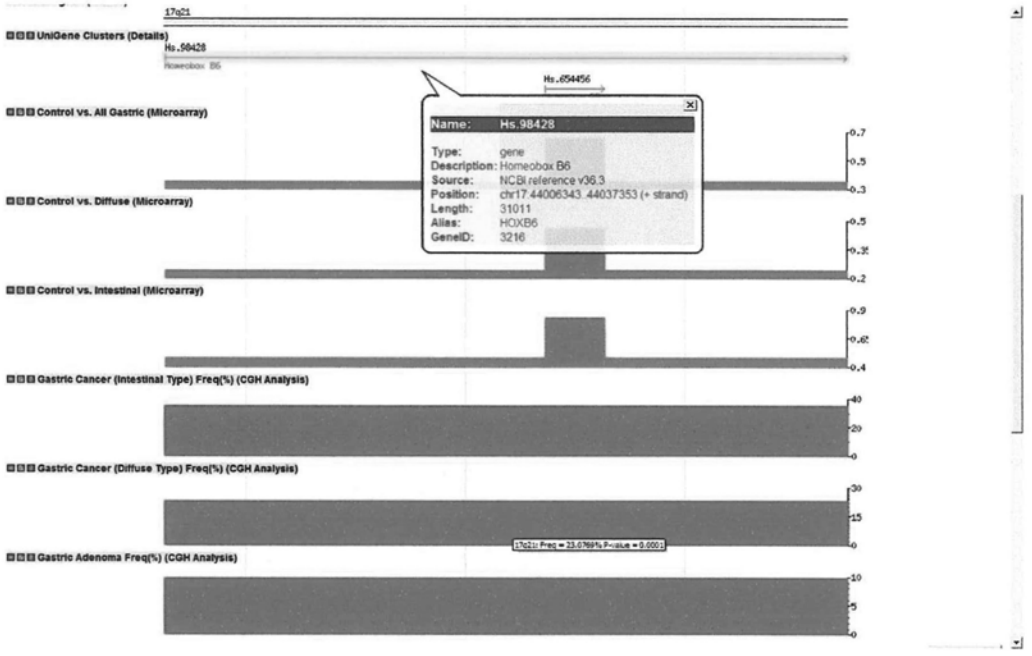
-22Q12	1.00000	0.00%	0.03990	10.10%	0.98313	1.71%
-22Q13	0.62212	2.44%	0.10785	9.60%	0.98805	1.71%
-XP22	1.00000	0.00%	0.99995	3.03%	0.81442	5.13%
-XP21	1.00000	0.00%	0.99985	3.03%	0.82795	5.13%
-XP11	1.00000	0.00%	1.00000	2.53%	0.84932	5.13%
-XQ11	1.00000	0.00%	1.00000	2.53%	0.84242	5.13%
-XQ12	1.00000	0.00%	1.00000	2.53%	0.87210	5.13%
-XQ13	1.00000	0.00%	1.00000	2.53%	0.88560	5.13%
-XQ21	0.75053	2.44%	0.99997	2.53%	0.88260	5.13%
-XQ22	0.74832	2.44%	1.00000	1.52%	0.88023	5.13%
-XQ23	0.74823	2.44%	1.00000	1.52%	0.88318	5.13%
-XQ24	0.75130	2.44%	1.00000	1.52%	0.88500	5.13%
-XQ25	0.75580	2.44%	1.00000	1.52%	0.88978	5.13%
-XQ26	0.75652	2.44%	1.00000	1.01%	0.89523	5.13%
-XQ27	0.75190	2.44%	1.00000	1.52%	0.89123	5.13%
-XQ28	0.74960	2.44%	1.00000	1.52%	0.88318	5.13%

[a] As false positive results on the Y-chromosome occurred at high frequencies, all data for the Y-chromosome were excluded from analysis.

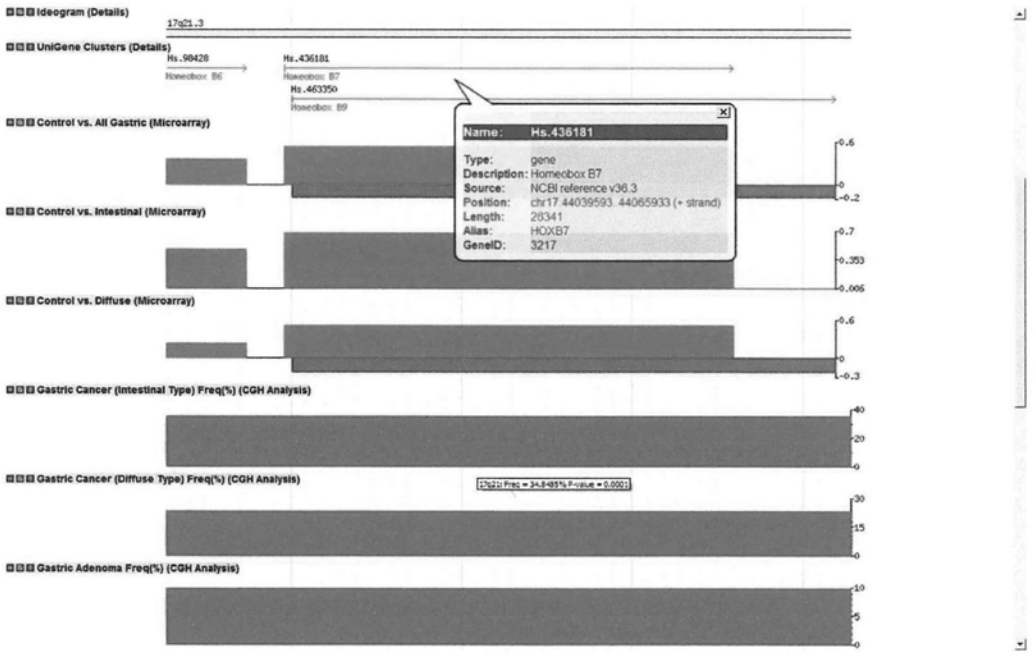
Appendix 6 Screenshots from Gbrowse displaying the cytogenetic and transcriptomic data of target genes



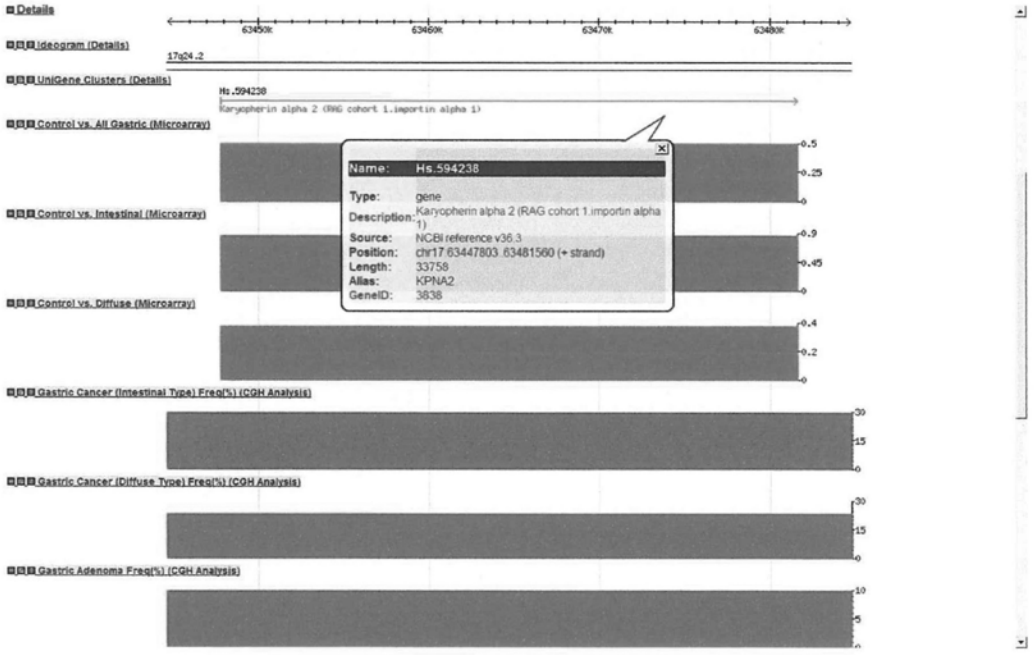
Appendix 6.1 Visualization of over-expression of UBE2C in both GA, IGC and DGC cases with the chromosomal gain at 20q13 where UBE2C located.



Appendix 6.2 Visualization of over-expression of HOXB6 in both GA, IGC and DGC cases with the chromosomal gain at 17q21 where HOXB6 located.



Appendix 6.3 Visualization of over-expression of HOXB7 in both GA, IGC and DGC cases with the chromosomal gain at 17q21 where HOXB7 located.



Appendix 6.4 Visualization of over-expression of KPNA2 in both GA, IGC and DGC cases with the chromosomal gain at 17q24 where KPNA2 located.