Recurrent Copy Number Alteration Analysis Identifies Risk Genes in Young Women with Breast Cancer

by

Chen Chi

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Biochemistry and Medical Genetics
University of Manitoba
Winnipeg

Abstract

Breast cancer (BC) diagnosis in young women (<45 years old) has come forth as an independent factor with higher recurrence risk and death than their older counterparts, and it has been suggested that it may exhibit its own unique biology. Copy number alterations (CNAs) have led many to consider them as an alternate paradigm for the genetic basis of human diseases, as these large alterations may encompass key genes that contribute to carcinogenesis and disease progression. Although many complex diseases have been linked to CNAs in the genomic DNA, prior studies have yet to document age-related changes in somatic CNAs for young women with BC. We hypothesize recurrent somatic CNA regions uniquely found in young women with BC harbor cancer susceptibility genes that modulate the survival of young women with BC. We aim to find recurrent somatic CNA regions identified from BC microarray data and associate the CNA status of the genes harbored in the regions to the survival of young women with BC.

We have developed a new interval graph-based algorithm for identifying recurrent somatic CNAs in cancer using a maximal clique detection technique. The algorithm guarantees that the identified CNA regions are the most frequent and identifies the delineated minimal regions. By using the algorithm on the Molecular Taxonomy of Breast Cancer International Consortium CNA data consisting of 2000 breast tumour samples (equally divided into a Discovery set with 130 young women and a Validation set with 125 young women), a total of 38 validated recurrent CNA regions with 39 protein encoding genes have been identified, along with 68 validated recurrent CNA regions that did not encompass any protein encoding genes. CNA gain regions encompassing genes *CAPN2*, *CDC73* and *ASB13* are the top 3 with the highest

occurring frequencies in both the young Discovery and Validation dataset, while gene *SGCZ* ranked top for the recurrent CNA loss regions. The mutation status of 9 out of the 39 genes has significant associations with survival outcome. Of particular interest, the expression level of 2 of the 9 genes, *ASB13* and *SGCZ*, also show significant association with survival outcome. Patients with CNA mutations in both of these genes resulted in a worse survival outcome when compared to patients without the gene mutations. Association and survival analyses demonstrated that the mutated CNA status in *ASB13* seems to lead to correspondingly higher gene expression, which is able to predict patient survival outcome. Together, identification of the deletion and amplification events that may be prognostic in young women with breast cancer can be used in genomic-guided treatment.

Acknowledgement

I would like to express my sincere thanks to my supervisor Dr. Pingzhao Hu for his full support, encouragement, and guidance for my research throughout the years I spent at the University of Manitoba. His insightful guidance, passion and enthusiasm for the research has been an invaluable motivation for me.

It is an honour for me to extend my gratitude to my supervisory committee members Dr. Leigh Murphy and Dr. Kevin Coombs for their patient guidance, helpful suggestions and encouragement which have kept me on track during my research and have been particularly instrumental during key academic milestones including the scholarship applications and the completion of this thesis.

Special thanks to my colleagues and friends at the Hu-Lab for their friendship, encouragement, relaxing coffee breaks and the abundant laughters: Qin Kuang, Kaiqiong Zhao, Anna Cheng, Jiaying You, Svetlana Frenkel, Rasif Ajwad, Ye Tian and Md Mohaiminul.

I would also like to thank all the faculty and staff members in the Department of Biochemistry and Medical Genetics who have helped me in various ways.

Finally, I would like to dedicate this thesis to my parents, whose constant encouragement, understanding and love have helped me through the ups and downs, and have contributed immeasurably to everything I have achieved.

Publications

C Chi, R Ajwad, Q Kuang, P Hu (2016). A graph-based algorithm for detecting recurrent copy number variants in cancer studies. *Cancer Informatics* Suppl.2, 43-50.

C Chi, R Ajwad, Q Kuang, LC Murphy, P Hu (2016). Recurrent somatic copy number variation analysis identifies risk genes that modulate the survival of young women with breast cancer. *The American Society of Human Genetics (ASHG) Annual Meeting*.

Table of Contents

List of Tables

List of Figures

Chapter 1: Background and Introduction

1.1 Biology of Breast Cancer in Young Women

Breast cancer is mostly known as an aging disease, with only 7% of the patients being diagnosed at a young age (<40 years of age) [1]. Yet, breast cancer diagnosis in young women has come forth as an independent factor with higher recurrence risk and death than their older counterparts in various studies [2-6]. Young women (<40) with breast cancer have been described to depict more biologically aggressive tumours (basal and HER2-enriched subtypes) than their older counterparts (>40), which has been associated with a poorer prognosis (Table 1, modified from [6]).

Table 1. Clinical information for young and older patients. The combined data set includes 2,562 arrays/patients. Breast cancer patients are classified to young (< 40 years old) and older (>=40 years old) groups with each of the PAM50 tumor subtypes.

Characteristic	All (n=2562)		Younger (<40, n=210, 8.2% of total n)		Older (>=40,n=2352)	
	No.	%	No.	%	No.	%
Lum A	991	38.7	36	17.1	***955***	***40.6***
Lum B	662	25.8	31	14.8	***631***	***26.8***
HER2	380	14.8	***45***	***21.4***	335	14.3
Basal	529	20.7	***98***	***46.7***	431	18.3

Several factors influence poor prognosis in the young subgroup, such as higher tumour grade at diagnosis, high tumour proliferation, increased expression of HER-2 (*ERB-B2*) and reduced expression of both estrogen (ER) and progesterone receptor (PR) [7]. These women often struggle with life issues that are either absent or much less severe in older women, such as the possibility of early menopause and effects on fertility. While clinicopathologic differences point to underlying biological differences between the breast tumours derived in younger versus older women, limited studies have documented age-related changes at the molecular level. At the molecular level, it is now known that breast cancers exhibit at least 5 different intrinsic subtypes via PAM50 gene-clustering analysis on gene expression patterns: normal-like, luminal A, luminal B, HER2 enriched and basal-like subtypes [8]. Young women with breast cancer have been shown to present with more biologically aggressive tumour subtypes (basal and HER2-enriched) than their older counterparts, which has been associated with a poorer prognosis. In addition, heterogeneity at both the clinicopathologic and molecular level impact treatment and prognosis. For example, a study looking at 2929 young breast cancer patients showed that their clinical outcomes varied greatly despite being administered same treatment regimes for older patients with similar clinicopathologic factors [9].

It has been suggested that breast cancer in young women may be unique from their older counterparts via both immunohistochemical (IHC) and molecular classification studies [2,10]. Previous research has identified four subtypes: luminal A (less-aggressive subtype), luminal B, HER-2 enriched and triple negative (more-aggressive subtypes), which have prognostic relevance [11,12]. Evaluation of the previously identified four subtypes (luminal A/B, HER2 enriched and triple-negative/basal) in a cohort of 2970 young patients revealed that there were significantly more triple-negative subtypes and significantly fewer luminal A subtypes in the

"very young" sub-cohort (<35) when compared with the "less young" and postmenopausal women [13].

Similarly, Thomas *et al.* have reported that the "core basal" subtype was more commonly seen among younger women under 40 years old as compared to older women. Specifically, 18% of the breast cancers were "core basal" (ER-/PR-/HER2-/CK5+) among patients aged ≤40 as opposed to only 7% among patients aged >40 [11,12]. Other studies have also identified luminal subtypes in older patients, with triple-negative subtypes (ER-/PR-/HER2-) being overrepresented in women under 40 years of age [3,14].

Finally, Azim *et al.* [4] evaluated the prognostic significance of previously published gene signatures relating to stroma, immunity and proliferation in breast cancers arising in young women (<40 years of age) compared to older women. Stromal gene signatures had prognostic value only for young women with ER-negative, HER2-negative breast cancer, but not for older women, suggesting a role for the microenvironment in mediating breast cancer growth and proliferation in young women, leading to a more aggressive phenotype [4]. These findings with significantly more of the aggressive types of breast cancers (i.e. HER2 enriched and triple negative) in young age group results in an increased risk of relapse when compared to older counterparts with the same subtypes [13], suggesting that younger patients with breast cancer may exhibit a unique biology.

Therefore, it is warranted to mine genome-wide expression profiles to find other breast cancer genes and pathways with strong potential for prognostic significance as a function of age. Age-specific genomic signatures could guide not only young women with breast cancer who have received little benefit from adjuvant chemotherapy but also those who might be more

suitable for other types of adjuvant therapies. Given that approximately 40–50% of young breast cancer patients relapse after 5 years [15], these age-specific signatures could also serve as a tool to identify young patients that would gain more benefit from particular adjuvant therapies.

1.2 Sequence variations: SNPs

So far, amongst the numerous insights gained from the completion of the Human Genome Project, single nucleotide polymorphisms (SNPs) have been recognized as a major source of genetic variation, which led to the speculation that the majority of phenotypic variations in humans are due to SNPs [16]. Consequently, an immense amount of research has focused on developing genotyping assays and methodologies for SNP analysis, making it the most frequently assayed type of intraspecific genetic variation and the most centered gene-mapping studies with regards to the relationship between SNPs and human diseases.

However, growing number of studies have confirmed that more nucleotide bases are affected by CNVs as opposed to SNPs between any two individuals [17]. Aided by the advancement of high-throughput sequencing technologies and array comparative genome hybridization (aCGH) to detect the magnitude and location of genomic alterations within a single human genome, it is clear that large fragments of our genome have been deleted or duplicated.

1.3 Structural variations: CNVs

It was commonly thought that individuals of the same species have very similar genomes. However, human genomes are not the stable places people once thought they were; the traditional perception that genes always occur in two copies per genome has been defied by the growing evidence for structural variations in the genomes of different individuals.

Structural variation includes rearrangements (inversions and translocations) and copy number variation (CNV), consisting of duplications, deletions, insertions and multi-allelic variations ranging from one kilobase (kb) to several megabases of DNA, leading to possible dosage imbalances. In other words, genes that were originally understood to be present in two copies have now been revealed to sometimes exist in one, three, more than three copies or even missing altogether. We define a CNV as a DNA segment that is 1 kilobase or larger at variable copy numbers when referred to a reference genome [16].

These structural variations can change the genes copy number that encompass the affected regions and alter gene regulation. In particular, one group of scientist conducted a pilot study in mapping the CNVs in the complete human genome [17], which demonstrated the remarkable extent of large structural variation in the human genome, both within closely related people and between the global populations due to sizeable duplications and deletions of genomic segments. Likewise, similar studies have shown that despite the existence of powerful repair mechanisms in the human genome, CNV occurrence frequency is 100–10,000 times higher than point mutations [18].

These copy number changes may be germline variants or somatic mutation in the case of cancer samples. Studies have shown that some types of genes are more prone to be copy number variable than others, such as the genes involved in immunological and neurological development, possibly due to the rapid human evolution in these two functions [19,20]. On the other hand, genes that play a role in early development and mitosis, which are fundamental to life, are likely to be spared [20].

In some cases, having CNVs in the genome can be advantageous. For instance, extra copy number of genes may offer redundancy in the sequence, thereby liberating those genes to evolve de novo or altered functions for adaption, while other copies can sustain the original function [21]. Proteins might also gain new domains and acquire new or adapted functions. However, the disadvantage seems to outweigh the advantages, especially when there are too many CNVs in the genome [22]. CNVs have been reported to play a role in cancer susceptibility, formation and progression [23]. In many cases, a change in the copy number of any genes is not well accepted, leading to possible genomic disorders. Importantly, a few association studies have already elucidated the significance of CNVs as disease-susceptibility variants [24-27]. In one study, CNV detected in 1,400 regions have overlaps with 14.5% of human disease-causal genes [24]. Recent research concluded that CNV is frequently found in susceptible individuals who were predisposed to diseases such as colour blindness in Mendelian diseases, as well as in complex traits diseases such as autism, inflammatory bowel disease and cancers. [25-27].

1.4 Copy Number Alterations (CNAs)

Cancer progression is impelled by the accumulation of somatic genetic mutations, which consist of single nucleotide substitutions, translocations and somatic mutations [28]. Somatic mutations are non-heritable alterations to the human genome that occur spontaneously in somatic cells, which is often due to DNA replication error or chemical/UV radiation. Copy number alterations (CNA) are somatic changes in the copy numbers of a DNA sequence that arise during the process of cancer development. This results in changes to the chromosome structure in the form of gain or loss in copies of DNA segments, and are prevalent in many types of cancer [29].

Investigating these genomic alterations in breast cancer patients not only can offer valuable insights into breast cancer pathogenesis and discover potential biomarkers, but can also provide novel drug targets for better therapeutic treatment options [30]. Several cytogenetic and array-based studies have detected recurrent alterations linked with certain cancer types, and have found copy number alterations (CNA) to be a particularly common genetic mutation in cancer [31, 32]. In addition, some of these CNAs have resulted in the discovery of disease causal genes and novel therapeutic targets, and have been strongly associated with clinical phenotypes [33-36]. For example, the use of vemurafenib to inhibit BRAF V600E mutation has shown remarkably improved survival in melanoma patients [37]. In another study, treatment with tyrosine kinase inhibitors for EGFR lung cancer has also shown great success [38].

It has been reported that 85% of the variation in gene expression of breast tumours is due to somatic CNAs at gene loci [39]. Importantly, CNAs often encompass genes including oncogenes and tumour suppressors (i.e. driver genes), and this can directly affect both cancer development and disease progression. For instance, one study discovered CNAs in the gene ZNF703 on chromosome 8p as independent prognostic factors for luminal B breast cancer, as they are associated with a worse outcome in breast cancer patients who have acquired it than without [40]. Similarly, CNAs in 3q26.2-q29, 3p26.3-p11.1, 17p13.3-p11.2 and 9p13.3- p13.2 have been deemed as risk predictors of lung cancer [41]. Consequently, incorporating CNA analysis of breast tumours can offer important insights for disease survival outcome and the molecular profile of breast cancer.

One study in glioblastoma has shown that CNAs can be used to define the key pathways of a tumour [42], where an integrative analysis has reported that over 70% of the tumours have copy number variations in the retinoblastoma, p53 and receptor tyrosine kinase pathways. This demonstrates that copy number profiles, when incorporated with other high-throughput data such as mRNA expression levels and methylation changes, can help unravel a bigger portion of the cancer's genetic basis. Furthermore, tumour cells accrue genetic aberrations including CNA in DNA, which leads to defectiveness of major regulators. This presence of altered DNA copy number may provide growth advantages to cells, which will be selected for and lead to cancer formation. Changes in gene dosage are crucial in cancer development, as oncogenes may be triggered by DNA amplification and tumor suppressor genes may be inactivated due to a DNA deletion.

Since CNAs often encompass genes, it is suspected that they may greatly influence gene expression of the CNA regions. Indeed, several studies have reported a correlation between CNA and the average global expression levels of genes in the CN variable chromosomal region. For instance, one group has shown that in the tumour formation from an immortalized prostate epithelial cell line, 51% of up-regulated genes were mapped to DNA gain regions and 42% of down-regulated genes were mapped to DNA loss regions [43]. This was further supported by another group working with breast tumour cell lines, noting that DNA copy number influences gene expression across a range of CNAs, with 62% of amplified genes resulting in moderately or highly elevated expression [44].

1.5 Breast Cancer Signature

The complexity in breast cancer prognosis shows that the traditional "one size fits all" approach is inadequate in managing breast cancer treatments. Personalized medicine can incorporate the use of genetic profiles in the tumours of young women along with the clinical-pathologic factors to enhance diagnosis and treatment decisions. In particular, molecular profiles such as gene expression signatures have been introduced as clinical tools to help predict prognosis and guide treatment decisions for certain breast cancer patients, by means of risk group stratification.

Development of molecular signatures are mainly derived from two different approaches: top-down or bottom-up. The top-down approach is data-driven, in which global expression patterns consisting of thousands of genes are being examined without a priori biological knowledge [45]. First, gene expression levels are being measured by several probe sets. Through stringent filtering criteria, genes that have significantly differential expression patterns amongst case and control groups are being selected for, whilst genes that show little or no expression levels are being omitted. Subsequent association testing such as survival analysis is carried out to further select candidate genes that show prognostic significance in the subgroup of breast cancer patients. For example, Mammaprint is a 70-gene expression signature developed by measuring the mRNA expression profiles of patient groups with distinct clinical outcomes [46]. It remains a robust signature for predicting distant metastasis-free survival. Likewise, the Rotterdam 76-gene expression signature [47] and Endopredict [48] were also developed using the top-down approach, which could predict breast cancer patients at high risk for distant metastases for all age groups.

On the other hand, the bottom-up approach is hypothesis-driven, which involves extensive literature reviews to identify candidate genes that are disease- or functionally-relevant and correlating them with survival data. As an example, the Oncotype DX 21-gene signature was developed by selecting from 250 candidate genes based on published literatures and databases [49]. It provides a recurrence score that can be used to stratify ER-positive, node-positive or negative breast cancer patients into high, intermediate or low risk groups to better determine who may or may not benefit from aggressive chemotherapy treatments. The Genomic Grade Index (GGI) was another well-known signature using the "bottom-up" strategy [50]. Validation study has demonstrated its ability to stratify intermediate-grade tumors into prognostic subgroups via large-scale gene expression profiling data [51].

Ultimately, these transcriptome signatures can provide additional information beyond standard parameters to enable doctors and patients to be more confident in the treatment decision-making process. It illustrates a more accurate risk assessment based on understanding individual tumour biology, which contributes critical information and certainty to make informed, personalized treatment decisions. However, it should be noted that none of these signatures have reached 100% accuracy in their ability to predict patient outcomes and treatment decisions, more data and validation studies are needed for this purpose [52]. It should also be noted that age is not taken into account as an important factor in most of these breast cancer signatures. Furthermore, RNA signatures rely on the RNA integrity and the possibility of stromal contamination [53]. Therefore, it is worthwhile to investigate the copy number alterations occurring at the DNA level as a way to study the underlying genetic composition of breast tumours, as it is more stable than RNA and also less likely to be swayed by physiological conditions.

1.6 CNA Detection Methods

This intriguingly frequent and dynamic type of genomic variation has challenged the concept of analyzing the genome through the breakdown of a single diploid human reference genome, and has also triggered more research on the techniques of detecting CNAs. Array comparative genome hybridization (aCGH) is a popular experimental technique for detecting copy number variants in genomes [54-56]. aCGH requires the hybridization of fluorescently tagged differential DNA fragments from both a test genome and a reference genome to a set of probes originated from the reference genome sequence. The proportion of the test vs. reference fluorescence intensity at each probe will identify the positions in the test genome that have fewer, more, or similar copy as the reference genome, which can generate a copy number profile of the test genome [54-56]. These CNA profiles are generally compared across individuals in a group of interest to identify common CNAs that are shared amongst a portion of the group.

Another method that is popular for CNA detection and analysis is by using high-throughput array technologies for SNP genotyping from commercial companies such as Affymetrix and Illumina, due to their ability to perform a dual role for both SNP-based and CNA-based association studies. One widely used array is the Affymetrix genome-wide human SNP array 6.0 (Affy6), which is an array platform that aims to perform both high-density SNP genotyping and high-resolution CNA discovery simultaneously. Aside from the 906,600 SNP probe sets, the Affy6 array also contains 946,000 copy number probe sets that can be used to assess chromosomal copy-number changes in regions of the genome that are not well covered by SNPs.

Over the recent decade, whole genome sequencing technologies, particularly the Next-Gen Sequencing (NGS), have evolved and revolutionized the way to deal with the complexities of genomes. It is a comprehensive technique to generate millions of sequences simultaneously, which rapidly reduces the cost and time required to generate large amounts of genomic data. Unlike the microarrays which contain a specific set of known targets, NGS technology is capable of interrogating the entire genome or transcriptome, independent of pre-chosen targets. It is ideal for discovery studies such as identifying disease-causal variants and novel genome variations.

1.7 CNA Calling Methods

A number of CNA calling methods for individual patients have been implemented, which either follows the Circular Binary Segmentation (CBS) method or the Hidden Markov Model (HMM) method. CBS is a segmentation-based method that scans for change points in an ordered sequence of values to delineate segments with different distribution of values (measured by having different means). In other words, it will recursively divide up the chromosome until segments that have probe distribution different than neighbours have been identified [57]. For the HMM method, the aim is to uncover the hidden copy number states (0,1,2,3 copies etc) by searching the data point by point to determine the most probable copy number states based on observation, and transitions between states correspond to changes in copy number [58,59].

These methods share a common feature in which the CNA regions are segmented by individual-specific breakpoints, and detection is carried out sample-by-sample. However, it is much more likely for shared/common CNA regions (i.e. recurrent CNA) to occur at the same genomic positions across different individuals in a homogeneous group of people. As a result, recurrent CNA regions are more likely to harbor disease-causal genes, as it is more probable to

encompass “driver” alterations (functionally significant for disease initiation or progression), while individual sample CNAs are subject-specific and would be more likely to contain “passenger” alterations (random somatic events irrelevant to pathological events) than disease-relevant alterations [60]. Several methodologies have been proposed for recurrent CNA detection [61-66]. These methods mainly differ in the type of input data and the algorithm models being implemented. For the input, most of the recurrent CNA detection methods fall under two categories: continuous (log2 ratio) [62,64,65] and discrete (Gains/Losses) [61,63,66]. For the algorithms, they can be categorized into three main models: permutation [61,64], probabilistic null model [62,65,66] or none [63].

Chapter 2: Motivation, Hypothesis and Research Objectives

2.1 Motivation

From a biological point of view, our motivation comes from the fact that not every mutation in cancer drives cancer. This is something that was not well understood in the past but has now become a fundamental issue in understanding cancer genomes. That is: how do cancers arise? Cancers arise because as cells in the body divide, they acquire mutations at random throughout the genome. Most of these mutations are going to be neutral or deleterious to the cell, and not lead to cancer. But occasionally one cell may undergo a series of events that transform into cancer, then that cell is going to divide and all its daughter cells will share those driver mutations along with any passenger/neutral mutations that preceded it. Therefore, how to distinguish driver mutations from passenger mutations is a major challenge. The approach we take is to apply statistics. From a single tumour sample one could not distinguish between the drivers and passengers, but since every cancer requires driver events, the driver events tend to recur. Therefore, when many samples are available, in order to distinguish functional driver mutations from passenger mutations, we can identify those regions that are recurrent across multiple samples.

However, this is a very difficult problem since CNAs vary widely in length and position across different genomes, and this forms a complex pattern of overlapping CNAs, making it challenging to detect which gene of interest is the true target of the alteration. Early methods for detecting recurrent CNAs all have a similarity in their methods: compute a score at each genomic locus indicating the recurrence score, examine the correlations between recurrence scores of nearby loci to separate true target regions from the rest. The problem with that is that closely

located driver mutations lead to correlation between the recurrence scores, and these methods are unable to address how to separate these regions into independent copy number events.

From a clinical point of view, breast cancer treatment regimens are based largely on tumor characteristics such as tumor size and grade, ER, PR and HER2 status. Yet, traditional pathological assessment is inadequate in making accurate predictions of outcomes, and incorporation of molecular profiles may provide better prognostic evaluation. For instance, the PAM50 gene signature is able to identify intrinsic subtypes of breast cancer based on gene expression profiles and distinguish the subgroups with different biological features such as tumour stage, tumour grade, recurrence rates and sites of metastasis. However, it is speculated that even within the subtypes, patients can differ drastically in their outcome, such as response to homogeneous treatments and the rate of recurrence. Therefore, it is important to investigate whether tumours in young women share commonalities in genetic alterations first, regardless of subtypes.

2.2 Hypothesis

Copy number alterations, which consist of duplications and deletions of DNA segments, may encompass key genes that contribute to carcinogenesis and disease progression. Therefore, we hypothesize that recurrent CNA regions uniquely found in young women with breast cancer may harbor cancer susceptibility genes that may modulate their survival.

2.3 Research Aims

We have four aims: (1) develop a graph-based algorithm to identify recurrent CNA regions from breast cancer genomic data; (2) identify young-specific prognostic genes based on

CNA status and/or gene expression for the genes in the identified CNA regions; (3) explore the relationship between CNA mutation status and gene expression; (4) identify potential pathways enriched in the young-specific prognostic candidate genes.

Chapter 3: Materials and Methods

3.1 Data source

All breast cancer data were retrieved from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [67], which is a novel dataset consisting of comprehensive clinical features such as breast cancer-specific survival data, PAM50 subtypes, ER/PR/HER2 status, tumour grade and tumour sizes. It also contains corresponding whole gene expression profiles (Illumina HT-12 v3 platform), SNPs and individual patient level DNA copy-number profiling data (Affymetrix Human SNP 6.0 platform). The genome build for the CNA data is hg18, as well as all of the coordinates being stated throughout the thesis. Treatments for the patients are homogeneous among each clinically relevant group: almost all ER-positive/LN-negative patients did not receive chemotherapy, while ER-negative/LN-positive patients did receive chemotherapy. Furthermore, none of the HER2 patients received trastuzumab [67].

All samples are derived from ~2000 clinically annotated primary fresh-frozen breast cancer specimens from tumour banks in the UK and Canada (a discovery set of 997 primary tumours and a validation set of 995 tumours). All genomic and clinically annotated data is publicly available at the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/), under accession number EGAS00000000083 [67].

3.2 Experimental assay and genotype calling

Please note that the procedures stated in **sections 3.2-3.4** have been carried out by the original METABRIC study [67]. DNA was extracted from each tumour specimen and subject to copy number analysis on the Affymetrix Human SNP 6.0 platform. Data from Affymetrix SNP 6.0 arrays was pre-processed and genotyped using the SNP-RMA (Robust Multichip Analysis) algorithm, available in the crlmm Bioconductor R package [67]. This quantifies raw intensity values into proportional amount of DNA in the target sample associated with each of the alleles, A and B, for each SNP. Feature intensities were corrected for fragment length and sequence effects, followed by quantile normalization to a predefined reference distribution. Intensities were then summarized by median polish, with a single value for each allele. A mixture model was then used to adjust for remaining fragment length and intensity-dependent biases on the log-ratio of the summarized intensities. Samples with a signal-to-noise ratio < 5 were flagged in downstream analyses [67].

3.3 Pre-processing of copy number alteration analysis

Affymetrix SNP 6.0 arrays were pre-processed for copy number segmentation using aroma.affymetrix. Both tumour and normal samples were independently normalised using Robust Multichip Analysis [68] (CRMAv2), along with a publicly available SNP 6.0 dataset consisting of 270 HapMap individuals. For each sample, allelic-crosstalk calibration, probe sequence effects normalization, probe-level summarization, and PCR fragment length normalization were performed in order to obtain log2 intensity values for total copy number

estimation. Afterwards, probes were sorted by their genomic position, replicate probes were summarized by their median value, and missing values (generated by negative intensities in the normalization) were imputed using the loess procedure included in the snapCGH Bioconductor R package.

Two pooled references were generated, one using the median intensities across the HapMap individuals and another for the normals and tumours, using the median intensity values from a set of 473 normals. Next, log2 ratios were generated for the HapMap samples by subtracting the pooled value from the log2 intensities. Similarly, log2 ratios were obtained for the 473 normals using the corresponding pool. For the 997 tumour samples, two data sets were produced: one using the normal pool as the reference for all the tumours and another using the matched normal for each tumour when available, and the normal pool for the remainder. A similar approach was taken for the validation set.

3.4 Calling patient-level CNAs

The HapMap and normal datasets were used to estimate the frequency of germline CNVs in the cohort, while the tumour samples were used for estimating somatic CNAs [67]. After computing the log2 ratios for each probe, samples were segmented using the circular binary segmentation (CBS) algorithm implemented in the DNAcopy R Bioconductor package and individual patient level CNVs were called. For the tumour samples, any segmented mean that fell within a region included in the HapMap+Normals CNV list was labeled as an inherited CNV. In order to remove all possible germline CNVs, the frequencies of somatic CNAs in the tumour

samples were obtained after removing the germline CNVs from the normalized pool reference. For the Discovery dataset, a total of 111,460 individual patient level CNA gains and 54,316 individual patient level CNA losses were detected. For the Validation dataset, a total of 113,213 individual patient level CNA gains and 60,536 individual patient level CNA losses were detected.

For calling alterations, the thresholds for gains and losses were set to +2 σ and -2.5 σ (σ is the standard deviation of the log2 ratio for each array) respectively [67]. The asymmetry in the thresholds results from the assumption that one copy gain is 3/2 whereas one copy loss is 1/2. Upon log2 transformation and obtaining the absolute value of the gain and loss, the threshold for the gain is smaller in magnitude than for the loss ($|\log2(3/2)| < |\log2(1/2)|$). In order to identify common CNAs, at least five samples in a particular population were required to exhibit an alteration in the same probe. Consecutive probes were merged to call copy number variable regions and a list of CNAs was generated after merging the HapMap populations with the set of 473 normals. For tumour samples, the data were smoothed and DNAcopy was run with default parameters [69]. The MergeLevels algorithm [70] was then applied to the segmented data. Since it is often to see a mixture of normal (benign) and tumour cells in a given sample, the dependence between cellularity and the alteration proportions needs to be removed. In order to achieve this, different thresholds were used for calling alterations and high-levels events according to the cellularity of each sample, resulting in the following somatic copy number states:

KCNA = {HOMD; HETD; NEUT;GAIN; AMP}

For tumours with high cellularity, the median of the log2 ratio +2 σ or +6 σ was computed for the 50% of the central probes to call gains (GAIN) and amplifications (AMP), respectively. For losses, the median of the log2 ratio -2.5 σ or -7 σ was used to call heterozygous (HETD) and homozygous (HOMD) losses, respectively A median in between -2.5 σ and +2 σ is considered to be neutral (NEUT).

For tumours with moderate cellularity, the median of the log2 ratio of each array +2 σ or +6 σ was computed for the 45% of the central probes and used to call gains and amplifications, respectively. For losses, the median of the log2 ratios -2.5 σ or -7 σ was used to call heterozygous and homozygous losses, respectively.

3.5 Identification of recurrent CNAs

3.5.1 Representing CNAs as an interval graph

We denote a CNA segment as $R_j=(l_j,r_j)$, where j is the j^{th} CNA and l_j,r_j are its left and right chromosome positions. For a CNA set, we have $R=\{ R_1, R_2 ,\ldots, R_n\}$. When j is infinite, we call R a right-censored univariate data set. An intersection graph can easily be constructed from R as follows: Each member in R corresponds to a vertex which we denote by its index. Hence, R_j corresponds to vertex j. We denote the set of vertices as V. Two vertices j and k are linked by an edge if the corresponding members R_j and R_k in R are intersected. We denote the edge as jk and the set of edges as E. When the R is a linearly ordered set, the intersection graph is called an interval graph and all interval graphs are triangulated [71].

Figure 1. Representing CNAs as interval graphs.

A) A, B, C, D, E, F are individual patient level CNAs on a specific chromosome. Each of the CNAs has chromosome start and end positions. **B)** This is an interval graph where A, B, C,D,E,F are the individual patient level CNAs in **A)**. The edge between each of two vertices in the graph represents the two individual patient level CNAs share a piece of common regions on the chromosome.

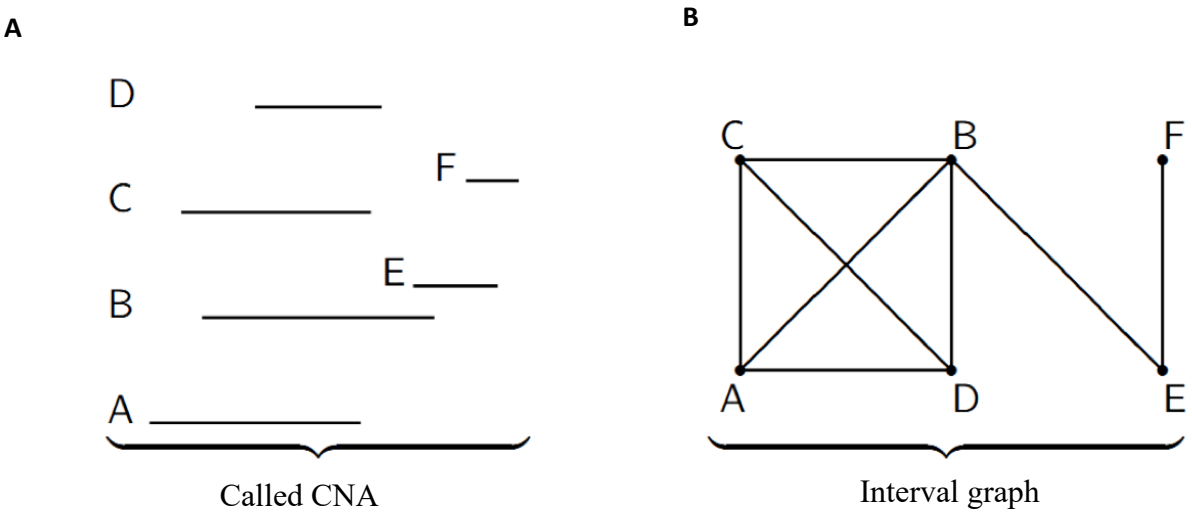


Figure 1A shows the examples of six individual patient level CNA segments (A, B, C, D, E, F) on the same chromosome. Each of the six CNAs contains chromosomal-specific start (left) and end (right) positions. To identify the common regions of individual patient level CNAs on the same chromosome, the intersection among the individual patient level CNAs can be represented as an interval graph, treating each called individual patient level CNA as a vertex of the graph and connecting two vertices only if the corresponding intervals have an intersecting region. Thus, the constructed interval graph $G(V, E)$ is comprised of a set of vertices V, where each

vertex ($v \in V$) corresponds to a specific interval of the individual patient level CNA and each edge ($\{u, v\} \in E$) connects two intersecting intervals u and v. In **Figure 1B**, an example of the interval graph is shown where A through E are the intervals (nodes of the graph or individual patient level CNAs) and an edge connects two nodes (individual patient level CNAs) if the intervals overlap. For instance, interval F only connects with interval E in **Figure 1B** because it overlaps with E only, and does not overlap with intervals A-D in **Figure 1A**.

3.5.2 Finding maximal cliques from an interval graph

A clique is a set of vertices in which any two vertices are connected by an edge in the interval graph. A maximal clique is a clique which cannot be a sub-clique of a larger clique. In the context of a CNA set R a clique can be viewed as a set of CNA segments whose regions intersect. For example, **Figure 1B** shows that {A,B,C,D} is a maximal clique, as it cannot be extended by adding any other vertices. However, {A,C,D} is not a maximal clique but a clique, as it can be extended by adding vertex B to it.

To find maximal cliques in an interval graph constructed from individual patient level CNAs, we applied Gentlemen and Vandal's algorithm [72]. The main idea of the algorithm is to sort the vertices based on their chromosomal end positions. The ordering is important because it allows the algorithm to discard vertices in each iteration without losing the triangulation property of an interval graph. The input of the algorithm is the individual patient level CNAs on a specific chromosome, which include two parameters for each CNA segment: start and end positions (base pair). It should be noted that we need to analyse CNA gains and losses separately. The algorithm adapted for processing individual patient level CNA data is summarized as follows:

Algorithm Finding the maximal cliques M

1. Sort all the vertices in terms of their chromosomal end positions
2. Initialize $M = \{\}$ and $k = 0$, where k is the k^{th} maximal clique
3. For each vertex v, initialize $S(v) = 0$, where $S(v)$ is the number of neighbours of v
4. For each vertex v, check

   if adjacent neighbor of v, $adj(v)$ is empty,

   $k = k + 1, M_k = \{v\}, M = M \cup M_k$

   else

   $X_v = adj(v)$, where X_v is the set of neighbours of vertex v

   $S(u) = max\{S(u), |X_v| - 1\}$, where u is the next vertex to be eliminated

   if $S(v) < |X_v|$

   $k = k + 1, M_k = \{v\} \cup X_v, M = M \cup M_k$

   else eliminate v

5. return M

The output of the algorithm will be a list of maximal cliques. We implemented the algorithm using R package Icens [71], which implemented the algorithm to find maximal cliques of a triangulated graph based on the fact that all interval graphs are triangulated [72]. The

method is efficient and the time complexity of the maximal clique detection algorithm is $O(n + e)$, where n is the number of individual CNAs and e is the total number of edges in the corresponding graph.

3.5.3 Analysing recurrent CNAs from the maximal cliques

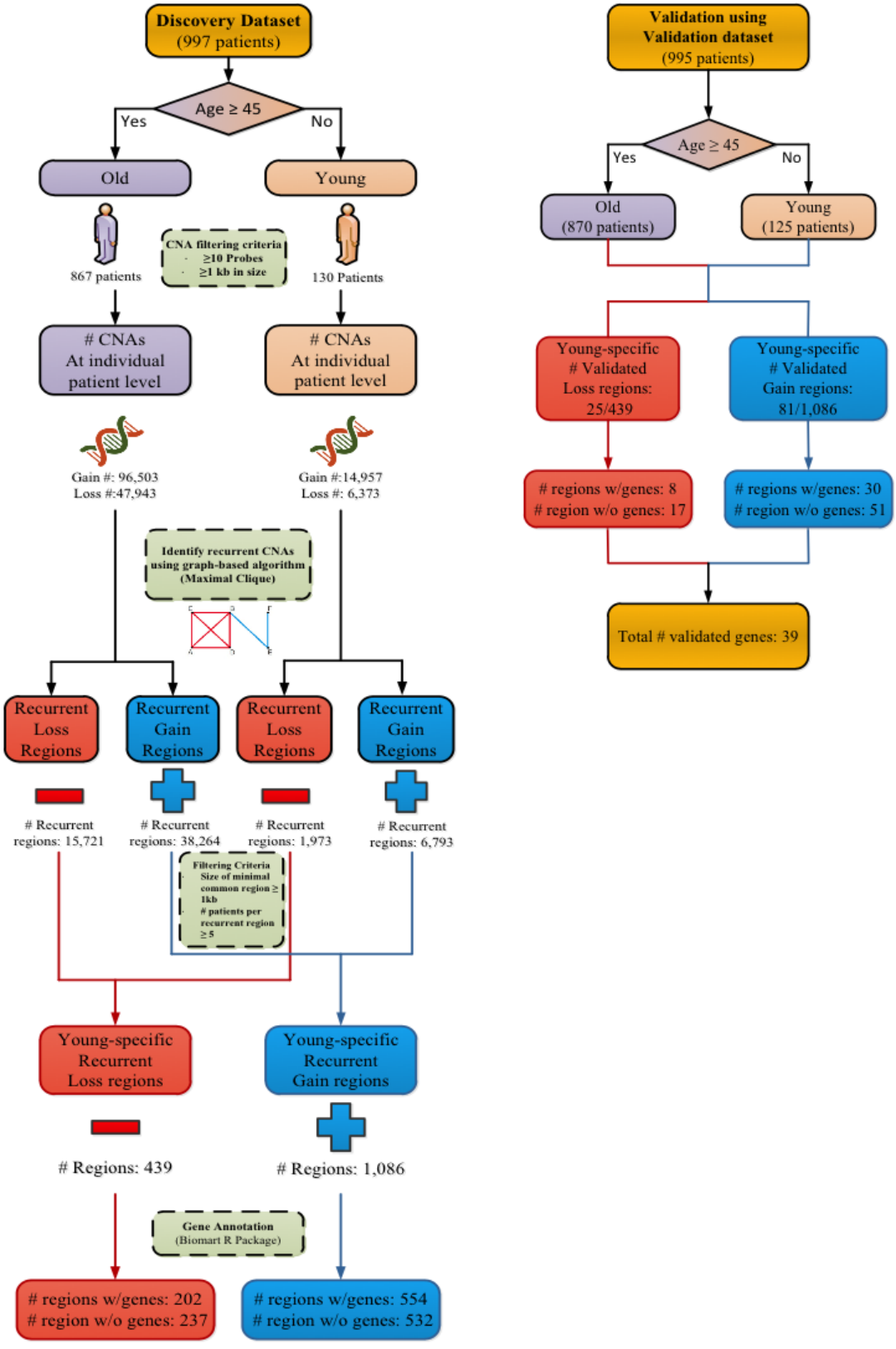
Each of the identified maximal cliques is a recurrent CNA, which is common in multiple patients. The shared region of the recurrent CNA across multiple patients is the minimal common region (MCR) of the CNA, which has the potential to harbour cancer-causing genes. In practice, the size of the maximal cliques should be at least 2 and the size of the MCRs should be at least 1kb.

Unlike Gentlemen and Vandal's algorithm to identify maximal cliques, Wu et al. also proposed an algorithm [73] to identify maximal cliques for detecting recurrent CNAs. However, this algorithm is based on a scoring scheme where blocks of consecutive maximal cliques were scored, defining a *pivot* within the block and calculating the number of left and right end position *pairs* that crosses that pivot.

Figure 2 shows our analysis flowchart using the maximal clique-based recurrent CNA detection. The individual patient level CNA data in Discovery dataset containing 997 patient samples was separated into two CNA types: gain and loss. Changes in gene dosage is crucial in cancer development, as oncogenes may be triggered by DNA amplification (i.e. CAN gain) and tumor suppressor genes may be inactivated due to a DNA deletion (i.e. CAN loss). Filtering criteria include retaining CNA data that was generated by ≥10 probes and having a CNA size of at least 1 kb. Amongst the total 997 patients, 867 patients were classified into the old age group (≥45

years old) and 130 patients into the young age group. For the old age group, there are 96,503 and 47,943 individual patient level CNA gain and loss regions respectively. For the young age group, there are 14,957 and 6,373 individual patient level CNA gain and loss regions respectively. The recurrent CNA calling algorithm was ran separately for the CNA gains and CNA losses, and analysis was done chromosome by chromosome. Further filtering at the recurrent CNA level includes retaining those that have a minimal region of at least 1 kb, and the number of patients per recurrent CNA region to be at least 5. The recurrent gain and loss regions detected in the old and young age group were then compared side by side, and regions that were uniquely found in the young age group were then extracted to form the young-specific recurrent gain and loss regions.

Figure 2. Analysis Flowchart for implementation and application of proposed algorithm to the METABRIC CNA data.



3.6 Statistical Analysis

3.6.1 Survival analysis

Survival analysis was performed for both the mutation status (CNA gain, CNA loss) by the product-limit method and the expression level of the corresponding genes that are encompassed in the validated recurrent CNA regions. Gene expressions were evaluated by using the Cox proportional hazard regression model. This enables the difference between survival times of groups if patients to be tested while allowing for other factors. The response/dependent variable is survival time and event, whereas the independent/predictor variable is gene expression data. The Survival R package is used for this analysis. It takes in the function of Surv(time, event)~predictors, where time is either the exact time of death or the censored data, event is coded as 1 (dead) and 0 (alive) depending on whether the event is dead or alive, and predictors being the gene expression and other potential confounding factors.

The cox proportional hazard regression model can be written as follows:

$$h(t) = h_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)$$

Hazard Ratio (HR) = $e^{-\beta}$

Where

$h(t)$ = the expected hazard at time t (rate of death in the next instant),

$h_0(t)$ = the baseline hazard when all of the predictors X_1, X_2 , X_p are equal to zero ($\beta = 0$).

β = regression coefficients for the corresponding predictor variables, where X1 is the gene expression while others X_2 , X_p are covariates adjusted in the model.

HR = ratio of the hazard in the intervention group / Hazard in the control group

The product-limit method of estimating a survival function, also known as the Kaplan-Meier method, is a nonparametric technique that uses the exact survival time for each individual in a sample instead of grouping the times into intervals. This method for estimating a survival curve is used in order to account for the partial information about survival times that is available from censored observations, such as when participants drop out of the study or if the event of interest has not been experienced in the study time period.

The Kaplan-Meier estimator can be written as follows:

For $t \in [t_j, t_{j+1})$,

$$\widehat{S}(t) = \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right)\cdots\left(1 - \frac{d_j}{n_j}\right)$$

$$= \prod_{i=1}^{j}\left(1 - \frac{d_i}{n_i}\right)$$

Where $\widehat{S}(t)$ = Probability of surviving at least to time t

t = actual time of death; t_j = actual time of death at time interval j

d_j = # of deaths that occur at time interval j

n_j = # of patients remaining at time interval j

3.6.2 eQTL analysis

An expression quantitative trait locus (eQTL) is a locus that explains a portion of the genetic variation of a gene expression phenotype. An eQTL analysis tests for direct associations between markers of genetic variation with gene expression levels; that is, to evaluate the

association between gene expression and CNA mutation status. Logistic regression is used to estimate the probability p associated with a dichotomous response for various values of an explanatory variable. The reason for choosing to do a logistic regression over a linear regression is that when the outcome is categorical, it is much easier to interpret and make sense of the result. In this case, the response (dependent) variable is gene expression (binarized-by-mean) and the predictor (independent) variable is CNA status.

The logistic regression model can be written as follows:

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Odds Ratio (OR) = $e^{-\beta}$

Where P = the expected probability that the outcome will occur

$X_1 \ldots X_k$ = predictor/independent variables

$\beta_0 \ldots \beta_k$ = regression coefficients for the corresponding predictor variables

OR = the odds that an outcome will occur given X, compared to the odds of the outcome occurring without X

3.7 Functional analysis

Functional analysis such as enrichment and annotations were carried out using softwares: Enrichr, ANNOVAR and EnhancerAtlas to determine whether the identified CNA regions with

genes are enriched in any interesting pathways or functional regions. Enrichr software [74] contains a diverse and up-to-date collection of over 100 gene set libraries available for analysis and download. It is used to perform pathway enrichment analysis on the identified young-specific genes to identify which pathways are over represented in the geneset. ANNOVAR [75] is a perl command line program for genome annotations. This region-based annotation is used to identify affected genomic regions that lie outside protein-coding regions. EnhancerAtlas [76] provides annotation of enhancers (distal regulatory elements) in the human genome, which are based on at least three independent high throughput experimental datasets (e.g. histone modification, eRNA, transcription factor binding and DNase I hypersensitive sites (DHS)) with the relative weights derived from a cross-validation approach. It is used to detect any non-coding regions that may overlap with nearby enhancers that are associated with genes.

3.8 Biological Visualization

In order to have a clearer interpretation of the results, softwares Oncoprint [77] and CIMminer [78] have been used to generate heatmap visualizations for the identified candidate regions. Oncoprint is included in the R package ComplexHeatMap, and it is a way to visualize multiple genomic alteration events in the format of a heatmap. This is used to visualize the frequencies of CNA mutation for each of the young-specific regions with genes in Discovery and Validation dataset. CIMminer generates color-coded Clustered Image Maps (CIMs) to portray "high-dimensional" data sets such as gene expression profiles. It is used to visualize the relative expression levels in terms of colour intensity for each of the identified young-specific genes.

Chapter 4: Results and Discussions

4.1 Clinical Characteristics

Based on the Fisher's exact test which compares the similarity of categorical variables, Discovery and Validation Data sets have very similar distribution in age, menopausal status, tumour grade, tumour size, ER, PR expression and HER2 expressions ($p>0.05$) (**Table 2)**.

On the other hand, the two sets have statistically significant differences in the tumour stage, with young patients in the Discovery set having a much higher prevalence in stage 0 compared to the Validation set (43.1% vs 0.8%) and PAM50 subtypes ($p<0.05$). However, the overall pattern of the basal subtype being the most abundant amongst the young patients is apparent in both the Discovery and Validation datasets.

Table 2. Clinical characteristics table comparing the METABRIC Discovery and Validation datasets for young patients only

Characteristic	Discovery Young	Validation Young	P-value †
Age*	40 (36,43)	40 (37,43)	1
Menopausal Status			0.5
Pre	127 (97.7%)	125 (100%)	
Post	2 (1.5%)	0 (0.0%)	
Subtype			<0.001
Normal	11 (8.5%)	25 (20%)	
LumA	41 (31.5%)	18 (14.4%)	
LumB	20 (15.4%)	9 (7.2%)	
Her2	16 (12.3%)	21 (16.8%)	
Basal	42 (32.3%)	52 (41.6%)	
Grade			0.98
1	7 (5.4%)	6 (4.8%)	
2	37 (28.5%)	34 (27.2%)	
3	86 (66.2%)	81 (64.8%)	
Stage			<0.001
0	56 (43.1%)	1 (0.8%)	
1	25 (19.2%)	28 (22.4%)	
2	42 (32.3%)	35 (28.0%)	
3	7 (5.4%)	11 (8.8%)	
4	0 (0.0%)	0 (0.0%)	
ER-expr			0.09
+	74 (56.9%)	57 (45.6%)	
-	56 (43.1%)	68 (54.4%)	
PR-expr			0.78
+	55 (42.3%)	56 (44.8%)	
-	75 (57.7%)	69 (55.2%)	
Her2-expr			0.27
+	22 (16.9%)	29 (23.2%)	
-	108 (83.1%)	96 (76.8%)	
Tumour Size* (mm)	22 (16,30)	23 (17,30)	1

* For continuous variables (Age, Tumour size), quantiles (50^{th} percentile (25^{th} percentile, 75^{th} percentile)) were presented.
† P-values were determined by Wilcoxon rank sum test for continuous variables and Fisher's exact test for categorical variables

4.2 Identification of Recurrent CNA Regions

Upon filtering for CNA regions of at least 1kb in size and having at least 5 patients per recurrent region, a total of 1,086 recurrent CNA gain regions (554/1,086 gain regions encompassing protein encoding genes) and 439 recurrent CNA loss regions (202/439 loss regions encompassing protein encoding genes) were identified.

Validation testing was then performed using the Validation set, which contains 995 patients. All filtering criteria and algorithm implementations followed the same procedure as the Discovery dataset analysis. For recurrent CNA gain regions, a total of 81 regions have been validated (found in both the Discovery and Validation datasets), in which 30 regions have encompassed 29 unique protein encoding genes (**Table 3.1**). For recurrent CNA loss regions, a total of 25 regions have been validated, in which 8 regions have encompassed 10 unique protein encoding genes (**Table 3.2**).

Table 3.1. Validated recurrent gain CNA regions with genes. The first four columns represent the chromosome #, start and end coordinates of the recurrent CNA region, and the size of the region in base pairs. The last three columns are the genes encompassed in each CNA region, followed by the sample size in both Discovery and Validation datasets.Inner start and end refers to the commonly shared region (100% overlap) among the members in each recurrent cluster, while outer refers to the region with the outermost start and end position within each recurrent cluster.

Chromosome	InnerStart	InnerEnd	Inner CNV Size	OuterStart	OuterEnd	Outer CNV Size	Genes	Discovery Cluster Size	Validation Cluster Size
1	222004315	222004925	610	221990859	222005526	14667	CAPN2	48	47
1	143607802	143609034	1232	143607067	143609055	1988	PDE4DIP	20	30
1	84551640	84565561	13921	84481190	84729446	248256	SAMD13	5	6
1	191374290	191385577	11287	191359529	191405183	45654	CDC73	40	50
1	191385797	191402004	16207	191359529	191405183	45654	CDC73	40	50
3	176423680	176427601	3921	176415397	176428705	13308	NAALADL2	18	22
3	176428607	176428705	98	176415397	176470832	55435	NAALADL2	18	22
5	22246497	22346803	100306	22194503	22414425	219922	CDH12	12	11
6	34625387	34634997	9610	34624907	34656516	31609	SPDEF	9	9
7	134782461	134787038	4577	134782461	134792291	9830	CNOT4	11	13
7	142150844	142154230	3386	142150819	142154515	3696	PRSS1	7	10
8	40695071	40697114	2043	40693570	40699795	6225	ZMAT4	17	16
9	93166194	93261927	95733	93157333	93373420	216087	NFIL3	5	7
10	14598341	14600566	2225	14477106	14629604	152498	FAM107B	19	27
10	5737990	5742226	4236	5736767	5744158	7391	ASB13	24	32
11	4931741	4932834	1093	4931741	4932966	1225	MMP26;OR51A2	7	11
12	7895693	7897774	2081	7895693	7897774	2081	SLC2A14	13	23
12	7899067	7905082	6015	7899067	7909593	10526	SLC2A14	13	23
12	180797	191614	10817	180797	195197	14400	SLC6A12	14	24
13	112354883	112363586	8703	112333434	112363586	30152	C13orf35	10	7
13	113356126	113365589	9463	113345036	113371998	26962	ATP4B	10	9
17	45136676	45139395	2719	45134175	45142242	8067	SLC35B1	21	18
17	43751830	43753351	1521	43722185	43756717	34532	SKAP1	9	14
18	43704001	43707399	3398	43703934	43707399	3465	SMAD2	6	6
18	9549925	9575313	25388	9417006	9594232	177226	PPP4R1	5	7
19	40629439	40647918	18479	40620640	40702499	81859	FFAR2	13	13
19	60890917	60901410	10493	60890917	60904859	13942	EPN1	8	11
20	14741416	14743670	2254	14741416	14743754	2338	MACROD2	8	10
20	41215727	41219453	3726	41202818	41220578	17760	PTPRT	12	18
22	20146692	20170596	23904	20145867	20170766	24899	PI4KAP2;TMEM191C	7	15

Table 3.2. Validated recurrent loss CNA regions with genes. The first four columns represent the chromosome #, start and end coordinates of the recurrent CNA region, and the size of the region in base pairs. The last three columns are the genes encompassed in each CNA region, followed by the sample size in both Discovery and Validation datasets.

Chromosome	InnerStart	InnerEnd	Inner CNV Size	OuterStart	OuterEnd	Outer CNV Size	Genes	Discovery Cluster Size	Validation Cluster Size
2	97507180	97517476	10296	97507180	97520698	13518	ANKRD36B	5	5
3	62243538	62257523	13985	62242606	62277516	34910	PTPRG	6	6
4	59521	61566	2045	59521	64435	4914	ZNF718;ZNF595	13	6
7	38296343	38297866	1523	38295506	38297939	2433	TRGV11	18	22
8	14388851	14391732	2881	14385622	14391732	6110	SGCZ	23	18
9	5027454	5029342	1888	5027454	5030334	2880	JAK2	7	11
10	89710114	89713882	3768	89708179	89713882	5703	PTENP1;PTEN	10	10
17	21245986	21253816	7830	21227031	21271210	44179	KCNJ12	15	9

In total, 38 validated recurrent CNA regions with 39 protein encoding genes have been identified, along with 68 validated recurrent CNA regions that did not encompass any protein encoding genes (**Table 3.3, Table 3.4**).

Table 3.3. Validated recurrent gain CNA regions without genes. The first four columns represent the chromosome #, start and end coordinates of the recurrent CNA region, and the size of the region in base pairs. The last two columns show the sample size for each of the CNA regions in both Discovery and Validation datasets.

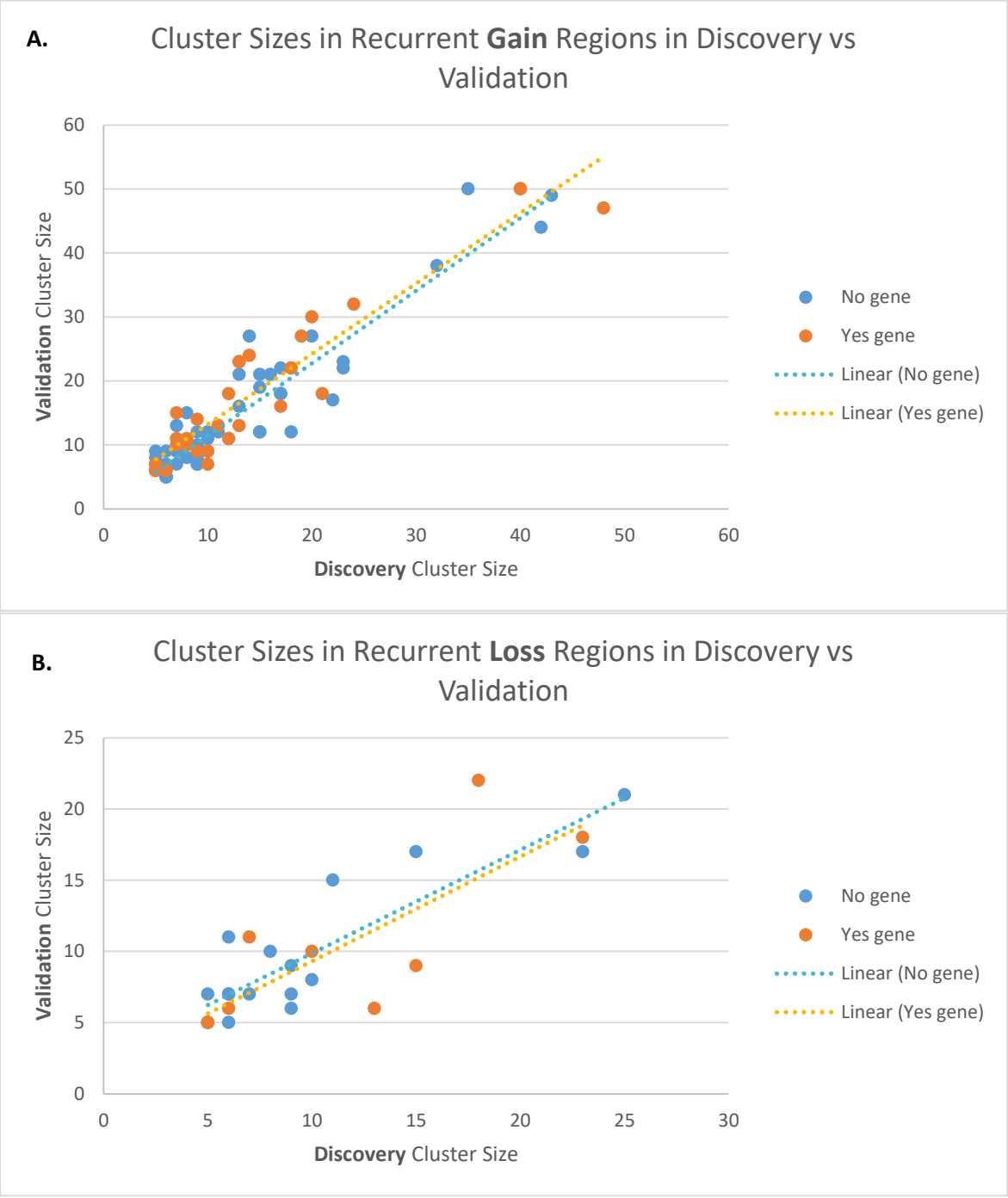
Chromosome	InnerStart	InnerEnd	Inner CNV Size	OuterStart	OuterEnd	Outer CNV Size	Discovery Cluster Size	Validation Cluster Size
1	58820476	58846018	25542	58803399	58846018	42619	6	7
1	147475947	147496433	20486	147475521	147496468	20947	35	50
2	4197415	4200525	3110	4195432	4201388	5956	11	13
2	29028479	29029800	1321	29010022	29040695	30673	5	8
2	52618925	52634800	15875	52618897	52634917	16020	17	22
2	64271883	64282835	10952	64247921	64315987	68066	9	7
2	64283559	64306614	23055	64271883	64315987	44104	9	7
2	65560324	65677884	117560	65463975	65678403	214428	5	6
3	9660712	9664562	3850	9640519	9688008	47489	10	12
3	109226991	109234844	7853	109226991	109235111	8120	7	7
3	131289615	131290967	1352	131289615	131290967	1352	18	12
3	185642866	185653686	10820	185633993	185656645	22652	11	12
3	199319836	199322976	3140	199318744	199332531	13787	12	11
3	196914905	196916045	1140	196907470	196918723	11253	23	22
3	196916080	196918723	2643	196914905	196940632	25727	23	23
5	18541955	18547895	5940	18514272	18554326	40054	15	12
6	32539531	32556611	17080	32539531	32557028	17497	6	5
6	32557501	32560908	3407	32557028	32562254	5226	6	5
6	79037167	79044205	7038	79037167	79044205	7038	16	21
7	54152334	54153709	1375	54152334	54153757	1423	6	5
7	61794167	61799563	5396	61794167	61818546	24379	10	11
7	76141447	76146160	4713	76141447	76146160	4713	8	8
7	109229197	109241095	11898	109228866	109241095	12229	9	12
7	143108242	143108276	34	143104624	143118598	13974	14	27
8	1349638	1350270	632	1348245	1350652	2407	5	6
8	3989394	3992254	2860	3985016	3992622	7606	5	7
8	8027374	8027446	72	8015467	8046615	31148	7	9
8	86730480	86740395	9915	86720993	86740395	19402	32	38
8	96278448	96284249	5801	96258231	96312735	54504	42	44
8	135232619	135235139	2520	135232619	135235713	3094	43	49
9	67699737	67709343	9606	67678161	67743272	65111	6	9
9	135177441	135178844	1403	135173914	135178844	4930	6	5
10	6703994	6745967	41973	6703994	6797043	93049	20	27
10	19676387	19685158	8771	19656598	19750338	93740	15	19
10	20731174	20741934	10760	20726541	20747050	20509	17	18
10	20747419	20761390	13971	20747068	20779396	32328	17	18
10	22772784	22815216	42432	22764104	22849622	85518	15	21
10	66980962	66984437	3475	66978024	66984437	6413	7	13
10	81439210	81449106	9896	81439210	81452222	13012	6	6
12	8253318	8257868	4550	8253318	8261982	8664	13	21
12	11634190	11638369	4179	11606108	11670489	64381	8	15
14	23564490	23564891	401	23559232	23570486	11254	7	9
14	25054809	25064593	9784	24846340	25069117	222777	6	5
14	25065074	25069117	4043	24846340	25075027	228687	6	5
15	84140524	84146976	6452	84140524	84157332	16808	5	9
16	14961825	14966336	4511	14961806	14966847	5041	15	12
17	31460821	31463472	2651	31460821	31479995	19174	13	16
17	69474254	69478126	3872	69429096	69496533	67437	22	17
18	55813854	55816005	2151	55813796	55819506	5710	9	10
21	43341103	43343023	1920	43329959	43356005	26046	9	8
21	46828863	46834725	5862	46828497	46846083	17586	10	9

Table 3.4. Validated recurrent loss CNA regions without genes. The first four columns represent the chromosome #, start and end coordinates of the recurrent CNA region, and the size of the region in base pairs. The last two columns show the sample size for each of the CNA regions in both Discovery and Validation datasets.

Chromosome	InnerStart	InnerEnd	Inner CNV Size	OuterStart	OuterEnd	Outer CNV Size	Discovery Cluster Size	Validation Cluster Size
1	16885044	16885329	285	16884038	16889666	5628	11	15
3	26589981	26589990	9	26588205	26591799	3594	7	7
3	64855417	64863915	8498	64731448	64923169	191721	6	5
4	104455633	104467384	11751	104414905	104467384	52479	5	5
4	157100338	157107691	7353	157100332	157107969	7637	5	7
5	60888563	61017766	129203	60887533	61460364	572831	8	10
8	3552110	3554053	1943	3551271	3564930	13659	25	21
8	5633227	5636521	3294	5625845	5636589	10744	23	17
9	1490330	1490755	425	1456070	1504392	48322	6	7
9	1500170	1503092	2922	1490330	1504392	14062	6	7
13	22518343	22523101	4758	22512167	22523373	11206	10	8
13	56656317	56661149	4832	56656291	56661535	5244	15	17
18	36514797	36519362	4565	36514489	36519388	4899	9	6
18	61883719	61883744	25	61878473	61885427	6954	5	5
22	20816553	20820788	4235	20816118	20833270	17152	9	9
15	22859182	22873331	14149	22846930	22875304	28374	6	11
17	33672360	33694008	21648	33664567	33752586	88019	9	7

Figure 3 shows an overview of how similar the cluster sizes (i.e. number of patients) are in the discovery set versus the validation set for all the identified young-specific recurrent CNA regions. It can be seen that for both gain (**Figure 3A**) and loss regions (**Figure 3B**), cluster sizes in the Discovery and Validation dataset have a fairly linear relationship. For example, if 30% of the young patients in the discovery set harbour a CNA region, it is likely that around 30% of patients in the validation set will harbour that region as well.

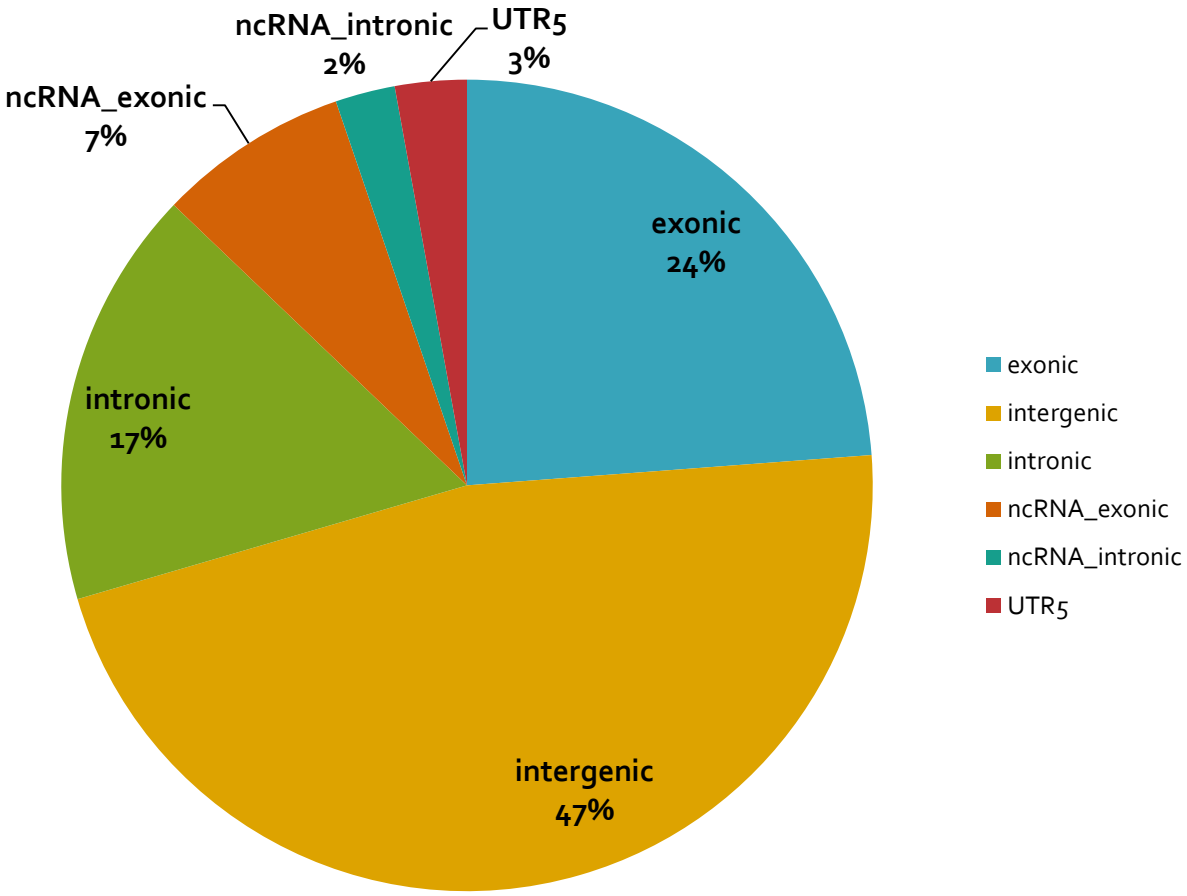
Figure 3. Scatter plot showing the cluster sizes identified for the A) gain recurrent young-specific regions and B) loss recurrent young-specific regions. Each point on the plot represent a young-specific recurrent CNA region. Blue represents regions without genes, and orange represents regions with genes.



4.3 Functional Annotation of the identified CNA regions

In order to better understand how these recurrent CNAs can influence the regulation of gene expression, we performed region-based functional annotation on these CNA regions using the software ANNOVAR (See Appendix). **Figure 4** shows the distribution of our recurrent CNAs with respect to the encompassed regions. Majority of the CNAs are in the non-coding region (59%), 24% in the exonic regions and 17% in the intronic regions.

Figure 4. Distribution of the identified young-specific recurrent CNA regions with respect to the genome structure.



4.3.1 Non coding regions

Genome-wide association studies (GWAS) have demonstrated that majority of the CNAs associated with cancer predisposition are preferentially located outside the coding regions and are enriched in putative regulatory elements. Recently, various SNPs, known as eQTL SNPs, have been identified in the non-coding regions of the genome to influence the expression of genes [79]. Therefore, it is of importance to determine if there exist eQTL CNAs for the non-coding regions we have identified. Since CNAs are fairly large in size, it would be interesting to see whether any of the non-coding CNA regions are bound by regulatory elements, such as promoters or enhancers, which have the ability to regulate gene expression levels in either a cis or trans way.

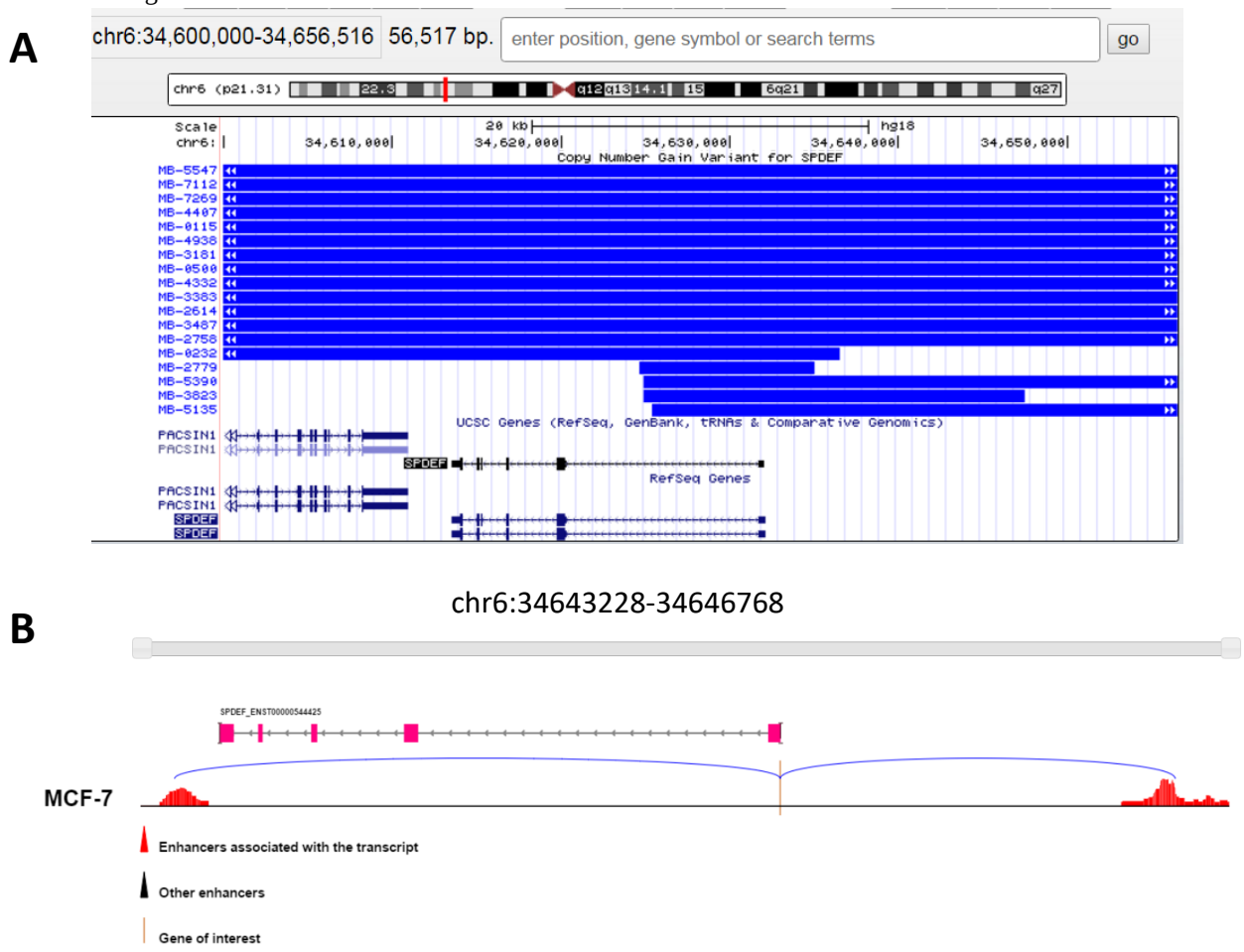
Promoters act as regulatory regions of DNA transcription initiation and the recruitment of RNA polymerases. Ranging from 100 to 1000 bp in size, promoters are located close to the transcription start site of the gene towards the 5' end of the same strand. Mutations in promoters may dysregulate gene transcription by altering the chromatin state, DNA looping or even result in gene silencing in the case of DNA methylation. In many cancers, it has been shown that a large number of carcinogenic genes (e.g. DNA repair genes) have a loss in gene expression, typically when multiple CpG islands are present in the promoters of the genes [80].

Enhancers are distal regulatory elements that are often located >10kb upstream or downstream from the transcription start site. However, enhancer activity is independent of location and orientation. In other words, an enhancer is not required to be physically near the transcription start site to have an effect on transcription [81]. Structural variants such as CNAs may disrupt enhancer action by altering chromatin conformation and shifting the enhancers further away from the target, hindering its interactions with other activator proteins. The role of

enhancers in cancer studies was first revealed in the MYC oncogene which is often translocated near the IGH's enhancer region in lymphoma cells, resulting in a significant overexpression of the MYC protein [82]. It was later suggested that large deletions in the regulatory elements region lead to abnormal expression of oncogenes (e.g. super-enhancers) as well as a loss in tumour suppressors.

Of those CNA regions found outside of the exonic and intronic regions (see Appendix), the recurrent gain CNA region (chr6:34624907-34656516, **Figure 5A**) was detected to be near the gene SPDEF (Sam-pointed domain containing Ets transcription factor) (chr6:34613557-34632088). An enhancer (chr6:34643228-34646768) found in breast cell line MCF-7 is located in the CNA region from the enhancer atlas database (http://www.enhanceratlas.org/) (**Figure 5B**), seemingly a downstream regulator.

Figure 5. Non-coding region overlapped in enhancer region. A) The recurrent CNA region (chr6:34624907-34656516) was found to close gene SPDEF (chr6:34613557-34632088). **B)** An enhancer found in breast cell line MCF-7 is located in the CNA region from the enhancer atlas database (http://www.enhanceratlas.org/) which would be downstream regulator.



SPDEF is a member of the ETS family of transcription factors which control a variety of developmental processes such as cell lineage specification, differentiation, proliferation and apoptosis [83]. The SPDEF gene has been shown to be frequently overexpressed in human breast cancer and induces accelerated tumor growth and survival in ER-positive tumours [83]. Increase in the SPDEF protein expression has been frequently observed in the progression from benign breast tissue to carcinoma, and has also been correlated with poor overall survival of patients with estrogen positive receptors (ER+) [84]. It has been shown to be a downstream target of the

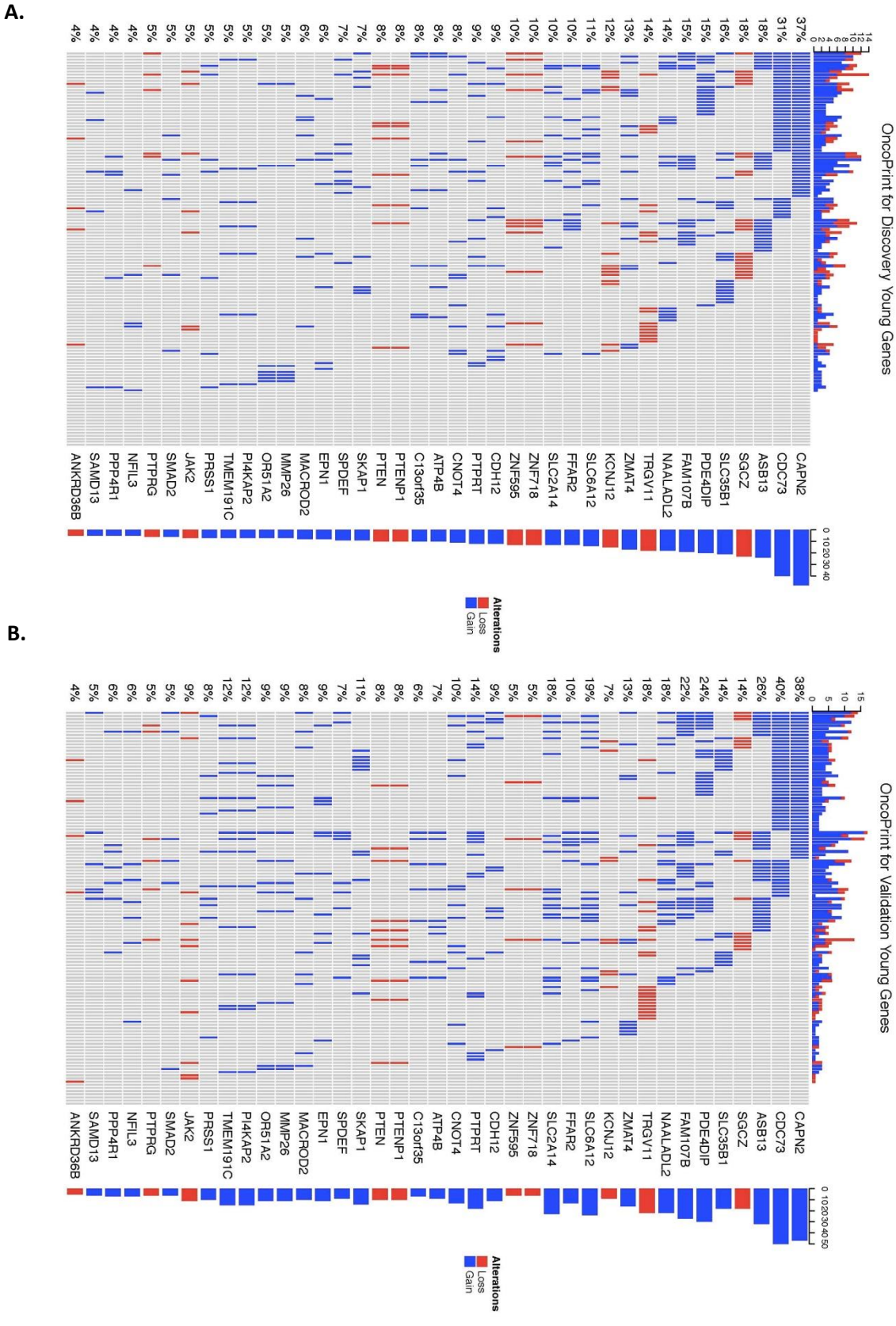
ER/FOXA1/GATA3 pathway and involved in the tumour suppressive action of GATA3, which induces invasive tumour behaviour and dissemination [83].

4.4 Coding regions

4.4.1 Mutation status heatmap

In order to better comprehend the multiple recurrent CNA mutation events identified in both the Discovery and Validation young women group, an R package called the ComplexHeatmap was used to generate a mutation distribution heatmap (**Figure 6**). From the heatmap, it can be observed that CNA gain regions encompassing genes *CAPN2, CDC73* and *ASB13* are the top 3 with the highest occurring frequencies in both Discovery and Validation dataset (young women age group), while gene *SGCZ* ranked top for recurrent CNA loss regions in the two datasets.

Figure 6. Mutation distribution heatmap for genes identified in the recurrent young-specific CNA gain and loss regions in A) Discovery dataset and B) Validation dataset. Rows are sorted based on the frequency of the alterations in all young-specific samples and columns are sorted to visualize the mutual exclusivity across genes. Barplots at both sides of the heatmap show numbers of different alterations for each sample and for each gene. Red represents CNA loss mutations and blue represents CNA gain mutations.



4.4.2 eQTL analysis

The gene expression heatmap (**Figure 7**) provides an overview of the expression levels for each of the identified young-specific genes across all the young patients samples in the Discovery (**Figure 7A**) and Validation dataset (**Figure 7B**). Further interrogation using logistic regression was performed to evaluate the statistical association between gene expression and CNA mutation status (**Table 4**). In total, 16 gain regions and 1 loss region show significant associations with gene expression changes. However, the directionality of the association is ambiguous. 14 out of the 16 gain regions correspond to having high gene expressions while the other 2 gain regions (encompassing MMP26 and SPDEF) were associated with low gene expressions. For example, mutated gain CNA status in *ASB13* seems to lead to higher gene expression. On the other hand, the loss region encompassing PTEN was found to be associated with having high gene expression level.

Figure 7. Gene expression heatmap for young breast cancer patients in A) Discovery dataset B) Validation dataset. Rows represent the gene expression levels for the genes identified in the recurrent young-specific CNA gain and loss regions (same order as in Figure 6 for comparison). Columns represent the young-specific samples in Discovery and Validation datasets. The higher the intensity of the red colour, the higher the gene expression level.

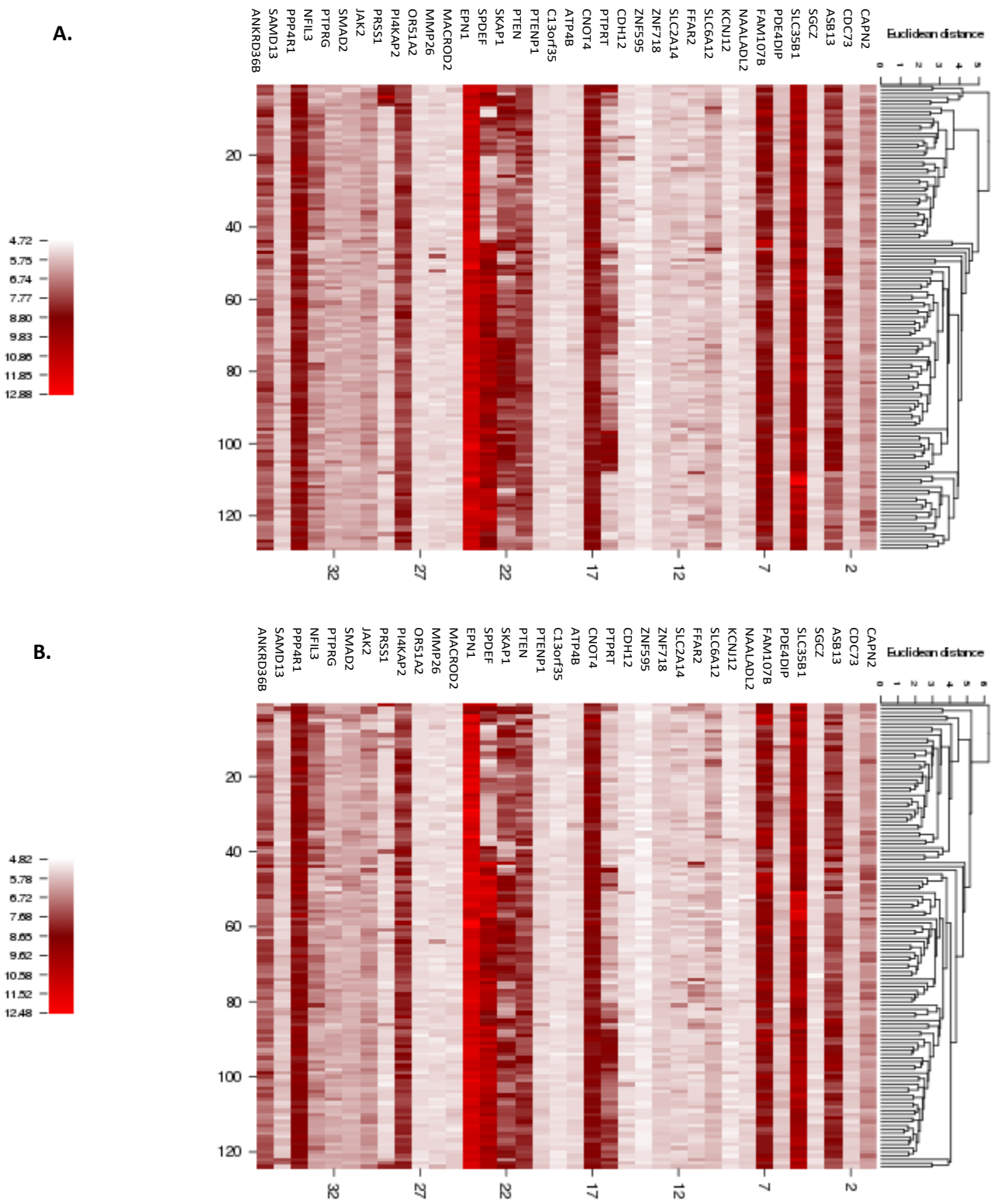


Table 4. Logistic regression analysis between CNA mutation status and gene expression (binarized by mean) in combined dataset.

	Copy Number State	P-value	Odds Ratio†	Discovery Sample Size	Validation Sample Size	Chromosome	InnerStart	InnerEnd
ASB13	Gain	0.049	1.83 (1.00–3.34)	24	32	10	5737990	5742226
ATP4B	Gain	0.021	3.51 (1.21–10.17)	10	9	13	113356126	113365589
CAPN2	Gain	0.00002	3.21 (1.89–5.45)	48	47	1	222004315	222004925
CDH12	Gain	0.026	2.87 (1.14–7.24)	12	11	5	22246497	22346803
CNOT4	Gain	0.0004	7.45 (2.46–22.49)	11	13	7	134782461	134787038
EPN1	Gain	0.004	9.57 (2.15–42.56)	8	11	19	60890917	60901410
FAM107B	Gain	0.083	1.79 (0.93–3.46)	19	27	10	14598341	14600566
MACROD2	Gain	0.085	2.73 (0.87–8.55)	8	10	20	14741416	14743670
MMP26	Gain	0.098	0.43 (0.15–1.17)	7	11	11	4931741	4932834
NFIL3	Gain	0.068	3.46 (0.91–13.08)	5	7	9	93166194	93261927
PDE4DIP	Gain	0.002	2.73 (1.45–5.15)	20	30	1	143607802	143609034
PI4KAP2	Gain	0.031	2.81 (1.10–7.14)	7	15	22	20146692	20170596
PPP4R1	Gain	0.016	6.7 (1.44–31.23)	5	7	18	9549925	9575313
SLC35B1	Gain	0.001	19.19 (6.54–56.27)	21	18	17	45136676	45139395
SMAD2	Gain	0.022	6.06 (1.30–28.22)	6	6	18	43704001	43707399
SPDEF	Gain	0.057	0.39 (0.15–1.03)	9	9	6	34625387	34634997
PTEN	Loss	0.002	29.12 (3.83–221.25)	10	10	10	89710114	89713882

† Odds ratio is followed by its corresponding 95% CI in brackets

4.4.3 Survival Analysis

We further evaluate whether the expression levels of these genes are associated with survival outcome. (**Table 5**). The expression levels of 8 out of the 39 young-specific genes are significantly associated with survival outcome. A higher gene expression (a unit increase) in genes *CAPN2*, *NFIL3* and *SLC35B1* corresponds to a moderately worse survival outcome.

Table 5. Cox Proportional Hazard survival analysis of gene expression in combined dataset (discovery and validation).

	Copy Number State	P-value	Hazard ratio†	Discovery Sample Size	Validation Sample Size	Chromosome	InnerStart	InnerEnd
ASB13	Gain	0.0001	0.54 (0.39-0.73)	24	32	10	5737990	5742226
CAPN2	Gain	0.091	1.54 (0.93-2.54)	48	47	1	222004315	222004925
NFIL3	Gain	0.009	1.58 (1.13-2.23)	5	7	9	93166194	93261927
PDE4DIP	Gain	0.027	0.37 (0.16-0.89)	20	30	1	143607802	143609034
PTPRT	Gain	0.0002	0.68 (0.55-0.83)	12	18	20	41215727	41219453
SKAP1	Gain	0.0002	0.66 (0.53-0.82)	9	14	17	43751830	43753351
SLC35B1	Gain	0.00005	1.88 (1.39-2.55)	21	18	17	45136676	45139395
JAK2	Loss	0.007	0.43 (0.24-0.79)	7	11	9	5027454	5029342

†Hazard ratio is followed by its corresponding 95% CI in brackets

Of particular interest, the mutation status and expression levels of two of these genes, *ASB13* (**Figure 8A)** and *SGCZ* (**Figure 9D),** were also significantly associated with patients' survival outcomes in the Kaplan Meier survival analysis, which allows estimation of a survival curve over time. Patients with a mutated status in both of these genes resulted in a worse survival outcome when compared to patients without the gene mutations. The mutation status of other genes found to be significant in the survival analysis include *ATP4B, FFAR2* and *PTPRT* in the encompassed CNA gain regions (**Figure 8)**; *PTENP1, PTEN, ZNF718* and *ZNF595* in the encompassed CNA loss regions (**Figure 9).**

Figure 8. Kaplan Meier survival analysis for genes with significant CNA gain mutations in the young women group. Genes showing with statistical significance ($p<0.05$) are A) *ASB13*, B) *ATP4B*, C) *FFAR2* and D) *PTPRT*. Survival curve in red represents patients without CNA mutation in the corresponding gene (CN=2) while the curve in blue represents patients with CNA gain mutations in the corresponding gene. Y-axis is the cumulative survival probability and X-axis is the survival time in years.

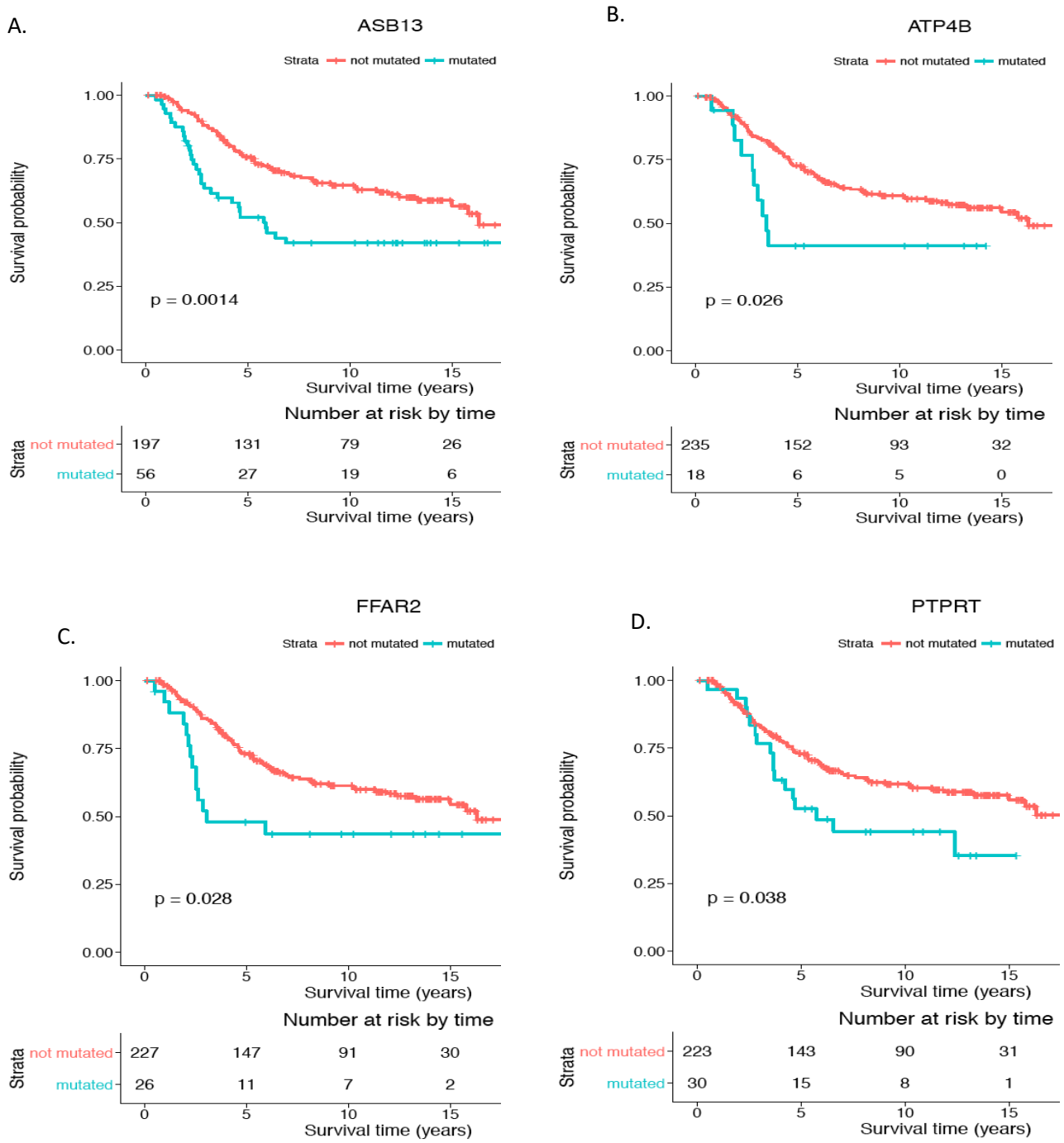
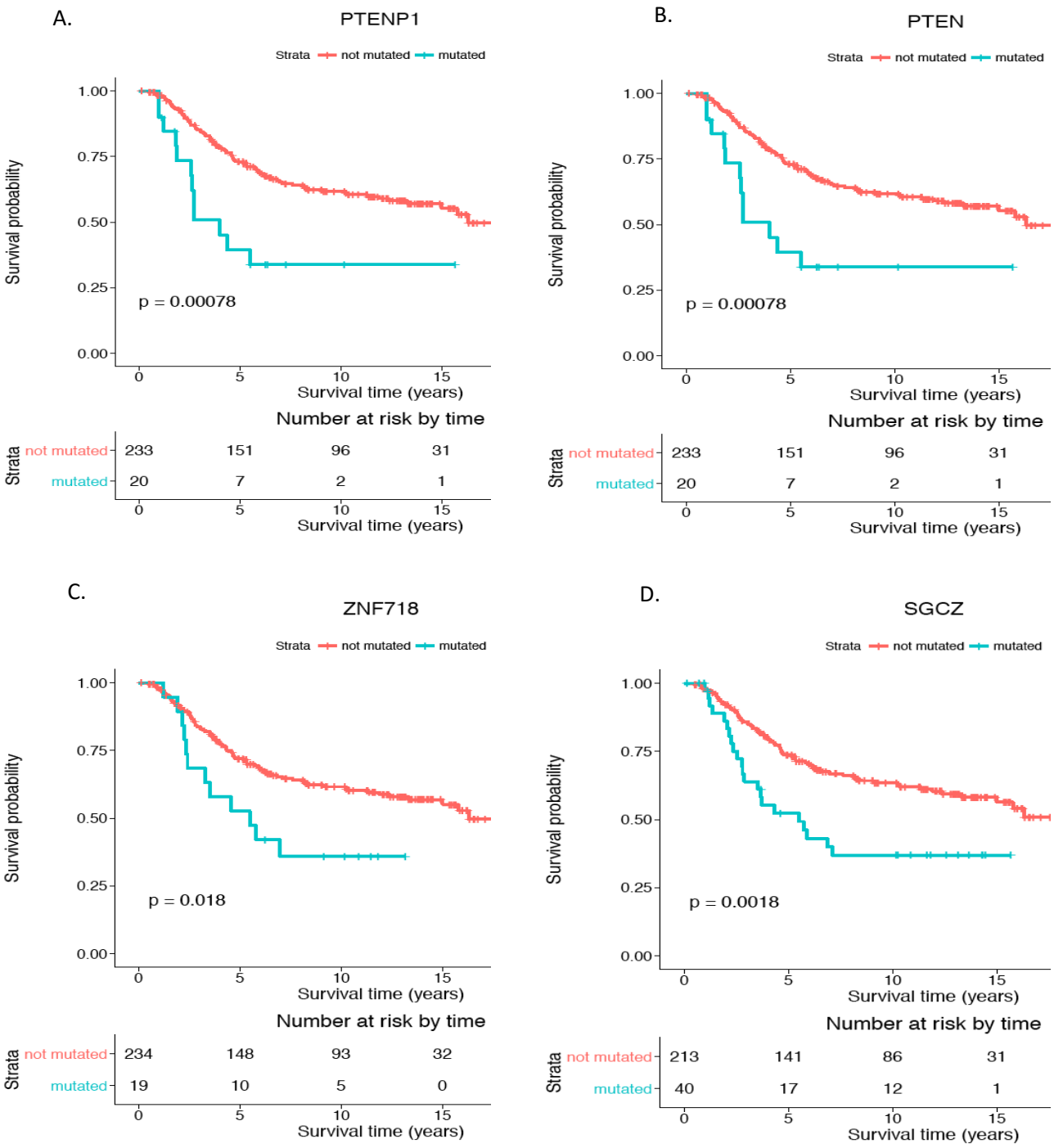
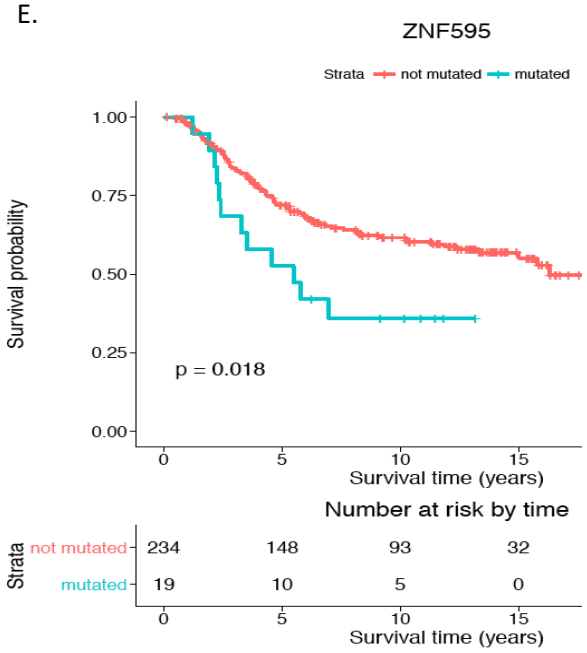


Figure 9. Kaplan Meier survival analysis for genes with significant CNA loss mutations in the young women group. Genes showing with statistical significance ($p<0.05$) are A) *PTENP1*, B) *PTEN*, C) *ZNF718*, D) *SGCZ* and E) *ZNF595*. Survival curve in red represents patients without CNA mutation in the corresponding gene (CN=2) while the curve in blue represents patients with CNA loss mutations in the corresponding gene. Y-axis is the cumulative survival probability and X-axis is the survival time in years.



E.



4.4.4 Cancer-relevant candidate genes

PTEN (Phosphatase and tensin homolog)

Results from our study show that the median survival time (i.e. half of the patients are expected to be alive) for young patients with a copy number loss in the *PTEN* gene region is ~4 years as opposed to ~15 years for those without. *PTEN* (cytoband 10q23.31) has been identified as a tumour suppressor which negatively regulates the PI3K/Akt/mTOR signalling pathway [85]. It has been shown to be one of the most frequently mutated genes in cancers, including that of breast, ovary, prostate, glioblastoma and lymphoma. Previous studies have observed that 40% of invasive breast cancers have a loss in *PTEN* heterozygosity, and that the loss of one gene copy is sufficient to disrupt cell signalling and induce cell growth. It has also been suggested that carriers of the *PTEN* mutation are at higher risk of developing early onset breast cancers at a relatively young age [86]. However, it should be noted that the significance of *PTEN* mutations with

regards to cancer have been controversial due to the identification of *PTEN* mutations in normal healthy tissues.

SGCZ (Sarcoglycan Zeta)

Our study shows that ~16% of all young patients present a CNA loss mutation encompassing *SGCZ*, with a median survival time for young patients with this mutation ~6 years as opposed to ~15 years for those without. *SGCZ* (8p22) encodes a protein that is part of the sarcoglycan complex, which plays a role in connecting the inner cytoskeleton to the extracellular matrix, possibly maintaining membrane stability [87]. Although the role of *SGCZ* has not been well established in terms of cancer implications, the loss of the region chr8p has been associated with several factors in cancer development, such as having an aggressive histology, increased cell proliferation, increased mortality and recurrence rate in early breast cancer, large tumour size and poor survival in young women. This region also contain the gene *DLC1* (deleted in liver cancer 1), which has been suggested to act as a tumour suppressor in inhibiting cell growth [88]. DLC1 encodes a GAP protein that inhibits the activation of the Rho-GTPases, which are often associated with a loss in cell adhesion. *DLC1* expression has been reported to be frequently lost in tumour cells, leading to a constitutive activation of the Rho-GTPases.

CAPN2 (Calpain 2)

CAPN2 (cytoband 1q41) was the most frequent CNA gain mutation in our study, with ~37-38% of all young patients harbouring a CAPN2 gain mutation. Calpains are calcium-activated intracellular proteases that have the ability to cleave cytoskeletal proteins, possibly playing a role in regulating cell invasion and migration [89]. A knockdown study of CAPN2 in

breast tumour cells has shown correlations with reduced cell migration, proliferation, as well as reduced Akt activation, increased FoxO localization and p27 expression [89]. It was suggested that CAPN2 promotes cell proliferation through the Akt-FoxO-p27 signalling pathway.

***NAALADL2* (N-acetyl-L-aspartyl-L-glutamate peptidase-like 2)**

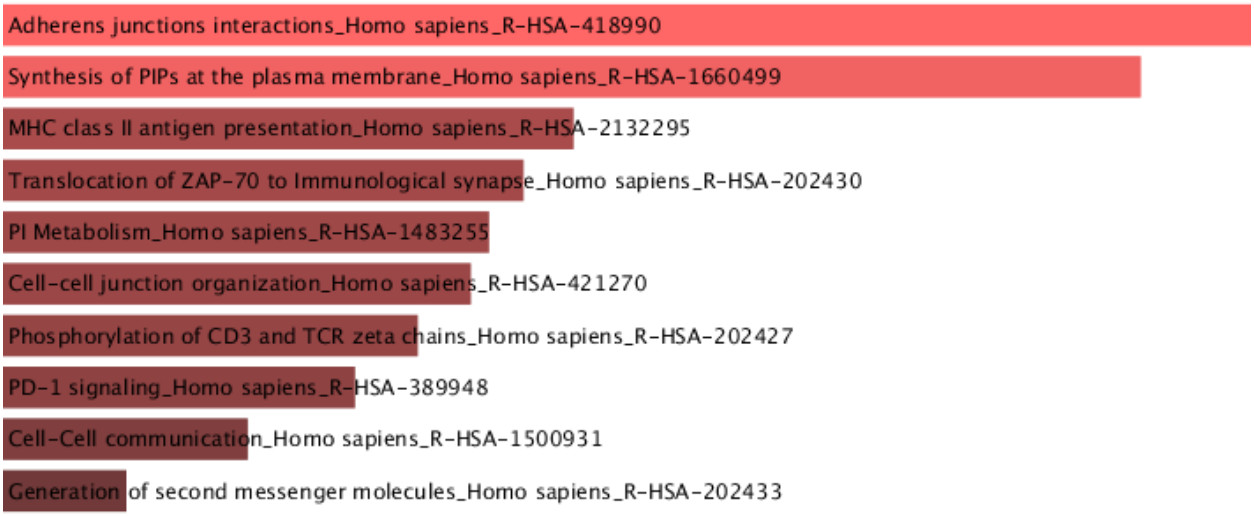
Our study shows that ~16% of all young patients present a CNA gain mutation encompassing NAALADL2. NAALADL2 is a member of the NAALADase protein family which act as matrix metalloproteases and have the ability to alter the tumour environment. Microarray studies have shown that NAALADL2 is often overexpressed in prostate and colon cancers and stimulates a migratory and metastatic phenotype. A proposed mechanism is that since NAALADL2 has been found to be basal-localized, it may enhance the interaction of tumour cells via the extracellular matrix of the tumour gland, and provide a mechanism for the cells to escape [90]. Subsequent survival analysis shows that patients with NAALADL2 overexpression have a 45% chance of surviving up to 5 years as opposed to 93% for patients with low NAALADL2 expression. It remained to be prognostic for recurrence rate even after correction for clinical variables such as tumour stage and grade. Expression arrays also elucidated changes to the epithelial-to-mesenchyme transition (EMT) and cell adhesion pathways.

4.5 Enrichment Analysis

A pathway enrichment analysis was performed using the ANNOVAR gene list (see **Appendix**)**.** The ANNOVAR genelist was selected due to its comprehensive coverage of genes: not only does it contain the genes that were encompassed in the recurrent CNA regions (Tables 3.1,3.2), but it also includes nearby genes found in the regions which did not encompass any

genes. The Enrichr REACTOME database reveals a significant overrepresentation in the regulation of phospholipid signalling (*MTMR14*,*PTEN*,*PIP4K2A*) and adherens junction (*CDH12*, *CDH18*, *CDH7*) pathways ($p < 0.05$) in the identified young-specific recurrent CNA regions with genes **(Figure 10)**. It is evident that both enriched pathways are highly relevant with regards to cancer development.

Figure 10. Enrichment analysis of all the identified young-specific recurrent gain and loss regions with genes. Graph bars are sorted by p-value ranking. The length of the bar represents the significance of that specific pathway. Light red coloured bars have a p-value < 0.05.



4.5.1 Phospholipid Signalling

Aside from playing an important role in structural components, lipids also have a role in signalling processes [91, 92]. These lipid molecules aggregate to form lipid rafts as highly specific platforms for cell signalling, carrying signals from activated growth factor receptors to the cellular machinery [93]. These receptors may recruit signalling effectors that induce cell proliferation and failure in cell death, all contributing toward cancer development. The phosphatidylinositol (3,4,5)-trisphosphate molecule, also known as PIP3, is generated by the

PI3K and take on the task of activating downstream signaling components. One of the most well-known consequence is recruitment and activation of protein kinase Akt, which may phosphorylate a variety of substrates, thereby disrupting the regulation in cell growth, apoptosis and cell cycle processes. PIP3 acts as a substrate for phosphatase and tensin homologue (*PTEN*), which is required for the dephosphorylation of PIP3 into PIP2, essential for inhibition of the AKT pathway.

4.5.2 Cell adhesion

Cellular adhesion plays a major role in maintain the integrity of normal cell-cell connections, and disruption in this pathway has been strongly associated with metastasis in cancers. Adherens junctions, which are the sites of intracellular signalling and anchoring, provide a strong bond between the adjacent cell membranes. The molecular processes governing the cell-cell adhesion is very finely controlled, such as the epithelial-mesenchymal transition (EMT) mechanism that is normally present during embryogenesis and tissue repair. Characteristics of EMT include a loss in intercellular adhesion and enhancement of cell migration, leading to a more motile phenotype [94]. Notably, the adherens junctions are lost during the process of EMT, which in the case of dysregulation, bear the potential to become pathological and lead to oncogenic development such as metastasis. In normal tissues, epithelial cells are tightly bound to one another. However, it has been reported that in the state of advanced cancer, many epithelial tumour cells are present with a loss in cell-cell adhesion and increase in tissue invasion. Tumours featuring local spreading and invasion are suggested to result in a more aggressive phenotype and higher mortality rate of the patient. This phenomenon has been widely implicated in various cancer types, including breast, colon, prostate, ovarian and other types of cancer [95].

4.6 Significance and Conclusion

A graph-based algorithm is proposed to call recurrent CNAs in breast cancer patients. The algorithm has an optimal solution, which means all maximal cliques can be identified. Additionally, it guarantees that the identified CNA regions are the most frequent and that the minimal regions have been delineated. From application to the METABRIC breast cancer dataset, we have identified 81 validated recurrent CNA gain regions and 25 validated recurrent CNA loss regions, and have located the corresponding candidate genes that were encompassed in these regions.

It is becoming progressively clear that genetic studies of complex diseases must heed to the involvement of recurrent CNAs. Therefore, investigation into recurrent CNAs could provide significant contributions to the understanding of the basis of genetic variations in biological functions and disease predisposition. At present, CNA studies with regards to cancer are still in their infancy, but it is an area that is growing rapidly due to denser microarrays and next generation sequencing technologies. As we lean towards personalized structural genomic analysis and diagnostics, the conventional genomic definition of what is “normal” vs “diseased” will start to blur. There is much to take in from previous studies on genomic disorders and by incorporating the knowledge of the vast amount of CNAs present in our genome. Identification of molecular alterations associated with disease outcome may improve risk assessment and treatments for aggressive breast cancer, especially for young women. It can give new insights into the role of CNAs in cancer predisposition and development, and contribute to a more accurate and complete human genome sequence reference. We hope that the algorithm can be adapted to other complex trait diseases such as IBD and other cancer types, and that the results of

this study will in the future facilitate the development of screening methods for breast cancer biomarker discovery as more prospective samples become available.

Chapter 5: Limitations and Future Directions

Limitations to our project include small sample sizes and limited validation. Out of the 2000 patient samples, there are only ~300 young patient samples. Validation was performed on a separate subset of samples within the same study. This may cause an issue with overfitting, meaning that the parameter estimates may work well for the current dataset but not as well in another dataset. Therefore, validation using an independent dataset can help avoid the issue with overfitting by ensuring that the parameter estimates are not unique to the dataset being used. By doing so, we may also increase our statistical power by increasing our young patient sample size. One solution is by pooling data. This may provide a better representation of the patients' clinical features, making subgroup analysis possible (e.g. analyze by PAM50 subtypes). A possible dataset we can use is from the TCGA database, which contains whole exome sequencing data including CNA and gene expression profiles of breast tumour samples. It should be noted that some of the samples may not be feasible to use due to major differences in important variables, such as treatment and population differences. However, if the data are rich enough and that most of the parameters are comparable, then it may be more robust to include than just using data from a single study.

Functional experimental validation is also critical in providing the proof of the biological significance and relevance of the identified genes with regards to phenotype. With the widespread genetic variation and the accumulation of somatic mutations, candidate genes identified via genomics approach may be found to be enriched due to chance; therefore, experimental validation is required to further assess which of the candidates have a true effect on the phenotype of interest, demonstrating how a defect of the gene may influence specific pathways such as cell proliferation, apoptosis or cell adhesion. A few options include performing

a gene knockdown or transfection in a human breast cell line to observe its phenotypic change at the cellular level, or using an animal model such as mice or zebrafish to study the gene effect in vivo.

In addition, further interrogation into the functional impact of the non-coding regions are needed to explore potential overlaps with regulatory elements such as promoter region or enhancer regions, which may play a role in regulating gene expression levels in a cis or trans way.

Bibliography

1. Anders CK, Johnson R, Litton J, Phillips M, Bleyer A: Breast cancer before age 40 years. Semin Oncol 2009; 36:237-249.
2. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406:747-52.
3. Bauer KR, Brown M, Cress RD, et al. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry. Cancer 2007;109:1721-172.
4. Azim HA Jr, Michiels S, Bedard PL, et al.: Elucidating prognosis and biology of breast cancer arising in young women using gene expression profiling. Clin Cancer Res 2012; 18:1341-1351.
5. Gnerlich JL, Deshpande AD, Jeffe DB, Sweet A, White N, Margenthaler JA: Elevated breast cancer mortality in women younger than age 40 years compared with older women is attributed to poorer survival in early-stage disease. J Am Coll Surg 2009; 208:341–347.
6. Johnson RH, Hu P, Fan C, Anders CK. Gene expression in “young adult type” breast cancer: a retrospective analysis. Oncotarget. 2015;6(15):13688-702.
7. Bharat A, Aft RL, Gao F, et al. Patient and tumor characteristics associated with increased mortality in young women (≤ 40 years) with breast cancer. J Surg Oncol 2009;100:248-51.
8. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*2009;**27**:1160–1167.
9. Gort M, Broekhuis M, Otter R, Klazinga NS. Improvement of best practice in early breast

cancer: actionable surgeon and hospital factors. Breast Cancer Res Treat 2007;102: 219–226.

10. Kim EK, Noh WC, Han W, et al. Prognostic significance of young age (< 35years) by subtype based on ER, PR, and HER2 status in breast cancer: a nationwide registry-based study. World J Surg 2011;35:1244-53.
11. Thomas GA, Leonard RC: How age affects the biology of breast cancer. Clin Oncol 2009; 21:81-85.3083
12. Benz CC: Impact of aging on the biology of breast cancer. Crit Rev Oncol Hematol 2008; 66:65-74.
13. Cancello G, Maisonneuve P, Rotmensz N, et al. Prognosis and adjuvant treatment effects in selected breast cancer subtypes of very young women (< 35years) with operable breast cancer. Ann Oncol 2010;21:1974-81.
14. Carey LA, Perou CM, Livasy CA, et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA 2006;295:2492-2502. 28.
15. Copson E, Eccles B, Maishman T, Gerty S, Stanton L, Cutress RI, Altman DG, Durcan L, Simmonds P, Lawrence G, Jones L, Bliss J, Eccles D, POSH Study Steering Group: Prospective observational study of breast cancer treatment outcomes for UK women aged 18–40 years at diagnosis: the POSH study. J Natl Cancer Inst 2013; 105:978–988.
16. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. Nature 2006;444:444-54
17. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 2009; 10: 451–481.

18. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. Nat Genet 2004; 36:949–951.
19. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics 2006; 7:85-97.
20. Perry GH, Yang F, Marques-Bonet T, et al. Copy number variation and evolution in humans and chimpanzees. Genome Research 2008;18:1698-1710.
21. Ohno S. Evolution by Gene Duplication. Springer-Verlag; Berlin, New York: 1970.
22. Volik S, et al. Decoding the fine-scale structure of a breast cancer genome and transcriptome. Genome Res. 2006;16:394–404.
23. Brodeur GM, Hogarty MD. In: The Genetic Basis of Human Cancer. Vogelstein B, Kinzler KW, editors. McGraw-Hill; New York: 1998; 161–172.
24. Redon R, Ishikawa S, Fitch K, Feuk L, Perry G, Andrews T, et al. Global variation in copy number in the human genome. Nature 2006;444:444–454.
25. Crespi BJ, Crofts HJ. Association testing of copy number variants in schizophrenia and autism spectrum disorders. Journal of Neurodevelopmental Disorders 2012; 4:15.
26. Saadati HR, Wittig M, Helbig I, Häsler R, Anderson CA, Mathew CG, Kupcinskas L, Parkes M, Karlsen TH, Rosenstiel P, Schreiber S. Genome-wide rare copy number variation screening in ulcerative colitis identifies potential susceptibility loci. BMC medical genetics. 2016;17(1):26.
27. Lupski J. Structural variation in the human genome. N Engl J Med 2007;356:1169–1171.
28. Futreal PA, et al. A census of human cancer genes. Nat Rev Cancer. 2004; 4:177–183.
29. Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. Nature, 2012; 463, 889.

30. Stuart D, Sellers WR. Linking somatic genetic alterations in cancer to therapeutics. Curr Opin Cell Biol. 2009; 21:304–310

31. Baudis M. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. BMC Cancer. 2007; 7:226.

32. Mitelman F, Johansson B, Mertens F. Mitelman Database of Chromosome Aberrations in Cancer. 2009 Weir BA, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007; 450:893–898.

33. Eder AM, et al. Atypical PKCiota contributes to poor prognosis through loss of apical-basal polarity and cyclin E overexpression in ovarian cancer. Proc Natl Acad Sci U S A. 2005; 102:12519–12524.

34. Zender L, et al. Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. Cell. 2006; 125:1253–1267.

35. Chitale D, et al. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR mutant tumors. Oncogene. 2009; 28:2773–2783.

36. The Cancer_Genome_Atlas_Research_Network (TCGA). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068.

37. Chapman, P.B. et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N. Engl. J. Med. 2011; 364, 2507.

38. Nguyen, K-S.H. and Neal, J.W. First-line treatment of EGFR-mutant non-small-cell lung cancer: the role of erlotinib and other tyrosine kinase inhibitors. Biologics. 2012;6, 337.

39. Srihari S, Kalimutho M, Lal S, Singla J, Patel D, Simpson PT, Khanna KK, Ragan MA. Understanding the functional impact of copy number alterations in breast cancer using a network modeling approach. Molecular BioSystems. 2016;12(3):963-72.
40. D. G. Holland, et al., ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium, EMBO Mol. Med., 2011; (3):167–180.
41. van Boerdonk RA, Sutedja TG, Snijders PJ, Reinen E, Wilting SM, van de Wiel MA, et al. DNA copy number alterations in endobronchial squamous metaplastic lesions predict lung cancer. Am J Respir Crit Care Med. 2011;184(8):948–956
42. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;455:1061–1068.
43. Phillips JL, Hayward SW, Wang Y, Vasselli J, Pavlovich C, Padilla-Nash H,Pezullo JR, Ghadimi BM, Grossfeld GD, Rivera A, et al. The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. *Cancer Res*. 2001;61:8143-8149.
44. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R,Botstein D, Borresen-Dale AL, Brown PO. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA* 2002;99:12963-12968.
45. Chibon F. Cancer gene expression signatures–the rise and fall? European journal of cancer. 2013;49(8):2000-9.

46. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ. Gene expression profiling predicts clinical outcome of breast cancer. nature. 2002 Jan 31;415(6871):530-6.
47. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet. 2005;365(9460):671-9.
48. Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, Dietze O, Greil R, Jelen A, Sevelda P, Freibauer C. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clinical Cancer Research. 2011;17(18):6012-20.
49. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, Cronin M, Baehner FL, Watson D, Bryant J, Costantino JP. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. Journal of clinical oncology. 2006;24(23):3726-34.
50. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute. 2006;98(4):262-72.
51. Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Ellis P, Harris A, Bergh J, Foekens JA, Klijn JG. Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. Journal of clinical oncology. 2007;25(10):1239-46.

52. Ellsworth RE, Seebach J, Field LA, Heckman C, Kane J, Hooke JA, Love B, Shriver CD. A gene expression signature that defines breast cancer metastases. Clinical & experimental metastasis. 2009;26(3):205-13.

53. Edén P, Ritz C, Rose C, Fernö M, Peterson C. "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. European journal of cancer. 2004;40(12):1837-41.

54. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet. 1998;20:207–11.

55. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West JA, Rostan S, Nguyen KC, Powers S, Ye KQ, Olshen A, Venkatraman E, Norton L, Wigler M. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res. 2003;13:2291–305.

56. Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, Kincaid R, Tsang P, Curry B, Baird K, Meltzer PS, Yakhini Z, Bruhn L, Laderman S. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. Proc Natl Acad Sci USA.2004;101:17765–70.

57. Erdman C, Emerson JW: A fast Bayesian change point analysis for the segmentation of microarray data. Bioinformatics 2008; 24: 2143–2148.

58. Colella S. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Research 2007;35:2013–2025.

59. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant S, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research 2007; 17:1665-1674.

60. Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. Proc Natl Acad Sci USA 2007; 104: 20007- 20012.

61. Diskin S, Eck T, Greshock J, et al. STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. Genome Res 2006; 16: 1149-1158.

62. Shah S, Lam W, Ng R, Murphy K. Modeling recurrrent CNA copy number alterations in array CGH data. Bioinformatics 2007; 23: i450-i458.

63. Rouveirol C, Stransky N, Hupé P, et al. Computation of recurrent minimal genomic alterations from array-CGH data. Bioinformatics 2006; 22: 2066-2073.

64. Kim TM, Jung YC, Rhyu MG, Jung MH, Chung YJ. GEAR: genomic enrichment analysis of regional DNA copy number changes. Bioinformatics 2008; 24: 420-421.

65. Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhinim Z. Efficient calculation of interval scores for DNA copy number data analysis. J Comput Biol 2006; 13: 215-228.

66. Ben-Dor A, Lipson D, Tsalenko A, et al. Framework for identifying common aberrations in DNA copy number data. Proc RECOMB’07 2007; 4453: 122-36.

67. Curtis C, Shah SP, Chin S-F, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486:346–352.

68. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003*; 19: 185–93*

69. Seshan VE, Olshen A. DNAcopy: DNA copy number data analysis. R package version. 2010;1(1).

70. Habib M, Paul C, Viennot L. Lex-BFS, a Partition Refining Technique. Application to Transitive Orientation, Interval Graph Recognition and Consecutive 1's Testing. 1996; 1–21

71. Zhang P, Schon EA, Fischer SG, Cayanis E, Weiss J, Kistler S, Bourne PE. An algorithm based on graph theory for the assembly of contigs in physical mapping of DNA. Bioinformatics 1994; 10: 309–317.

72. Gentleman R, Vandal C. Computational Algorithms for Censored-Data Problems Using Intersection Graphs. J. Comput. Graph. Stat. 2001; 10: 403–421.

73. Wu HT, Hajirasouliha I, Raphael BJ. Detecting independent and recurrent copy number aberrations using interval graphs. Bioinformatics 2014; 30: i195–i203.

74. Chen Y. E, Christopher M. T, Kou Y, *et al*. Enrichr: interactive and collaborative HTML5gene list enrichment analysis tool. BMC Bioinformatics 2013;14:128

75. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 2010;38(16)

76. Gao T, He B, Liu S, Zhu H, Tan K, Qian J. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics. 2016;32(23):3543-51.

77. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.

78. Weinstein JN, Myers TG, O'connor PM, Friend SH, Fornace AJ, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK. An information-intensive approach to the molecular pharmacology of cancer. Science. 1997;275(5298):343-9.

79. Knight JC. Regulatory polymorphisms underlying complex disease traits. Journal of molecular medicine. 2005;83(2):97-109.

80. Lahtz C, Pfeifer GP. Epigenetic changes of DNA repair genes in cancer. *Journal of Molecular Cell Biology*. 2011;3(1):51-58.

81. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009 Sep 10;461(7261):199-205.

82. Dalla-Favera R, Bregni M, Erikson J, Patterson D, Gallo RC, Croce CM. Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. Proceedings of the National Academy of Sciences. 1982;79(24):7824-7.

83. Buchwalter G, Hickey MM, Cromer A, Selfors LM, Gunawardane RN, Frishman J, Jeselsohn R, Lim E, Chi D, Fu X, Schiff R. PDEF promotes luminal differentiation and acts as a survival factor for ER-positive breast cancer cells. Cancer Cell. 2013; 23(6):753-67.

84. Sood AK, Saxena R, Groth J, et al. Expression characteristics of PDEF support a role in breast and prostate cancer progression. *Hum Pathol* 2007;**38**:1628–38.

85. Bose S, Wang SI, Terry MB, Hibshoosh H, Parsons R: Allelic loss of chromosome 10q23 is associated with tumor progression in breast carcinomas. Oncogene. 1998; 17 (1): 123-127.

86. Feilotter HE, Coulon V, McVeigh JL, Boag AH, Dorion-Bonnet F, Duboue B, Latham WC, Eng C, Mulligan LM, Longy M: Analysis of the 10q23 chromosomal region and the PTEN gene in human sporadic breast carcinoma. Br J Cancer. 1999; 79 (5–6): 718-723.

87. Zhou X, Thorgeirsson SS, Popescu NC. Restoration of DLC-1 gene expression induces apoptosis and inhibits both cell growth and tumorigenicity in human hepatocellular carcinoma cells. Oncogene. 2004;23(6):1308-13.

88. Emi M, Utada Y, Yoshimoto M, Sato T, Minobe K, Matsumoto S, Akiyama F, Iwase T, Haga S, Kajiwara T, Sakamoto G. Correlation of allelic loss with poor postoperative survival in breast cancer. Breast Cancer. 1999;6(4):351-6.

89. Ho WC, Pikor L, Gao Y, Elliott BE, Greer PA. Calpain 2 regulates Akt-FoxO-p27Kip1 protein signaling pathway in mammary carcinoma. Journal of Biological Chemistry. 2012;287(19):15458-65.

90. Whitaker HC, Shiong LL, Kay JD, Grönberg H, Warren AY, Seipel A, Wiklund F, Thomas B, Wiklund P, Miller JL, Menon S. N-acetyl-L-aspartyl-L-glutamate peptidase-like 2 is overexpressed in cancer and promotes a pro-migratory and pro-metastatic phenotype. Oncogene. 2014;33(45):5274-87.

91. Menendez JA, Lupu R. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. Nature Reviews Cancer. 2007;7(10):763-77.

92. Wymann MP, Schneiter R. Lipid signalling in disease. Nature Reviews Molecular Cell Biology. 2008;9(2):162-76.

93. Lingwood D, Simons K. Lipid rafts as a membrane-organizing principle. *Science* 2010; **327**: 46–50.

94. Talbot LJ, Bhattacharya SD, Kuo PC. Epithelial-mesenchymal transition, the tumor microenvironment, and metastatic behavior of epithelial malignancies. Int J Biochem Mol Biol. 2012;3(2):117-36.

95. Martin TA, Mansel RE, Jiang WG. Loss of occludin leads to the progression of human breast cancer. International journal of molecular medicine. 2010;26(5):723.

Appendix

ANNOVAR gene list

Chr	Start	End	Size	Discovery Cluster	Validation Cluster	CNA	Type	Func.refGene	Gene Symbol	GeneDetail.refGene
1	58820476	58846018	25542	6	7	GAIN	INNER	intergenic	TACSTD2,MYSM1	dist=4722;dist=46981
1	84551640	84565561	13921	5	6	GAIN	INNER	exonic	SAMD13	.
1	1.44E+08	1.44E+08	1232	20	30	GAIN	INNER	ncRNA_intronic	LOC100996724	.
1	1.47E+08	1.47E+08	20486	35	50	GAIN	INNER	intergenic	NBPF25P,LOC388692	dist=99598;dist=57321
1	1.91E+08	1.91E+08	11287	40	50	GAIN	INNER	exonic	CDC73	.
1	1.91E+08	1.91E+08	16207	40	50	GAIN	INNER	exonic	CDC73	.
1	2.22E+08	2.22E+08	610	48	47	GAIN	INNER	intronic	CAPN2	.
1	16885044	16885329	285	11	15	LOSS	INNER	intergenic	MIR3675,ESPNP	dist=4634;dist=4971
1	58803399	58846018	42619	6	7	GAIN	OUTER	exonic	TACSTD2	.
1	84481190	84729446	248256	5	6	GAIN	OUTER	exonic	DNASE2B,RPF1,SAMD13	.
1	1.44E+08	1.44E+08	1988	20	30	GAIN	OUTER	ncRNA_intronic	LOC100996724	.
1	1.47E+08	1.47E+08	20947	35	50	GAIN	OUTER	intergenic	NBPF25P,LOC388692	dist=99172;dist=57286
1	1.91E+08	1.91E+08	45654	40	50	GAIN	OUTER	exonic;exonic	CDC73	.
1	1.91E+08	1.91E+08	45654	40	50	GAIN	OUTER	exonic;exonic	CDC73	.
1	2.22E+08	2.22E+08	14667	48	47	GAIN	OUTER	exonic	CAPN2	.
1	16884038	16889666	5628	11	15	LOSS	OUTER	downstream	ESPNP	.
2	4197415	4200525	3110	11	13	GAIN	INNER	intergenic	LOC105373394,LINC01249	dist=196670;dist=453158
2	29028479	29029800	1321	5	8	GAIN	INNER	intergenic	WDR43,FAM179A	dist=3889;dist=27868
2	52618925	52634800	15875	17	22	GAIN	INNER	intergenic	LOC730100,MIR4431	dist=130366;dist=148364
2	64271883	64282835	10952	9	7	GAIN	INNER	ncRNA_exonic	LINC00309	.
2	64283559	64306614	23055	9	7	GAIN	INNER	ncRNA_exonic	LINC00309	.
2	65560324	65677884	117560	5	6	GAIN	INNER	intergenic	SPRED2,MIR4778	dist=47164;dist=761001
2	97507180	97517476	10296	5	5	LOSS	INNER	exonic	ANKRD36B	.
2	4195432	4201388	5956	11	13	GAIN	OUTER	intergenic	LOC105373394,LINC01249	dist=194687;dist=452295
2	29010022	29040695	30673	5	8	GAIN	OUTER	exonic	WDR43	.
2	52618897	52634917	16020	17	22	GAIN	OUTER	intergenic	LOC730100,MIR4431	dist=130338;dist=148247
2	64247921	64315987	68066	9	7	GAIN	OUTER	ncRNA_exonic	LINC00309,LOC100507006	.
2	64271883	64315987	44104	9	7	GAIN	OUTER	ncRNA_exonic	LINC00309,LOC100507006	.
2	65463975	65678403	214428	5	6	GAIN	OUTER	exonic	SPRED2	.
2	97507180	97520698	13518	5	5	LOSS	OUTER	exonic	ANKRD36B	.
3	9660712	9664562	3850	10	12	GAIN	INNER	intergenic	LHFPL4,MTMR14	dist=90226;dist=1555
3	1.09E+08	1.09E+08	7853	7	7	GAIN	INNER	intergenic	LINC00636,CD47	dist=96548;dist=9787
3	1.31E+08	1.31E+08	1352	18	12	GAIN	INNER	intronic;intronic	ALG1L2	.
3	1.76E+08	1.76E+08	3921	18	22	GAIN	INNER	intronic	NAALADL2	.
3	1.76E+08	1.76E+08	98	18	22	GAIN	INNER	intronic	NAALADL2	.
3	1.86E+08	1.86E+08	10820	11	12	GAIN	INNER	intergenic	LOC105374250,EIF2B5-AS1	dist=2736;dist=93510
3	1.97E+08	1.97E+08	1140	23	22	GAIN	INNER	ncRNA_intronic	LINC00969	.
3	1.97E+08	1.97E+08	2643	23	23	GAIN	INNER	ncRNA_intronic	LINC00969	.
3	1.99E+08	1.99E+08	3140	12	11	GAIN	INNER	intergenic	ANKRD18DP,FAM157A	dist=27897;dist=40658
3	26589981	26589990	9	7	7	LOSS	INNER	intergenic	LINC00692,LRRC3B	dist=699791;dist=49311
3	62243538	62257523	13985	6	6	LOSS	INNER	exonic	PTPRG	.
3	64855417	64863915	8498	6	5	LOSS	INNER	ncRNA_intronic	ADAMTS9-AS2	.
3	9640519	9688008	47489	10	12	GAIN	OUTER	exonic	MTMR14	.
3	1.09E+08	1.09E+08	8120	7	7	GAIN	OUTER	intergenic	LINC00636,CD47	dist=96548;dist=9520
3	1.31E+08	1.31E+08	1352	18	12	GAIN	OUTER	intronic;intronic	ALG1L2	.
3	1.76E+08	1.76E+08	13308	18	22	GAIN	OUTER	intronic	NAALADL2	.
3	1.76E+08	1.76E+08	55435	18	22	GAIN	OUTER	exonic	NAALADL2	.
3	1.86E+08	1.86E+08	22652	11	12	GAIN	OUTER	exonic	LOC105374250	.
3	1.97E+08	1.97E+08	11253	23	22	GAIN	OUTER	ncRNA_exonic	MIR570	.
3	1.97E+08	1.97E+08	25727	23	23	GAIN	OUTER	exonic	MUC20	.
3	1.99E+08	1.99E+08	13787	12	11	GAIN	OUTER	intergenic	ANKRD18DP,FAM157A	dist=26805;dist=31103
3	26588205	26591799	3594	7	7	LOSS	OUTER	intergenic	LINC00692,LRRC3B	dist=698015;dist=47502
3	62242606	62277516	34910	6	6	LOSS	OUTER	exonic	PTPRG	.
3	64731448	64923169	191721	6	5	LOSS	OUTER	ncRNA_exonic	ADAMTS9-AS2	.
4	59521	61566	2045	13	6	LOSS	INNER	intronic	ZNF595,ZNF718	.
4	1.04E+08	1.04E+08	11751	5	5	LOSS	INNER	intergenic	CENPE,LOC101929448	dist=116618;dist=98264
4	1.57E+08	1.57E+08	7353	5	7	LOSS	INNER	intergenic	CTSO,PDGFC	dist=5840;dist=794522
4	59521	64435	4914	13	6	LOSS	OUTER	intronic	ZNF595,ZNF718	.
4	1.04E+08	1.04E+08	52479	5	5	LOSS	OUTER	intergenic	CENPE,LOC101929448	dist=75890;dist=98264
4	1.57E+08	1.57E+08	7637	5	7	LOSS	OUTER	intergenic	CTSO,PDGFC	dist=5834;dist=794244
5	18541955	18547895	5940	15	12	GAIN	INNER	intergenic	LOC646241,CDH18	dist=575600;dist=961017
5	22246497	22346803	100306	12	11	GAIN	INNER	UTR5	CDH12	
5	60888563	61017766	129203	8	10	LOSS	INNER	exonic	C5orf64	.
5	18514272	18554326	40054	15	12	GAIN	OUTER	intergenic	LOC646241,CDH18	dist=547917;dist=954586
5	22194503	22414425	219922	12	11	GAIN	OUTER	UTR5	CDH12	
5	60887533	61460364	572831	8	10	LOSS	OUTER	exonic	C5orf64	.
6	32539531	32556611	17080	6	5	GAIN	INNER	intergenic	HLA-DRA,HLA-DRB5	dist=18729;dist=36521
6	32557501	32560908	3407	6	5	GAIN	INNER	intergenic	HLA-DRA,HLA-DRB5	dist=36699;dist=32224
6	34625387	34634997	9610	9	9	GAIN	INNER	UTR5	SPDEF	

6	79037167	79044205	7038	16	21	GAIN	INNER	intergenic;intergenic	MEI4,IRAK1BP1	dist=343689;dist=589775
6	32539531	32557028	17497	6	5	GAIN	OUTER	intergenic	HLA-DRA,HLA-DRB5	dist=18729;dist=36104
6	32557028	32562254	5226	6	5	GAIN	OUTER	intergenic	HLA-DRA,HLA-DRB5	dist=36226;dist=30878
6	34624907	34656516	31609	9	9	GAIN	OUTER	UTR5	SPDEF	
6	79037167	79044205	7038	16	21	GAIN	OUTER	intergenic;intergenic	MEI4,IRAK1BP1	dist=343689;dist=589775
7	54152334	54153709	1375	6	5	GAIN	INNER	intergenic	LINC01446,HPVC1	dist=305216;dist=82702
7	61794167	61799563	5396	10	11	GAIN	INNER	intergenic	NONE,ZNF733P	dist=NONE;dist=589542
7	76141447	76146160	4713	8	8	GAIN	INNER	intergenic;intergenic	LOC100133091, DTX2P1-UPK3BP1-PMS2P11	dist=46212;dist=301915
7	1.09E+08	1.09E+08	11898	9	12	GAIN	INNER	intergenic	C7orf66,EIF3IP1	dist=917317;dist=145425
7	1.35E+08	1.35E+08	4577	11	13	GAIN	INNER	intronic	CNOT4	.
7	1.42E+08	1.42E+08	3386	7	10	GAIN	INNER	intergenic	PRSS1,PRSS3P2	dist=10343;dist=4102
7	1.43E+08	1.43E+08	34	14	27	GAIN	INNER	intergenic	CTAGE6,TCAF2P1	dist=22466;dist=29740
7	38296343	38297866	1523	18	22	LOSS	INNER	intergenic	TARP,TRG-AS1	dist=16570;dist=49837
7	54152334	54153757	1423	6	5	GAIN	OUTER	intergenic	LINC01446,HPVC1	dist=305216;dist=82654
7	61794167	61818546	24379	10	11	GAIN	OUTER	intergenic	NONE,ZNF733P	dist=NONE;dist=570559
7	76141447	76146160	4713	8	8	GAIN	OUTER	intergenic;intergenic	LOC100133091, DTX2P1-UPK3BP1-PMS2P11	dist=46212;dist=301915
7	1.09E+08	1.09E+08	12229	9	12	GAIN	OUTER	intergenic	C7orf66,EIF3IP1	dist=916986;dist=145425
7	1.35E+08	1.35E+08	9830	11	13	GAIN	OUTER	intronic	CNOT4	.
7	1.42E+08	1.42E+08	3696	7	10	GAIN	OUTER	intergenic	PRSS1,PRSS3P2	dist=10318;dist=3817
7	1.43E+08	1.43E+08	13974	14	27	GAIN	OUTER	intergenic	CTAGE6,TCAF2P1	dist=18848;dist=19418
7	38295506	38297939	2433	18	22	LOSS	OUTER	intergenic	TARP,TRG-AS1	dist=15733;dist=49764
8	1349638	1350270	632	5	6	GAIN	INNER	intronic	DLGAP2	.
8	3989394	3992254	2860	5	7	GAIN	INNER	intronic	CSMD1	.
8	8027374	8027446	72	7	9	GAIN	INNER	intergenic	MIR548I3,FAM86B3P	dist=43353;dist=96056
8	40695071	40697114	2043	17	16	GAIN	INNER	intronic	ZMAT4	.
8	86730480	86740395	9915	32	38	GAIN	INNER	intergenic	CA2,REXO1L2P	dist=149507;dist=13685
8	96278448	96284249	5801	42	44	GAIN	INNER	intergenic	PLEKHF2,LINC01298	dist=40359;dist=4162
8	1.35E+08	1.35E+08	2520	43	49	GAIN	INNER	intergenic	LOC101927822,ZFAT	dist=248787;dist=324074
8	3552110	3554053	1943	25	21	LOSS	INNER	exonic	CSMD1	.
8	5633227	5636521	3294	23	17	LOSS	INNER	intergenic	CSMD1,LOC100287015	dist=793491;dist=611964
8	14388851	14391732	2881	23	18	LOSS	INNER	intronic	SGCZ	.
8	1348245	1350652	2407	5	6	GAIN	OUTER	intronic	DLGAP2	.
8	3985016	3992622	7606	5	7	GAIN	OUTER	intronic	CSMD1	.
8	8015467	8046615	31148	7	9	GAIN	OUTER	intergenic	MIR548I3,FAM86B3P	dist=31446;dist=76887
8	40693570	40699795	6225	17	16	GAIN	OUTER	intronic	ZMAT4	.
8	86720993	86740395	19402	32	38	GAIN	OUTER	intergenic	CA2,REXO1L2P	dist=140020;dist=13685
8	96258231	96312735	54504	42	44	GAIN	OUTER	ncRNA_exonic	LINC01298	.
8	1.35E+08	1.35E+08	3094	43	49	GAIN	OUTER	intergenic	LOC101927822,ZFAT	dist=248787;dist=323500
8	3551271	3564930	13659	25	21	LOSS	OUTER	exonic	CSMD1	.
8	5625845	5636589	10744	23	17	LOSS	OUTER	intergenic	CSMD1,LOC100287015	dist=786109;dist=611896
8	14385622	14391732	6110	23	18	LOSS	OUTER	intronic	SGCZ	.
9	67699737	67709343	9606	6	9	GAIN	INNER	intergenic	ANKRD20A3,MIR4477B	dist=139624;dist=195785
9	93166194	93261927	95733	5	7	GAIN	INNER	exonic	NFIL3	.
9	1.35E+08	1.35E+08	1403	6	5	GAIN	INNER	intergenic	ABO,SURF6	dist=36990;dist=8520
9	1490330	1490755	425	6	7	LOSS	INNER	intergenic	DMRT2,SMARCA2	dist=442776;dist=514464
9	1500170	1503092	2922	6	7	LOSS	INNER	intergenic	DMRT2,SMARCA2	dist=452616;dist=502127
9	5027454	5029342	1888	7	11	LOSS	INNER	intronic	JAK2	.
9	67678161	67743272	65111	6	9	GAIN	OUTER	intergenic	ANKRD20A3,MIR4477B	dist=118048;dist=161856
9	93157333	93373420	216087	5	7	GAIN	OUTER	exonic	AUH,NFIL3	.
9	1.35E+08	1.35E+08	4930	6	5	GAIN	OUTER	intergenic	ABO,SURF6	dist=33463;dist=8520
9	1456070	1504392	48322	6	7	LOSS	OUTER	intergenic	DMRT2,SMARCA2	dist=408516;dist=500827
9	1490330	1504392	14062	6	7	LOSS	OUTER	intergenic	DMRT2,SMARCA2	dist=442776;dist=500827
9	5027454	5030334	2880	7	11	LOSS	OUTER	intronic	JAK2	.
10	5737990	5742226	4236	24	32	GAIN	INNER	intronic	ASB13	.
10	6703994	6745967	41973	20	27	GAIN	INNER	ncRNA_exonic	LOC101928150,MIR4454	.
10	14598341	14600566	2225	19	27	GAIN	INNER	UTR3	FAM107B	
10	19676387	19685158	8771	15	19	GAIN	INNER	exonic	MALRD1	.
10	20731174	20741934	10760	17	18	GAIN	INNER	intergenic	PLXDC2,MIR4675	dist=112384;dist=138971
10	20747419	20761390	13971	17	18	GAIN	INNER	intergenic	PLXDC2,MIR4675	dist=128629;dist=119515
10	22772784	22815216	42432	15	21	GAIN	INNER	intergenic	LOC100499489,PIP4K2A	dist=5920;dist=48556
10	66980962	66984437	3475	7	13	GAIN	INNER	intergenic	LOC101928887,LINC01515	dist=626680;dist=16752
10	81439210	81449106	9896	6	6	GAIN	INNER	ncRNA_exonic	NUTM2B-AS1	.
10	89710114	89713882	3768	10	10	LOSS	INNER	exonic	PTEN	.
10	5736767	5744158	7391	24	32	GAIN	OUTER	intronic	ASB13	.
10	6703994	6797043	93049	20	27	GAIN	OUTER	ncRNA_exonic	LOC101928150,MIR4454	.
10	14477106	14629604	152498	19	27	GAIN	OUTER	exonic	FAM107B	.
10	19656598	19750338	93740	15	19	GAIN	OUTER	exonic	MALRD1	.
10	20726541	20747050	20509	17	18	GAIN	OUTER	intergenic	PLXDC2,MIR4675	dist=107751;dist=133855
10	20747068	20779396	32328	17	18	GAIN	OUTER	intergenic	PLXDC2,MIR4675	dist=128278;dist=101509
10	22764104	22849622	85518	15	21	GAIN	OUTER	ncRNA_exonic	LOC100499489	.

10	66978024	66984437	6413	7	13	GAIN	OUTER	intergenic	LOC101928887,LINC01515	dist=623742;dist=16752
10	81439210	81452222	13012	6	6	GAIN	OUTER	ncRNA_exonic	NUTM2B-AS1	.
10	89708179	89713882	5703	10	10	LOSS	OUTER	exonic	PTEN	.
11	4931741	4932834	1093	7	11	GAIN	INNER	exonic	OR51A2	.
11	4931741	4932966	1225	7	11	GAIN	OUTER	exonic	OR51A2	.
12	180797	191614	10817	14	24	GAIN	INNER	exonic	SLC6A12	.
12	7895693	7897774	2081	13	23	GAIN	INNER	intronic;intronic	SLC2A14	.
12	7899067	7905082	6015	13	23	GAIN	INNER	intronic	SLC2A14	.
12	8253318	8257868	4550	13	21	GAIN	INNER	intergenic	FAM66C,FAM90A1	dist=8455;dist=7255
12	11634190	11638369	4179	8	15	GAIN	INNER	intergenic	LINC01252,ETV6	dist=25588;dist=55686
12	180797	195197	14400	14	24	GAIN	OUTER	exonic	SLC6A12	.
12	7895693	7897774	2081	13	23	GAIN	OUTER	intronic;intronic	SLC2A14	.
12	7899067	7909593	10526	13	23	GAIN	OUTER	exonic	SLC2A14	.
12	8253318	8261982	8664	13	21	GAIN	OUTER	intergenic	FAM66C,FAM90A1	dist=8455;dist=3141
12	11606108	11670489	64381	8	15	GAIN	OUTER	ncRNA_exonic	LINC01252	.
13	1.12E+08	1.12E+08	8703	10	7	GAIN	INNER	intronic	ATP11AUN	.
13	1.13E+08	1.13E+08	9463	10	9	GAIN	INNER	exonic	ATP4B	.
13	22518343	22523101	4758	10	8	LOSS	INNER	intergenic	LINC00621,SGCG	dist=129835;dist=129959
13	56656317	56661149	4832	15	17	LOSS	INNER	intergenic	PRR20D,PCDH17	dist=13964;dist=442641
13	1.12E+08	1.12E+08	30152	10	7	GAIN	OUTER	UTR5	ATP11AUN	
13	1.13E+08	1.13E+08	26962	10	9	GAIN	OUTER	exonic	ATP4B,GRK1	.
13	22512167	22523373	11206	10	8	LOSS	OUTER	intergenic	LINC00621,SGCG	dist=123659;dist=129687
13	56656291	56661535	5244	15	17	LOSS	OUTER	intergenic	PRR20D,PCDH17	dist=13938;dist=442255
14	23564490	23564891	401	7	9	GAIN	INNER	intronic	DHRS4L1	.
14	25054809	25064593	9784	6	5	GAIN	INNER	intergenic	STXBP6,NOVA1	dist=465466;dist=920336
14	25065074	25069117	4043	6	5	GAIN	INNER	intergenic	STXBP6,NOVA1	dist=475731;dist=915812
14	23559232	23570486	11254	7	9	GAIN	OUTER	intronic	DHRS4L1	.
14	24846340	25069117	222777	6	5	GAIN	OUTER	intergenic	STXBP6,NOVA1	dist=256997;dist=915812
14	24846340	25075027	228687	6	5	GAIN	OUTER	intergenic	STXBP6,NOVA1	dist=256997;dist=909902
15	84140524	84146976	6452	5	9	GAIN	INNER	intergenic	KLHL25,AGBL1	dist=1331;dist=277067
15	22859182	22873331	14149	6	11	LOSS	INNER	ncRNA_exonic	SNORD116-10,SNORD116-11,SNORD116-12, SNORD116-2,SNORD116-3, SNORD116-5,SNORD116-6,SNORD116-7, SNORD116-8,SNORD116-9	.
15	84140524	84157332	16808	5	9	GAIN	OUTER	intergenic	KLHL25,AGBL1	dist=1331;dist=266711
15	22846930	22875304	28374	6	11	LOSS	OUTER	ncRNA_exonic	SNORD116-1,SNORD116-10,SNORD116-11, SNORD116-12,SNORD116-13,SNORD116-2, SNORD116-3,SNORD116-4,SNORD116-5, SNORD116-6,SNORD116-7, SNORD116-8,SNORD116-9	.
16	14961825	14966336	4511	15	12	GAIN	INNER	intergenic	NPIPA1,PDXDC1	dist=8393;dist=9757
16	14961806	14966847	5041	15	12	GAIN	OUTER	intergenic	NPIPA1,PDXDC1	dist=8374;dist=9246
17	31460821	31463472	2651	13	16	GAIN	INNER	intergenic	CCL4,CCL3L1	dist=3694;dist=82909
17	43751830	43753351	1521	9	14	GAIN	INNER	intronic	SKAP1	.
17	45136676	45139395	2719	21	18	GAIN	INNER	exonic	SLC35B1	.
17	69474254	69478126	3872	22	17	GAIN	INNER	intergenic	LINC00469,RPL38	dist=137983;dist=233264
17	21245986	21253816	7830	15	9	LOSS	INNER	UTR5	KCNJ12,KCNJ18	
17	33672360	33694008	21648	9	7	LOSS	INNER	intergenic	LOC440434,MRPL45	dist=5278;dist=12500
17	31460821	31479995	19174	13	16	GAIN	OUTER	intergenic	CCL4,CCL3L1	dist=3694;dist=66386
17	43722185	43756717	34532	9	14	GAIN	OUTER	ncRNA_exonic	THRA1/BTR	.
17	45134175	45142242	8067	21	18	GAIN	OUTER	exonic	SLC35B1	.
17	69429096	69496533	67437	22	17	GAIN	OUTER	intergenic	LINC00469,RPL38	dist=92825;dist=214857
17	21227031	21271210	44179	15	9	LOSS	OUTER	exonic	KCNJ12,KCNJ18	.
17	33664567	33752586	88019	9	7	LOSS	OUTER	exonic	GPR179,MRPL45	.
18	9549925	9575313	25388	5	7	GAIN	INNER	exonic	PPP4R1	.
18	43704001	43707399	3398	6	6	GAIN	INNER	intronic	SMAD2	.
18	55813854	55816005	2151	9	10	GAIN	INNER	intergenic	PMAIP1,MC4R	dist=91336;dist=373539
18	36514797	36519362	4565	9	6	LOSS	INNER	intergenic	LINC01477,KC6	dist=581602;dist=794872
18	61883719	61883744	25	5	5	LOSS	INNER	intergenic	CDH7,CDH19	dist=183537;dist=435660
18	9417006	9594232	177226	5	7	GAIN	OUTER	exonic	PPP4R1,RALBP1	.
18	43703934	43707399	3465	6	6	GAIN	OUTER	intronic	SMAD2	.
18	55813796	55819506	5710	9	10	GAIN	OUTER	intergenic	PMAIP1,MC4R	dist=91278;dist=370038
18	36514489	36519388	4899	9	6	LOSS	OUTER	intergenic	LINC01477,KC6	dist=581294;dist=794846
18	61878473	61885427	6954	5	5	LOSS	OUTER	intergenic	CDH7,CDH19	dist=178291;dist=433977
19	40629439	40647918	18479	13	13	GAIN	INNER	exonic	FFAR2	.
19	60890917	60901410	10493	8	11	GAIN	INNER	exonic	EPN1	.
19	40620640	40702499	81859	13	13	GAIN	OUTER	exonic	DMKN,FFAR2,KRTDAP	.
19	60890917	60904859	13942	8	11	GAIN	OUTER	exonic	EPN1	.
20	14741416	14743670	2254	8	10	GAIN	INNER	intronic	MACROD2	.
20	41215727	41219453	3726	12	18	GAIN	INNER	intronic	PTPRT	.
20	14741416	14743754	2338	8	10	GAIN	OUTER	intronic	MACROD2	.
20	41202818	41220578	17760	12	18	GAIN	OUTER	intronic	PTPRT	.
21	43341103	43343023	1920	9	8	GAIN	INNER	intergenic	PKNOX1,CBS	dist=13993;dist=3347
21	46828863	46834725	5862	10	9	GAIN	INNER	intergenic	DIP2A,S100B	dist=14509;dist=8234
21	43329959	43356005	26046	9	8	GAIN	OUTER	exonic	CBS,CBSL	.
21	46828497	46846083	17586	10	9	GAIN	OUTER	exonic	S100B	.
22	20146692	20170596	23904	7	15	GAIN	INNER	exonic	TMEM191C	.

22	20146692	20170596	23904	7	15	GAIN	INNER	exonic	TMEM191C	.
22	20816553	20820788	4235	9	9	LOSS	INNER	intergenic	PRAMENP,VPREB1	dist=88221;dist=108404
22	20145867	20170766	24899	7	15	GAIN	OUTER	exonic	TMEM191C	.
22	20816118	20833270	17152	9	9	LOSS	OUTER	intergenic	PRAMENP,VPREB1	dist=87786;dist=95922