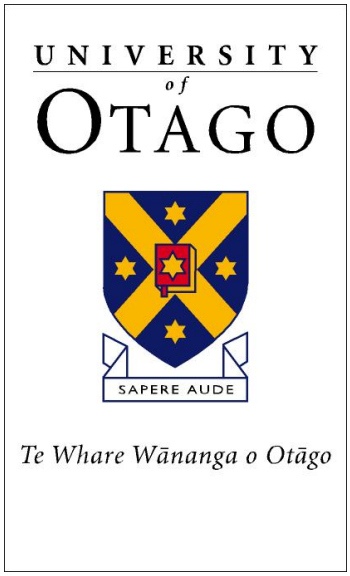
Body Mass Index and Pathway Dysregulation in Cancer

Riku Takei



A Thesis Submitted For The Degree Of

Master Of Science

At The University Of Otago, Dunedin, New Zealand

March 31, 2017

Abstract

Obesity has been a major global problem for more than a decade, associated with many noncommunicable diseases such as cancer. The number of obese people, both adults and children, has risen in every country of the world and the trend will likely to continue. Cancers are caused by dysregulation of various molecular pathways that allow tumour cells to proliferate, survive and migrate. One of the difficulties associated with the treatment of cancers is the identification of the underlying biological pathways that drive tumorigenesis. This research aims to determine whether gene expression signatures exist that are specific to obesity across multiple cancer types, and to investigate whether there are any common pathways being dysregulated in cancers based on these genetic signatures. In this work no genetic signatures or differentially expressed genes were found between obese and non-obese patients that were common across multiple cancer types. However, the Akt, epidermal growth factor receptor (EGFR), transforming growth factor-β (TGFβ) and Src pathways may have a role in promoting the tumour progression in patients that are obese. It is likely that there is some complex mechanism underlying the relationship between obesity and cancer. A better understanding of the pathways being dysregulated in cancer cells in obese patients may lead to improved clinical decisions, and contribute towards personalised treatment in the future.

Acknowledgement

First and foremost, I would like to thank my supervisor Mik, for giving me this project and the opportunity to learn and develop the bioinformatics skills required to complete it. You have been a great supervisor; I have learnt a lot from you over the past two years, and I have enjoyed being a part of the Black lab.

I would like to thank my committee members, Dr. Anita Dunbier, Dr. Miriam Sharpe, and Dr. Peter Mace (temporary member), for being patient during my presentations that were full of statistics and heatmaps. I am grateful for all the suggestions and input you have given me to make this a better project.

I would also like to thank both past and present members of the Black lab, for useful tips and tricks to get me through my Masters. Especially James Boocock, for introducing me to a variety of useful tool sets (by the way, everyone should start using vim); Murray Cadzow, for organising the fortnightly "Shit You Should Know About" (a.k.a. SYSKA) sessions to teach us more about R and other useful tools; and Tom Kelly, for being my companion in the Fish-bowl and helping me out with some of the code in my project.

I would like to thank all of my friends and family, for all their support and fun memories throughout the years that I have been at University. I wish I could name everyone, but extending this thesis any longer is probably not a good idea... Long story short, my life would not be nearly as interesting or enjoyable without you guys.

Lastly, a special thanks to Gerry and Tessa who generously proof-read my introduction.

Table of Contents

List of Figures

List of Tables

Glossary

Adipocyte	Adipose tissue cell. 23–25
Allele	Variant form of a given gene. 19
Apoptosis	The death of cells which occurs as a normal and controlled part of an organism's growth or development. 13
Carcinogen	A substance capable of causing cancer. 20
Chromothripsis	An all-at-once mutational pattern in which a large number of rearrangements are confined to local regions of one or more chromosomes (Stephens *et al.*, 2011). 15, 16
Driver mutation	Mutations that helps tumorigenesis and its subsequent progression. 19
Germline mutation	Mutations that are passed on from the parents. 19, 20
Immunoediting	Process where the immune system successfully eliminates the highly immunogenic cancer cells during surveillance, but fails to eliminate the weakly immunogenic cells. 17
Kataegis	Showers of single nucleotide changes confined to local regions, often near sites of chromosome rearrangements (Stephens *et al.*, 2011). 15, 16
Limited duration prevalence	Number of patients diagnosed with a disease within a fixed period of time in the past (Bray *et al.*, 2013). 9

Oncogene	Genes that promote or "accelerate" the process of tumorigenesis and/or tumour progression. 18–20
Ontology	A set of concepts and categories in a subject area or domain that shows their properties and the relations between them. 49
Passanger mutation	Mutations that have neither a selective advantage nor contribute to tumour progression. 19
Power	Probability that it will reject a false null hypothesis; it is inversely related to the probability of making a Type II error. 45
R	Statistical programming language. iv, 32, 34, 37, 38, 41–43, 49–52, 58, 59, 200
RNA-seq	Sequence data using the transcriptome. 32, 36, 37, 40, 41, 63, 65, 118, 119
Somatic mutation	Mutations that are acquried by the cell over the course of the lifetime of an individual. 19
Stability gene	Genes that are involved in DNA repair mechanisms. 19, 20
Tumour suppressor gene	Genes that prevent or "decelerate" the process of tumorigenesis and/or tumour progression. 18–20
Type I error	Incorrect rejection of a true null hypothesis; false positive. 44, 45, 48, 111, 112
Type II error	Incorrect retaining of a false null hypothesis; false negative. 45

Acronyms

μ	Mean. 41
σ	Standard deviation. 41
AID	Activation-induced cytidine deaminase. 16
ANOVA	Analysis of variance. 58, 62, 68, 78, 84, 109, 135, 139, 141–143, 196, 200
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like. 16
ATP	Adenosine triphosphate. 17
BCAT	β-catenin. 30, 34, 54, 56, 57, 89, 91, 92, 96, 97, 102, 103, 157, 179–187
Bcl-2	B-cell lymphoma 2. 13, 16
BER	Base-excision repair. 19
BIC	Bayesian Information Criterion. 57
BLCA	Bladder urothelial carcinoma. vi, viii, ix, xii, 36, 63, 64, 68, 71, 78, 81, 82, 84, 85, 108–110, 121, 123, 124, 131, 147, 148
BMI	Body mass index. v, vii–x, xii–xv, 1, 2, 6, 23, 24, 33, 35–39, 41, 51, 53, 56, 57, 60, 62–72, 77–82, 84, 87, 96–99, 102, 107–109, 111–119, 124, 126, 131, 135, 139, 143, 144, 151, 179–200
CAMERA	Correlation adjusted mean rank gene set test. iv, 48, 49
cDNA	Complimentary DNA. 26
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma. xiii, 37, 84, 85, 112, 135, 147, 149
COAD	Colon adenocarcinoma. xiii, 37, 84, 85, 135, 147, 149
cpm	Counts per million. 40
CYP	Cytochrome P450. 20

DALY Disability-Adjusted Life Year. 1

DEG Differentially expressed gene. viii, xii, 43–46, 59, 72, 74, 81, 83–85, 109, 111–113

DNA Deoxyribonucleic acid. xvii, 13, 17, 19, 20, 26, 27, 55

EGFR Epidermal growth factor receptor. i, 30, 34, 55, 57, 89, 92, 94, 102–104, 115, 116, 121, 173

ER Estrogen receptor. vii, viii, x, 29, 30, 33, 34, 36, 41, 52–57, 73, 89, 92, 94, 96, 97, 102, 104, 108, 110, 111, 115, 116, 120, 121, 151–154, 157, 179–187

ERK Extracellular signal regulated kinase. 25

ETC Electron transport chain. 17

FC Fold change. 29, 48, 72

FCS Functional class scoring. iv, 46–49

FDR False discovery rate. xii, 45, 74, 85, 143

FTO Fat mass and obesity associated. 6

FWER Family-wise error rate. 45

GEO Gene Expression Omnibus. 32–34

GO Gene Ontology. xiii, 49, 85, 111, 112, 148

GSEA Gene Set Enrichment Analysis. 47

GWAS Genome wide association study. 6

HER2 Human epidermal growth factor receptor 2. 30, 33, 34, 36, 41, 55, 57, 73, 89, 92, 94, 102, 108, 173

ICGC International Cancer Genome Consortium. iv–vi, viii–x, xii, xiii, 32, 36–39, 42, 60, 63–65, 67, 68, 71, 77, 78, 81–86, 108, 109, 111, 112, 118, 119, 123–126, 131, 135, 138, 139, 141–143, 147, 148

IFNα Interferon-α. 30, 34, 55–57, 89, 92, 96, 97, 102, 157, 179–187

IFNγ Interferon-γ. 25, 30, 34, 55–57, 89, 92, 96, 97, 102, 157, 179–187

IGF Insulin-like growth factor. 23, 25

IGFBP Insulin-like growth factor binding protein. 25, 26

IL-10 Interleukin-10. 16

IL-6 Interleukin-6. 16

IRS Insulin receptor substrate. 4

JAK Janus kinase. 4, 5, 24, 25

KEGG Kyoto Encyclopedia of Genes and Genomes. xiii, 50, 83, 111, 112, 147

KIRP Kidney renal clear cell carcinoma. xiii, 37, 84, 85, 141, 147, 149

LIHC Liver hepatocellular carcinoma. xiii, 37, 84, 85, 142, 147, 149

LN Lymph node. 33, 36, 41, 73, 108

MAS5 Microarray analysis suite 5.0. ix, x, 38, 40, 42, 87–89, 155, 157, 158, 167

MET Mesenchymal-endothelial transition. 15

MM Mismatch. 38

MMR Mismatch repair. 19

mRNA Messenger RNA. 26

mTOR Mammalian target of rapamycin. 29, 95, 113

NER Nucleotide-excision repair. 19, 20

NF-κB Nuclear factor κB. 16, 21, 25

NGS Next generation sequencing. 27, 28, 118

NZ New Zealand. 3, 9, 10, 32, 33, 35

NZBC New Zealand Breast Cancer. v–xv, 32, 33, 35, 56–58, 65–68, 70, 77, 78, 80, 89, 92, 96–102, 104–106, 108, 110, 114–116, 118, 126, 130, 131, 134, 135, 151, 153, 167, 170, 172, 173, 176, 178–193, 196

ORA Over representation analysis. iv, 46–48

PEP Phosphoenol pyruvate. 17

PI3K Phosphatidylinositol 3-kinase. 4, 5, 22, 24, 25, 30, 34, 55, 57, 85, 89, 92, 102, 173

PM Perfect match. 38, 40

PR Progesterone receptor. vii, x, xiii–xv, 30, 33, 34, 36, 41, 55–57, 73, 88, 89, 92, 94, 96–99, 102–104, 108, 110, 111, 114–116, 120, 121, 151–154, 157, 167, 178–200

PRR Pattern recognition receptor. 25
RB Retinoblastoma-associated. 13
READ Rectum adenocarcinoma. xiii, 37, 84, 85, 142, 147, 149
RMA Robust multi array. viii–xii, 38, 40, 42, 53, 61, 63, 65, 67, 72–74, 76, 78, 87–94, 96, 98, 110, 122, 129, 151, 155–158, 167, 168, 170–172, 174, 176
RNA Ribonucleic acid. 85
SKCM Skin cutaneous melanoma. xiii, 37, 84, 85, 142, 147, 150
SNP Single nucleotide polymorphism. 26, 27
STAT Signal transducer and activator of transcription. 4, 5, 24
STAT3 Signal transducer and activator of transcription 3. 16, 25, 30, 34, 55, 57, 89, 92, 93, 102, 103, 173
SVD Singular value decomposition. iv, vi, vii, x, xiii, 50–52, 56, 58, 67, 76, 78, 91–93, 110, 111, 151, 173, 174
TCGA The Cancer Genome Atlas. 32, 36, 37
TGFβ Transforming growth factor-β. i, 25, 30, 34, 55, 57, 88, 89, 92, 94, 102–104, 115, 116, 121, 157, 167, 173
TLR Toll-like receptor. 21, 25
TNFα Tumour necrosis factor-α. 30, 34, 55, 57, 89, 92, 102, 103, 173
TSP-1 Thrombospondin-1. 14
UCEC Uterine corpus endometrial carcinoma. xiii, 37, 84, 85, 143, 147, 150
USA United States of America. 7
UV Ultraviolet. 20
VEGF Vascular endothelial growth factor. 14, 16
VEGF-A Vascular endothelial growth factor-A. 14
VIF Variance inflation factor. 48
WHO World Health Organization. xii, 2, 38, 39
WMW Wilcoxon-Mann-Whitney. 48

Chapter 1

Introduction

1.1 Obesity

Obesity has been a major global problem for more than a decade, associated with many noncommunicable diseases such as diabetes, cardiovascular diseases and certain types of cancers (World Health Organisation, 2014). In fact, the risk of comorbidities increases with an increase in body mass index (BMI), where the risk becomes severe as the BMI level approaches the obese category (World Health Organisation, 2000). The number of overweight and obese people, in both adults and children, has risen in every country of the world and the trend will likely to continue in the future. It is estimated to account for 3.4 million deaths per year and 93.6 million Disability-Adjusted Life Years (DALYs) in 2010, obesity is a serious disease that continues to grow in our society (Lim *et al.*, 2012).

1.1.1 Definition of obesity

Obesity is defined as an abnormal or excessive fat accumulation that may impair the health of an individual (Garrow, 1988). One common and widely used approach to categorise obesity is to measure the BMI of the individual. BMI is a measurement based on the weight-to-height ratio of an individual. It is often used by clinicians and epidemiologists to classify adults into underweight, normal weight, overweight, or obese categories. The unit of BMI is defined as the

weight in kilograms per square of the height in metres (kg/m^2).

The World Health Organization (WHO) (2014) categorise overweight and obese adults with a BMI of ≥ 25 kg/m^2 and ≥ 30 kg/m^2 respectively. For children under the age of 5, overweight and obesity is categorised as a weight-for-height ratio greater than 2 standard deviation and 3 standard deviation above the WHO Child Growth Standards median respectively. For children aged between 5 to 19 years, overweight and obesity is defined as BMI-for-age greater than 1 standard deviation and 2 standard deviation above the WHO Growth Reference Standards median respectively.

However, the use of BMI as an accurate measure of body fat composition and/or representation of the metabolic state of an individual remains controversial. In fact, many studies have suggested that the use of other measurements, such as waist-to-height and waist-to-hip ratios, to classify the patients into different categories, as these measurements were better predictors of the diseases related to obesity than BMI (Dalton *et al.*, 2003; Gelber *et al.*, 2008; Lee *et al.*, 2008). With that said, BMI is still widely used due to the ease of data collection compared to other measurements. As a result, most of the publicly available clinical data has height and weight data of the patients from which the BMI can be calculated.

1.1.2 Prevalence of obesity

From the latest global status report on noncommunicable diseases by the World Health Organisation (2014):

> *"In 2014, 39% of adults aged 18 years and older (38% of men and 40% of women) were overweight. The worldwide prevalence of obesity nearly doubled between 1980 and 2014. In 2014, 11% of men and 15% of women worldwide were obese. Thus, more than half a billion adults worldwide are classed as obese."*

Approximately 1 in every 14 people in the world were classed as obese in 2014; therefore, 1 in every 14 people around the world were also at severe health risks associated with obese people. The highest prevalence of overweight and obesity was in the American regions, where 61% and 27% of the population were overweight and obese respectively (World Health Organisation, 2014). It was also

noted in the report that women were more likely to be obese than men, and that the prevalence of overweight and obesity increased as the income level of the countries increased (World Health Organisation, 2014).

In addition to this, the worldwide prevalence of childhood obesity has been increasing steadily since 2000 (World Health Organisation, 2014). The prevalence of overweight children in those under 5 was estimated to rise from 6.3% in 2013 to 11% by 2025 if the current trend continued (World Health Organisation, 2014). The current prevalence of obesity in the younger population poses an even greater risk to the health of the population in the future than the current prevalence of obesity in the adult population.

In New Zealand (NZ), obesity is a serious problem as approximately one third (31.6%) of adults in NZ were classified as obese in 2015/2016 (New Zealand Ministry of Health, 2016a). The Pacific and Māori populations were affected the most, where 67% and 47% of the adults, respectively, were obese (New Zealand Ministry of Health, 2016a). NZ not only has a high prevalence of obesity compared to many other countries, but also has a unique health problem that affects Pacific and Māori populations disproportionately to the rest of the NZ population.

1.1.3 Physiological mechanism of obesity

Obesity is caused by continuous intake of food without expending all of the energy gained from the food. The key to the problem is the control of food intake and energy expenditure, and how they are regulated in the body. The role of the hypothalamus is central to the control of satiety and many of the mutations that cause obesity affects the satiety signals to and/or from hypothalamus. The hypothalamus receives both neuronal and hormonal inputs from the body and propagates the signal to the downstream neurons which ultimately affects the food intake of the individual (Bell *et al.*, 2005; Spiegelman and Flier, 2001).

There are two types of neurons in the arcuate nucleus (located in the hypothalamus): orexogenic and anorexogenic neurons. Orexogenic neurons are responsible for the promotion of food intake and reducing energy expenditure, whereas the anorexogenic neurons have the opposite effect (Barsh and Schwartz, 2002). When stimulated by, for example, endocrine signals, these two different types of

neurons act on the downstream neurons to promote or reduce food intake and energy expenditure.

Leptin-melanocortin pathway is the main pathway in which the level of satiety is controlled and regulated (Spiegelman and Flier, 2001). In this pathway, both the orexogenic and anorexogenic neurons (located in the arcuate nucleus of the hypothalamus) have crucial roles in controlling the satiety of the individual, and therefore their eating behaviour (Barsh and Schwartz, 2002; Bell *et al.*, 2005). When a satiety-controlling hormone such as leptin reaches the arcuate nucleus region of the hypothalamus, it can stimulate or inhibit orexogenic, anorexogenic, or both orexogenic and anorexogenic neurons. Depending on which neurons are stimulated or inhibited, the hormone will ultimately alter the satiety level of the individual.

Though there are many hormones that regulate satiety, only leptin and insulin will be covered here. Leptin is a hormone that signals satiety and reduces food intake, and is secreted mainly by the white adipose tissue (El-Sayed Moustafa and Froguel, 2013; Zhang *et al.*, 1994). The circulating concentration of leptin in the body is proportional to the body fat stores (Barsh and Schwartz, 2002; El-Sayed Moustafa and Froguel, 2013). Leptin signals through the janus kinase (JAK)/signal transducer and activator of transcription (STAT) pathway via the long form of the leptin receptor that is present in high expression in the hypothalamus (Ghilardi *et al.*, 1996; Lee *et al.*, 1996). Depending on the type of neurons the leptin receptor is expressed on, leptin stimulates or inhibits these neurons and ultimately produces an anorexogenic signal to the downstream effector neurons (Bell *et al.*, 2005).

Insulin, perhaps the most well-known hormone for its role in glucose homeostasis, also has an important role in satiety and food intake regulation. In brief terms, when insulin binds to the insulin receptor, insulin receptor substrate (IRS) protein family is recruited and phosphorylated by the intrinsic tyrosine kinase activity of the insulin receptor (Saltiel and Pessin, 2002). The phosphorylated IRS protein then activates phosphatidylinositol 3-kinase (PI3K), which in turn activates its downstream targets, causing a rapid signal cascade through the cell (Saltiel and Pessin, 2002).

The exact role of insulin in the satiety control was not particularly clear even

though there was evidence of increased food intake with neuron-specific deletion of insulin receptor (Barsh and Schwartz, 2002; Brüning *et al.*, 2000). The fact that the insulin level, like leptin, was proportional to the fat mass also suggested the role of insulin in satiety and body weight control (Barsh and Schwartz, 2002; Brüning *et al.*, 2000; Woods *et al.*, 1979). This was perhaps due to the fact that leptin and insulin used two different signalling pathways in the cell: leptin used the JAK/STAT signalling pathway, whereas insulin used the PI3K pathway for signalling (Ghilardi *et al.*, 1996). However, leptin, and subsequently insulin, was shown to activate ATP-dependent K^+ channel in glucose-responsive neurons in the hypothalamus (Spanswick *et al.*, 1997, 2000). The activation of the K^+ channel provided evidence for the two pathways to act similarly on the cell, at least for an acute response. This suggests a convergence in the pathway response by the two hormones, and therefore their role in satiety control.

Satiety control is crucial in order to manage and maintain the body weight in a healthy state. As described above, there are many components that regulate satiety. Disruption of any part of the satiety controlling pathway can encourage the development of obesity.

1.1.4 Factors that affect the prevalence of obesity

Individual-level (micro-level) factors

There is much evidence of obesity being caused by individual-level genetic factors that make some individuals more predisposed to obesity than others. Generally speaking, there are two classes of genetically-driven obesity: monogenic obesity and polygenic obesity (El-Sayed Moustafa and Froguel, 2013).

Monogenic obesity is where a single mutation in one of the genes that is directly involved in the satiety controlling pathway causes severe obesity in that individual. Since these mutations are usually directly related to the key constituents in the pathway mentioned earlier (Section 1.1.3), the resulting phenotype from the mutation causes severe early-onset obesity. However, although these mutations cause a devastating effect on those who carry the mutation, these mutations occur very rarely in the population.

An example of a mutation that causes severe early-onset obesity is a mutation

in the leptin gene. Montague *et al.* (1997) found a single nucleotide deletion mutation at position 133 of the leptin gene in children with early-onset obesity. They further showed that the mutation caused a frameshift mutation in the protein, resulting in an inactive form of leptin that was not able to be secreted (Montague *et al.*, 1997). As a result, children with this mutation were constantly hungry due to the lack of anorexogenic signals and became obese at a very early stage of their lives.

Polygenic obesity is where a combination of variants in multiple genes or multiple regions of the genome (together with the effect from the environmental factors) causes obesity in the individual (El-Sayed Moustafa and Froguel, 2013). Unlike in monogenic obesity, where a single mutation in a gene causes severe obesity, the variants in some of the genes or genomic regions involved in polygenic obesity have little effect on the obesity status of the individual by itself. However, in combination with many other variants and together with environmental factors, it has a serious obesogenic effect on the individual.

There have been many studies that have investigated the association of genes and/or genomic regions with obesity. Of those, many of the genome wide association studies (GWAS) published in 2007 showed strong evidence of the fat mass and obesity associated (FTO) gene with BMI and obesity (Dina *et al.*, 2007; Frayling *et al.*, 2007; Gerken *et al.*, 2007; Scuteri *et al.*, 2007). Until these studies came out, there has been no common genes or genomic regions found to be associated with obesity that was reproducible and reliable (Frayling *et al.*, 2007). Thus, these studies were the first evidence of common variants that contributed to the polygenic obesity phenotype in the population.

These individual-level genetic factors contribute to the overall increase in the proportion of the population that are obese. However, it is likely that the environmental factors are the most significant contributor to the current global obesity epidemic (Malik *et al.*, 2013).

Population-level (macro-level) factors

It is clear that the individual-level factors cannot be accounted for the global epidemic of obesity that the world currently observes. For this reason, obesity is

thought to be caused mainly by the environmental factors that act on a population-level. The first key factor that contributes to the global rise in the proportion of the population that is obese is the availability and abundance of certain foods in the country. With the help from government free trade policies, food and other goods have become easier to trade between countries, helping the economic growth of many countries around the world (Kearney, 2010). This has led to a greater abundance of food in general, a greater number of food choices and a change in the overall nutritional status of countries (Malik *et al.*, 2013). This shift in food choice and availability (and the economic benefit associated with the trade) had a positive impact on countries, but the growth in the food market also resulted in the population-wide rise in obesity.

As an example, in the United States of America (USA), the cost of corn and soy are low due to the fact that they are the raw ingredients for most processed food and beverages, such as high fructose corn syrups used to sweeten many soft drinks in the USA (Malik *et al.*, 2013). Moreover, corn and soy are also the main food for livestock, which results in lower price for meat (Malik *et al.*, 2013). On the other hand, the prices of fruits and vegetables remain expensive due to the lack of support by the government to lower the cost associated with the production (Malik *et al.*, 2013). As a result, more people are inclined to consume food that are cheap, have little nutritional value and high in energy, than food that are expensive, nutritional and healthy alternatives. Thus, the population is more likely to become obese by making poor food choice.

The second factor to consider is the behavioural changes led by urbanisation. As more people take up urban lifestyle, they are exposed to new environments with a better range of food selection, as well as technological and mechanical advancement in transport and other daily chores that may not be available in rural areas. Although urbanisation has many advantages, such as access to developed health care systems, education and advanced technologies, the lifestyle change also imposes negative health behaviour that ultimately can lead to positive energy balance and obesity (Malik *et al.*, 2013). Lack of physical activity is one of the clear components linked with urbanisation. In many urban cities, it is often difficult to achieve and maintain the recommended level of physical exercise due to the shift towards sedentary behaviours that are encouraged by developed trans-

port systems, automated household chores and indoor entertainment (Malik *et al.*, 2013).

The last factor to be mentioned here is the income and socioeconomic status of the population. As the average income increases, more people are able to buy food and are likely to take up a sedentary lifestyle. From this, one would expect the high-income group to have the largest impact from obesity, but on the contrary, the biggest impact from obesity is seen in the low- and middle-income groups (Malik *et al.*, 2013). Access to a developed health care system will allow weight maintenance, but this is likely to be of most benefit to high-income groups. The reason for this is the cost associated with a visit to the health care system, as high-income groups are more likely to make a consistent and periodic visit than a low- or middle-income groups (Malik *et al.*, 2013).

Therefore, for the low- and middle-income groups, the rise in the average income in these groups allows people to have access to food and activities that promotes sedentary behaviours, but the limited access to the health care system, education, and recreational activities (that promote physical activity) ultimately leads to positive energy balance (Malik *et al.*, 2013). In contrast, the high-income group of the population have better access to facilities that promote weight maintenance, thus have less of an impact compared with the low- and middle-income groups of the population. All of these population-level environmental factors in synergy with the individual-level genetic factors have contributed to the global epidemic of obesity.

1.2 Cancer

> *"Cancer is a generic term for a large group of diseases that can affect any part of the body ... One defining feature of cancer is the rapid creation of abnormal cells that grow beyond their usual boundaries, and which can then invade adjoining parts of the body and spread to other organs ... referred to as metastasizing. Metastases are the major cause of death from cancer." (World Health Organisation, 2016)*

1.2.1 Prevalence and mortality of cancer

Cancer is considered as one of the leading causes of deaths around the world, accounting for 8.2 million deaths in 2012 (World Health Organisation, 2014). More than half of cancer deaths are caused by lung, breast, colorectal, stomach and liver cancers (World Health Organisation, 2014). The leading cause of cancer deaths in high-income countries is lung cancer for both men and women, then colorectal cancer and breast cancer for men and women, respectively (World Health Organisation, 2014). Although the levels of mortality vary among different countries, cervical, liver and stomach cancers make up a large proportion of cancer deaths among low- and middle-income countries (World Health Organisation, 2014).

Bray *et al.* (2013) estimated the global prevalence of cancer in the adult population in 2008. Since there was no clear definition for cancer prevalence, Bray *et al.* (2013) used limited duration prevalence as an estimate for cancer prevalence, and measured the incidence of cancer during the period between 2004 and 2008 (5 years). From these estimates, Eastern Asia had the greatest 5-year prevalence of cancer, with approximately 7 million people with cancer in 2008, followed by North America and Western Europe with approximately 4.7 and 3 million people with cancer (Bray *et al.*, 2013). When the population of these regions was taken into account and the proportion of cancer prevalence calculated, Western Europe had the highest prevalence proportion of 1.9%, followed closely by Australia/New Zealand with 1.82% and North America with 1.7% (Bray *et al.*, 2013).

Looking at the prevalence of cancer by specific types, breast cancer was the most prevalent type of cancer with 5.2 million cases in 2008 (Bray *et al.*, 2013). Colorectal and prostate cancers were the second and third most prevalent cancer types, with approximately 3.2 million cases (Bray *et al.*, 2013). Bray *et al.* (2013) noted that female-specific cancer types were the most prevalent types of cancer in 75% of countries in the world. Furthermore, the prevalence of breast, cervix or prostate cancer is the highest in almost 95% of countries (Bray *et al.*, 2013). Although these estimates provide a relatively crude measure of cancer prevalence, Bray *et al.* (2013) provided a useful overview of the prevalence of cancer around the world.

In NZ, the most common cancer types diagnosed for males are prostate, col-

orectal, melanoma, lung and non-Hodgkin lymphoma, which all together accounted for 65% of newly registered cancers in males in 2013 (New Zealand Ministry of Health, 2016b). For females, breast (which accounted for 28.3% of all new female cancer cases), colorectal, melanoma, lung and uterine cancers were the most common cancer types diagnosed in 2013 (New Zealand Ministry of Health, 2016b). The cancer registration rates were higher in females aged 25-54 years than males of the same age group, but the registration rates were significantly higher in males for people who were 60 years or older (New Zealand Ministry of Health, 2016b). Like with obesity, there is a significant difference in the cancer registration rates between Māori and non-Māori. The rates were 1.2 and 1.4 times greater in the Māori population compared to the non-Māori population for males and females, respectively (New Zealand Ministry of Health, 2016b). Again, this difference between Māori and non-Māori is unique to NZ, and improved understanding of the underlying problem is essential for better treatment plans and outcomes for Māori.

The total number of deaths by non-communicable diseases (including cancer) is predicted to increase, reaching 52 million deaths by 2030 (World Health Organisation, 2014). With the prevalence of cancer at such high levels around the world, and with cancer mortality increasing in the future, cancer is a serious challenge that the world has to face.

1.2.2 Mechanisms of cancer development

It is now commonly understood that cancers are caused by dysregulation of various molecular mechanisms or pathways that allow tumour cells to proliferate, survive and migrate. By taking advantage of these pathways, cancer cells are able to grow and eventually metastasise to other parts of the body. The environment in which the tumours arise can also be an important factor for their evolution and growth.

Hallmarks of Cancer

Extensive research in cancer biology has uncovered many pathways and regulatory mechanisms involved in cancer. With more than 100 distinct types and sub-

types of cancers (each with its own distinct development mechanisms), it became difficult to focus on the overarching mechanisms of cancer development (Hanahan and Weinberg, 2000). To resolve this, Hanahan and Weinberg (2000) proposed "six essential alterations" that cancer cells usually take advantage of during development and progression. These six key mechanisms were coined as being "The Hallmarks of Cancer", and have been used extensively by many researchers as their guide to target their research in specific areas of cancer biology.

Ten years after the first review where the concept of Hallmarks of Cancer was proposed, Hanahan and Weinberg revisited these hallmarks and added four more on top of the six hallmarks presented in 2000 (Hanahan and Weinberg, 2011). The ten Hallmarks of Cancer are: sustaining proliferative signalling, evading growth suppressors, enabling replicative immortality, activating invasion and metastasis, inducing angiogenesis, resisting cell death, avoiding immune destruction, tumour-promoting inflammation, genome instability and mutation, and deregulating cellular energetics (Figure 1). Each of these hallmarks will be described briefly to provide an overview of the significance of these mechanisms in tumour biology.

Sustaining proliferative signalling

Cancer is a disease of uncontrolled cell proliferation and division. Normal cells cannot proliferate without any external mitogenic growth signals, but cancer cells are able to deregulate these signals to maintain a steady proliferative signal (Hanahan and Weinberg, 2011).

There are many ways in which cancer cells can maintain this signal. Firstly, cancer cells can acquire the ability to produce growth factors and express the corresponding receptor, creating a positive feedback mechanism that continuously stimulates the cells to proliferate (Hanahan and Weinberg, 2000). Secondly, they can express more receptors for a given growth signal, and thus become "hyper-responsive" to the signal (Hanahan and Weinberg, 2000, 2011). Thirdly, cancer cells can stimulate the neighbouring normal cells within the tumour microenvironment to produce more growth factors that further assist growth (Bhowmick *et al.*, 2004; Liotta and Kohn, 2001; Wiseman and Werb, 2002). Finally, cancer cells are able to maintain proliferative signalling by mutating the growth factor receptor

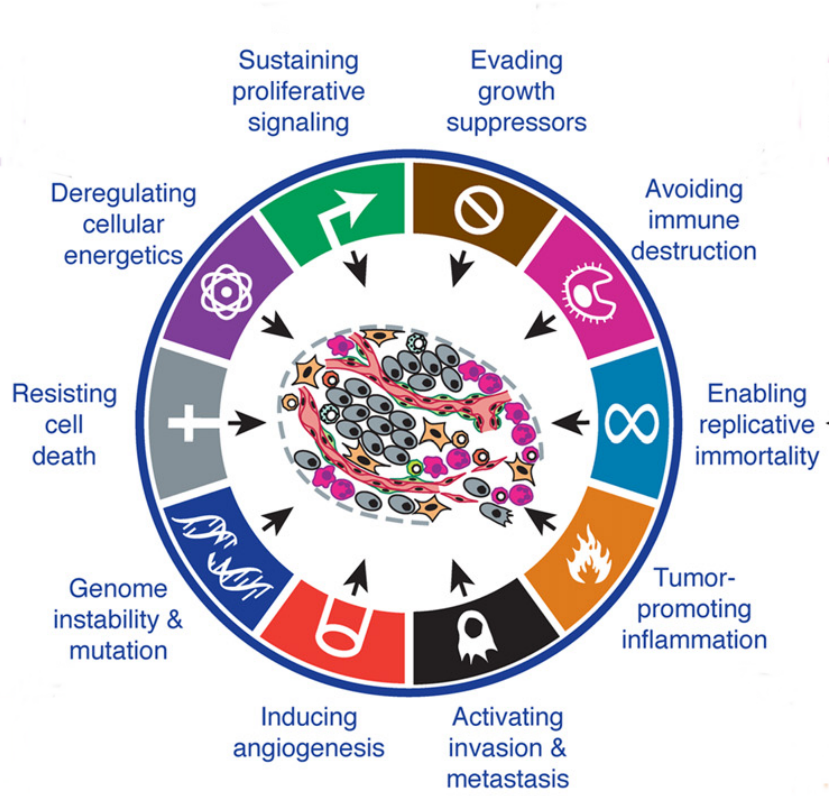


Figure 1: Hallmarks of Cancer. The ten Hallmarks of Cancer as proposed by Hanahan and Weinberg in 2011. (Figure adapted from Hanahan and Weinberg (2011))

itself, or the downstream signalling targets of the growth factor receptor pathway (Fuqua *et al.*, 1991; Su Huang *et al.*, 1997; Satyamoorthy *et al.*, 2003). By maintaining the proliferative signals, cancer cells are able to continuously grow in their environment.

Evading growth suppressors

As well as maintaining proliferative signals, cancer cells must also evade signals that inhibit cell growth. The p53 protein is a tumour suppressor protein that has a role in cell cycle arrest (Hanahan and Weinberg, 2011; Levine, 1997). As Hanahan and Weinberg (2011) mention in their review, "TP53 receives inputs from stress and abnormality sensors that function within the cell's intracellular operating systems". The level of expression and its stability is crucial for the function of

p53 protein, as the abundance of p53 protein in the cell triggers cell cycle arrest or cell apoptosis, depending on the extent of the cell signal (Fridman and Lowe, 2003; Hanahan and Weinberg, 2011; Levine, 1997). When abundant and active, p53 protein transcribes the p21 protein which, through series of events in the p16-cyclin D_1-cdk4-retinoblastoma-associated (RB) pathway, triggers cell cycle arrest (Levine, 1997). In many cancers, the p53 protein is mutated and selected against to allow continuous growth of the tumour cells.

Resisting cell death

Apoptosis, the natural and controlled death of a cell and is one of many defence mechanisms in the body to prevent the development of cancer. The triggering of apoptosis is controlled and maintained by the balance between the pro- and anti-apoptotic signals that affect the downstream apoptotic proteins (Hanahan and Weinberg, 2011).

The precise cellular conditions that trigger apoptosis are yet to be uncovered, but there is a clear evidence of the p53 tumour suppressor playing a key role in apoptosis initiation. In short, when DNA damage is sensed, the p53 protein transcribes the pro-apoptotic members of the B-cell lymphoma 2 (Bcl-2) family proteins, such as Bax, Puma and Noxa (Fridman and Lowe, 2003; Hanahan and Weinberg, 2011). The fact that the p53 protein is involved not only in the regulation of the cell cycle, but also in the apoptotic pathway shows its importance in the control of cancer development, and perhaps that is the reason why more than 50% of tumours have a mutation in the *TP53* gene (Levine, 1997).

Enabling replicative immortality

When acquired by cancer cells, the previous three hallmarks allow the cells to continuously grow in an uncontrolled manner. However, this is not the case, as the disruption in cell signalling on its own does not trigger the rapid expansion of the tumour; this occurs only when the cells have the capability of replicating infinitely (Hanahan and Weinberg, 2000, 2011).

There are two barriers in achieving infinite replicative capacity: senescence and crisis (Hanahan and Weinberg, 2011). When cells are grown in a culture, the

cells are able to replicate and grow until a certain point, and the cells reach the stage of senescence where the cells stop growing but are still viable (Hanahan and Weinberg, 2011). The cells that manage to bypass this senescence phase by mutating the tumour suppressor proteins such as the p53 protein, undergoes the crisis phase where the majority of the cells in the culture dies (Hanahan and Weinberg, 2011). Very rarely, some cells are able to survive the crisis phase and acquire the ability to replicate infinitely, and this transition has been termed "immortalisation" (Hanahan and Weinberg, 2011; Wright *et al.*, 1989). Without immortalisation, cancer cells are not able to generate macroscopic tumours that are capable of killing the host organism (Hanahan and Weinberg, 2000, 2011).

Inducing angiogenesis

Like normal cells, cancer cells also require access to oxygen and nutrient supplies from the blood. However, new blood vessel formation, or angiogenesis, only occurs in a handful of occasions, and thus the tumour needs to acquire the ability to stimulate angiogenesis in the microenvironment in which they live (Hanahan and Weinberg, 2011).

It is thought that the angiogenic ability of the tumour is gained at an early stage of tumour development by disrupting the "angiogenic switch" (Hanahan and Weinberg, 2011). Vascular endothelial growth factor-A (VEGF-A) is the major growth factor that triggers angiogenesis by signalling through the VEGF receptors (though some receptors have an inhibitory effect) (Yancopoulos *et al.*, 2000). Opposing the effect of VEGF-A is the thrombospondin-1 (TSP-1), a protein that is known to inhibit the process of angiogenesis by affecting the bioavailability of VEGF, as well as promoting apoptosis (Kazerounian *et al.*, 2008). In general, many cancers stimulate the expression of VEGF-A protein and reduces the expression of TSP-1 (Kazerounian *et al.*, 2008).

Activating invasion and metastasis

One of the prime reasons that cancers are so difficult to treat and have a high mortality rate is due to their ability to metastasise to other parts of the body. Once metastasised to different parts of the body, the clinician not only has to consider

the primary tumour site, but also the second, third, or even fourth tumour sites. These tumours may or may not have similar tumour biology as the primary tumour, and thus has to consider multiple treatment plans to cure the patient.

Metastasis is a multi-step process that requires the selection and survival of cancer cell(s) that is able to metastasise ("seed"), and is also compatible with the specific target tissue where they prosper ("soil") (Talmadge and Fidler, 2010). Once mobile and invasive, the metastatic cancer cell is able to circulate in the vascular system and eventually invade the target organ, where it must undergo mesenchymal-endothelial transition (MET) and colonise the site (Hanahan and Weinberg, 2011; Kalluri and Weinberg, 2009). There are numerous signals, growth factors and pathways that initiate the metastatic phenotype in the cancer cell, and it remains an active field of research (Hanahan and Weinberg, 2011; Kalluri and Weinberg, 2009).

Genome instability and mutation

Many of the hallmarks are acquired by the cancer cells during tumour growth, through many genomic changes and mutations of the essential genes that allow cancer cells to survive, proliferate, invade and metastasise. Thus, it is only logical to induce that many, if not all, cancers are more prone to genetic mutations and restructuring, which allows the cancers to subsequently acquire the already described Hallmarks of Cancer (Hanahan and Weinberg, 2011).

Until recently, the evolutionary progress of cancers have thought to have been gradual and stepwise, where each driver mutation is acquired over many years, or even decades (Stephens *et al.*, 2011). There is now evidence that shows otherwise, whereby a single catastrophic event (termed chromothripsis) occurs in the chromosome structure, resulting in major structural rearrangements such as inversions, deletions, duplications and translocations (Leibowitz *et al.*, 2015; Stephens *et al.*, 2011). As it is a one-off chromosomal disruption, chromothripsis (usually) cause disruptions in tumour suppressor genes rather than duplications of oncogenes (Leibowitz *et al.*, 2015).

In addition to this, areas of the chromosome where major structural rearrangements have occurred can be prone to a regional hypermutation, termed kataegis

(Leibowitz *et al.*, 2015; Nik-Zainal *et al.*, 2012). Although the full mechanism in which kataegis occurs is not yet defined, activation-induced cytidine deaminase (AID)/APOBEC family proteins, a family of proteins involved in gene editing, may play a major role in kataegis (Leibowitz *et al.*, 2015; Nik-Zainal *et al.*, 2012). It is evident that both chromothripsis and kataegis contribute to the genomic aberrations and mutations that ultimately encourage the acquisition of many of the tumorigenic phenotypes.

Tumour-promoting inflammation

While inflammation and immune response may assist in the control of diseases and injuries, they have been implicated in the progression of tumour and their acquisition of many of the cancer hallmarks (Hanahan and Weinberg, 2011). There are two key transcription factors of inflammation that ultimately leads to the promotion of many, if not all, of the hallmarks discussed so far: nuclear factor κB (NF-κB) and signal transducer and activator of transcription 3 (STAT3) (Mantovani *et al.*, 2008).

The study by Pikarsky *et al.* (2004) showed that NF-κB was an essential component for the progression of hepatocellular carcinoma, providing the first evidence as to how inflammation may affect the tumour biology. NF-κB regulates the expression of genes that encode the inflammatory cytokines, angiogenic factors, and many others, including anti-apoptotic genes (through the activation of STAT3) (Elinav *et al.*, 2013; Mantovani *et al.*, 2008). STAT3 is a transcription factor that is activated through a variety of receptors and pathways (Yu *et al.*, 2007, 2014). By increasing the transcription of anti-apoptotic proteins such as Bcl-2, STAT3 promotes tumour survival and proliferation (Yu *et al.*, 2007). Furthermore, STAT3 is able to increase the expression of inflammatory cytokines (such as interleukin-6 (IL-6) and IL-10) and VEGF (Yu *et al.*, 2007). These cytokines provide a positive feedback mechanism for further inflammation (and therefore survival and proliferation signals) and advancement of angiogenesis, respectively (Yu *et al.*, 2007).

Deregulating cellular energetics

In order to grow and proliferate, cancer cells must adjust their energy metabolism to match their energy requirement (Hanahan and Weinberg, 2011). Wardburg (1956) found that even if cancer cells have an abundance of oxygen in the cell, they utilised the glycolytic pathway instead of the more efficient aerobic respiration, and was later termed "aerobic glycolysis" (Hanahan and Weinberg, 2011). It seems counterintuitive to limit its full potential of creating adenosine triphosphate (ATP) by utilising glycolysis over electron transport chain (ETC), as the latter produces almost 18 times more ATP (Hanahan and Weinberg, 2011; Vander Heiden *et al.*, 2009). However, it is hypothesised that this inefficiency helps the cell to generate and direct the molecules required for biosynthetic pathways during cell growth and division (Cairns *et al.*, 2011; Vander Heiden *et al.*, 2009). By deliberately relying purely on glycolysis (known as the "Wardberg effect") and slowing the conversion of phosphoenol pyruvate (PEP) to pyruvate, the cell is able to direct the carbon molecules into reaction pathways (Cairns *et al.*, 2011; Vander Heiden *et al.*, 2009). These produce the constituents used to synthesise macromolecules, such as DNA and lipids (Cairns *et al.*, 2011; Vander Heiden *et al.*, 2009).

The fact that the deregulation of energy metabolism is mediated somewhat by the genes involved in many of the other hallmarks (such as p53 and Myc) suggests that the disruption in energy metabolism may be a phenotype caused in conjunction with those core hallmarks (Hanahan and Weinberg, 2011).

Avoiding immune destruction

Our immune system is constantly monitoring and maintaining normal cell biology within our body, preventing the emergence of cancer (Hanahan and Weinberg, 2011). With that said, cancers are frequently observed in clinics. This highlights how the cancer cells can evade the diligent monitoring by the immune system (Hanahan and Weinberg, 2011).

Immunoediting is a process where the immune system successfully eliminates the highly immunogenic cancer cells during surveillance, but fails to eliminate the weakly immunogenic cells and therefore poses a selective advantage on these cells (Hanahan and Weinberg, 2011; Teng *et al.*, 2008). In fact, there have been

reports where the cancer cells in the immunodeficient mice were not able to initiate tumour growth in the immunocompetent mice, but the cancer cells from the immunocompetent mice were able to do so in the immunodeficient mice (Hanahan and Weinberg, 2011). One possible reason for this was that the cancer cells from the immunodeficient mice were not immunoedited to the extent in which the cells from immunocompetent mice were, and thus unable to evade the immune system as efficiently as those cells in the immunocompetent mice (Hanahan and Weinberg, 2011). Though more experiments must be carried out to confirm its definitive role and mechanisms in which immunoediting contributes to cancer biology, immune system evasion is no doubt a significant factor that contributes to tumour progression.

By utilising many, if not all, of these major cancer hallmarks, cancer cells are able to initiate growth, proliferate, invade, metastasise and eventually kill the host if untreated. There is an evident need for an efficient way to monitor and treat cancer patients, and these Hallmarks of Cancers will prove to be the best target for effective treatment of cancer in the future.

1.2.3 Causes and risk factors of cancer

There are many causes and risk factors associated with the development and progression of cancer. Like obesity, cancers are dependent on both genetic and environmental factors.

Genetic mutations

Any one of the ten Hallmarks of Cancers has the potential to initiate cancer growth and progression, and every hallmark has their own complex mechanistic pathways. Furthermore, every pathway has many proteins associated with its function or role in the cell and, though their impacts may differ from protein to protein, mutations in any one of the proteins involved in a pathway have the potential to cause cancer.

Oncogenes are genes that promote or "accelerate" the process of tumorigenesis and/or tumour progression, whereas tumour suppressor genes prevent or

"decelerate" those effects (Vogelstein and Kinzler, 2004). Cancers usually mutate oncogenes in such a way that they become constitutively activated, and mutate the tumour suppressor genes so that these genes are inactivated and/or have reduced expression (Vogelstein and Kinzler, 2004). Another thing to note for these genes is that the oncogenes usually only require a "one-hit" mutation that affects a single allele, whereas the tumour suppressor genes generally require mutations in both maternal and paternal alleles to have any effect (Stratton *et al.*, 2009; Vogelstein and Kinzler, 2004).

Stability genes, also known as caretakers, are genes that are involved in, for example, mismatch repair (MMR), nucleotide-excision repair (NER) and base-excision repair (BER); in other words, genes that are involved in DNA repair mechanisms (Vogelstein and Kinzler, 2004). These caretaker genes help prevent the cell from developing new mutations during DNA replication and mitosis. However, only the oncogenes or the tumour suppressor genes are able to directly affect the biology of the cell, and therefore mutations in stability genes tend to affect the mutation rates of the other classes of genes, ultimately leading to tumorigenesis (Vogelstein and Kinzler, 2004).

There are two types of mutations that occur in an individual: germline mutation and somatic mutation. Germline mutations are the mutations that are passed on from the parents of an individual, and somatic mutations are the mutations acquired by the cell over the course of the lifetime of an individual.

Unfortunately, there are germline mutations that make the individual more predisposed to cancer (Vogelstein and Kinzler, 2004). These individuals have a mutation in one or more of the classes of cancer-related genes, which makes them more likely to develop a tumour than individuals without the mutation (Vogelstein and Kinzler, 2004). In contrast, somatic mutations are acquired during the lifetime of an individual, and is not passed down to their children. Somatic mutations can either be a driver or a passanger mutation, where a driver mutation helps tumorigenesis and its subsequent progression (in other words, mutations in oncogene, tumour suppressor, or stability gene), and passanger mutation have neither a selective advantage nor contribute to tumour progression (Stratton *et al.*, 2009).

Since cancers require multiple driver mutations, somatic mutations are an essential part of tumour development. Perhaps that is one of the reasons why,

in some cancers, germline mutation in the stability genes (such as *BRCA1* and *BRCA2* genes) cause greater susceptibility to cancers than mutations in oncogenes or tumour suppressor genes (Vogelstein and Kinzler, 2004).

Environmental factors

Cancers are developed through many cellular pathways (Hallmarks of Cancers) that are not fully uncovered yet, with either too many or very few causes associated with those pathways. Therefore it is difficult to assess how, rather than whether, a given environmental factor causes a specific cancer. With that said, several environmental factors have a strong evidence of causing cancer.

Smoking is a known risk factor for many cancers; the main one being lung cancer (Gandini *et al.*, 2008; Hecht, 1999). For current smokers, the relative risk is greater than 1.5 in many cancers, including liver, cervix and kidney, and it can rise up to 8.96 in lung cancer (Gandini *et al.*, 2008). Tobacco smoke contains many carcinogens which are processed first by cytochrome P450 (CYP) enzymes (termed metabolic activation) (Hecht, 1999). The resulting product from this reaction can either be processed further to be detoxified completely and cause no harm to the cell, or alternatively, the product can covalently bind to the DNA, forming DNA adducts (Hecht, 1999). Unless these adducts are resolved correctly by the DNA repair mechanisms, it can result in a permanent mutation within the genome and lead to oncogenesis (Hecht, 1999).

Exposure to ultraviolet (UV) radiation has been significantly associated with an increased risks of skin cancers (Armstrong and Kricker, 2001; Gallagher and Lee, 2006). Similar to smoking, the mechanism of how UV radiation may cause skin cancers is related to the DNA repair system. UV radiation causes the formation of a dimer in the adjacent pyrimidines, resulting in a covalent bond formation between these nucleotides (Friedberg, 2003; Hoeijmakers, 2001). These dimers introduce bulky structures in the DNA that prevent the DNA replication and transcription enzymes from functioning properly, and it may cause lasting mutations if not repaired (Friedberg, 2003; Hoeijmakers, 2001). The damage caused by UV radiation can be relieved by the NER system or the photoreactivation process, where a photoreactivating enzyme uses light to monomerize the dimeric pyrim-

idines (Friedberg, 2003).

Diet and obesity are also known risk factors for many types of cancers (Ames *et al.*, 1995; Calle and Kaaks, 2004). Epidemiological evidence of obesity being a risk factor for many cancers has been well established, yet the mechanisms of its contribution to cancer is still speculative (Calle *et al.*, 2003; Kelesidis *et al.*, 2006). Although there are a lot of possible mechanisms, there is growing evidence of the role of chronic inflammation caused by increased fat mass in the body (Kelesidis *et al.*, 2006; Lumeng and Saltiel, 2011; Rodríguez-Hernández *et al.*, 2013). Free fatty acids from excess adipose tissue and/or diet can be recognised by Toll-like receptor (TLR) (specifically TLR4), which can then activate the NF-κB pathway, leading to the activation of one of the Hallmarks of Cancer (tumour-promoting inflammation) (Lumeng and Saltiel, 2011). There are many other environmental factors associated with cancers, thus limiting the exposure to detrimental environmental factors is crucial for the prevention of many cancers.

1.2.4 Treatment of cancer

Traditionally, the treatment of cancer has been surgery, radiation therapy and chemotherapy. Though surgery is an effective method to remove the tumour from the patient, there is always risk of metastasis and recurrence. This is why a combination of surgery with other treatments like chemotherapy is usually considered. Previously, chemotherapeutic drugs and radiation therapy had a limited effect on tumours due to their resistance against these treatments (Wilhelm *et al.*, 2006). However, a better understanding of the pathways and mechanisms used by the tumour cells have shed light on the development of drugs that are more effective against tumours. RAF inhibitors are among one of the many chemotherapeutic drugs.

The RAS-RAF-MEK-ERK signalling pathway is involved in cell survival, growth and proliferation (Samatar and Poulikakos, 2014; Wilhelm *et al.*, 2006). The components of this pathway are often mutated in many types of cancer (approximately a third of cancers contain a mutation in the RAS oncogene), causing an overactivation of this pathway (Samatar and Poulikakos, 2014). Development of drugs that target the downstream RAF and its mutant forms (such

as $BRAF^{V600E}$) have been successful, and many RAF inhibitor drugs such as sorafenib and vemurafenib have been approved for clinical use (Samatar and Poulikakos, 2014; Wilhelm *et al.*, 2006). Even though the signalling pathways are well understood and the quality of the drugs that are developed based on this knowledge is improving, there is always the possibility of the tumour to acquire resistance to those drugs (Samatar and Poulikakos, 2014).

In fact, there has been evidence of pathway cross-talks between the RAS-ERK pathway and the PI3K/Akt pathway, which may help the tumour cells achieve resistance to the drugs that block the RAS-ERK pathway (Moelling *et al.*, 2002; Zimmermann and Moelling, 1999). A study in a *Drosophila* model showed that the inhibition of just a single pathway does not necessarily stop the effect of the pathway, as the signal can flow through other pathways (Dar *et al.*, 2012). Pathway cross-talk and increased flux through alternative pathways could be one of the ways in which tumour cells acquire resistance to drugs. The effectiveness of targeting multiple pathways in humans is yet to be confirmed and further investigations are required to solidify the mode of action and its relevance to tumour biology.

Further understanding of the tumour biology, its mechanism of resistance to drugs, and the effect of targeting multiple pathways (or targeting effective pathways for a patient) will become important in the near future, both for novel drug development and for better clinical treatment plans for the patients.

1.3 Obesity and cancer

As mentioned briefly in Section 1.2.3, obesity is considered a major risk factor for many types of cancer. The prevalence of obesity is extremely high around the world and it is important to establish the link between the risks associated with cancer.

1.3.1 Cancer risks associated with obesity

Although the mechanisms through which obesity causes cancer are still under debate, the association between obesity and cancer has long been established.

Perhaps the most convincing evidence was first presented by Calle *et al.* (2003). In this study, the link between BMI and variety of cancers was investigated in 900,000 adults between 1982 and 1998 (Calle *et al.*, 2003). The results from this study showed that morbidly obese men and women (BMI > 40) had a 52% and 62% higher chance of dying from all cancers compared to normal weight adults (18.5 < BMI < 25) (Calle *et al.*, 2003). BMI was also associated with higher rates of death from esophageal, colorectal, liver, gallbladder, pancreas, kidney, and some sex-specific cancers (Calle *et al.*, 2003). Furthermore, the death rates had a positive linear correlation with increasing BMI in all cancers (Calle *et al.*, 2003).

Renehan *et al.* (2008) showed in their meta-analysis study that an increase in BMI by 5 kg/m^2 significantly increased the risks of esophageal, thyroid, colon and renal cancers in men, and endometrial, gallbladder, esophageal and renal cancers in women. The summary of the risks associated with various cancers from this study are shown in Table 1. This meant that the association between BMI and cancers were sex-specific, applicable to wide range of cancer types, and generally consistent across different geographic populations (Renehan *et al.*, 2008; Roberts *et al.*, 2010). It is undeniable that obesity is a major risk factor for many types of cancer and it may even be the cause for some of these cancers.

1.3.2 Mechanisms of cancer progression in obese patients

The exact mechanisms by which obesity causes cancer has still not been fully elucidated yet. There are several hypotheses that link obesity with cancer: disruption of various hormone levels, chronic and sustained inflammation, and the activation of the insulin/insulin-like growth factor (IGF) pathway (Lumeng and Saltiel, 2011; Roberts *et al.*, 2010).

Adipose tissue has long known to have been a group of cells to store excess fats, but it has recently become apparent that adipose tissue is also a metabolically active endocrine organ (Roberts *et al.*, 2010). Increased adiposity is known to influence the bioavailability of endogenous sex hormones, including estrogens, androgens and progesterone (Calle and Kaaks, 2004). Adipocytes synthesise enzymes, such as aromatase, that convert precursor molecules into different forms

Table 1: Summary of the risk ratios[1] associated with each cancer type from the meta-analysis study by Renehan *et al.* (2008)[2].

	Men	Women
Colon	1.24 (1.20, 1.28)	1.09 (1.05, 1.13)
Rectum	1.09 (1.06, 1.12)	NA[3]
Gallbladder	NA	1.59 (1.02, 2.47)
Leukemia	1.08 (1.02, 1.14)	1.17 (1.04, 1.32)
Malignant melanoma	1.17 (1.05, 1.30)	NA
Multiple myeloma	1.11 (1.05, 1.18)	1.11 (1.07, 1.15)
Non-Hodgkin lymphoma	1.06 (1.03, 1.09)	1.07 (1.00, 1.14)
Esophageal adenocarcinoma	1.52 (1.33, 1.74)	1.51 (1.31, 1.74)
Pancreatic	NA	1.12 (1.02, 1.22)
Renal	1.24 (1.15, 1.34)	1.34 (1.25, 1.43)
Thyroid	1.33 (1.04, 1.70)	1.14 (1.06, 1.23)
Prostate	1.03 (1.00, 1.09)	NA
Post-menopausal breast	NA	1.12 (1.08, 1.16)
Endometrial (<27 kg/m^2)	NA	1.221 (1.084, 1.376)
Endometrial (>27 kg/m^2)	NA	1.729 (1.598, 1.872)

[1] All risk ratios are per increase in 5 kg/m^2 BMI; 95% confidence intervals in brackets.
[2] Table adapted from Roberts *et al.* (2010), where the original data in the table was from Renehan *et al.* (2008).
[3] Not available.

of androgens and estrogens (Calle and Kaaks, 2004). Many studies have confirmed the mitogenic effect of sex hormones on certain cell types, mainly breast and endometrial cell types (Roberts *et al.*, 2010).

On top of the effects that adipocytes have on circulating concentration of sex hormones, adipocytes also produce adipose tissue-specific hormones called adipokines (Roberts *et al.*, 2010). Of the many adipokines identified so far, leptin and adiponectin have been studied the most in terms of cancer development (Renehan *et al.*, 2006; Roberts *et al.*, 2010). Leptin is known to activate not only the satiety signalling pathway, but also the JAK/STAT and PI3K/Akt signalling pathways (Garofalo and Surmacz, 2006; Renehan *et al.*, 2006). Activation of these pathways leads to the activation of the target genes that are involved in cell proliferation and cell survival, as well as pro-angiogenic factors (Garofalo and Surmacz, 2006).

Adiponectin is a hormone that is synthesised and secreted only by adipo-

cytes (Kelesidis *et al.*, 2006). Unlike leptin, where the concentration of leptin is higher in those who are obese, adiponectin levels are lower in people that are obese compared to those with a normal weight (Kelesidis *et al.*, 2006; Renehan *et al.*, 2006). Adiponectin have anti-angiogenic and anti-proliferative effects on the cells, and is secreted by mature adipocytes but not by the premature pre-adipocytes (Gilbert and Slingerland, 2013). During chronic inflammatory state in obese patients, pro-inflammatory proteins such as transforming growth factor-β (TGFβ) and interferon-γ (IFNγ) can block the maturation of the pre-adipocytes, thus decreasing the level of adiponectin and the accompanying beneficial effects (Gilbert and Slingerland, 2013).

Alternatively, or possibly in conjunction with the disruption of hormone levels, inflammation caused by obesity may be associated with a tumour-promoting environment. TLR4 is a type of pattern recognition receptor (PRR) and, as mentioned earlier, is able to recognise circulating free fatty acids and trigger immune response via the NF-κB pathway (Lumeng and Saltiel, 2011). In addition to NF-κB pathway, recent evidence has shown that TLR4 could also activate the JAK/STAT3 pathway (Yu *et al.*, 2014). Activation of these pathways will further enhance the development of the tumour-promoting microenvironment caused by these unwanted inflammatory responses brought upon by obesity. Furthermore, since it is likely that people who are obese have less adiponectin concentration, obesity-induced inflammation may affect the patients that are obese more than those with a normal weight.

Finally, the insulin/IGF pathway may have a major role in encouraging tumorigenesis in obese patients. IGF signalling results in the activation of PI3K/Akt and extracellular signal regulated kinase (ERK) pathways, resulting in cell proliferation and reduced apoptosis (Roberts *et al.*, 2010). Though insulin is also capable of activating these pathways, it is thought that the mitogenic effects are controlled mainly through the IGF receptors (Roberts *et al.*, 2010). Since insulin reduces the level of insulin-like growth factor binding protein (IGFBP)-3, the main circulating IGFBP, hyperinsulinemia has been hypothesised to promote tumorigenesis through the mitogenic effect of insulin and by increasing the level of free IGFs (Giovannucci, 1995; McKeown-Eyssen, 1994; Roberts *et al.*, 2010). In fact, many studies have consistently associated the increased concentration of free IGF-I with

obesity and increased risk of some cancers; however the results for IGFBP have been inconsistent (Basen-Engquist and Chang, 2011).

It is evident that there are many possible mechanisms in which obesity contributes to the progression of cancer by activating many of the Hallmarks of Cancer mentioned in Section 1.2.2. Furthermore, these mechanisms are related to one another and are part of a larger, more complex physiological network (Renehan *et al.*, 2006). Due to this complexity, it is difficult to confidently explain how obesity contributes to the tumour development, and therefore the biological link between obesity and cancer is of great interest.

1.4 Genetic signatures

One of the difficulties associated with the treatment of cancers is the identification of the underlying biological mechanisms or hallmarks that drive the cancer progression. As one can imagine, it is difficult for clinicians to treat a patient if they do not know what is causing the disease. It is difficult to identify the underlying biological mechanisms because there are very few reliable biological markers (or "genetic signatures") that signify the presence or absence of the causal pathways. With that said, significant improvements have been made in recent years due to the advancement in the sequencing technologies. This has allowed researchers to study the genetic signatures that represent the hallmarks in much greater detail.

1.4.1 Microarray and Next Generation Sequencing technologies

Microarray technology

Microarray technology was first developed by Schena *et al.* (1995) and has been used extensively to study the gene expression patterns and single nucleotide polymorphism (SNP) identification for over two decades. In short, microarray technology uses a glass slide with hundreds of thousands of short DNA sequences attached to it, where fluorescently labelled complimentary DNAs (cDNAs) taken from the sample messenger RNAs (mRNAs) are hybridised to these immobilised DNA sequences (Schena *et al.*, 1995; Schulze and Downward, 2001). The fluo-

rescent probes are then detected to measure the gene expression (Schena *et al.*, 1995; Schulze and Downward, 2001).

One disadvantage of the microarray technology is that it requires previous knowledge of the genes and SNPs of the samples being investigated (Hurd and Nelson, 2009). This is due to the fact that the detection of sequences relies on the hybridisation of the sample sequences with the microarray sequences. Therefore, the sequences under investigation must be known prior to the investigation to attach the correct sequences to the microarray. Another disadvantage of microarrays is the potentially erroneous detection of the sample sequences, both in terms of its sequence and its expression level (Hurd and Nelson, 2009). Again, because the microarray slides require pre-defined sequences to be attached to it, there is always a possibility of hybridisation of similar but different sequence to the slide. Even if the sequences correctly hybridise to the slide, it may be difficult to measure the exact amount of particular sequences in the samples, since the amount of sequences in the sample is measured with the relative strength of the fluorescence of the dyes (Hurd and Nelson, 2009). Therefore, it is difficult to detect the exact level of expressions of the sequences that are either very rare or highly abundant in the sample (Hurd and Nelson, 2009).

Next generation sequencing technology

Next generation sequencing (NGS) technology refers to the high throughput sequencing techniques that allows the sequencing of a large quantity of genetic material (Metzker, 2010). NGS technology has been developed and improved in the last decade at a revolutionary pace. This has made it possible for researchers to study gene expressions and variations in days to weeks, where the same experiment would have taken years to complete before the technology was developed.

Every NGS technology uses the same principle process to sequence DNA samples. First, DNA samples are fragmented into smaller sizes and universal primers are attached to the ends of the fragments. These fragments are then amplified into clusters to make the signals stronger for detection (Metzker, 2010). After the amplification step, the clusters of DNA fragments are sequenced and "imaged" using fluorescently labelled DNA bases, where the incorporation of different bases are

detected by its unique fluorescence (Metzker, 2010).

NGS technology has many advantages over microarray technology. Firstly, NGS does not require any prior knowledge of the sample like in the microarray technology, as all genetic materials are attached to universal primers when they are sequenced; there is no need to pre-define the sequences to be searched and thus allows the detection of novel genes or transcripts (Hurd and Nelson, 2009). This also prevents the introduction of experimental bias, as the sequences are not altered in any way and no specific sequence is searched in any greater detail over another (Hurd and Nelson, 2009). Another advantage is that NGS produces count data of the genetic material being sequenced instead of relying on the intensity of the fluorescence to detect the presence and amount of certain sequence. This allows an absolute quantification of the amount of transcripts from the sample, making it possible for a more accurate comparison of the expression levels between samples. Lastly, many NGS techniques are capable of sequencing multiple samples simultaneously with high throughput, generating a tremendous amount of genetic data in a single run.

The use of NGS technology has significantly improved the workflow of modern genetic studies and its application in a variety of research areas has made it possible to dig deeper into the causes of complex diseases, such as cancer. With the cost of sequencing declining, personalised genome sequencing and treatment based on the sequence will be a reality in the near future.

1.4.2 Obesity associated genetic signatures

There is clear epidemiological evidence of obesity being a major risk factor for many cancer types (Section 1.3), but the genes or pathways that directly contribute to tumorigenesis in patients that are obese are yet to be clarified. Two studies in particular, one by Creighton *et al.* (2012) and another by Fuentes-Mattei *et al.* (2014), have investigated whether there were any differences in the genetic composition between obese and non-obese patients, specifically in a group of breast cancer patients. In both studies, the authors were able to identify a set of genes that were up or downregulated in breast tumours from the patients that were obese.

Creighton* et al. *(2012) study

Creighton *et al.* (2012) used 103 breast tumour samples and analysed these samples using microarrays. Of the 103 samples, 35% were from normal weight, 28% were from overweight and 37% were from obese patients. It was also noted that African-American patients were more likely to be obese than Caucasian patients. They compared the gene expression between the tumours from obese patients with the tumours from non-obese patients and identified 799 gene probe sets (662 unique genes) with $p < 0.01$ and with a fold change (FC) > 1.2. The majority of these genes (725 gene probes) were downregulated and 74 gene probes were upregulated in the breast tumour samples from the obese patients compared with the breast tumour samples from the non-obese patients.

Fuentes-Mattei* et al. *(2014) study

Fuentes-Mattei *et al.* (2014) studied 137 breast tumour samples from estrogen receptor (ER) positive patients and analysed their transcriptome using microarrays. Of the 137 patients, 43 patients were obese and 94 patients were not obese. Comparison of the gene expressions of the tumour samples from the obese patients with the tumour samples from the non-obese patients revealed 112 genes that were statistically significantly dysregulated (62 genes upregulated and 50 genes downregulated). Furthermore, functional transcriptomic analysis suggested that the insulin signalling and inflammation affected human ER^+ breast cancer.

On top of this, Fuentes-Mattei *et al.* (2014) validated their results using a mouse model, where they compared the expression level in the breast tumours from the obese mice with those from the lean mice. Results from this animal model provided strong evidence that obesity significantly accelerated tumour growth through Akt/mammalian target of rapamycin (mTOR) signalling pathway.

1.4.3 Pathway associated genetic signatures

Like obesity associated genetic signatures, there has also been a focus on the identification of pathways and genes that affect tumour growth. In 2006, Bild *et al.* demonstrated that a collection of genes that are involved in a specific pathway

could be used to determine the status of those pathways in tumour samples (Bild *et al.*, 2006). These pathway specific genes were identified experimentally by altering the expression of one specific gene that is central to the pathway in a model cell culture system. For example, all of the genes that have altered expression in the cells that had the *MYC* gene induced were noted as genes involved in the Myc pathway.

Gatza* et al. *(2010) study

Building on the work of Bild *et al.* (2006), Gatza *et al.* (2010) used 18 pathway signatures to classify human breast tumours into more detailed subtypes. The 18 pathways used in this study were: Akt, β-catenin (BCAT), E2F1, epidermal growth factor receptor (EGFR), ER, human epidermal growth factor receptor 2 (HER2), interferon-α (IFNα), IFNγ, Myc, p53, p63, PI3K, progesterone receptor (PR), Ras, STAT3, Src, TGFβ, and tumour necrosis factor-α (TNFα) pathways. This was a significant improvement from their previous work on pathway signatures (Bild *et al.*, 2006, 2009). Using more pathways in their analysis has allowed them to identify certain pathways that are coactivated (or cluster) together, suggesting that some of the pathways are related to one another and may have relevance in identifying the biological causes of cancer.

One thing to note here is that most of these pathway associated genetic signatures are involved in one or more of the Hallmarks of Cancer (Section 1.2.2). This means that these pathway associated signatures can be used to provide insight into the underlying biological mechanisms of the samples used in this project and help understand the underlying biology of cancers in obese patients.

1.5 Aims of the project

This research aims to determine whether gene expression signatures exist that are specific to obesity across multiple cancer types, and to investigate whether there are any common pathways being dysregulated in cancers based on these genetic signatures. A better understanding of the pathways being dysregulated in cancer cells in obese patients may lead to improved clinical decisions, and contribute towards personalised treatment in the future.

1.6 Thesis outline

In the following chapter, the detailed methods of all the analyses carried out for this project is outlined. The results from these analyses are split into two parts, each presented in a separate chapter. The first part of the results clarify the relationship between the obesity associated genetic signatures with various cancer types. The second part of the results consist of the links between the pathway associated genetic signatures and various cancer types, as well as the previously identified obesity associated signatures. Lastly, the results will be summarised and conclusions will be drawn from these results.

Chapter 2

Methods

2.1 R – statistical programming language

All statistical analyses and data manipulation were carried out with R (version 3.3.2 – "Sincere Pumpkin Patch"), a free open-source programming language and software environment for statistical computing and graphics (R Development Core Team, 2016).

2.2 Cancer data

The raw microarray data from the Creighton *et al.* (2012), Fuentes-Mattei *et al.* (2014) and Gatza *et al.* (2010) studies were downloaded from the Gene Expression Omnibus (GEO) website (Edgar *et al.*, 2002; NCBI, 2016). In addition to these breast cancer data sets, the raw microarray data set from a cohort of NZ breast cancer patients used in the Muthukaruppan *et al.* (2016) study was downloaded from the GEO website (abbreviated as New Zealand Breast Cancer (NZBC) data hereafter). RNA-seq and clinical data of multiple different cancer types were downloaded from the International Cancer Genome Consortium (ICGC) (ICGC, 2016; Zhang *et al.*, 2011) and The Cancer Genome Atlas (TCGA) (NCI and NHGRI, 2016) websites, respectively.

2.2.1 Breast cancer data

Creighton et al. *(2012) data*

The raw Affymetrix HGU133A microarray gene expression data files from the Creighton *et al.* (2012) study were downloaded from the GEO database (GEO accession ID: GSE24185). Clinical data of the samples (age, ethnicity, tumour grade, menopause status, BMI, ER status, PR status, HER2 status, and lymph node (LN) status) were obtained from the Supplementary Table 1 from the Creighton *et al.* (2012) study. 799 obesity associated gene probes identified in the Creighton *et al.* (2012) study were obtained from the Supplementary Data File 1 from the Creighton *et al.* (2012) study.

Fuentes-Mattei et al. *(2014) data*

The raw Affymetrix HGU133A microarray gene expression data files from the Fuentes-Mattei *et al.* (2014) study were downloaded from the GEO database (GEO accession ID: GSE20194). Clinical data for the samples (age, ethnicity, tumour grade, ER/PR/HER2 statuses, and treatments used) in this study was also downloaded from the GEO database (same GEO accession ID). Patient height, weight and BMI information were not available in this data set. 130 obesity associated gene probes identified by Fuentes-Mattei *et al.* (2014) were taken from the Supplementary Table 3 from their paper.

New Zealand Breast Cancer (NZBC) data

To provide an insight into the relevance of the obesity associated genetic signatures in NZ patient cohort, the NZBC data set used by Muthukaruppan *et al.* (2016) were used. The raw Affymetrix HGU133A microarray gene expression data files from the Muthukaruppan *et al.* (2016) study were downloaded from the GEO database (GEO accession ID: GSE36771). Clinical data for the samples (age, ethnicity, tumour grade, breast cancer subtype, ER/PR statuses, LN status, BMI and treatments used) in this study was obtained from Prof. Cris Print (personal communication with Assoc. Prof. Mik Black).

Gatza et al. *(2010) data*

The raw microarray gene expression data files from the Gatza *et al.* (2010) study were downloaded from the GEO database (GEO accession ID: GSE1456, GSE-1561, GSE2034, GSE3494, GSE4922 and GSE6596). Only the Affymetrix HGU-133A microarray samples were included in this project, as the other microarray data from the Creighton *et al.* (2012), Fuentes-Mattei *et al.* (2014) and Muthukarup-pan *et al.* (2016) studies were analysed using the Affymetrix HGU-133A platform. Clinical data for the samples in Gatza *et al.* (2010) study was not available, as these samples were a combination of many different datasets. The Akt, BCAT, E2F1, EGFR, ER, HER2, IFNα, IFNγ, Myc, p53, p63, PI3K, PR, Ras, STAT3, Src, TGFβ, and TNFα pathway associated genetic signatures were used by Gatza *et al.*, and the gene probes of all the pathway associated genetic signatures were obtained from the Supplemental Table 10 from their paper.

Gene probe ID conversion

All of the microarray data used gene probe IDs to refer to the genes, and there-fore these probe IDs had to be converted into their corresponding gene symbols. The gene probe IDs in the raw data were converted into their corresponding gene symbols using the *hgu133a.db* package in R (Carlson, 2016b). Since multiple gene probe sets match back to a single gene of interest in a microarray chip, there were conflicting expression data for some of the genes after the conversion of the gene probe sets into gene symbols. For the gene symbols that had multiple probe set entries, a single probe set was chosen to represent the gene symbol by using the `collapseRows` function in the *WGCNA* package in R, using the default parameters (which chose the probe set with a maximum mean expression value) (Langfelder and Horvath, 2008). Likewise, any obesity associated or pathway associated gene probes were converted into gene symbols.

Table 2: Number of samples in each of the breast cancer microarray data

Data set	Number of samples
Creighton *et al.*	103
Fuentes-Mattei *et al.*	278
NZBC	99
Gatza *et al.*	1060

Table 3: Summary of the clinical variables in the Creighton *et al.*, Fuentes-Mattei *et al.* and NZBC microarray data

Clinical variable	Creighton *et al.*	Fuentes-Mattei *et al.*	NZBC
Age			
Min.	30	26	31
Max.	72	79	92
Mean	49	52	59
Median	48	51	60
Ethnicity			
Caucasian	77	176	71
African-American	16	29	0
NZ Māori	0	0	10
Pacific Islands	0	0	14
Asian	10	18	4
Other	0	55	0
BMI status			
Normal weight	36	NA[1]	28
Overweight	29	NA	28
Obese	38	NA	43
Menopause status			
Premenopause	49	NA	NA
Perimenopause	9	NA	NA
Postmenopause	45	NA	NA
Tumour grade			
1	14	13	11
2	35	104	39

Table 3 (continued)

3	54	150	49
ER status			
ER^+	58	164	72
ER^-	42	114	27
Unknown	3	0	0
PR status			
PR^+	48	121	62
PR^-	50	157	37
Unknown	5	0	0
HER2 status			
$HER2^+$	12	59	NA
$HER2^-$	54	219	NA
Unknown	37	0	NA
LN status			
LN^+	14	NA	56
LN^-	89	NA	40
Unknown	0	NA	3

[1] Not available.

2.2.2 ICGC cancer data

RNA-seq and clinical data

The clinical data for all available cancer types (33 types in total) were downloaded from TCGA database (last accessed 1 April 2015) and were checked for both the height and weight data for each sample. Any cancer type with no height and/or weight data for the samples was excluded from the project, as no BMI information can be obtained without these data. Out of these 33 cancer types with clinical data, 14 cancer types had both height and weight data. However, only 8 cancer types out of these 14 types had RNA-seq data available from the ICGC database (last accessed 7 September 2015), so only those 8 cancer types were downloaded and used in this project. The selected cancer types were: bladder

urothelial carcinoma (BLCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), rectum adenocarcinoma (READ), skin cutaneous melanoma (SKCM), uterine corpus endometrial carcinoma (UCEC).

Data formatting and processing

The raw RNA-seq data from the ICGC database were formatted so that the count data of all the genes were listed for one sample, then the count data for all the genes for the next sample, and so on. This data format was highly inconvenient for later analyses so the data were reformatted into a "gene by sample" matrix format using the *dplyr* package in R.

Another problem with the data was the sample ID in the RNA-seq data. Though similar to the TCGA sample IDs, the ICGC IDs in the RNA-seq data had an extra identification code in the sample names. To associate each sample in the raw RNA-seq data from the ICGC database with the correct samples in the clinical data from the TCGA database, the ICGC IDs were re-coded to match the TCGA IDs. After the IDs were re-coded, samples were checked to see if there were any duplicates in either the ICGC RNA-seq data or the TCGA clinical data. Where there was a duplicate in the sample ID, a single sample with the highest mean expression value was chosen to represent that particular ID by using the `collapseRows` function in the *WGCNA* package in R (Langfelder and Horvath, 2008).

Since some of the samples did not have either height or weight data, these samples were removed from the analyses. In each of the eight cancer types, the TCGA sample IDs in the clinical data were cross-checked with the sample IDs in the RNA-seq data, and vice versa. Any samples that did not have either the patient BMI information or the RNA-seq data were removed from the analyses. This ensured that both BMI and RNA-seq data were available for all of the samples that were included in this project. See Table 4 for a summary of the total number of samples included in the analyses for each cancer type. Where missing, Equation 1 was used to calculate the BMI of the samples in all of the cancer data with height and weight data available (Section 2.2).

$$BMI = \frac{Weight(kg)}{Height^2(m^2)} \quad (1)$$

Samples were classified as normal weight (BMI < 25), overweight (25 $\leq$ BMI < 30) or obese (BMI $\geq$ 30) based on the WHO definition (Table 5). Equation 1 and Table 5 were used to calculate and group the patients from the ICGC cancer data into appropriate BMI classes (Table 6).

2.3 Data processing

2.3.1 Data normalisation

Any experimental procedure is prone to errors due to differences in experimental conditions, machinery used to measure signals and technical procedures in different laboratories, just to name a few. All of the data were normalised to remove such experimental bias, errors and noise, so only the true biological signals are considered in the analyses.

Microarray data

In Affymetrix HGU133A microarray experiments, there are two types of probes present on the microarray chips: perfect match (PM) and mismatch (MM) probes (Irizarry *et al.*, 2003). As the name suggests, PM probes represent the probes that should perfectly match the gene of interest, whereas MM probes have there 13th base pair intentionally altered to measure non-specific binding of the gene (Irizarry *et al.*, 2003). Various normalisation methods make use of PM and MM, or "probe pairs", to identify true signals from the noise.

Microarray analysis suite 5.0 (MAS5) and robust multi array (RMA) normalisation methods are both part of the *affy* package in R (Gautier *et al.*, 2004). MAS5 uses a difference-based method where it subtracts a value derived from MM from PM, but this approach may introduce additional noise and/or errors in some cases

Table 4: Summary of the total number of samples included in the analyses for each cancer type

Cancer Type	Initial number of samples	Number of samples after exclusion
BLCA	295	261
CESC	259	224
COAD	428	226
KIRP	222	124
LIHC	294	264
READ	154	73
SKCM	430	218
UCEC	508	482

Table 5: WHO defined BMI classification

Classification	BMI Value
Underweight	< 20.0
Normal weight/lean	20.0$\sim$24.9
Overweight	25.0$\sim$29.9
Obese	$\geq$ 30.0

Table 6: Summary of the patient BMI status in the ICGC cancer data

	Number of samples			
Cancer type	Normal weight	Overweight	Obese	Total
BLCA	103	97	61	261
CESC	83	64	77	224
COAD	75	81	70	226
KIRP	26	53	45	124
LIHC	139	76	49	264
READ	22	36	15	73
SKCM	71	74	73	218
UCEC	88	108	286	482

(Irizarry *et al.*, 2003). The RMA method is based on the observation that PM is a mixture of background and true signal, and uses mathematical models to estimate the expression while correcting for background signals from PM only (Irizarry *et al.*, 2003). In fact, the RMA method showed better identification of the true signals than the other methods, including MAS5 method (Irizarry *et al.*, 2003).

The raw microarray data were normalised using both RMA and MAS5 methods, each done separately on a copy of the data. The reason why the MAS5 method was used as well as the RMA method was because the MAS5 normalisation method was used in the Gatza *et al.* (2010) study. Though the normalisation methods were not specified in the Creighton *et al.* (2012) and Fuentes-Mattei *et al.* (2014) studies, to allow for data comparison between different datasets and accurate result validation of some of the studies, both RMA and MAS5 methods were considered for normalising the microarray data.

RNA-seq data

RNA-seq data are fundamentally different to microarray data, as RNA-seq data is a count data of transcript fragments, while microarray data is a continuous data based on the intensity of the probe pairs. This means that RNA-seq data must be processed in a different manner to microarray data. However, by analysing RNA-seq data as a count data, it limits the range of statistical tools and types of analyses that can be done on the data, as many tools are designed for normally distributed data like the light intensity measures from the microarray data (Law *et al.*, 2014).

To overcome this limitation, Law *et al.* (2014) developed a method called "variance modelling at the observational level", or "voom", that allows any RNA-seq data to be used in "any statistical pipeline for microarray data that is precision weight aware". In brief, voom first constructs a standard deviation trend from the logged counts per million (cpm) value of the genes (experimental design, treatment conditions and other factors are taken into account). This trend is then used to interpolate the standard deviation of the observation based on its predicted count size, and the inverse square of the predicted standard deviation is used as the weight for that observation. The weights of the observations and the logged cpm can then be used in other statistical pipelines that allow the input of

quantitative weights (Law *et al.*, 2014).

The raw RNA-seq data were normalised in two different ways depending on the analysis. For gene expression analyses (Section 2.4) and pathway enrichment analyses (Section 2.5), voom normalisation from the *limma* package in R was used to normalise the data (Ritchie *et al.*, 2015). For the purposes of data visualisation (Section 2.8) or the application of the metagene transformation matrices (Section 2.6.1) on the RNA-seq data, the raw data had 1 added (to prevent logging of 0) and then logged to the base of 10.

2.3.2 Data standardisation

For metagene creation and data visualisation that used heatmaps, the data were standardised so that each gene in the data had a mean (μ) of 0 and a standard deviation (σ) of 1. Since the expression levels of the genes could vary substantially (some may have very low expression, whereas another may have very high expression), the direct comparison of the raw expression values between different genes was not feasible. Standardisation of each gene allowed the expression levels to be on the same scale, thus allowing better visualisation with heatmaps and comparison between different genes was made possible.

2.3.3 Residual data creation

In some analyses, the Creighton *et al.* data was adjusted for clinical variables to remove any confounding effects from the variables. For example, a gene from the raw data might have a strong association with BMI, but it is possible that other clinical variables such as ER status and/or tumour grades are also associated with the gene. To prevent such confounding effect, all of the clinical variables except those that are of interest were adjusted with the use of linear models.

A linear model was created from the Creighton *et al.* data with age, ethnicity, menopause status, tumour grade, hormone (ER, PR, and HER2) statuses and LN status included in the model. In addition to this model, another linear model was constructed with the Creighton *et al.* data with only the Caucasian patients included in the data to completely remove the effect of ethnicity (same clinical variables were controlled for, except ethnicity). Once a linear model was fitted to

the Creighton *et al.* data, the remaining data, or the "residual data", represented the data that had been corrected for the effects of the unwanted variables, and was used in the analyses that required focus on certain variables without the other variables affecting the result.

2.3.4 Batch correction

It is most likely that experiments are done at different time period, location and laboratory environment. These differences between experiments introduces systematic non-biological differences or "batch effects" into the data, making it difficult to directly compare the data from different batches (Johnson *et al.*, 2007). The problem with the batch effect is that the normalisation methods do not control and adjust for the effect (Johnson *et al.*, 2007). Fortunately, Johnson *et al.* (2007) developed a method to correct for batch effects in the data, using an empirical Bayesian framework.

The data used in Gatza *et al.* (2010) study was a combination of multiple microarray data from various studies, so the batch effect had to be corrected before any analysis was carried out. Each microarray dataset was normalised with either RMA or MAS5 separately then combined together into a single dataset, and the batch effect was corrected with the `ComBat` function (an implementation of the batch correcting method by Johnson *et al.* (2007)) from the *sva* package in R (Leek *et al.*, 2012).

ICGC cancer data were also corrected for batch effect for the pathway enrichment analysis (Section 2.5). Two combined data sets were created with the ICGC data. The first data set was created by voom normalising each cancer data set separately, combine all of the normalised cancer data sets into one, then applied batch correction to the data set (Section 2.3.4). The second data set was created by first combining the cancer data sets into one, then the combined data was voom normalised, and then the batch effect was corrected to obtain the final data set.

2.3.5 Sample randomisation in simulation analysis

In some cases, an analysis was repeated a number of times to provide statistical evidence that the observed results did not occur by chance. All of the simulated

results must be derived from a randomised data, and by comparing the observed results with the results that occurred by chance, simulation analysis provides statistical evidence for the observed results. When a simulation was required, none of the values in the original data were modified, but instead the values in the original data were assigned randomly to different samples; in other words, the samples were shuffled. This provided a random set of data in which a simulation could be conducted with. The samples were shuffled with the `sample` function from the *base* package in R, and were reshuffled for each simulation.

2.4 Gene expression analysis

For gene expression analysis, or differential expression analysis, the *limma* package in R was used (Ritchie *et al.*, 2015). Since there were thousands of genes to be hypothesis tested, adjustment for multiple hypothesis testing had to be considered in order to identify the truly differentially expressed genes (DEGs).

2.4.1 Limma

limma is a package that contains variety of tools to analyse gene expression data using linear models-based methods, developed by Ritchie *et al.* (2015). For gene expression analysis, *limma* package fits a linear model in a gene-wise manner and produce test statistics that allow the assessment of whether the gene is differentially expressed or not between certain groups of samples (Ritchie *et al.*, 2015). In this project, the gene expression analysis was carried out to identify DEGs in the samples that were obese compared to the samples that were non-obese group.

Before the data was analysed using the *limma* analysis pipeline, it was normalised as described in Section 2.3.1: Data normalisation. In addition to this, a design matrix that describes the experimental design (obese versus non-obese) was created from the clinical data. To create the design matrix, the samples were divided into two groups, the obese group and the non-obese group, and the constructed group information was used in the `model.matrix` function (in *limma* package) to form the design matrix.

The `lmFit` function (*limma*) was used to fit a linear model to the normalised

data, using the experimental design information from the design matrix. The output of this function was used in the `eBayes` function (*limma*) to identify the DEGs from the data. In the `eBayes` function, statistical parameters are estimated from the data and these parameters are used in the empirical Bayes approach to calculate the summary statistics used for the ranking and identification of DEGs (Smyth, 2004).

The summary statistics from the `eBayes` function can be displayed with the `topTable` function (*limma*) and includes the estimate of the fold change of the gene expression in $\log_2$, average gene expression level in $\log_2$, moderated *t*-statistic, raw p-value of the gene, multiple hypothesis testing adjusted p-value, and *B*-statistic. The first value shows the estimate of the $\log_2$ fold change of the gene expression compared with the reference group, so this represents the $\log_2$ fold change of the gene expression in the obese samples relative to the non-obese group. The second value presents the $\log_2$ value of the average expression of the gene across all of the arrays/samples. The moderated *t*-statistic is the same as normal *t*-statistic, but its standard error has been adjusted (or "shrunk") toward a common value by a simple Bayesian method (Smyth, 2005). Raw p-value and adjusted p-value represent the p-value of the gene before and after it has been corrected for multiple hypothesis testing, respectively. Lastly, the value of *B*-statistic represents the log-odds that the gene is differentially expressed, where *B*-statistic of 0 shows that there is a 50% chance the gene is differentially expressed (Smyth, 2005).

From these summary statistics, the most likely DEGs were chosen from the list of significant genes by setting the threshold of the p-value to be either less than 1% or 5%. In the case where there were more than 1000 DEGs, the top 799 probe sets were picked from the list, as this was the number of genes found in the Creighton *et al.* (2012) study; otherwise, as many significant genes identified were taken from the list.

2.4.2 Multiple hypothesis testing correction

In an experiment where there are multiple hypotheses being tested, the rate or proportion of Type I error appearing from the experiment must be controlled. The

probability of Type I error occurring for a single hypothesis is usually controlled at some significance threshold α, which is usually set at 0.05 (Shaffer, 1995). In a typical microarray experiment, there are over 20,000 gene probes to be tested for differential expression, and with a significance threshold of $\alpha = 0.05$, this would yield approximately 1,000 Type I errors. In other words, 1,000 gene probes would be identified as differentially expressed, when in fact they are not.

There are two broad classes of methods to correct for this problem: family-wise error rate (FWER) control and false discovery rate (FDR) control. With FWER control, the method primarily aims to set the α-value for each hypothesis testing (α_i) such that the sum of all the α_i is equal to α (Hochberg and Tamhane, 1987; Shaffer, 1995). Usually, α_i is set to $\frac{\alpha}{n}$, where n is the number of hypothesis tests carried out in the experiment (Shaffer, 1995). This highly conservative approach significantly improves the certainty of the result from the experiment, but at the same time it significantly increases the likelihood of missing the truly DEGs, or Type II errors.

In contrast to the conservative FWER control methods, the FDR method developed by Benjamini and Hochberg (1995) control the Type I errors while maintaining statistical power. The FDR method controls the "expected proportion of errors among the rejected hypotheses" by adjusting the α-level (denoted as q^* in FDR) depending on the rank of the p-value (Benjamini and Hochberg, 1995). With FDR, the p-values are ordered and ranked from the lowest to the highest p-value, and for each hypothesis the p-value is compared with the adjusted threshold value:

$$P_{(i)} \leq \frac{i}{m} q^* \tag{2}$$

where $P_{(i)}$ is the p-value of the ordered and ranked ith hypothesis, m is the total number of hypotheses, and q^* is the threshold value (Benjamini and Hochberg, 1995). From Equation 2, the adjusted p-value from FDR control is the product of the p-value with the number of hypotheses, divided by its rank. Since FDR method controls Type I error in a less conservative manner than the FWER methods, the FDR method was used to identify the DEGs for gene expression analyses.

2.5 Pathway enrichment analysis

Identification of DEGs provides a list of genes that may have a role in a particular experimental setting. With that said, it is often difficult for the investigators to provide a plausible biological explanation with just a long list of DEGs, as the list lacks the link between the genes and the biological cause (Khatri *et al.*, 2012). Therefore, given a list of DEGs, it is important for any researcher to undertake pathway enrichment analysis in order to obtain useful insights into the underlying biological mechanisms the DEGs may be involved in. Both over representation analysis (ORA) and functional class scoring (FCS) can be used to identify whether certain pathways are enriched, given a list of DEGs.

2.5.1 Over representation analysis (ORA)

The common approach taken by ORA is to count the DEGs that are part of a biological pathway and perform a statistical test to decide whether the pathway is over- or under-represented in the list of genes (Khatri *et al.*, 2012). Statistical test used in ORA includes χ^2-test, hypergeometric test, and binomial test (Khatri *et al.*, 2012).

There are several limitations to the ORA approach. Firstly, the statistical tests are independent of the measured changes and therefore ignores the values associated with the genes, such as intensities and significance of the change (Khatri *et al.*, 2012). Secondly, only the most significant genes are selected for the input list of genes, which means that the almost significant genes are discarded from the analysis and results in a loss of information (Khatri *et al.*, 2012). Lastly, by treating each gene individually, the analysis ignores the biological interaction and complexity of the genes with the other genes, as well as between pathways (Khatri *et al.*, 2012).

2.5.2 Functional class scoring (FCS)

In general, the FCS methods measure gene-level statistics from the given data, then aggregate these statistics into pathway-level statistics, which is then used to calculate the statistical significance of the pathway (Khatri *et al.*, 2012). There are

two major classes of methods when calculating the statistical significance: self-contained or competitive tests (Goeman and Bühlmann, 2007). Self-contained tests are more appropriate in assessing whether a biological process is significantly involved in an experiment, whereas the competitive tests are better for selecting the most relevant biological processes from those that are not (Wu and Smyth, 2012).

The main difference between the two classes is the definition of the null hypothesis. Letting G be the gene set of interest and G^c its compliment, then in the self-contained tests, the null hypothesis is formulated as follows:

H_0^{self}*: No genes in G are differentially expressed.*

whereas in the competitive tests, the null hypothesis is defined as:

H_0^{comp}*: The genes in G are at most as often differentially expressed as the genes in* G^c.

These hypotheses show that the self-contained tests care only about the genes defined in the gene set, but the competitive tests examine the genes in the defined set as well as the genes not present in the set (Wu and Smyth, 2012). Due to the fact that self-contained tests do not take into account the genes not defined in the gene set, self-contained tests are more restrictive compared to competitive tests. The restrictive property of the self-contained tests give it greater power, as they are able to reject the null hypothesis at a higher accuracy for more gene sets than the competitive tests (Goeman and Bühlmann, 2007). However, as Goeman and Bühlmann (2007) stated in their paper, "the competitive types of test can be said to voluntarily relinquish some power in order to make a stronger statement". In fact, the Gene Set Enrichment Analysis (GSEA) method (developed by Subramanian *et al.* (2005)), one of the first and perhaps the most popular FCS methods, uses the competitive test approach.

The FCS approach addresses some of the limitations presented by the ORA methods. By calculating the statistical significance per pathway, the FCS approach takes into account of all of the genes involved in the pathways, and not just the genes that are differentially expressed. Furthermore, the FCS approach also detects small but consistent and coordinated changes in the expression of the

genes, unlike in the ORA where molecular measurements are completely ignored (Khatri *et al.*, 2012).

Though the FCS approach improves on the ORA approach, there are still some limitations. Since the analyses that use FCS compare the pathways independently of one another, they ignores the interactive nature of biological pathways. Another limitation is that the FCS methods use the molecular measurements to rank the genes, but they do not consider these changes in further analysis (Khatri *et al.*, 2012). For example, if one gene was expressed with a 2-fold change, whereas another gene was expressed with a 20-fold change, the FCS method will rank these genes accordingly (first, second, and so on) and disregard the fact that the second gene is much more highly expressed (Khatri *et al.*, 2012). Although both approaches have their own limitations, it is clear that FCS approach have more advantages over ORA methods.

2.5.3 Correlation adjusted mean rank gene set test (CAMERA)

Correlation adjusted mean rank gene set test (CAMERA) is a competitive FCS-based method developed by Wu and Smyth (2012), implemented as the `camera` function in the *limma* package (Ritchie *et al.*, 2015). Briefly, CAMERA fits a linear model in a gene-wise manner and calculates the gene-wise test statistics using the logFC between the conditions (for example, obese and non-obese samples). These gene-wise test statistics are used in the Wilcoxon-Mann-Whitney (WMW) rank sum test to question whether a pathway is significantly enriched in the data (Wu and Smyth, 2012).

One problem with the other FCS-based methods is that these methods do not consider the inter-gene correlation of the gene set being tested (Wu and Smyth, 2012). Since other methods assume that all the genes are equivalent under the null hypothesis, inter-gene correlation of the genes in a gene set will violate this assumption (Wu and Smyth, 2012). As a consequence, the Type I error rate is increased in these methods. However, CAMERA accounts for this correlation by estimating the variance inflation factor (VIF) from the gene-wise correlation and the number of genes in the gene set (Wu and Smyth, 2012).

CAMERA was used in this project to identify the pathways that were enriched

in the samples from obese patients compared with the samples from non-obese patients, as CAMERA provided a competitive FCS-based method.

2.5.4 Pathway databases

Gene Ontology (GO) database

Ontology is defined as a set of concepts and categories in a subject area or domain that shows their properties and the relations between them. Gene Ontology (GO), as the name suggests, is an ontology of the genes and biological pathways, curated and maintained by the GO Consortium (Gene Ontology Consortium, 2000, 2004). The goal of the GO Consortium is to "produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism" (Gene Ontology Consortium, 2000). The GO database includes not only the human genetic information, but collates variety of information from eukaryotic cells, including *Arabidopsis thaliana* and *Drosophila melanogaster* (Gene Ontology Consortium, 2000, 2004).

GO provides three distinct ontologies to describe the role of a gene or its gene product in the cell: biological process, molecular function, and cellular component (Gene Ontology Consortium, 2000). Biological process categorises the genes or its gene products into their overarching biological purpose or goal in the cell; for example, "cell death" and "cell growth and maintenance". Molecular function describes the biochemical activity a gene or protein has, ignoring when or where the activity takes place. The definition of "biochemical activity" is very broad as it includes specific binding to ligands and structure. Examples of terms used in molecular function are "kinase activity", "transporter" and "ligands". Lastly, cellular components refers to the location in which the gene product is active, such as "Golgi apparatus" and "nuclear membrane".

The *GO.db* package was used to load the GO database into R (Carlson, 2016a). Only the human related GO terms were used in pathway enrichment analysis, and these terms were organised into a format so that the pathways were able to be queried using the gene symbol.

Kyoto Encyclopedia of Genes and Genomes (KEGG) database

KEGG is a genomic database constructed by Kanehisa and Goto (2000) that store information about genes, pathways and ligands for a variety of species. The KEGG database also curates other information such as the approved drugs in US and Japan and genes related to diseases, making it a great tool for exploring the biological significance of genes and pathways (Kanehisa *et al.*, 2008). The *KEGG.db* package was used to load the KEGG database into R (Carlson, 2016c). Only the human related KEGG terms were used in pathway enrichment analysis, and these terms were organised into a format so that the pathways were able to be queried using the gene symbol.

Reactome database

The Reactome database is a "curated, peer-reviewed resource of human biological processes" (Joshi-Tope *et al.*, 2005). The Reactome database is constructed based on the reactions that have direct evidence from the literature, which are connected into higher order pathway structures (Joshi-Tope *et al.*, 2005). The *reactome.db* package was used to load the Reactome database into R (Lightenberg, 2016). Only the human related Reactome terms were used in pathway enrichment analysis, and these terms were organised into a format so that the pathways were able to be queried using the gene symbol.

2.6 Metagene analysis

2.6.1 Singular value decomposition (SVD)

In order to determine whether a set of genes is up or down regulated by a sample, some sort of score based on these genes must be calculated. SVD is a mathematical method that splits a data matrix into several smaller matrices (Golub and Reinsch, 1970). These matrices, when multiplied together, are able to recreate the original data matrix. With SVD, a matrix X of size n genes by m arrays (or

samples) can be represented as:

$$X = UDV^T \tag{3}$$

where the columns of U and V contain the left and right singular vectors of X respectively, and D contains the singular values of X.

The term that is of the greatest importance is V, as this matrix contains the principal components of the original data matrix X. Previous studies have used the principal components to summarise the expression of the set of genes for each sample in the experiment (Alter *et al.*, 2000; West *et al.*, 2001). This allows direct comparison of the expressions of multiple different genes in different arrays or experiments (Alter *et al.*, 2000). Furthermore, this summary gene, or "metagene", can be used to sort the samples and provide a meaningful grouping of the data which may help understand the underlying biology (Alter *et al.*, 2000).

With respect to this project, SVD was used to assess whether the metagenes created from various obesity or pathway associated genes were significantly associated with certain clinical variables, such as BMI. To do this, the normalised original expression data was reduced in size so that it only included the genes of interest (either obesity or pathway associated) and the samples, then `svd` function from the *base* package in R was used to apply SVD to the matrix. The first principal component was taken from the V matrix and was used as the metagene scores for each samples in the data set.

Another important property of SVD is that the signal used to create the metagene in the first data set can be directly applied to other data sets to obtain the metagene scores for the samples in those data set. Rearrangement of Equation 3 for V^T gives the following equation:

$$V^T = U^T D^{-1} X \tag{4}$$

where $U^T D^{-1}$ will be referred to as the "transformation matrix" hereafter. Clearly, by substituting the data matrix X with another data matrix, the transformation matrix allows for the generation of the metagene for that data set. The reason why the transformation matrix is used instead of independently applying SVD to

the other data set is because the genetic signature, and therefore the metagene, from a single data set is dependent on the signal within its data set. Therefore, the metagenes created independently from different data sets using SVD do not have the same weighting as the metagene from the original data set, and consequently does not provide a fair comparison of the metagenes across different data sets. For this reason, the transformation matrix for a given set of genes was created in the data set in which the genes were first identified in, and then transformed into other data sets to obtain the metagenes.

2.6.2 Ranking of the metagene scores

Two different approaches were taken in order to rank the metagene scores of the samples. The first approach was to rank the metagene scores with the `rank` function from the *base* package in R, then divided the ranked scores with the total number of samples in the data set to obtain a ranked metagene scores between 0 and 1 (known as fractional ranking). In most analyses, fractional ranking was used to assess the association of the metagene with the sample gene expressions and clinical variables.

An alternative approach called probit transformation was used by Gatza *et al.* (2010). In this transformation, the sample metagene scores were transformed with the probit function. To do this, the sample metagene scores were scaled and centered using the `scale` function from the *base* package in R, then the `pnorm` function from the *stats* package was used to transform the metagene scores so that the scores were between 0 and 1. Unless stated otherwise, all of the metagenes in this project were fractionally ranked.

2.6.3 Metagene direction

One thing to consider when analysing data with metagenes is the direction of the metagene. When SVD creates the metagene from a given data set, it does not consider the direction of the signal, but only the magnitude of the signal in the data. Therefore, the metagene created by SVD can be either positively or negatively correlated to the phenotype that the set of genes reflect. For example, a metagene created from a set of genes that are associated with the ER pathway may

have higher metagene scores in the samples with low expression of the ER gene, and low metagene scores in the samples with high expression of the ER gene (Figure 2, left heatmap). In this case, the direction of the generated metagene must be corrected in order to reflect the phenotype with the metagene scores (low metagene score with low gene expression and high metagene score with high gene expression; Figure 2, right heatmap).

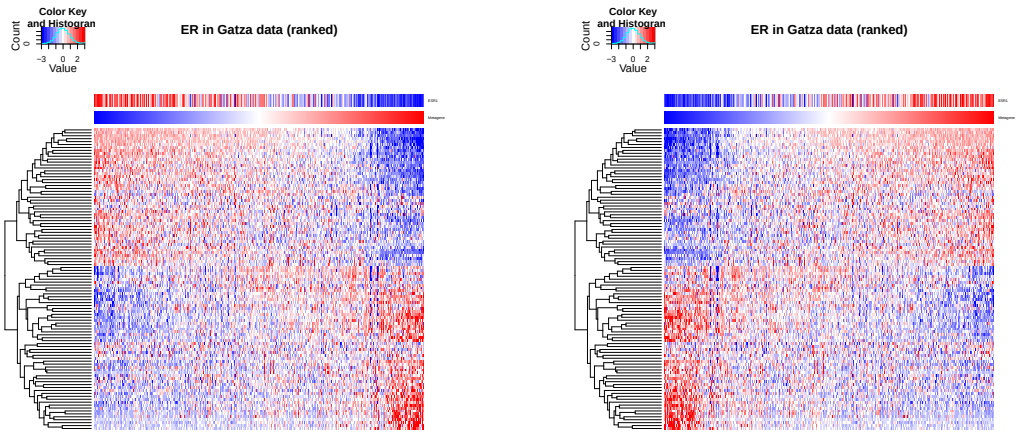


Figure 2: Heatmaps showing the direction of the uncorrected (left) and corrected (right) ER pathway metagene with the gene expression of the ER pathway genetic signature in the RMA-normalised Gatza *et al.* data. Level of gene expression is represented in the top right histogram, where low and high gene expression are colour-coded with blue and red, respectively. Each row of the heatmap represents a gene from the ER genetic signature, and each column of the heatmap represents a sample from the Gatza *et al.* data. The *ESR1* gene expressions of the samples are shown as a separate row at the top of the heatmap with the sample ER metagene scores shown below the *ESR1* gene expression, and the tree diagram of the hierarchical clustering of the genes is shown on the left of the heatmap.

In this project, two types of metagenes were created: obesity associated and pathway associated metagenes. For obesity associated metagenes, the direction of the metagenes were checked with the patient obesity status and BMI value from the data in which the metagene was derived from. When the data set did not have BMI or BMI status information for the patients (for example the Gatza *et al.* data set), a set of genes that were common in all obesity associated genetic signatures were taken and the expression of these common genes were compared with the metagene scores. The genes that are common to all of the obesity associated genetic signatures will be related to the metagenes produced by these signatures,

and therefore the expressions of these genes should be in line with the metagene scores.

Checking the direction of the pathway metagenes from Gatza *et al.* (2010) study is not as simple as the obesity associated metagenes, as the experimental design that identify which samples were treated to generate the pathway signature was not available. Due to this, it was not possible to validate whether the generated pathway metagene was in the "correct" direction. As an alternative, the pathway metagenes were visually compared with the gene that represent the pathway using a heatmap, as the expression of this gene would likely be affected the most by the pathway signature. For example, the ER pathway metagene was compared with the *ESR1* gene expression, and if the metagene was not in the same direction as the *ESR1* gene expression, the direction of the metagene was corrected (Figure 2). Table 7 summarises the genes used to check the direction of each of the pathway metagenes.

In addition to this, the correlation value of the pathway metagene with its respective gene was checked to further aid the decision on whether the metagene is going in the correct direction. This was due to the fact that some of the metagenes were visually difficult to determine whether it was in the right direction or not. Therefore, the direction as well as the magnitude of the correlation value was also taken into account in making the decision of whether to change the direction of the metagene. Lastly, the clustering of the pathway metagenes with the other pathway metagenes were also considered. It was known from the Gatza *et al.* (2010) study that some pathways clustered and grouped together with certain pathways, so the direction of some of the metagenes were altered in order to keep the grouping of the pathways matched to the results in their paper (Appendix B4).

Table 7: 18 pathways from Gatza *et al.* (2010) and its respective genes used to check the direction of the pathway metagene

Pathway	Representative gene	Brief pathway description
Akt	AKT1	Also known as protein kinase B; involved in cell survival
BCAT	CTNNB1	β-catenin; involved in the Wnt signalling pathway and cell-cell adhesion

Table 7 (continued)

E2F1	E2F1	Member of the E2F transcription factor family in higher eukaryotes; involved in cell cycle regulation
EGFR	EGFR	Also known as ERBB1 or HER1; involved in signalling the pathways for cell proliferation
ER	ESR1	Involved in cell cycle regulation, cell proliferation and cell survival
HER2	ERBB2	Also known as ERBB2; involved in cell proliferation and apoptosis
IFNα	IFNA1	Cytokine involved in innate immune response against viral infection
IFNγ	IFNG	Cytokine critical for innate and adaptive immune response
Myc	MYC	Transcription factor involved in cell proliferation, cell growth and apoptosis
p53	TP53	Tumour suppressor protein involved in DNA repair, cell cycle regulation and apoptosis
p63	TP63	Involved in development
PI3K	PIK3CA	Involved in cell growth, cell proliferation, cell differentiation and cell survival
PR	PGR	Signal transducer that activates other signalling pathways; involved in cell proliferation
Ras	HRAS	Involved in cell proliferation, cell differentiation, cell adhesion, apoptosis and cell migration
STAT3	STAT3	Involved in cell growth and apoptosis
Src	SRC	Involved in cell survival, angiogenesis, cell proliferation and cell invasion
TGFβ	TGFB1	Involved in cell proliferation, cell differentiation and apoptosis
TNFα	TNF	Involved in immune response, inflammation and apoptosis

2.7 Obesity metagene prediction with pathway metagenes

To determine whether the obesity metagene was associated with any of the pathway metagenes, a linear model was constructed with some clinical variables and pathway metagenes, which was then used to predict the obesity metagene. This was based on the idea that, if a clinical variable or a pathway metagene was a significant part of the model for the obesity metagene, then that clinical variable or pathway metagene may be associated with the obesity metagene. Furthermore, the model that was created based on the significant variables can be used to predict the obesity metagene, which can be used to compare the values with the original obesity metagene. If the predicted metagene was similar to the original metagene, then it provides an evidence that the variables used to construct the model was associated with the obesity metagene.

Since the NZBC data was the only data set (other than the Creighton *et al.* data set) that had the patient BMI information, all of the linear models were constructed in the NZBC data set. Eleven linear models were created in the NZBC to predict the obesity associated metagene from the Creighton *et al.* (2012) study. First seven models created were combinations of the sample BMI, BMI status and a selection of pathway associated metagene scores. The selected pathways used in the linear model construction were BCAT, ER, IFNα, IFNγ, Myc and PR. In deciding on the pathways to be used for the construction of the linear model, the correlation of the SVD- and transformation matrix-derived pathway metagene scores in different data sets were examined.

When a transformation matrix is applied to the data set in which it was derived from, the resulting metagene is identical to the one created from the direct application of SVD. Therefore, if the metagene derived from a transformation matrix is similar to the metagene created from SVD, then it shows that the genetic signature is consistent and reliable across different data sets. The above mentioned pathways had high correlation value between the SVD-derived and transformation matrix-derived pathway metagenes, and so it was used in the construction of the linear model.

The last four linear models were constructed with the different combinations of the sample BMI, BMI status and the PR pathway metagene. The reason that the PR pathway metagene was used separately was because in the initial analysis, the PR pathway metagene was the only pathway associated genetic signature that came up as significant in the other linear models (Section 4.3). All of the created linear models are summarised in Table 8.

Table 8: Summary of all the linear models constructed in the NZBC data

Linear models
BMI only
BMI status only
Selected pathways only[1]
BMI and BMI status
BMI and selected pathways[1]
BMI status and selected pathways[1]
BMI, BMI status and selected pathways[1]
PR pathway only
PR pathway and BMI
PR pathway and BMI status
PR pathway, BMI and BMI status

[1] BCAT, ER, IFNα, IFNγ, Myc and PR pathway metagene scores were used.

In addition to these eleven linear models, a linear model was created with the sample BMI, BMI status and all of the pathway associated metagenes (Akt, BCAT, E2F1, EGFR, ER, HER2, IFNα, IFNγ, Myc, p53, p63, PI3K, PR, Ras, STAT3, Src, TGFβ and TNFα) using a stepwise approach. The stepwise approach allows the construction of a linear model that best fits the data, given a set of variables. There are three approaches in creating a linear model in a stepwise method: forward, backward, and both.

In a backward (also known as "top-down") approach, all of the variables are first included in the linear model. At each step, a variable is removed from the model and the Bayesian Information Criterion (BIC) of the model is compared with the original model to determine whether the variable should remain in the model or not. This will result in a model with the variables that are jointly most predictive. In a forward, or "bottom-up", approach, the model begins with no

variable and adds one variable at a time to the model. If the variable increased the ability of the model to predict the response variable, then the variable is included in the model. Again, this will result in a model with only the variables that are contributing to predicting the response variable. Alternatively, both methods in combination can be used, where the model begins with all or none of the variables, and at each step a variable can be added to or removed from the linear model. In this project, the combination of forward and backward approaches (beginning with no variables in the model) was used to create the stepwise linear model.

The obesity metagenes that were predicted from the linear models in the NZBC data set were compared with the original obesity metagene scores. Where appropriate, p-values, analysis of variance (ANOVA) p-values and R^2-values were calculated to assess the relationship between the predicted and the original metagene scores. Scatter plots and box plots were also created to visually compare the two metagene scores (Section 2.8.1).

2.8 Plot creation

2.8.1 Bar, box and scatter plots

Bar plots were plotted to visualise the correlation between the SVD- and transformation matrix-derived obesity and pathway metagene scores in different data sets. The `barplot` function from the *graphics* package in R was used to plot the bar plots. Box plots and scatter plots were used to visualise the association of a metagene with some of the clinical variables of the data. Box plots and scatter plots were plotted using the `boxplot` and `plot` functions, respectively, from the *graphics* package in R. The p-value and ANOVA p-value in the box plots were calculated using the `t.test` and `aov` functions from the *stats* package in R, respectively. The line of best fit and the R^2-value of the line were taken from the summary statistics produced by the `lm` function from the *stats* package in R.

2.8.2 Heatmaps

To visualise the association of the sample metagene scores with the sample gene expression, heatmaps were used. Heatmap is an effective method to visualise such data, as the ordering of the genes and/or samples may be controlled based on certain values, for example metagene scores. Furthermore, heatmaps are able to cluster the samples and/or genes based on the "distance" between the data, or in other words, how closely related the data are. Heatmaps were created using the `heatmap.2` function from the *gplots* package in R (Warnes *et al.*, 2016). For the heatmaps that required two bars above the heatmap (like those that investigate the directionality of the pathway metagenes), the heatmaps were created with the `heatmap.2x` function developed by Tom Kelly (https://github.com/TomKellyGenetics/heatmap.2x).

2.8.3 Venn diagrams

To summarise the number of DEGs from the gene expression analyses of the Creighton *et al.* data that overlapped with the original Creighton *et al.* obesity associated genetic signature, Venn diagrams were used. Venn diagrams were plotted using the `venn` function from the *gplots* package in R (Warnes *et al.*, 2016). The `overLapper` function (from http://faculty.ucr.edu/~tgirke/Documents/R_BioCond/My_R_Scripts/overLapper.R) was also used to help plot the Venn diagram (Girke, 2016).

2.8.4 Additional colours

Some colour palettes such as the colours used by MATLAB are not installed in R by default. In the Gatza *et al.* (2010) study, Gatza *et al.* used MATLAB to generate all of their plots in their paper. Since a visual comparison had to be made with one of the plots from the Gatza *et al.* (2010) study, some of the heatmaps in this project required MATLAB-like colours for a more accurate comparison with their results. For a colour palette used in MATLAB, the `matlab.like` function from the *colorRamps* package in R was used (Keitt, 2012). For the pastel colour palette used in the bar plots, the `brewer.pal` function from the *RColorBrewer* package in R was used (Neuwirth, 2014).

Chapter 3

Obesity associated genetic signatures and cancer

The obesity associated genetic signatures are central to this project, as a means to clarify the relationship between BMI and tumour biology. In this chapter, the previously identified obesity associated genetic signatures from the studies conducted by Creighton *et al.* (2012) and Fuentes-Mattei *et al.* (2014) are examined in turn to judge the agreement of these signatures with the patient BMI and BMI status, as presented in their results. After this, novel obesity associated genetic signatures are identified in the Creighton *et al.* (CR) data set and was compared with the obesity associated genetic signatures from the Creighton *et al.* (2012) and the Fuentes-Mattei *et al.* (2014) studies. Lastly, the presence of common genes or pathways associated with obesity in multiple types of cancer is explored using the data sets from ICGC (ICGC, 2016; Zhang *et al.*, 2011).

3.1 Obesity associated genetic signature from the Creighton *et al.* (2012) study

An obesity metagene was created using the obesity associated genetic signature from the Creighton *et al.* (2012) study. There was no description about the normalisation method used by Creighton *et al.* when they first analysed their data, so the

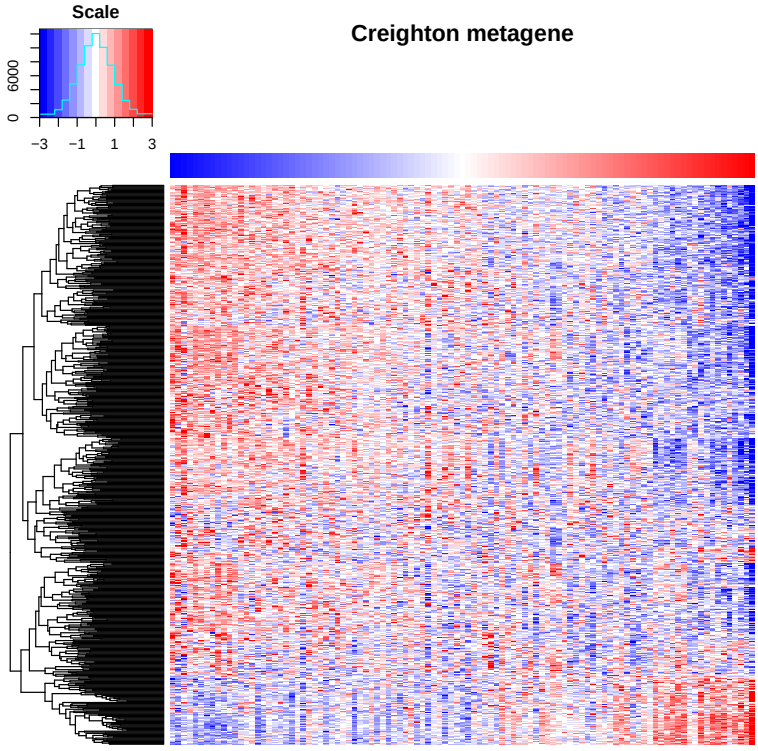


Figure 3: Heatmap showing the obesity metagene from the Creighton *et al.* (2012) study with the sample gene expressions of the obesity associated genes in the CR data. Level of expression is represented in the top right histogram, where low and high gene expression are colour-coded with blue and red, respectively. Each row of the heatmap represents a gene from the obesity associated genetic signature, and each column of the heatmap represents a sample from the CR data. The obesity associated metagene scores of the samples are shown as a separate row at top of the heatmap, and the tree diagram of the hierarchical clustering of the genes is shown on the left of the heatmap. For clarity, the sign of the metagene scores were reversed in order to match the results from the Creighton *et al.* (2012) study.

more popular RMA method was used to normalise the CR data set (Irizarry *et al.*, 2003). The obesity metagene was created in the standardised RMA normalised CR data set, and was plotted above a heatmap with the sample gene expression to check whether the metagene scores were in accordance with the overall gene expression of the samples (Figure 3).

As shown in Figure 3, a high obesity associated metagene score reflects low expression in majority of the genes in the signature, and in contrast, a low obesity

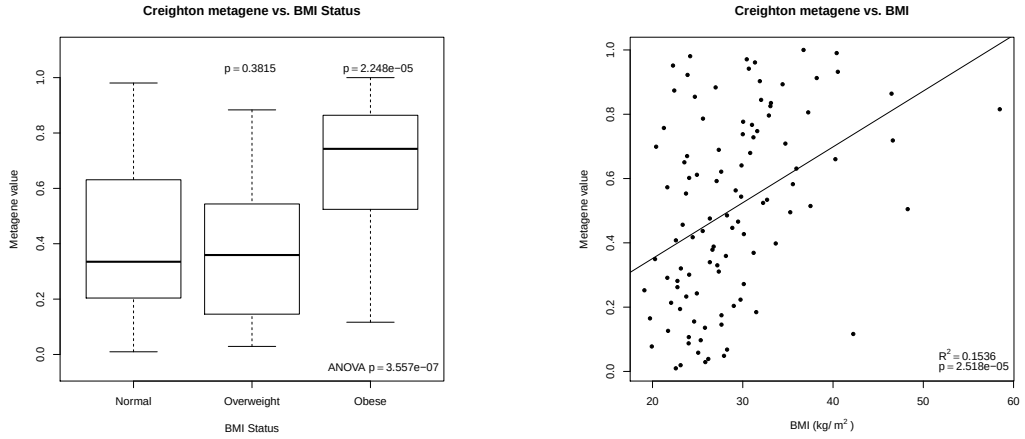


Figure 4: Box plot and scatter plot showing the association of the obesity metagene from the Creighton *et al.* (2012) study with the sample BMI status and BMI, respectively, from the CR data. In the box plot, the p-values above the groups represent the statistical significance of the association of the obesity metagene with the overweight or obese group compared with the normal weight group. The ANOVA p-value shows the statistical significance of the association of the obesity metagene with the sample BMI groups. In the scatter plot, R^2- and p-values describe the adjusted coefficient of determination of the regression line and the statistical significance of the linear model used to draw the regression line, respectively.

associated metagene score reflects high expression in the majority of the genes in the signature. This was consistent with the reported property of the obesity associated genetic signature by Creighton *et al.* (2012) (see Section 1.4.2). To provide further evidence that the obesity metagenes were in fact associated with both the BMI status and the BMI value of the patients, a box plot and a scatter plot were created, respectively (Figure 4).

Figure 4 clearly showed that the obesity metagene from the Creighton *et al.* (2012) study significantly associated with the obese group of patients, as well as patient BMI value. It should be noted here that the obesity metagene significantly associated with the samples from the patients that were obese, but not with the sample from the patients that were overweight. This was due to the fact that the obesity associated genetic signature was originally identified from the comparison of the patients that were obese with the patients that were not obese, and therefore the obesity metagene scores were significant with the obese group, but not with the overweight group. Another thing to note here was that even though the

regression line in the scatter plot showed statistically significant association with the patient BMI, patient BMI values seem to be randomly dispersed across the obesity metagene scores. This suggests that perhaps the obesity metagene from the Creighton *et al.* (2012) study may not be associated with the patient BMI as strongly as reported.

Now that the association of the obesity metagene from the Creighton *et al.* (2012) study was established in the CR data set, the Creighton *et al.* obesity metagene was generated in the ICGC cancer data. The direction of the obesity metagene was checked in the CR data first, so that high metagene scores reflected high patient BMI and low metagene scores reflected low patient BMI (Figure 4). The transformation matrix was then created in the RMA normalised CR data, as described in Section 2.6.1.

All of the ICGC RNA-seq data were normalised as described in Section 2.3.1. Before the transformation matrix was applied to the $\log_{10}$-normalised cancer data, the obesity metagene was created from the standardised data or untouched (non-standardised) ICGC data and compared with one another to determine which data format was better for the application of the transformation matrix (Appendix A2). From these results, the standardised data was found to be the most suitable format for the application of the transformation matrix. The transformation matrix was applied to each cancer data set in turn to generate an obesity metagene from each of the data sets. Each obesity metagene was plotted in a heatmap with the corresponding ICGC data set from which the obesity metagene was generated from (Figure 5; Appendix A7). These heatmaps confirmed that the obesity metagene was able to capture the overall gene expression pattern in all of the ICGC cancer data, where the metagene scores reflected the expression levels of the majority of the genes in the signature. As before, association of the obesity metagene with the patient BMI and BMI status was examined in their respective cancer data set (Figure 5; Appendix A7). Out of all the cancer types, only the BLCA data set showed a significant association with the obesity metagene, and only for the overweight group (not the obese group).

There could be several reasons for the apparent lack of association of the obesity metagene with patient BMI in most of the ICGC data. First, the transforma-

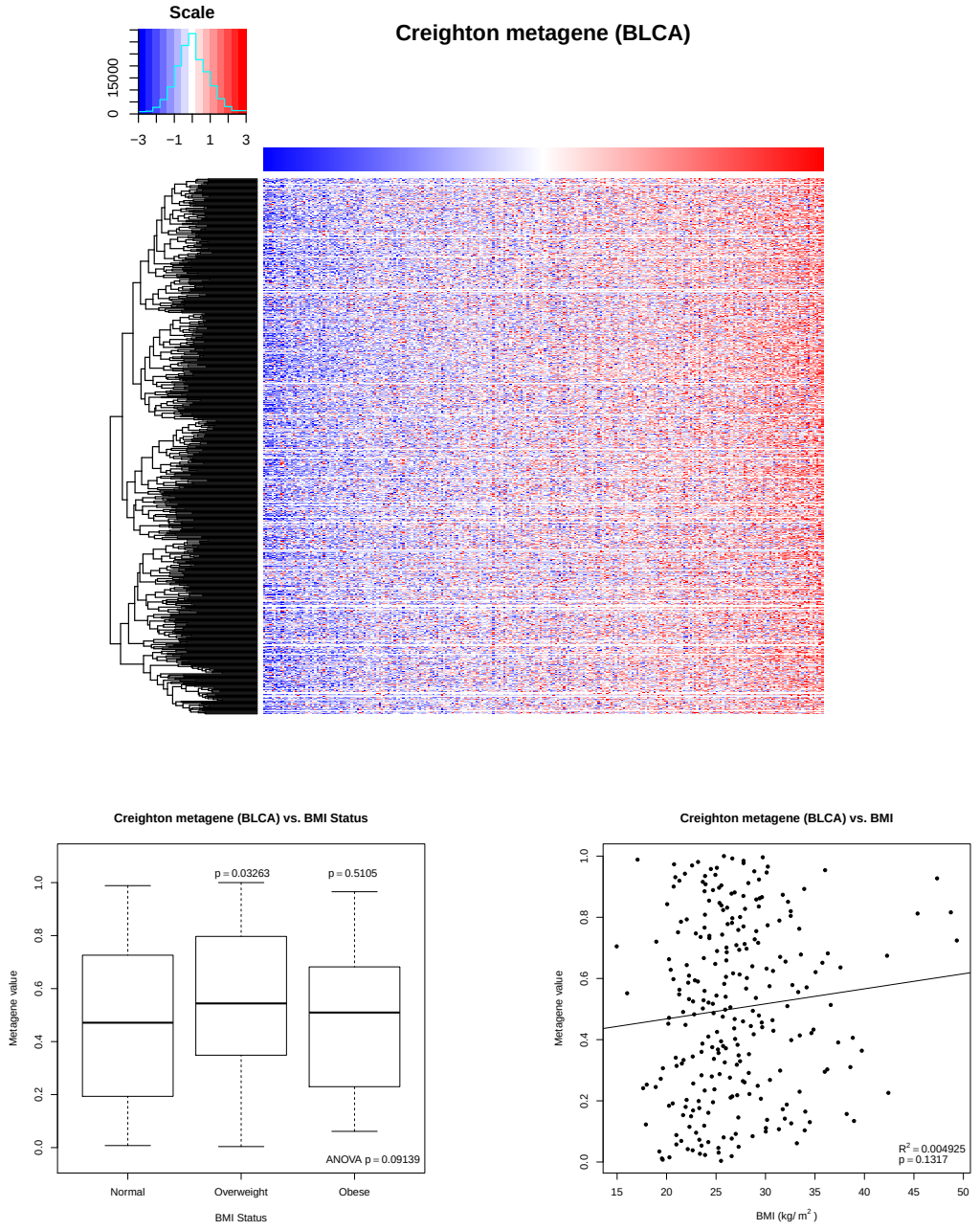


Figure 5: Heatmap, scatter plot and box plot showing the association of the obesity metagene from the Creighton *et al.* (2012) study with the sample gene expression, patient BMI and BMI status, respectively, from the ICGC BLCA data. The results for the other ICGC cancer types are shown in Appendix A7. Scales, p-values and R^2-value are as described in previous figures.

tion matrix was derived from the CR microarray data, but the ICGC cancer data were generated via RNA-seq. Though the $\log_{10}$-normalisation and standardisation of the data was the most appropriate adjustments to be made to the RNA-seq data, these adjustments were not equivalent to the RMA normalisation method that was used on the microarray data. Secondly, none of the ICGC cancer data in this project originated from breast tumours as in the CR data. Since the obesity associated genetic signature was identified in the breast cancer data, the signature may be specific to breast cancer and may not be generalisable to other cancer types.

To check whether the signature was specific to breast cancer microarray data, the same transformation matrix used in the ICGC data was applied to the breast cancer microarray data from the Muthukaruppan *et al.* (2016) study (referred to as NZBC data hereafter). NZBC data was normalised with the RMA method and the transformation matrix was applied to the normalised data to obtain the Creighton *et al.* obesity metagene in the NZBC data set. The generated obesity metagene was again compared with the gene expression of the samples with a heatmap and the association of the metagene with the patient BMI and BMI status was examined with box and scatter plots, respectively (Figure 6).

The Creighton *et al.* obesity metagene managed to reflect the overall gene expression of the samples in the NZBC data (Figure 6). However, as with the ICGC cancer data, the obesity metagene scores did not significantly associate with patient BMI or BMI status (Figure 6). These results confirmed that the lack of association of the obesity associated genetic signature from the Creighton *et al.* study was not due to the technology in which the data was gathered (microarray or RNA-seq), nor the cancer type in which the genetic signature was derived from.

Taken together, these results suggest that the obesity associated genetic signature identified in the study conducted by Creighton *et al.* (2012) correlated with patient BMI and BMI status only in the CR data set. The lack of association with patient BMI in other cancer data sets was not due to the type of technology platform in which the data was gathered, as neither the ICGC RNA-seq data nor the NZBC microarray data showed significant association of the obesity metagene with the patient BMI. Furthermore, the obesity associated genetic signature was

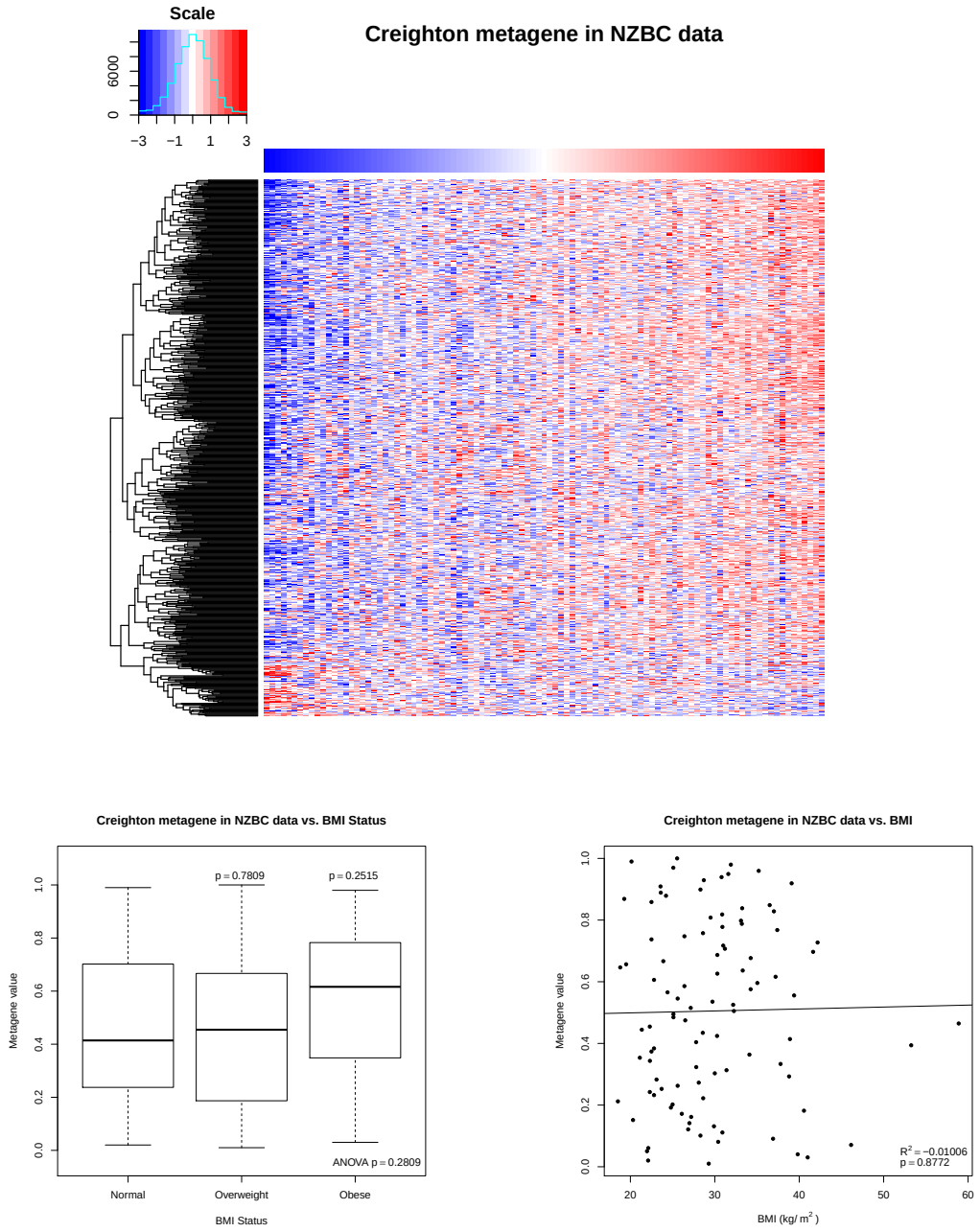


Figure 6: Heatmap, scatter plot and box plot showing the association of the obesity metagene from the Creighton *et al.* (2012) study with the sample gene expression, patient BMI and BMI status, respectively, from the NZBC data. Scales, p-values and R^2-value are as described in previous figures.

not dependent on the cancer type in which it was generated from, since the obesity metagene did not show significant association in either the ICGC cancer data or in the NZBC data.

One possible reason why the obesity metagene from the Creighton *et al.* (2012) study did not show significant association with other data sets could be because the genetic signature was in fact not an obesity specific signature, but a signature that was detected due to another clinical variable (investigated in Section 3.3). Another reason for this apparent lack of association could be that the obesity associated genetic signature was too specific to the CR data and was not a broad obesity associated genetic signature, but an obesity associated signature that was specific to the patient cohort that was profiled in the Creighton *et al.* (2012) publication.

3.2 Obesity associated genetic signature from the Fuentes-Mattei *et al.* (2014) study

The obesity associated genetic signature from the Creighton *et al.* (2012) study was not associated with patient BMI or BMI status in majority of the cancer data sets, so the obesity associated genetic signature from the Fuentes-Mattei *et al.* (2014) study (FM) was examined to see whether this obesity metagene was able to significantly associate with the patient BMI and BMI status across different data sets. Since the FM data set did not have patient BMI information, the FM obesity metagene was not able to be compared with the patient BMI or BMI status in the original FM data. Nevertheless, the transformation matrix was still generated in the FM data and applied first to the microarray data (CR and NZBC data sets) then to the ICGC data sets to see whether the FM obesity metagene associated with patient BMI and BMI status in these data sets.

The FM data was normalised with the RMA method and SVD was performed on the normalised FM data to obtain the transformation matrix. The transformation matrix was used to transform the RMA normalised CR data to extract the FM obesity metagene scores in the CR data. The FM obesity metagene scores were compared with sample gene expression levels, patient BMI and BMI status in the CR data (Figure 7). Clearly, as with the CR obesity metagene, the FM obesity

metagene was reflective of the overall gene expression of the samples, but did not associate with patient BMI or BMI status, suggesting that this signature also does not generalise to other data sets. The transformation matrix was then applied to the NZBC microarray data. Again, FM obesity metagene scores reflected the gene expression of the samples, but did not significantly associate with patient BMI or BMI status (Figure 8).

Next, the transformation matrix was applied to the ICGC cancer data and the resulting FM metagene was compared with the sample gene expression, patient BMI and BMI status in each of the ICGC cancer data. As evident in Figure 9 and Appendix A3, the FM obesity metagene scores appeared to reflect the overall gene expression of the FM obesity associated genetic signature. As with all the results in this chapter so far, the FM obesity metagene did not significantly associate with any of the ICGC cancer data, except in the BLCA data set (Figure 9; Appendix A3). The FM obesity metagene significantly associated with the overweight group (but not with the obese group), and also had a significant ANOVA p-value Figure 9. On the contrary to the association of the metagene with the patient BMI status, the FM obesity metagene was not associated with the patient BMI. These results suggested that the samples from the patients that were overweight in the BLCA data set had similar biological properties as the samples taken from the patients that were obese in the FM data set. However, due to the fact that the FM obesity metagene lacked association with the patient BMI in the BLCA data set, and that the metagene did not show any significant association in any of the other ICGC cancer types, it was difficult to determine whether the observed association of the FM obesity metagene with the overweight group was truly reflective of the quality of the FM metagene or observed by chance.

These results showed that the obesity associated genetic signature from the Fuentes-Mattei *et al.* (2014) study was not generalisable to other cancer data sets, similar to the obesity signature identified by Creighton *et al.*. This meant that both the Creighton *et al.* and Fuentes-Mattei *et al.* obesity metagenes may have been too specific to the original data set in which the signatures were identified in. Furthermore, there was a possibility that these obesity associated metagenes were not related to obesity, but associated with a different clinical variable that may be closely related to BMI (Section 3.3).

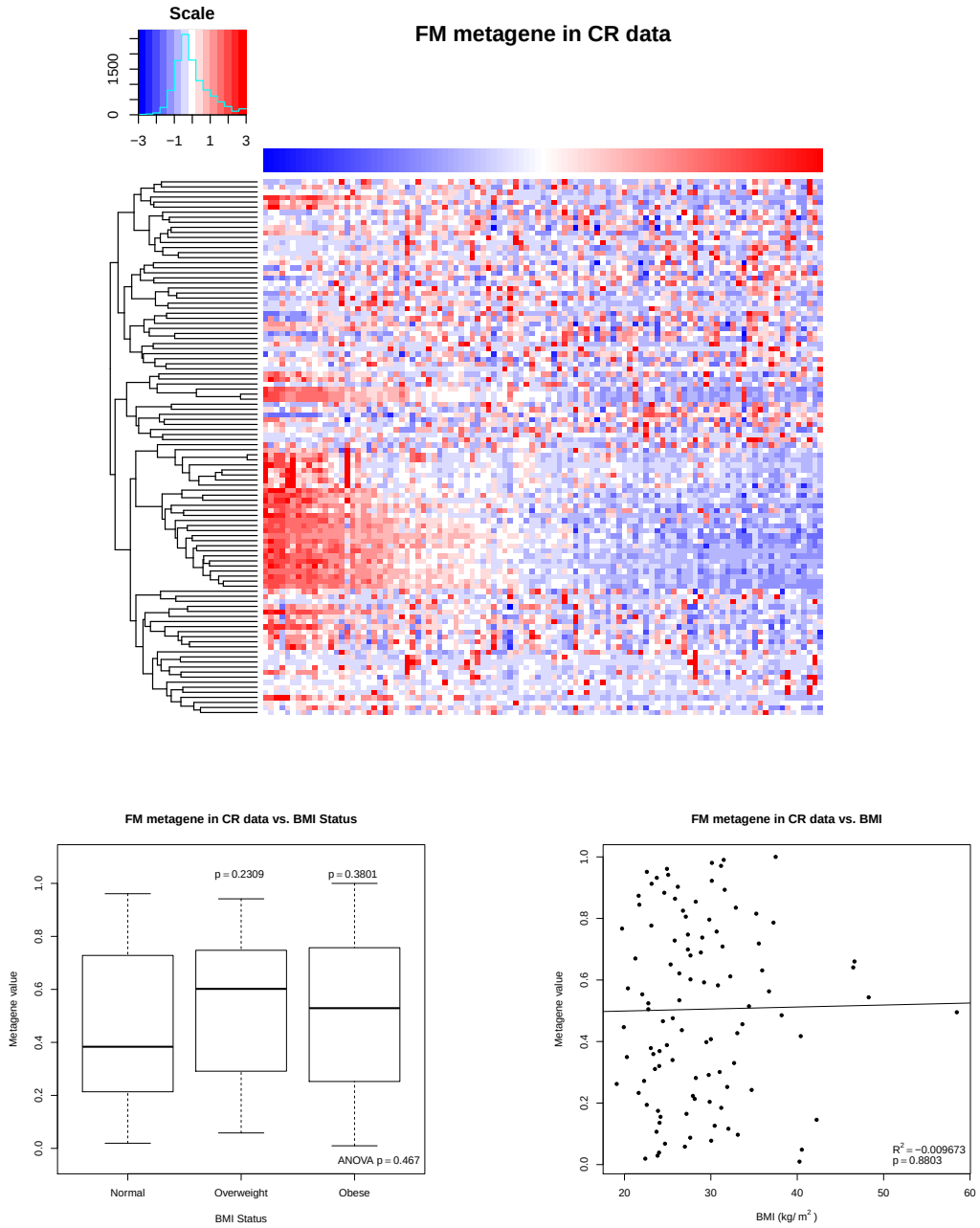


Figure 7: Heatmap, scatter plot and box plot showing the association of the FM obesity metagene with the sample gene expression, patient BMI and BMI status, respectively, from the CR data. Scales, p-values and R^2-value are as described in previous figures.

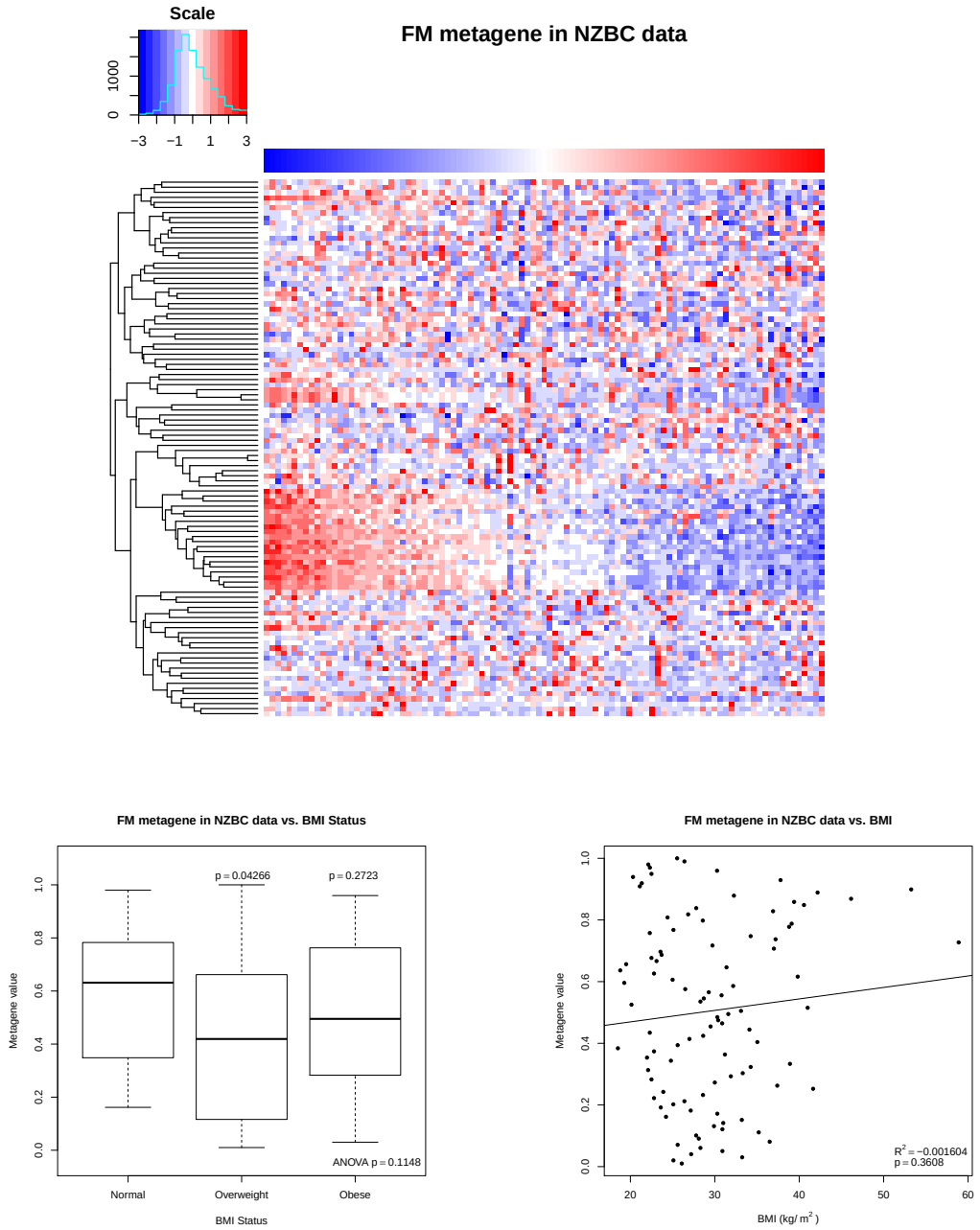


Figure 8: Heatmap, scatter plot and box plot showing the association of the FM obesity associated metagene with the sample gene expression, patient BMI and BMI status, respectively, from the NZBC data. Scales, p-values and R^2-value are as described in previous figures.

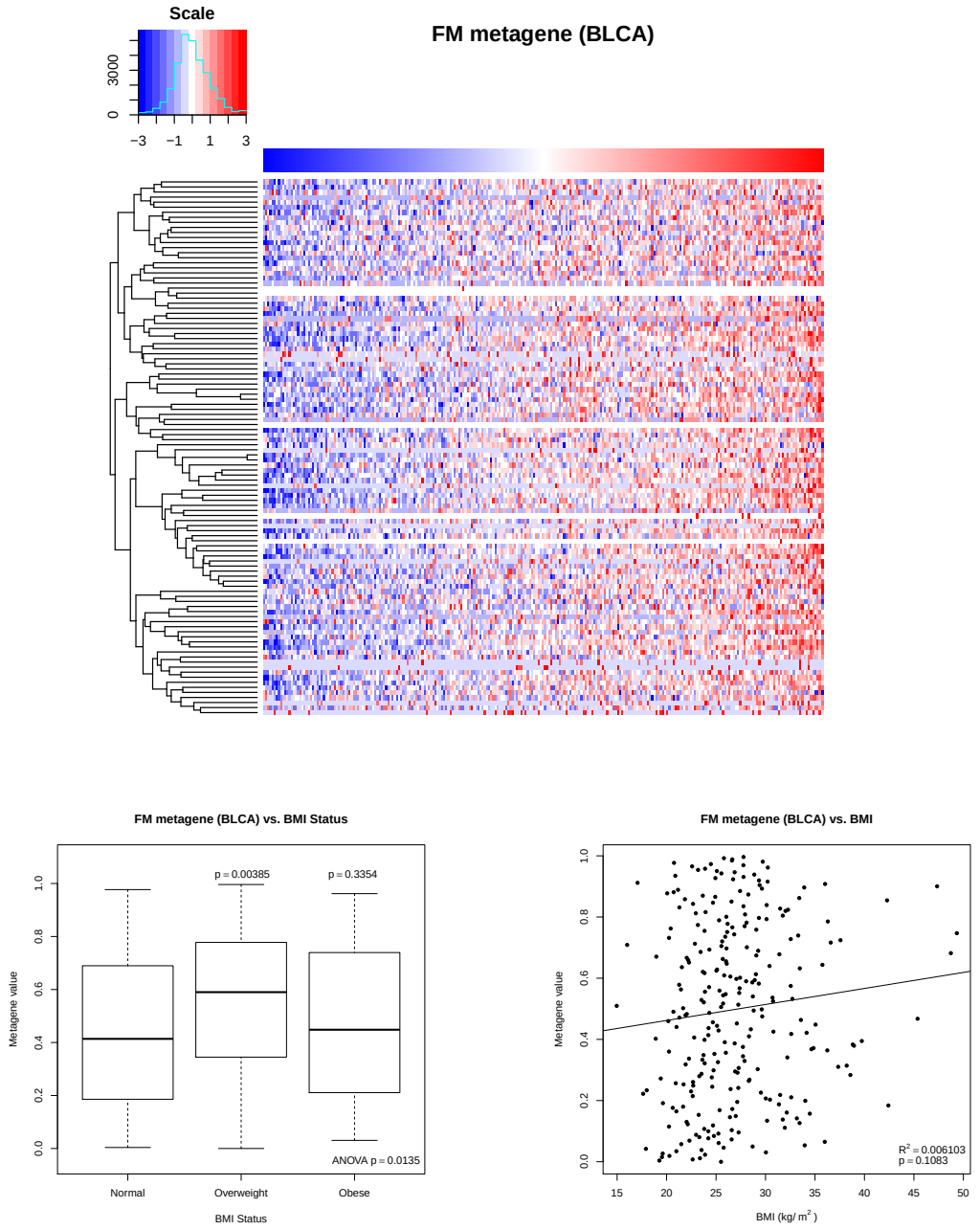


Figure 9: Heatmap, scatter plot and box plot showing the association of the FM obesity metagene with the sample gene expression, patient BMI and BMI status, respectively, from the ICGC BLCA data. The results for other ICGC cancer types are shown in Appendix A3. Scales, p-values and R^2-value are as described in previous figures.

3.3 Novel obesity associated genetic signatures from the Creighton *et al.* (2012) data set

3.3.1 Identification of novel obesity associated genetic signatures

Both the obesity associated genetic signatures identified from the Creighton *et al.* (2012) and Fuentes-Mattei *et al.* (2014) studies were able to capture the overall gene expression patterns of the genetic signatures of the samples, but did not associate with patient BMI or BMI status in majority of the cancer data sets. One possible reason for this result could be that the obesity associated genetic signatures from the Creighton *et al.* (2012) and Fuentes-Mattei *et al.* (2014) studies may not have been truly associated with patient BMI, but with another clinical variable. To investigate this possibility, novel obesity associated genetic signatures were identified in the CR data after controlling for all the clinical variables in the data set. The FM data set was not used to get the obesity associated genetic signatures, as no BMI information was available for the patients in the FM data.

Firstly, an attempt was made to replicate the Creighton *et al.* obesity associated genetic signature from the original CR data. Creighton *et al.* (2012) originally found their obesity associated genetic signature by doing a gene expression analysis between the breast tumour samples from the obese and non-obese patients, and from this list, Creighton *et al.* selected the 799 statistically significant DEGs ($p < 0.01$) with $\log_2$-fold change greater than 1.2. Therefore, gene expression analysis between the samples from obese and non-obese patients in the RMA normalised CR data was carried out (described in Section 2.4). Without adjusting the p-value to account for multiple hypothesis testing, 5278 gene probes and 1781 gene probes were significant at $p < 0.05$ and $p < 0.01$, respectively (Table 9). After adjustments were made for multiple hypothesis testing, there were only 9 gene probes significant at $p < 0.05$ and no gene was significant at $p < 0.01$, which suggest that the majority of the 799 probe sets reported by Creighton *et al.* may actually be false positives.

Furthermore, there were only 61 gene probes that were significantly differentially expressed at $p < 0.05$ with a $\log_2$-FC greater than 1.2. From these obser-

vations, the $\log_2$-fold change of the gene probes were ignored and the threshold p-value was set to 0.01 (unadjusted p-value) for the identification of the significant probe sets. Additionally, when there were more than 799 gene probes identified, only the most significant 799 gene probes were taken as the obesity associated genetic signature, as this many genes were originally identified by Creighton *et al.*. This was to include as many probe set as possible for the novel genetic signatures to be comparable with the original Creighton *et al.* obesity associated genetic signature. These criteria were applied for the identification of other genetic signatures as well.

The above analysis was repeated with the residual data (RMA normalised CR data that had been controlled for other clinical variables; see Section 2.3.3). The clinical variables controlled were age, ethnicity, menopause status, tumour grade, hormone (ER, PR, and HER2) statuses and LN status. In the residual data, 1104 gene probes were significantly differentially expressed with unadjusted p-value ($p < 0.01$; additional results shown in Table 9). Again, the most significant 799 gene probes were taken as the obesity associated genetic signature from this data.

In addition to the above two obesity associated genetic signatures, two more sets of genetic signatures were identified by taking the gene probes that were common between each of the above two signatures with the original obesity associated genetic signatures. There were 239 common gene probes between the original Creighton *et al.* obesity associated genetic signature and the gene probes identified from the unadjusted CR data, and 168 common gene probes between the original signature and the gene probes from the residual CR data. The genetic signatures identified were summarised as a Venn diagram, as shown in Figure 10.

There was a possible bias between the different ethnic groups, where African-American patients were more likely to be obese compared with the Caucasian patients (13 out of 16 were obese in African-American group, whereas 22 out of 77 were obese in Caucasian group; see Section 1.4.2 and Table 2 in Section 2.2.1). Though ethnicity was controlled in the residual CR data, the effect of ethnicity on the CR data was completely removed to prevent any possibility of ethnicity influencing the analysis. Therefore, the effect of ethnicity was ignored by considering only the Caucasian patients in the CR data, which left a total of 77 Caucasian

Table 9: Summary of the number of DEGs identified using different unadjusted and FDR-adjusted p-value thresholds in different versions of the RMA normalised CR data set

	Number of DEGs identified			
	Unadjusted p-value		FDR-adjusted p-value	
	0.01	0.05	0.01	0.05
CR data	1781	5278	0	9
Residual CR data	1104	4371	0	0
Caucasian-only CR data	2129	6029	0	0
Residual Caucasian-only CR data	1558	5427	0	0

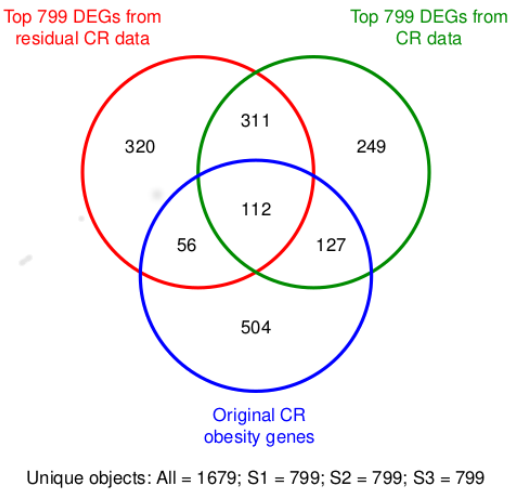


Figure 10: Venn diagram showing the common gene probes between the signatures obtained from the unadjusted and the clinical variable-adjusted CR data and the original obesity associated genetic signature from the Creighton *et al.* (2012) study

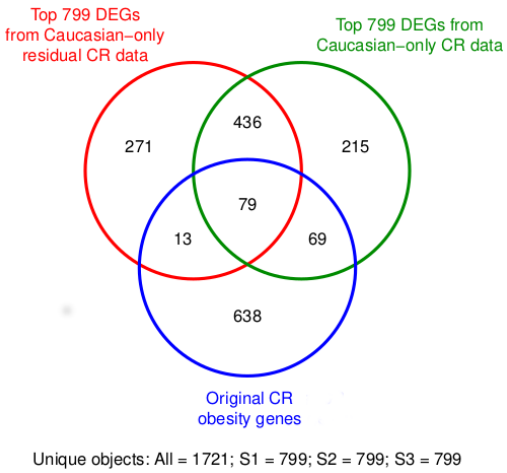


Figure 11: Venn diagram showing the common gene probes between the signatures obtained from the unadjusted and the clinical variable-adjusted CR data (Caucasian patients only) and the original obesity associated genetic signature from the Creighton *et al.* (2012) study

patients in the data set.

With the Caucasian-only CR data set, obesity associated genetic signatures were identified as described above. 2129 and 1558 gene probes were significantly differentially expressed (unadjusted p-value < 0.01) in the unadjusted and clinical variable-adjusted Caucasian patient data, respectively (Table 9). As before, the most significant 799 gene probes were selected from these gene probes. There were 148 and 92 common gene probes with the original Creighton *et al.* obesity associated genetic signatures and unadjusted or clinical variable-adjusted Caucasian patient data set, respectively. Again, Venn diagram was used to summarise the genes identified from the Caucasian patient data set (Figure 11).

Additionally, the correlation of all of the obesity metagenes were examined to see whether these metagenes were similar to one another. As clearly shown in Figure 12, there were two distinct groups within the eight metagenes: the first group contained the obesity metagenes that were not overlapped with the Or metagene (Cr, Res, Ca and CaRes), while the other group had the overlapped metagenes (CrOl, ResOl, CaOl and CaResOl). With that said, all eight metagenes showed high correlation with one another (lowest correlation approximately at 0.85), which suggested that all of these metagenes were detecting similar underlying biological mechanism from the data, as would be expected.

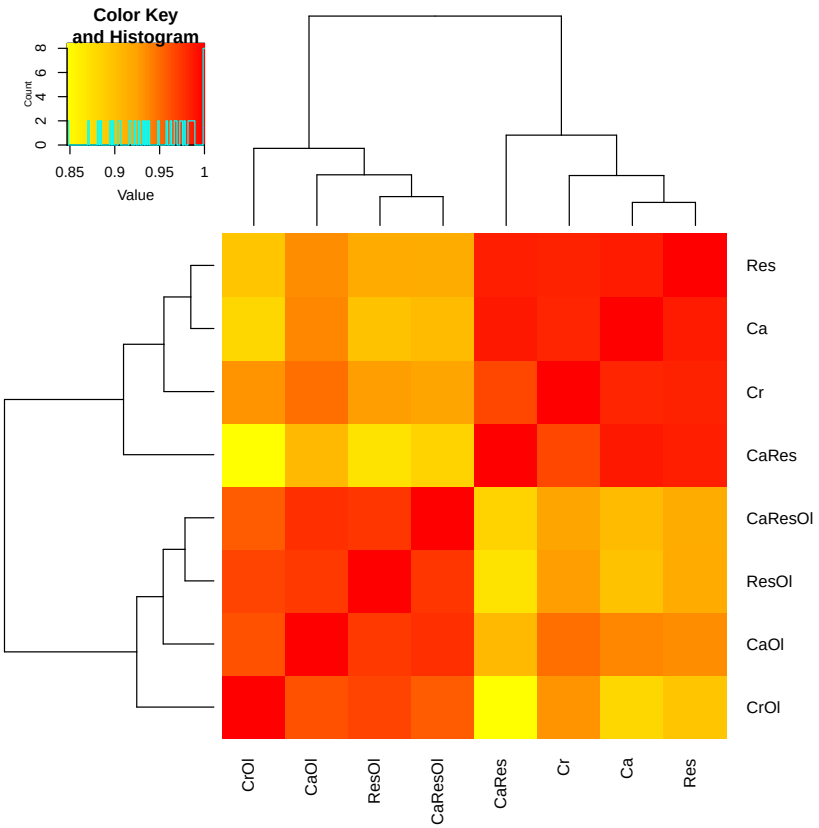


Figure 12: Heatmap showing the Pearson correlation of all eight obesity associated metagenes from the CR data with one another. SVD was applied to the RMA normalised CR data to generate each of the eight metagenes. High and low correlation were represented as red and yellow, respectively, where the colours were matched with the values on the scale shown in the top right histogram. The lowest correlation was 0.8469, between the CrOl and the CaRes metagenes.

Table 10: Summary of the abbreviations used to refer to the different obesity associated genetic signatures identified in the CR data

Abbreviation	Definition	No. of gene probes
Or	Original obesity associated genetic signature identified by Creighton *et al.* (2012)	799
Cr	Obesity associated genetic signature identified from unadjusted CR data	799
CrOl	Genes common between Or and Cr genetic signatures	239
Res	Obesity associated genetic signature identified from clinical variable-adjusted CR data	799
ResOl	Genes common between Or and Res genetic signatures	168
Ca	Obesity associated genetic signature identified from unadjusted Caucasian-only CR data	799
CaOl	Genes common between Or and Ca genetic signatures	148
CaRes	Obesity associated genetic signature identified from clinical variable-adjusted Caucasian-only CR data	799
CaResOl	Genes common between Or and CaRes genetic signatures	92

Since all of these obesity associated genetic signatures were derived based on the discrete values of the patient BMI status, an obesity associated genetic signature was searched for using the correlation of the gene probe expression with the patient BMI value (continuous variable) in the CR data with all of the samples included, but no significant genes were identified (Appendix A8). All of these obesity associated genetic signatures (eight in total) were checked to see whether these signatures had significant association with patient BMI or BMI status in the NZBC and ICGC cancer data sets (Section 3.3.2). For simplicity, the abbreviations in Table 10 will be used to refer to the appropriate genetic signatures.

3.3.2 Novel obesity associated signatures and patient BMI/BMI status

As with the Or signature, all eight obesity associated genetic signatures were validated in the CR data set first, and then compared in other cancer data sets by using the transformation matrix that was generated in the RMA normalised CR data. The CR data (unadjusted, all samples included) was normalised with the RMA method and SVD was applied to generate the all of the obesity metagenes. The direction of the metagenes were examined to make sure that all of the metagenes were in line with one another (see Section 2.6.3; Appendix A4). The comparison of the metagenes with the sample gene expression in the CR data were displayed as heatmaps (Figure 13; Appendix A5). It was clear from the heatmaps that all of the metagenes reflected the overall expression of the corresponding obesity associated genetic signatures. The association of the metagenes with patient BMI and BMI status was significant for all eight of the obesity associated genetic signatures identified (Figure 13; Appendix A5). These results confirmed that all of the obesity associated genetic signatures identified in Section 3.3.1 significantly associated with patient BMI and BMI status in the CR data in which the metagenes were derived from. In the NZBC data, the higher the metagenes scores of the samples were, the more expressed the genes were (and vice versa), but lacked the association with the patient BMI or BMI status (Figure 14; Appendix A6).

To confirm whether these metagenes showed significant association with patient BMI or BMI status in other cancer types, the transformation matrix was applied to the ICGC cancer data. In all of the ICGC cancer data sets, all eight metagene scores were reflective of the sample gene expression of the corresponding obesity associated genetic signatures (Figure 15; Appendix A7). In all cancer types but BLCA, none of the obesity metagenes significantly associated with the patient BMI or BMI status (Appendix A7). As shown in Figure 15, the BLCA data set showed significant association with the Cr metagene with the patient BMI and BMI status. Furthermore, Res and Ca metagenes also showed statistical significance in the overweight group, ANOVA and the regression line p-values; CaRes and Res metagenes were significant in overweight group and ANOVA p-value; CrOl and CaResOl metagenes were significant in overweight group; and ResOl

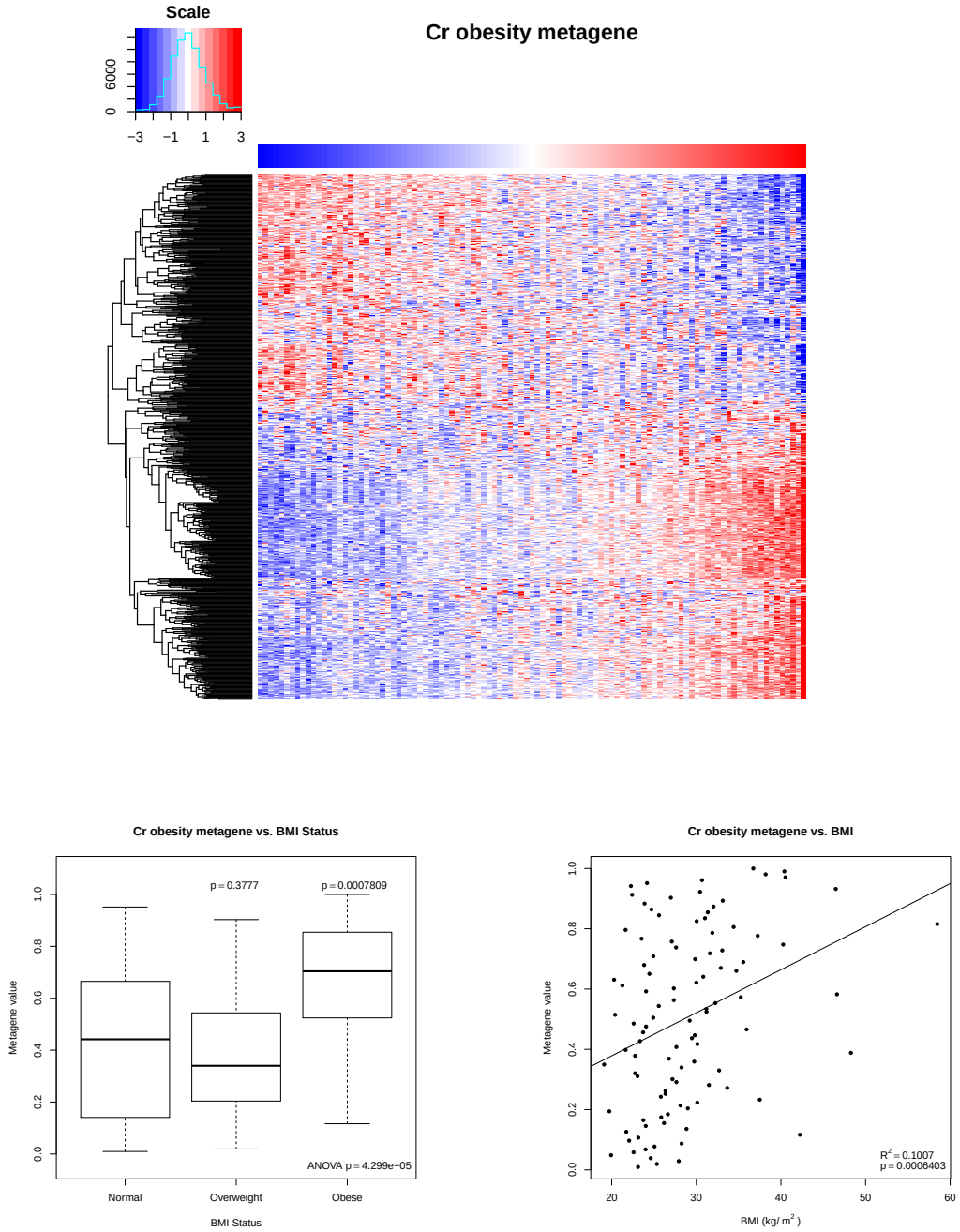


Figure 13: Heatmap, scatter plot and box plot showing the association of the Cr obesity associated metagene with the sample gene expression, patient BMI and BMI status, respectively, from the CR data. The results for the other obesity metagenes are shown in Appendix A5. Scales, p-values and R^2-value are as described in previous figures.

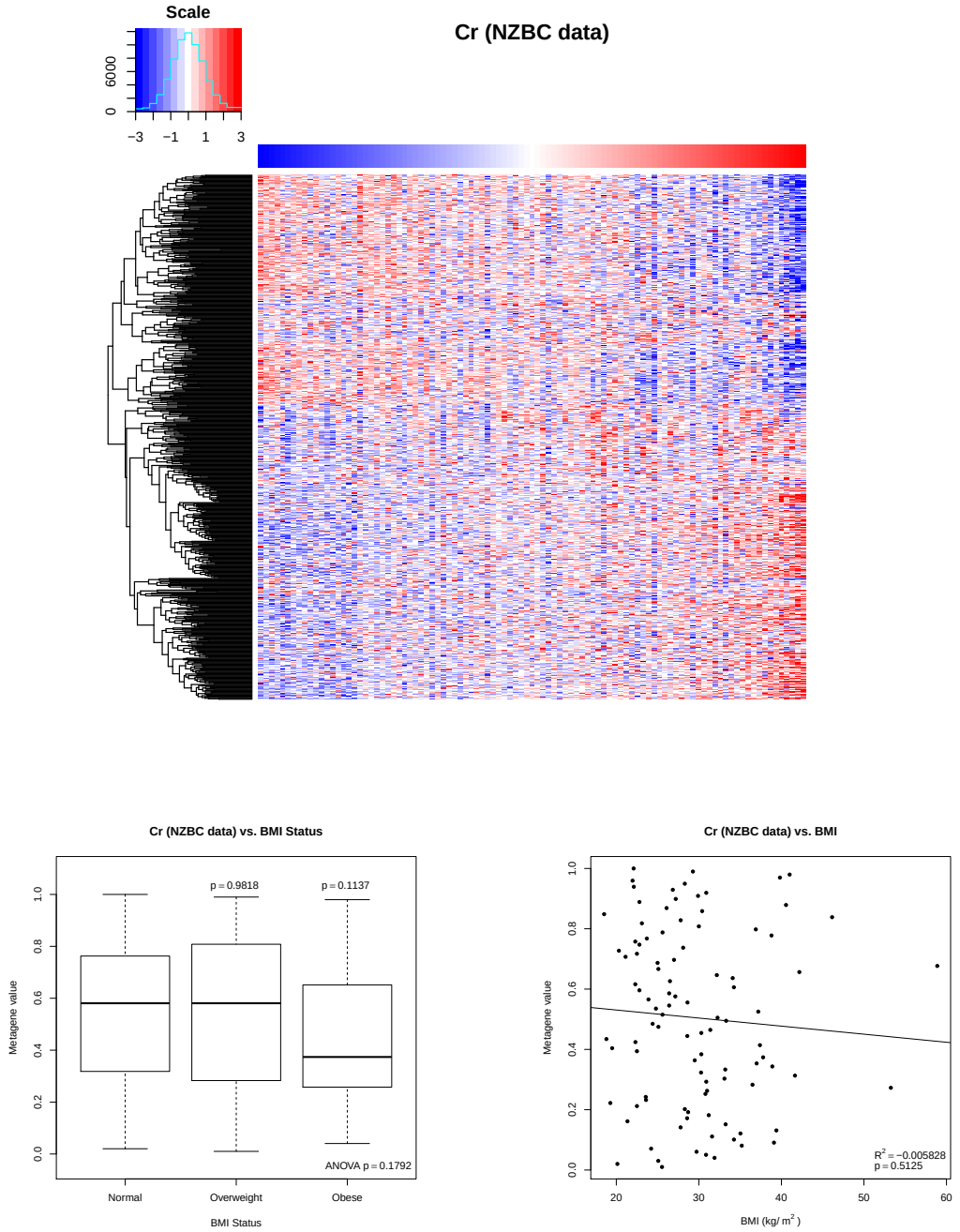


Figure 14: Heatmap, scatter plot and box plot showing the association of the Cr obesity associated metagene with the sample gene expression, patient BMI and BMI status, respectively, from the NZBC data. The results for the other obesity metagenes are shown in Appendix A6. Scales, p-values and R^2-value are as described in previous figures.

and CaOl metagenes were not significantly associated with the sample BMI or BMI status in BLCA data set (summarised in Table 11).

This was unexpected since all of the genetic signatures resulted from the gene expression analysis between the non-obese group and the obese group of the samples in the CR data, and yet all of the obesity metagenes showed significant association with the overweight group rather than the obese group in the BLCA data set. These results suggested that the samples from the patients that were overweight in the BLCA data were similar in genotype with the samples from the patients that were obese in the CR data. Again, due to the fact that all of the metagenes lacked association with the patient BMI and BMI status in all of the other ICGC cancer types, it was difficult to conclude whether the observed association of many of the obesity metagenes with the overweight group of the samples was truly reflective of the effect resulted from these metagenes.

Taken together, these results showed that, even though all of the obesity metagenes significantly associated with the patient BMI and BMI status in the original data where the genetic signatures were derived from, none of the metagenes were not generalisable in other cancer data sets. Furthermore, these results showed that the lack of association with patient BMI and BMI status were not due to other clinical variables in the CR data set. This raised a question of whether there was any obesity associated genetic signature that was common in multiple types of cancer (Section 3.4).

3.4 Common genes across multiple cancer types

It was clear that the all of the obesity associated genetic signatures created from the CR data set were not significantly associated with patient BMI across different ICGC cancer data sets. To determine whether there was any obesity associated genetic signature that was expressed in multiple different cancer types, gene expression analysis was carried out on each of the eight ICGC cancer types, and common genes were searched from the DEGs that were identified. All of the ICGC cancer data sets were normalised with voom (Section 2.3.1), which were then put through the gene expression analysis pipeline to identify the DEGs between the obese and non-obese groups of samples (Section 2.4). Table 12 summarised the number of

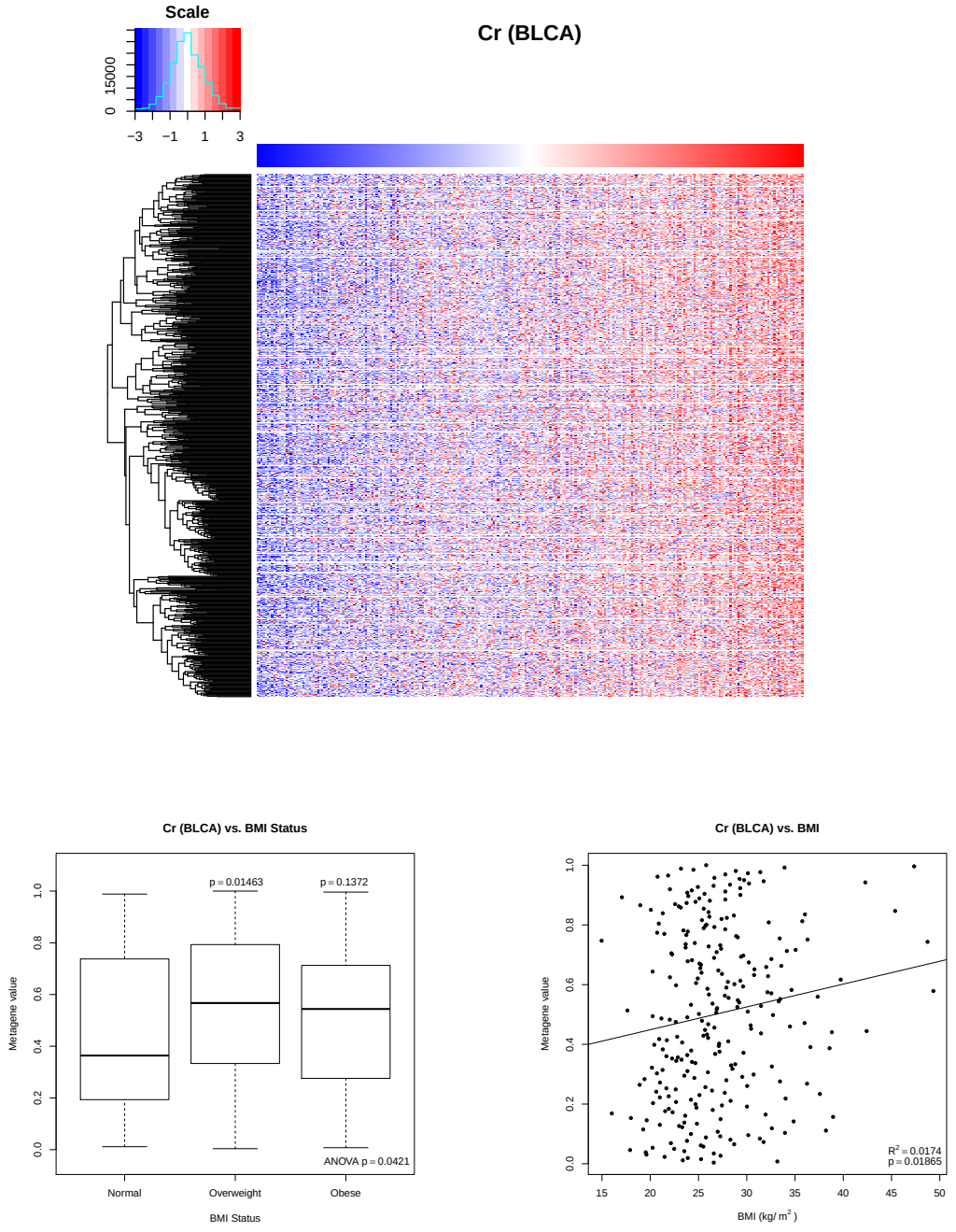


Figure 15: Heatmap, scatter plot and box plot showing the association of the Cr obesity metagene with the sample gene expression, patient BMI and BMI status, respectively, from the ICGC BLCA data. The results for other other metagenes and other ICGC cancer types are shown in Appendix A7. Scales, p-values and R^2-value are as described in previous figures.

DEGs (unadjusted $p < 0.05$) found from each cancer type.

There were 9695 unique DEGs across the eight different cancer types, and these genes were checked for any commonalities across the different cancer types. There were 7330 genes that were differentially expressed in one cancer type, 2024 genes differentially expressed in any two cancer types, 320 genes expressed in any three cancer types, and 21 genes expressed in any four cancer types. There were no genes differentially expressed by five or more cancer types (see Table 13 for a summary). To confirm that these results were statistically significant, the gene expression analysis was repeated 1000 times for each cancer type after the samples were randomised in each analysis (see Section 2.3.5). The results from the simulation were summarised together with the earlier results in Table 13.

The results from the simulation showed that, on average, 5732 genes were found to be differentially expressed in any single cancer type, 1057 in any two cancer types, 111 in any three cancer types, and 7 in any four cancer types. Unfortunately, there were no DEGs expressed in all eight cancer types, which confirmed that there was no common genes that were differentially expressed between the samples that were obese and normal weight. When the results from the ICGC gene expression analysis were compared with the simulation results, the number of DEGs found were statistically significant, as the numbers of genes identified exceeded the 95^{th} percentile values for up to four cancer types. This result confirmed that the DEGs from the ICGC cancer data were not identified by chance. However, this also showed that many of the DEGs identified in the gene expression analysis of the cancer types were observed by chance, and that the majority of these genes were likely to be false positives.

3.5 Pathways enriched in ICGC data sets

To check whether there were any significant pathways enriched in any of the ICGC cancer data sets that were associated with obesity, pathway enrichment analysis was carried out on each cancer type separately and then all cancer types combined (Section 2.5). Each cancer type was normalised with voom (Section 2.3.1) and pathway enrichment analysis was carried out as described in Section 2.5.

When the pathway enrichment analysis was carried out with the KEGG path-

Table 11: Statistics of all the obesity metagenes with the patient BMI and BMI status in the ICGC BLCA cancer data

	P-values			Regression line statistics	
Metagenes	Overweight	Obese	ANOVA	R^2	P
Cr	**0.0134**[1]	0.1365	**0.0387**	0.0171	**0.0195**
Res	**0.0055**	0.1974	**0.0196**	0.0117	**0.0446**
CrOl	**0.0383**	0.4583	0.1096	0.0047	0.1374
ResOl	0.0909	0.7092	0.2125	0.0018	0.2254
Ca	**0.0077**	0.1973	**0.0231**	0.0116	**0.0456**
CaRes	**0.0104**	0.2712	**0.0322**	0.0101	0.0572
CaOl	0.0575	0.6185	0.1487	0.0022	0.2126
CaResOl	**0.0463**	0.5820	0.1263	0.0036	0.1660

[1] All values in bold are statistically significant ($p < 0.05$).

Table 12: Summary of the number of DEGs identified in each of the ICGC cancer data set

Cancer type	No. of DEGs identified
BLCA	679
CESC	1229
COAD	974
KIRP	687
LIHC	3340
READ	796
SKCM	1137
UCEC	2934

Table 13: Summary of the number of DEGs identified by the gene expression analysis and simulation analysis in the ICGC cancer data

	No. of cancer types expressing the DEGs[1]							
	1	2	3	4	5	6	7	8
Results from gene expression analysis	7330	2024	320	21	0	0	0	0
Results from the simulation:[2]								
Mean no. of DEGs identified	5732	1057	111	7	0	0	0	0
95^{th} percentile	6965	1722	227	20	2	0	0	0

[1] The numbers represent the number of cancer types a gene was expressed in.

[2] The simulation was repeated 1000 times, each with randomised samples.

way database, only the "ABC transporter" pathway was significantly enriched (FDR-adjusted p-value < 0.05) in the CESC data, and no other pathways were enriched in any of the other ICGC cancer types (Appendix A9). With the Reactome database, "Phosphatase bond hydrolysis by NUDT proteins" (in BLCA) and "Mitochondrial ABC transporters" (in CESC) were significantly enriched (Appendix A9). With the GO database, there were 22 GO terms significantly enriched in BLCA, 17 terms in CESC, 3 terms in KIRP, 21 terms in READ, 10 terms in SKCM, 14 terms in UCEC, and no terms were significant in the COAD and LIHC cancer types (Appendix A9).

Though there were many GO terms identified as significantly enriched with the GO database in each of the cancer types, there were no terms that were common across all of the different cancer types. Furthermore, these terms were not similar in terms of the biological activities involved. For example, GO terms enriched in the BLCA data suggested possible activation of the PI3K pathway, but those enriched in the READ data were mainly involved in RNA processing (Appendix A9). One interesting result from this enrichment analysis was that the "ABC transport" pathway was identified as significantly enriched in the CESC data set by all three databases, suggesting that the "ABC transport" pathway may be a core component in the CESC tumour biology. All of these results indicated that the biological processes that drive tumour progression were unique to each cancer type, and no one pathway was associated with obesity and tumour biology across multiple cancer types.

Since there were no common pathway enriched in the ICGC cancer types

when analysed individually, all of the ICGC cancer data sets were combined into a single data set to see any enriched pathways could be identified in a collective analysis. The combined data set was generated in two ways. The first combined data set was created by normalising (via voom) each cancer data set separately, combine all of the normalised cancer data sets into one, then applying batch correction to the data set (Section 2.3.4). The second combined data set was created by combining the cancer data sets into one big data set first, then the combined data was normalised, and finally batch correction was applied to obtain the final data set. Each of the combined data sets were analysed for enriched pathways, but there were no pathways significantly enriched in either data sets. These results suggested that there were no common biological pathways associated with obesity in the ICGC cancer data sets.

Chapter 4

Obesity associated genetic signatures and pathway signatures

In this chapter, the underlying biological mechanism of the obesity associated signatures from the CR data were investigated. In doing so, the pathway genetic signatures from the Gatza *et al.* (2010) study (GT) were utilised to determine which biological pathways the obesity associated genetic signatures were most similar to. First, the direction of GT pathway associated genetic signatures were resolved, then the pathway associated metagenes were compared with the obesity associated metagenes, and lastly linear models were constructed based on the pathway metagenes and patient BMI to predict the obesity associated metagenes.

4.1 Pathway associated genetic signatures from the Gatza *et al.* (2010) study

4.1.1 Ranking and normalisation methods for pathway metagenes

In the Gatza *et al.* (2010) study, their data comprised of samples from different studies which were aggregated into a single data set that was MAS5 normalised, and the metagene scores were transformed to a scale between 0 to 1 using a probit function. However, the analyses so far have used the RMA normalisation method and ranked based on the number of samples present in the data (fractional ranking;

Section 2.6.2). To decide which normalisation methods or ranking approaches were suitable for the analysis, the different methods were compared in the GT data (Appendices B1 and B2). Results shown in Appendix B1 clarified that there was no substantive difference in the ranking approaches used.

Scatter plots were generated and the correlations were calculated for each of the GT pathway metagenes derived from either RMA or MAS5 normalised GT data (Appendix B2). The GT pathway metagenes were generated using the transformation matrices that were derived from either the RMA or MAS5 normalised GT data. From these results, it was clear the normalisation methods used on the data set had the most significant effect on the resulting metagenes, rather than the normalisation methods used to generate the transformation matrices (Appendix B2). This meant that all of the data sets had to be normalised with a single normalisation method, so RMA was chosen as the normalisation method to be used as this method is regarded as being more reliable than the MAS5 method (Irizarry *et al.*, 2003).

Another thing noted from these scatter plots was that there were some pathway signatures that were more variable than the others (Appendix B2). As an example, the TGFβ metagenes were significantly more dispersed compared to the PR metagenes, even though both of these metagenes were generated in the same data set through similar processes (Figure 16). These differences were seen in other data sets as well (Figure S22 in Appendix B2). This result provided evidence that some of the pathway genetic signatures from the Gatza *et al.* (2010) study were less variable across different data sets than the others. Taken together, it was decided that the GT data set should be batch corrected first (Section 2.3.4), then normalised with the RMA method, and the metagene scores were ranked with fractional ranking (Section 2.6.2).

4.1.2 Pathway metagene directionality

Before GT pathway metagenes were compared with the obesity associated metagenes, the direction of the GT pathway metagenes had to be checked to make sure that the metagenes were in the "correct" direction (Section 2.6.3). The correlation of all the pathway metagenes with one another were plotted as a heatmap in Fig-

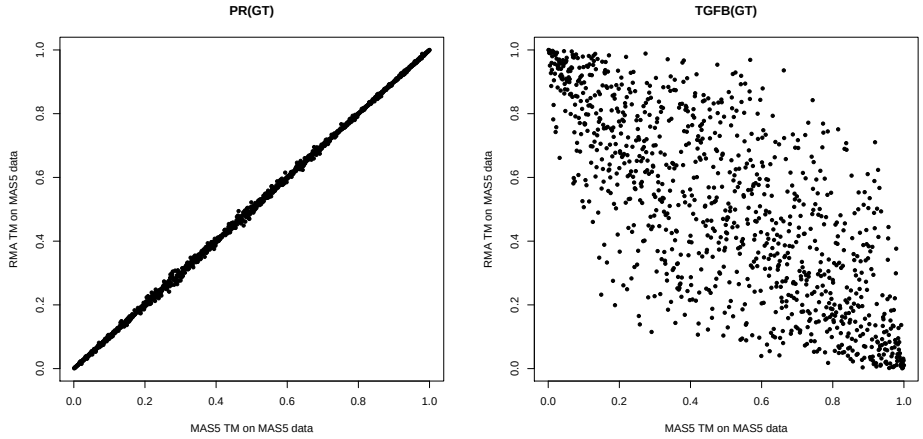


Figure 16: Comparison of the pathway metagenes generated in the MAS5 normalised GT data set from the application of the transformation matrices derived from either the RMA or MAS5 normalised GT data. Only the PR (left) and the TGFβ (right) pathway metagenes are shown. The scatter plots for the other pathway metagenes are shown in Appendix B2.

ure 17. The most prominent group had five pathways (E2F1, PI3K, Myc, BCAT and Ras) that clustered at the top right hand corner of the heatmap. Other groups included IFNα/IFNγ/TNFα pathways, ER/PR/p53 pathways and p63/HER2 pathways. In addition to these highly correlated groups, STAT3/TGFβ/Src/EGFR/Akt pathways showed little correlation with one another. Comparing these groups with the results presented by Gatza *et al.* (2010) (Appendix B4), the identified clusters approximately resembled the pathway groups identified in by Gatza *et al.*, which confirmed that the directions of GT pathway metagenes were consistent with those used by Gatza *et al.*.

To see whether the directionality of the GT pathway metagenes were being transferred across to other data sets properly, pathway metagenes were generated in other data sets with the same transformation matrices and the groupings of the metagenes were examined. CR, FM and NZBC data were RMA normalised and transformation matrices of the pathway genetic signatures (derived from the GT data set) were applied to the data sets. The metagenes were plotted in a heatmap (shown in Appendix B5), which showed similar groupings as seen in Figure 17. This result confirmed that the pathway metagenes were acting similarly in all of the data sets in which the transformation matrices have been applied to.

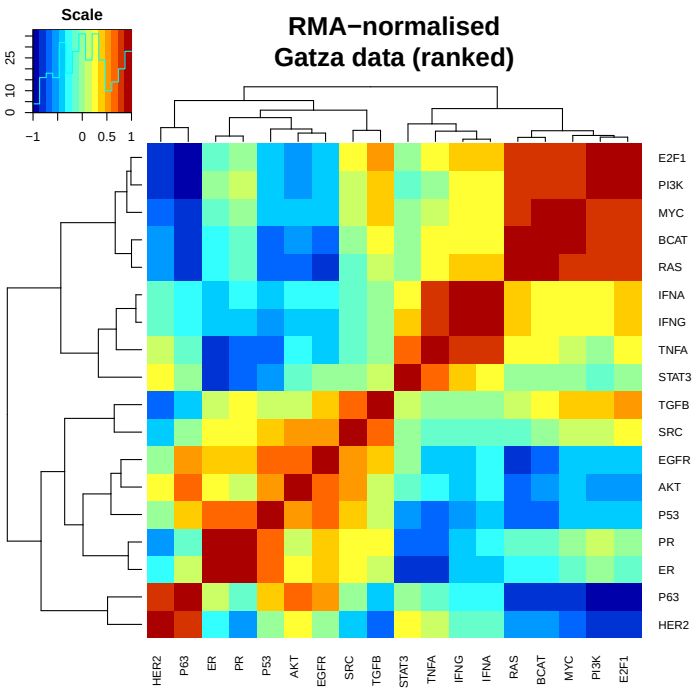


Figure 17: Heatmap showing the Pearson correlation of all the pathway metagenes from the Gatza *et al.* (2010) study with one another in the RMA normalised GT data set. High and low correlation were represented as red and blue, respectively, where the colours were matched with the values on the scale shown in the top left histogram. Refer to Table 7 for the detailed descriptions of the abbreviations used.

4.2 Pathway associated metagenes and obesity associated metagenes

Results from the previous section confirmed that the pathway metagenes from the Gatza *et al.* (2010) study were in the correct directions, and that the pathway metagenes were behaving as expected in the different data sets. Now that the directions of the pathway metagenes were established, these metagenes were ready for comparison with the obesity associated metagenes. However, all of the pathway metagenes were derived from the GT data, whereas the majority of the obesity associated metagenes were derived from the CR data, which presents a problem in terms of deciding which data set the transformation matrices should be generated in.

4.2.1 Obesity and pathway metagene transformation matrices

Metagenes generated from the application of SVD and metagenes generated from the transformation matrix are exactly the same if the transformation matrix is derived from the same data set. For example, an obesity associated metagene generated from the CR data with SVD would have the same values as the metagenes generated in the CR data with the transformation matrix (since the transformation matrix was derived from the CR data). Furthermore, if the SVD-derived metagenes and transformation matrix-derived metagenes were the same (or at least similar) in a different data set, this suggests that the metagene-related biology contained in the two data sets is highly similar. In other words, the transformation matrix can be made in either the original data set or in a different data set if there was no difference in the SVD-derived or transformation matrix-derived metagenes.

To decide which data set the transformation matrices for obesity and pathway associated genetic signatures should be generated in, the SVD- and transformation matrix(TM)-generated metagene scores were compared in all of the data sets. As in Section 4.1, all data sets were normalised with the RMA method and metagenes were ranked with fractional ranking. TMs for the obesity associated genetic signatures were made in the CR data set and the TMs for pathway associated genetic signatures were made in the GT data set. Metagenes for all of the obesity and pathway associated genetic signatures were generated in all of the data sets with SVD and TMs. The Spearman correlation of the SVD-generated and TM-generated metagene scores were calculated for each genetic signatures (Tables 14 and 15).

In Tables 14 and 15, the obesity and pathway associated genetic signatures had a correlation of 1 in CR and GT data sets respectively, as the obesity and pathway TMs were generated in the CR and GT data sets, respectively. The correlations of the obesity and pathway metagenes were variable across different data sets, which was as expected since all the data sets were different from one another. However, what was unexpected was the fact that there were some pathway genetic signatures that were highly correlated in all of the data sets, and there were others that varied considerably across different data sets. For example, the BCAT path-

Table 14: Summary of the Spearman correlations of the SVD- and transformation matrix[1]-derived pathway metagenes in the GT, CR, NZBC and FM data sets

	GT	CR	NZBC	FM
Akt	1.000	0.5190	0.5900	0.5563
BCAT	1.000	0.9897	0.9977	0.9905
E2F1	1.000	0.9646	0.8438	0.9193
EGFR	1.000	0.0430	0.3358	0.4040
ER	1.000	0.9978	0.9942	0.9966
HER2	1.000	0.9553	0.5817	0.9794
IFNα	1.000	0.9830	0.9991	0.9951
IFNγ	1.000	0.9086	0.9950	0.9718
Myc	1.000	0.9878	0.9852	0.9689
p53	1.000	0.3808	0.9981	0.7923
p63	1.000	0.8319	0.2951	0.8368
PI3K	1.000	0.9543	0.5989	0.9365
PR	1.000	0.9511	0.9887	0.9845
Ras	1.000	0.9078	0.9125	0.7229
Src	1.000	0.9575	0.7173	0.9548
STAT3	1.000	0.1902	0.9159	0.6167
TGFβ	1.000	0.9918	0.2543	0.9890
TNFα	1.000	0.6046	0.9365	0.4315

[1] Transformation matrices were derived from the RMA normalised GT data set.

Table 15: Summary of the Spearman correlations of the SVD- and transformation matrix[1]-derived obesity metagenes in the GT, CR, NZBC and FM data sets

	CR	GT	FM	NZBC
Cr	1.000	0.9985	0.9872	0.9161
Res	1.000	0.9987	0.9898	0.9652
CrOl	1.000	0.9982	0.9926	0.9715
ResOl	1.000	0.9981	0.9927	0.9571
Ca	1.000	0.9985	0.9893	0.9468
CaRes	1.000	0.9988	0.9939	0.9865
CaOl	1.000	0.9983	0.9937	0.9677
CaResOl	1.000	0.9984	0.9952	0.9642
Original	1.000	0.9928	0.9862	0.9344

[1] Transformation matrices were derived from the RMA normalised CR data set.

way metagene was highly correlated in all of the data sets (> 0.98), whereas the STAT3 pathway metagene was variable across different data sets, ranging from 0.1902 to 0.9159 (Table 14; Appendix B7). The fact that some pathway associated metagenes were consistent across different data sets suggested that some of these genetic signatures were reliable and did not depend on the data set the transformation matrices were derived from. On the other hand, the genetic signatures that were not consistent across the data sets were likely to be dependent on the data set in which they were derived from, and therefore the transformation matrices for these signatures must be derived from the GT data set. This also showed that the tumour biology may be different across the data sets, likely due to the difference in the patient cohorts that were selected for each of the studies.

In contrast to the pathway associated genetic signatures, all of the obesity associated genetic signatures showed high correlation for the SVD- and transformation matrix-generated metagenes across the different data sets. This suggested that the obesity associated genetic signatures were relatively consistent across all data sets, and the transformation matrices for these signatures could be made in any of the data sets. For simplicity, it was decided that the transformation matrices for all the genetic signatures would be made in the GT data set, since there were some pathway associated signatures that were specific to the GT data.

4.2.2 Comparison of the obesity and pathway associated signatures

To visually determine which pathway associated genetic signatures were most similar to the obesity associated genetic signatures, heatmaps were created with the metagenes for all of the genetic signatures. The directions of the pathway associated metagenes have already been determined in Section 4.1, and the directions of the obesity associated metagenes were checked in GT data set as described in Section 2.6.3. All of the metagenes were created in the GT data set (normalised with RMA) using SVD, and the metagene scores were plotted in a heatmap and clustered into groups (Figure 18).

In Figure 18, all of the obesity associated metagenes from the CR data set cluster together into a group, but none of the pathway associated metagenes had

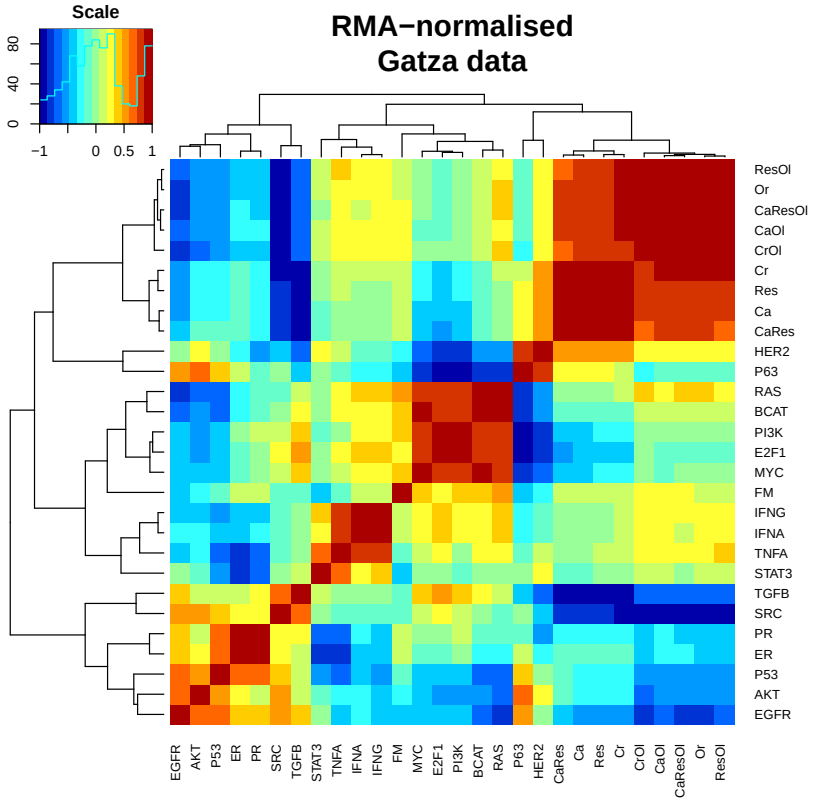


Figure 18: Heatmap showing the Pearson correlation of all the pathway and obesity metagenes with one another in the RMA-normalised GT data set. High and low correlation were represented as red and blue, respectively, where the colours were matched with the values on the scale shown in the top left histogram. Refer to Tables 7 and 10 for the detailed descriptions of the abbreviations used.

a strong positive correlation with the obesity metagenes. Though the HER2 and p63 pathway metagenes were grouped next to the cluster of obesity metagenes, the similarity was at most about 0.6 (Pearson correlation; Figure 18). With that said, these obesity associated metagenes seem to have a strong negative correlation with some of the pathway associated genetic signatures, such as the Akt, ER, EGFR, PR, p53, Src and TGFβ pathways, and this strong negative correlation was observed in the other data sets as well (Appendix B8). In fact, high Or metagene scores related to lower levels of gene expression for the genes in the Or signature (Figure 3; Section 3.1), and therefore it could be possible that many of the genes in the Or signature (and the other obesity associated signatures from the

CR data) may be positively correlated with the genes in these pathway signatures. This suggests that the obesity associated genetic signatures from the CR data may be related to these pathway signatures, rather than the obesity phenotype of the patients.

The FM obesity associated metagene did not cluster together with the obesity metagenes from the CR data set, nor with any of the other pathway metagenes. The fact that the FM metagene did not cluster with any of the CR metagenes showed that the signals captures by the FM metagene was different from those metagenes found in the CR data set. Furthermore, the FM metagene did not cluster with the Akt pathway in the heatmap, even though Fuentes-Mattei *et al.* (2014) provided clear evidence that their metagene specifically activated the Akt/mTOR signalling pathway in a mouse model. This may have been due to the lack of consistency of the Akt pathway signature in the different data sets (Table 14). Another likely reason could be that FM genetic signature was not comprised only of the genes from the Akt pathway, but genes from the other pathways as well, and so the metagene did not cluster together with the Akt pathway signature in the heatmap.

The results in this section suggested that none of the obesity associated genetic signatures were positively correlated with any of the pathway associated genetic signatures from the Gatza *et al.* (2010) study. However, some pathway signatures were negatively correlated with the obesity associated genetic signatures, which meant that the obesity signatures were likely capturing the signals from these signatures rather than the biological signals from the obesity phenotype of the patient.

4.3 Prediction of obesity associated metagenes with pathway associated metagenes

To confirm the biological relationship between the obesity associated metagenes with the pathway associated metagenes from the Gatza *et al.* (2010) study, linear models were created to predict the obesity metagene scores with the pathway metagene scores. If any of the pathway metagenes were significant in the linear model, this would provide evidence that the significant pathways in the linear

model were related to the obesity metagene. Since most of the obesity metagenes were found in the CR data set, and only the CR and the NZBC data sets had BMI information for the patients, NZBC data set was used as the training data set to create the linear models for obesity metagene score prediction.

4.3.1 Linear model prediction in the NZBC and CR data sets

Metagene scores used for the construction of linear models were generated from the application of the transformation matrices derived from the GT data (from Section 4.2) to RMA normalised NZBC data. From Table 14, it was evident that some pathway associated genetic signatures could be variable across different data sets. Therefore, only six pathways (BCAT, ER, IFNα, IFNγ, Myc and PR) that were "consistent" (correlation > 0.9) across all of the data sets were used to construct the linear models.

For each of the obesity metagenes identified, seven linear models were created: BMI-only, BMI status-only, BMI and BMI status, pathway metagenes-only, and the combinations of pathway metagenes with BMI and/or BMI status. The only variables that showed significance in any of the models that predicted the obesity metagenes taken from the CR data set were the PR metagene and whether the patients were obese or not (Table 16). In the model that predicted the FM metagene scores, the obesity status of the patients and Myc metagene score were significant, but the PR metagene score was not (Appendix B9).

The models constructed from the patient BMI, BMI status and/or GT pathway metagene scores confirmed that none of the pathways were significant, except for the PR and Myc pathway metagene scores in the models for the CR metagenes and the FM metagene, respectively. It was not surprising that the PR pathway metagene showed significance in the models generated in NZBC data set, as majority of the patients in the NZBC data set were PR^+ (Table 3). The significance of the Myc metagene in the model for FM metagene was difficult to interpret, as there was no reported association of the FM metagene with the Myc pathway by Fuentes-Mattei *et al.* (2014). *MYC* is an oncogene that encodes the Myc transcription factor that, when overexpressed, disrupts various pathways that directly promote tumour growth and proliferation, and has a central role in many of the cancer

Table 16: Description of the linear models constructed from the NZBC data to predict the Cr obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	-0.0016	0.717
BMI status only	Overweight	-0.0126	0.871
	Obese	-0.0981	0.166
BMI and BMI status	BMI	0.0109	0.123
	Overweight	-0.0687	0.420
	Obese	-0.2481	**0.040**[1]
Pathways only	BCAT	0.0916	0.677
	ER	0.0523	0.834
	IFNα	-0.4993	0.307
	IFNγ	0.4224	0.398
	Myc	-0.0038	0.987
	PR	0.5836	**0.016**
BMI and Pathways	BMI	-0.0024	0.527
	BCAT	0.1121	0.614
	ER	0.0540	0.829
	IFNα	-0.5383	0.276
	IFNγ	0.4661	0.358
	Myc	0.0168	0.941
	PR	0.5927	**0.015**
BMI status and Pathways	Overweight	-0.0120	0.864
	Obese	-0.1035	0.113
	BCAT	0.1495	0.501
	ER	0.0326	0.896
	IFNα	-0.6707	0.177
	IFNγ	0.5888	0.245
	Myc	0.0439	0.846
	PR	0.5799	**0.016**
BMI, BMI status and pathways	BMI	0.0091	0.161
	Overweight	-0.0576	0.455
	Obese	-0.2292	**0.040**
	BCAT	0.1460	0.509
	ER	0.0091	0.971
	IFNα	-0.6937	0.161
	IFNγ	0.5890	0.243
	Myc	0.0263	0.907
	PR	0.5449	**0.024**

[1] All values in bold are statistically significant ($p < 0.05$).

Table 17: Description of the linear models constructed from the NZBC data to predict the Cr obesity, using only the patient BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	0.477	$\mathbf{6.06 \times 10^{-7}}$[1]
BMI and PR	BMI	-0.002	0.586
	PR	0.478	$\mathbf{6.35 \times 10^{-7}}$
BMI status and PR	Obese	-0.085	0.176
	Overweight	-0.011	0.875
	PR	0.471	$\mathbf{8.25 \times 10^{-7}}$
BMI, BMI status and PR	BMI	0.007	0.237
	Obese	-0.188	0.081
	Overweight	-0.049	0.516
	PR	0.459	$\mathbf{1.54 \times 10^{-6}}$

[1] All values in bold are statistically significant ($p < 0.05$).

hallmarks (Coller *et al.*, 2000; Hanahan and Weinberg, 2000). "Sustained cell proliferation" was one of the cancer hallmarks that Fuentes-Mattei *et al.* (2014) had shown to be related to their obesity associated genetic signature, which could be the reason that the Myc pathway metagene showed up as a significant variable in the linear model. Since the PR pathway metagene showed significant association in many of the linear models, linear models were made with each of the combinations of BMI, BMI status and PR pathway metagene for all of the obesity associated metagenes (Table 17; Appendix B9). As expected, all of the linear models showed significant contribution of the PR metagene scores to the model for all obesity metagenes (except FM metagene; Appendix B9).

To determine whether any of the linear models were able to predict the corresponding obesity metagene scores, the predicted metagene scores were generated from the models and then compared with the original scores. The original obesity metagene scores from the NZBC were generated from the application of TM that were derived from the RMA-normalised GT data set. These NZBC obesity metagene scores were then compared with the metagene scores that were predicted by the linear models (Figure 19).

The predicted metagenes from all of the patient BMI-related linear models were not significantly related to the original metagene scores in many of the obe-

sity metagenes (Figure 19; Appendix B9). On the other hand, all of the predicted metagenes generated from the models that used the pathway metagene scores showed significant association with the original metagene scores (Figure 19; Appendix B9). However, the R^2 values for all of the predicted metagenes against the original metagenes were very low ($R^2 < 0.4$). All of the plots in Figure 19 and Appendix B9 showed clearly that the data points in the scatter plots were highly variable, suggesting that even though the association between the predicted and the true metagene scores were statistically significant, the variables in the linear models may not be strongly associated to the obesity metagene. Furthermore, the fact that all of the predictions made with the models that contained the PR metagene showed greater correlation with the true metagene values suggested that most of the variations in the Cr (and other) obesity metagenes were being explained by the PR metagene alone, and the patient BMI contributed very little to the prediction.

To confirm the results shown in the NZBC data set (the training set), the linear models were applied to CR data (the test set). As shown in Figure 20, all of the predicted Cr metagene generated from the linear models were statistically significantly associated with the true Cr metagenes in CR data, but there were some exceptions in the models for the other obesity metagenes (Appendix B11). Similar to the results from the NZBC data set, all of the R^2 values were low ($R^2 < 0.43$) in the CR data set and the data points were widely spread out in the plots. Again, these results suggested that the predicted metagene scores from the linear models may be statistically significant, but may not be relevant to the obesity metagenes which the models predict, as the predicted values were only weakly associated with the true values.

Another thing to note from Figure 20 was the apparent negative association of the predicted metagenes with the original metagenes in the BMI-only, the BMI status-only and the BMI and BMI status models. One possible explanation for this was that these BMI-based models were overfitted in the NZBC data set, and was not generalisable in the CR data set, and thus produced negative correlations. In fact, it was evident from the metagene analyses in Chapter 3 that these obesity associated genetic signatures were not generalisable with the patient BMI in different data sets, and overfitting of the model in the NZBC data was the likely cause of this negative correlation.

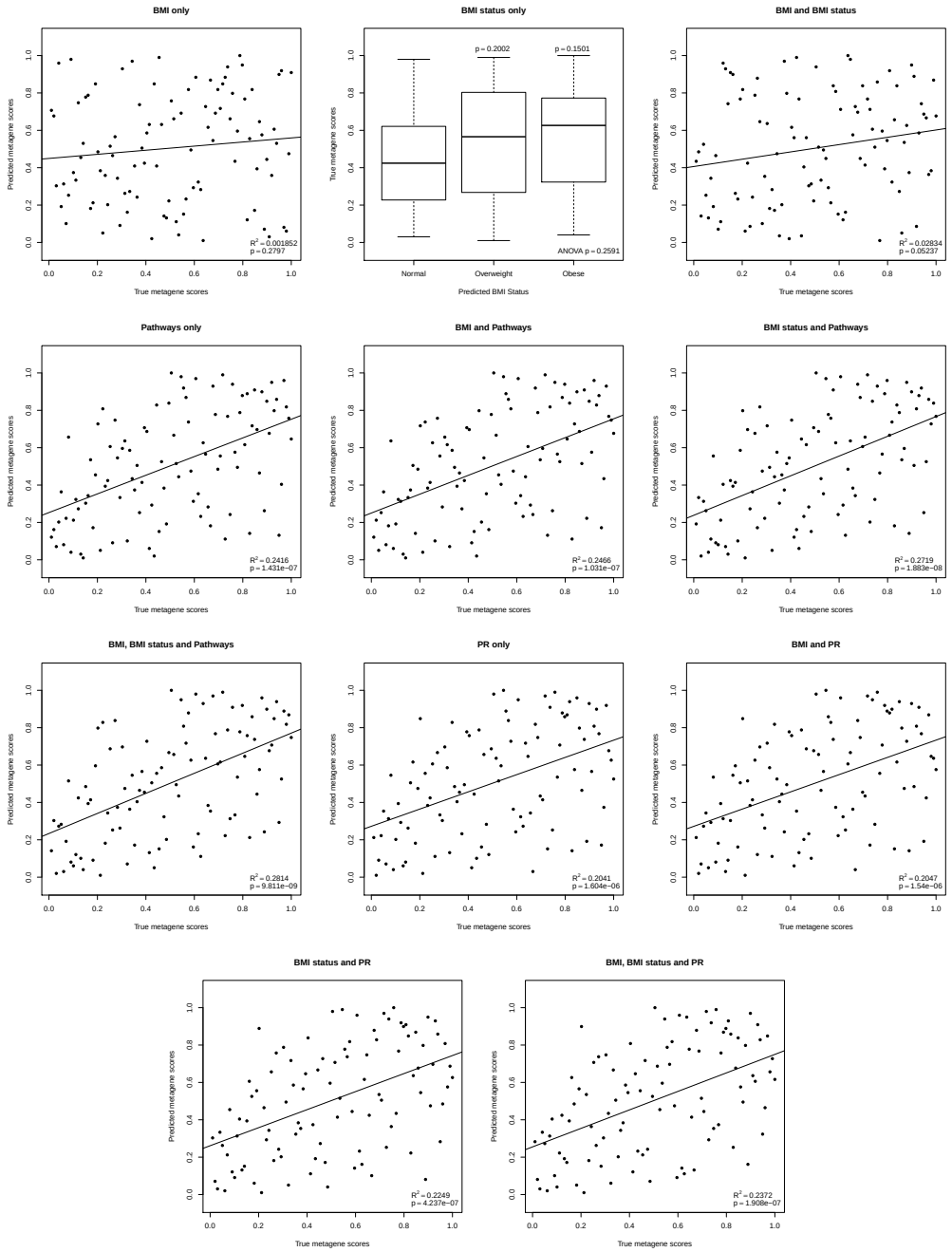


Figure 19: Box plot and scatter plots comparing the Cr metagene score predicted by the different linear models constructed from the NZBC data, with the Cr metagene score from the NZBC data. Only the results from the models used to predict the Cr metagene scores are shown. The summary of the statistics for the other obesity metagene predictions from the NZBC data are shown in Appendix B11. P-values and R^2-value are as described in previous figures.

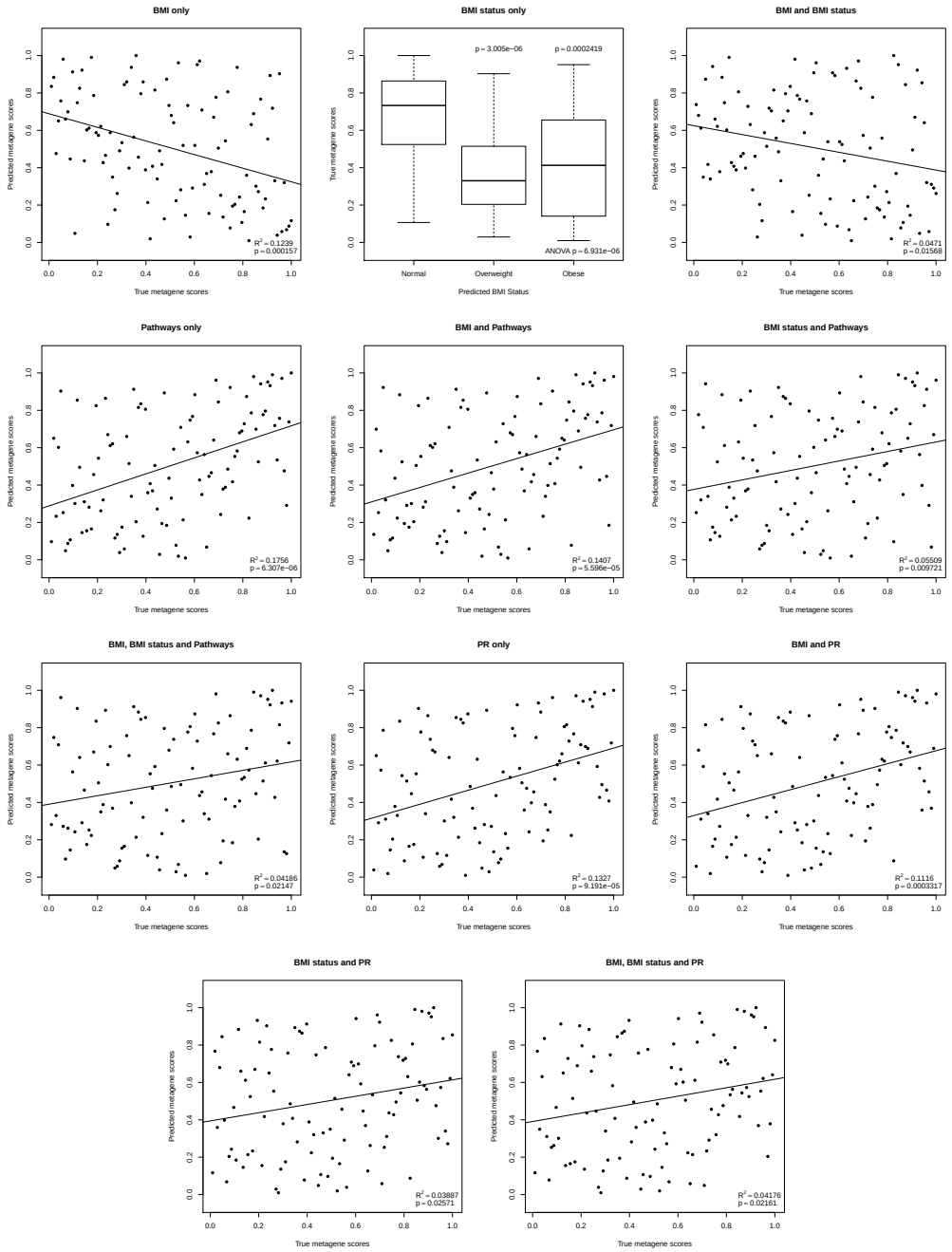


Figure 20: Box plot and scatter plots comparing the Cr metagene score predicted by the different linear models constructed from the NZBC data, with the Cr metagene score from the CR data. Only the results from the models used to predict the Cr metagene scores are shown. The summary of the statistics for the other obesity metagene predictions from the CR data are shown in Appendix B11. P-values and R^2-value are as described in previous figures.

4.3.2 Stepwise linear model prediction in NZBC and CR data sets

Lastly, since only the six most "consistent" pathway metagenes were picked from a total of eighteen pathway metagene scores, linear models were constructed with all eighteen pathway metagenes in a stepwise fashion. Step wise construction of a linear model is where the variables for a model are selected computationally depending on the contribution of the variable to the resulting linear model, and thus contain only the variables that "matter" to the final linear model Section 2.7. Patient BMI, BMI status and all of the pathway metagenes (Akt, BCAT, E2F1, EGFR, ER, HER2, IFNα, IFNγ, Myc, p53, p63, PI3K, PR, Ras, STAT3, Src, TGFβ and TNFα) were used to generate the stepwise linear model for each of the obesity metagenes. The linear models that resulted from the stepwise selection of the variables were summarised in Table 18.

These models were used to predict the obesity metagenes in the NZBC data to test the performance of the models. All of the models were able to make statistically significant prediction of the original obesity metagenes in the NZBC data set (Figure 21). Unlike the previous models, the stepwise models showed very high R^2 values, where most predictions had $R^2 > 0.80$. The statistical significance and the strong association of the predicted metagenes were observed in the CR data set as well (Figure 22). Though the R^2 values were slightly lower than those from the NZBC data set, the R^2 values from the CR data predictions were more promising than the predictions made by the non-stepwise linear models.

Table 18: Description of the stepwise linear models constructed from the NZBC data to predict all of the obesity metagenes

Obesity metagene	Variables retained in the model	Estimate	P-value
Cr	Src	-0.1900	**0.005150**[1]
	EGFR	0.1451	**0.002780**
	TGFβ	0.5105	$\mathbf{1.17 \times 10^{-15}}$
	Akt	-0.2821	$\mathbf{7.08 \times 10^{-6}}$
	IFNα	-0.06968	**0.02986**
CrOl	Akt	-0.3134	$\mathbf{1.76 \times 10^{-6}}$
	Src	-0.5386	$\mathbf{< 2 \times 10^{-16}}$
	p53	0.3372	$\mathbf{9.18 \times 10^{-7}}$

Table 18 (continued)

	PR	-0.1115	**0.02760**
Res	TGFβ	-0.5895	**1.53×10^{-15}**
	EGFR	-0.3597	**0.0005390**
	TNFα	0.1604	**0.0006970**
	BCAT	-0.2141	**0.01157**
	Akt	0.1527	**0.02206**
ResOl	Akt	-0.2802	**0.0001420**
	Src	-0.5214	**2.65×10^{-15}**
	p53	0.2861	**5.31×10^{-7}**
Ca	TGFβ	0.7470	**$< 2 \times 10^{-16}$**
	Akt	-0.2985	**2.83×10^{-10}**
CaOl	Src	-0.5595	**3.88×10^{-16}**
	Akt	-0.4331	**2.06×10^{-11}**
CaRes	TGFβ	0.7843	**$< 2 \times 10^{-16}$**
	EGFR	0.2418	**2.19×10^{-6}**
CaResOl	Src	-0.4553	**1.57×10^{-7}**
	EGFR	0.2776	**0.0009570**
	Akt	-0.3001	**0.0001640**
	E2F1	0.2484	**0.001300**
	Myc	0.1584	**0.025220**
Or	Src	-0.5165	**$< 2 \times 10^{-16}$**
	p53	0.3072	**2.60×10^{-12}**
	Akt	-0.3006	**1.440e-07**
	E2F1	0.09684	**0.01070**
	Obese	-0.02190	0.2166
	Overweight	0.03099	0.1223
FM	p63	0.3507	**0.0006760**
	STAT3	0.2969	**0.004191**
	p53	0.5368	**0.0008240**
	PR	-0.3462	**0.034070**

[1] All values in bold are statistically significant ($p < 0.05$).

Clearly, the stepwise linear models were able to predict the obesity associated metagenes better than the models created in Section 4.3.1. This could be due to the fact that many of the stepwise linear models included many pathways that had a strong negative correlation with the obesity metagenes (Akt, ER, EGFR, PR, p53, Src and TGFβ pathways; Figure 18). These pathway metagenes may have been better predictors of the obesity metagenes than the six pathways that were "consistent" across the data sets, and therefore generated better predictions than those linear models. With that said, some of the predicted obesity metagenes from the stepwise linear models were negatively correlated with the original metagene values in the CR data (Figure 22). Again, this could have been due to the model being overfit in the NZBC data, and thus making an incorrect prediction about the obesity metagenes in the CR data.

Taken together, these results suggest that the variables included in these stepwise models were relevant to the biology of the obesity metagenes identified in the CR and FM data sets. However, there were no common pathways that were present in all of the models; although Akt, EGFR, TGFβ and Src pathway metagenes were frequently included in many of the stepwise models. The role of the pathways included in the models and the possible biological links of these pathways to the obesity metagenes remain unclear, and should be the focus for future investigations.

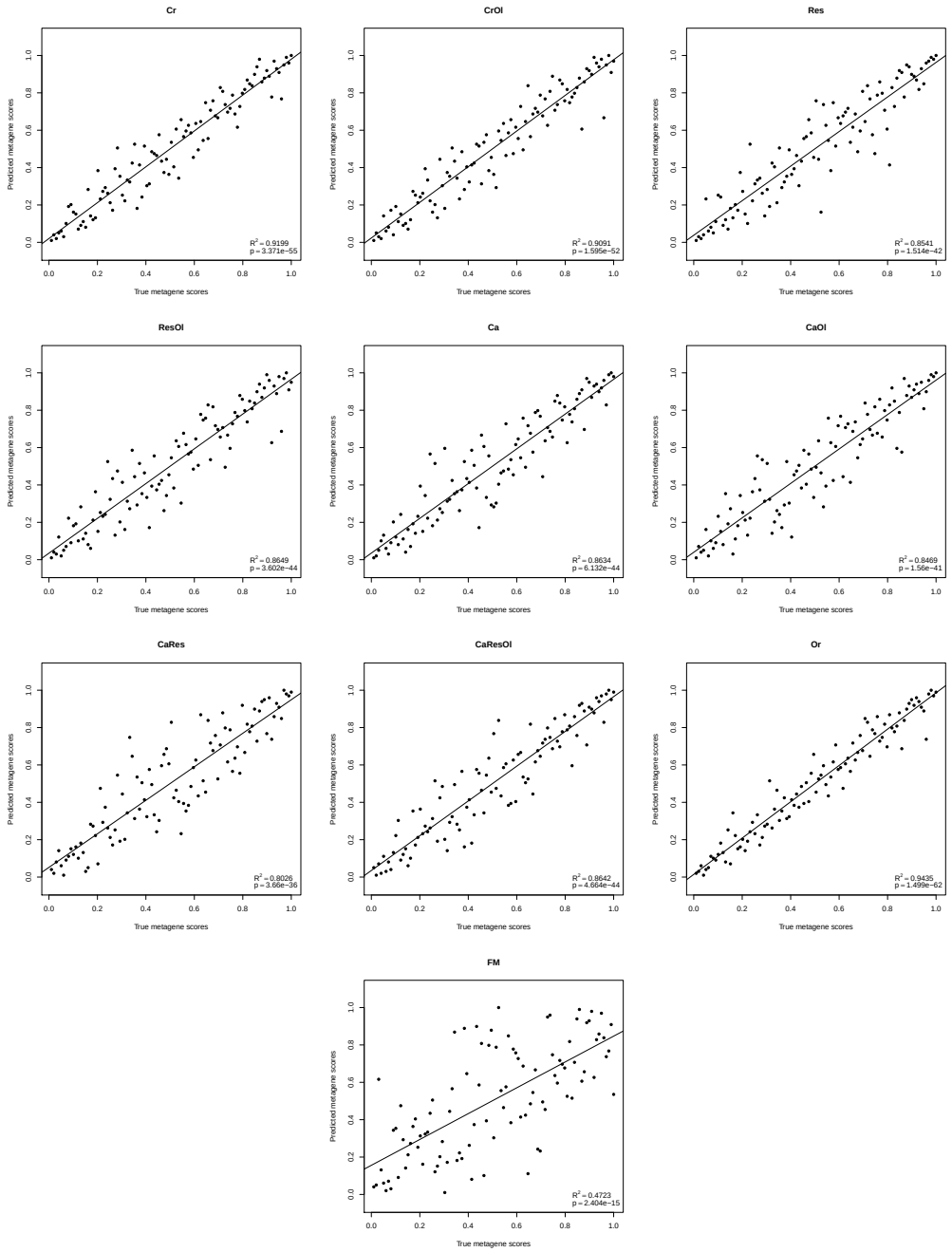


Figure 21: Scatter plots comparing all of the obesity metagene scores predicted by the stepwise linear models constructed from the NZBC data, with the corresponding obesity metagene scores from the NZBC data. P-values and R^2-value are as described in previous figures.

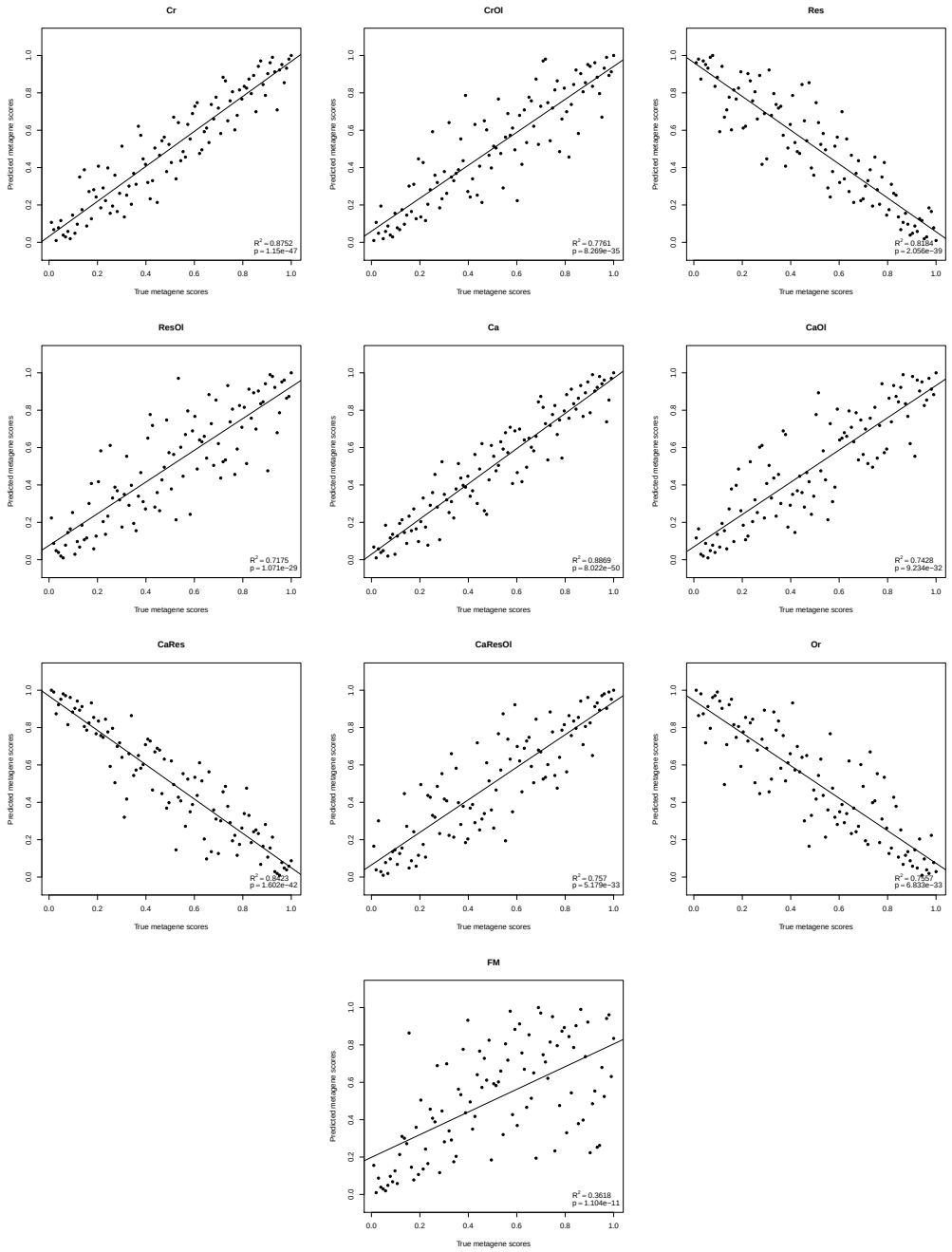


Figure 22: Scatter plots comparing all of the obesity metagene scores predicted by the stepwise linear models constructed from the NZBC data, with the corresponding obesity metagene scores from the CR data. P-values and R^2-value are as described in previous figures.

Chapter 5

Discussion

To conclude this thesis, all of the results were re-examined and are summarised here to highlight the key findings from this project. There were many limitations to this project and these limitations are addressed either in the summary in Section 5.1, or separately in Section 5.2. Lastly, conclusions are drawn from all of the evidence from this project and suggestions for future aims and experiments are made.

5.1 Summary of the results

5.1.1 Obesity associated genetic signatures

Chapter 3 focussed mainly on establishing the link between the obesity associated genetic signatures from the Creighton *et al.* (2012) and the Fuentes-Mattei *et al.* (2014) studies with the patient BMI and BMI status, as reported by their studies. The obesity associated genetic signature from the Creighton *et al.* (2012) study was examined first. The obesity metagene generated from the Creighton *et al.* obesity signature was able to "capture" the overall genetic expression patterns of the samples, where low and high metagene scores corresponded to low and high gene expressions, respectively. This result was in accordance with the characteristics of the obesity metagene provided by Creighton *et al.* (2012). Furthermore, the obesity metagene significantly correlated with the patient BMI and BMI status

in the CR data set.

To see whether this obesity signature was able to show similar association with the patient BMI and BMI status in other cancer data sets, metagenes were created in the NZBC and ICGC cancer data sets with the transformation matrix generated in the CR data set. Like in the CR data set, the obesity metagenes were reflective of the gene expression patterns of the genes in the genetic signature in all of the data sets. However, the obesity metagene was not significantly associated with the patient BMI or BMI status in any of the other cancer data sets, except for the BLCA data set which showed significant association only with the overweight group (discussed later in Section 5.1.2).

Initially, these results were thought to be due to the obesity associated genetic signature being specific only to the CR data set. Therefore, a different obesity associated genetic signature from the Fuentes-Mattei *et al.* (2014) study was used to see whether this signature was able to show significant association with the patient BMI in the CR, NZBC and ICGC cancer data sets. The results were similar to the obesity metagene from the Creighton *et al.* (2012) study; FM obesity metagene scores reflected the gene expressions of the signature but the scores were not significantly associated with the patient BMI or BMI status in any of the cancer data sets investigated (except the BLCA data set). Together, these results suggested that both CR and FM genetic signatures were only significant in the original data set in which the signature was derived in (though this was not confirmed in the FM data set, as no patient BMI information was available).

The fact that both CR and FM obesity associated genetic signatures showed no sign of significant association with the patient BMI in majority of the cancer data sets raised a question of whether these signatures were actually related to obesity, or a different clinical variable within the data set. This question led to the identification of various versions of obesity associated genetic signatures in Section 3.3.1. The CR data set was controlled for all the clinical variables (sex, age, ethnicity, menopause status, tumour grade, ER/HER2/PR statuses and LN status) to generate some of these signatures. All of the metagenes created from these novel obesity associated genetic signatures were consistent with the gene expression patterns, but none of the metagenes showed significant association with the patient BMI or BMI status in the data set other than the CR data set, with the

exception of the overweight group in the BLCA data set.

It was possible that all of these results were due to the sample quality (and hence due to the quality of the data), or perhaps many of the samples in these data sets had breast cancers due to a cause unrelated to their obesity status. As many epidemiological studies have pointed out, obesity is a strong risk factor for cancer (covered in Section 1.3), but this does not necessarily mean that all cancer in obese patients develop cancer as a result of obesity. In other words, the cancer identified in obese patients may have been caused by another, completely different biological mechanism. Thus the tumour samples taken from obese and overweight patients will be a mixture of samples that are driven by obesity-related mechanisms and mechanisms that are not related to obesity. These different subgroups of tumours (obesity-driven or not) cannot be easily identified and isolated. This may affect the detection of the truly obesity associated genes, as some of the tumour samples from the patients will have nothing to do with obesity and cancer, and hence misclassified into wrong groups.

5.1.2 BLCA data set and the obesity metagenes

In the BLCA data set, many of the obesity metagenes were associated with the overweight group with significant p-value and/or ANOVA p-value, but never with the obese group. This was unexpected, as all of the metagenes were generated based on the list of DEGs between the obese and the non-obese groups in the CR data set, and not the overweight group. In addition to the association of the metagenes with the overweight group, some metagenes showed significant association with patient BMI in the BLCA data set. It was difficult to conclude with confidence that those metagenes that showed significant association with the overweight group in the BLCA data set was truly due to the effect of the metagenes, as these metagenes were generated from the obese group and not the overweight group in the original data sets. Furthermore, there were no significant association of these metagenes in any other ICGC cancer data sets and provided no further evidence that supported the results presented in the BLCA data set.

With that said, there is a possibility that the genotypes of the patients that are overweight in the BLCA data set are similar to those patients that are obese in

the CR data set. In fact, Damrauer *et al.* (2014) have identified "basal-like" and "luminal" subtypes in bladder cancer that resembled the corresponding subtypes in breast cancer, and suggested that bladder and breast cancers may have common physiological properties. Since all of the obesity associated genetic signatures were derived from breast cancer data sets, it made sense that these signatures were significant in the BLCA data set, as bladder cancer and breast cancer could have been biologically similar to one another. However, the degree of similarity between these two cancer types are not yet known and further analyses are required for a definitive explanation of the results.

5.1.3 Use of SVD and transformation matrix as a valid method

From all of these results so far, one may question how robust or appropriate the use of SVD and transformation matrices are to generate metagenes in other data sets. It appeared as though the CR data-derived obesity associated genetic signatures correlated significantly in the CR data, but as soon as these metagenes were generated in the other data sets via the application of transformation matrices, the correlation became insignificant. One possible reason for this was that the obesity metagenes were not actually associated to the obesity phenotype in the other data sets. Another reason could be that the methodology used was not appropriate for investigating the association between the phenotype and its associated phenotype.

The appropriateness of the use of SVD and transformation matrices were examined with different genetic signatures: ER and PR pathway associated genetic signatures from the Gatza *et al.* (2010) study (covered in Appendix A10). Three data sets (CR, FM and NZBC) were used, as these data sets have both ER and PR status information available for the patients. SVD was first applied to the RMA normalised CR data to obtain the ER and PR metagenes, along with the transformation matrices for each of the genetic signatures. Both the ER and PR metagenes showed good concordance with the gene expression profiles of the genetic signature in the heatmap, and also showed significant association with patient ER and PR statuses (Appendix A10).

Transformation matrices were applied to the FM and NZBC data sets to generate the appropriate pathway metagenes in the data sets. Both the ER and PR path-

way metagenes showed similar overall gene expression patterning in the heatmap as the CR data (Appendix A10). Furthermore, both the ER and PR pathway metagenes from these data sets show significant association with patient ER and PR statuses, respectively (Appendix A10). These results clearly showed that the general approach of using SVD and transformation matrices was appropriate for examining the association of the genetic signatures with the patient phenotype.

5.1.4 Common obesity associated genes and pathways across multiple cancer types

The results so far have provided strong evidence that the obesity associated genetic signatures from the CR and the FM data sets were not associated with patient BMI. In a last attempt to identify any obesity associated genes that were common across multiple cancer types, gene expression analyses were carried out on the ICGC cancer data sets using samples for which patient BMI information was available (Section 3.4). There were many DEGs identified in each of the ICGC cancer types, but there were no genes in common for all eight cancer types, and this observation was supported by the results from the simulation.

The simulation results also showed that the number of DEGs identified in these cancer types were greater than one would expect by chance. However, as observed by the results from the simulation, there were many genes identified by chance alone, which suggested that majority of the DEGs found in the eight cancer types were false positives. This apparently high level of Type I errors was not only observed in the ICGC data sets, but also in the CR data set when gene expression analyses were carried out to generate various obesity associated genetic signatures in Section 3.3.1 (discussed in more detail in Section 5.1.5).

Finally, pathway enrichment analysis was carried out in the ICGC data sets (first with each cancer type separately, then all cancer types combined) to investigate whether there were any pathways that significantly associated with obesity (Section 3.5). There were not many pathways identified from the KEGG and the Reactome pathway databases that were significantly enriched in the ICGC cancer data sets. There were a substantial number of GO terms that were significantly enriched in the ICGC data, but there were no pathways that were common across all

cancer types. One thing noted from these results was that all three databases (GO, KEGG and Reactome) identified the "ABC transport" pathway as significantly enriched in the CESC data set. This suggests that the "ABC transport" pathway may have an important role in CESC tumour development in obese patients. There was no significantly enriched pathway that was associated with obesity in the combined ICGC data sets.

Although there was no common pathway that associated with obesity in the ICGC data sets, it was evident from the CR and FM obesity metagenes that the obesity metagenes were associated with the gene expression of the obesity signatures. This suggested that the metagenes were able to identify some sort of genetic pattern in the data, even though these metagenes were not significantly associated with obesity.

5.1.5 False positives in the gene expression analyses

The fact that there were so many false positives in the DEGs from the ICGC data sets indicated that the use of the patient BMI and BMI status as the "treatment" conditions (in other words independent variable) for gene expression analysis was prone to Type I errors. This meant that there were very little evidence of true obesity associated genetic signatures in any of the data sets that have been explored in this project, or perhaps the patient BMI and BMI status was not enough to pinpoint the underlying biological relationship between obesity and cancer (see Section 5.2.1 as well).

Evidence of large number of false positives was also seen in Section 3.3.1, where there were many DEGs identified from the gene expression analyses when obesity associated genetic signatures were sought for in the CR data set. This suggested that there was a significant number of genes that were not related to obesity (false positives) in the genetic signatures. Furthermore, the fact that about a third of the genes in any one of the obesity signatures were "unique" to each of the individual obesity signatures supported the apparent abundance of false positives in the signatures (Figures 10 and 11). From these results, it was likely that the obesity signatures that resulted from the CR data set had many genes that were false positives, perhaps due to the poor quality of the data set (Section 5.2.2).

Even if the patient BMI and/or BMI status was enough to identify the genes that were truly associated with obesity, it would not be a trivial task to identify the truly obesity associated genes from those that were not, given a list of DEGs. This raises an important question of whether the patient BMI was an appropriate clinical variable to uncover the true association between obesity and cancer (more in Sections 5.2 and 5.4).

5.1.6 Genetic signature captured by the obesity metagenes

In Chapter 4, analyses were carried out to determine what the genetic pattern the obesity metagenes from Chapter 3 were associating with. The pathway associated genetic signatures from the Gatza *et al.* (2010) study were used to establish the biological connection with the obesity associated genetic signatures. In order to make the obesity and pathway associated genetic signatures comparable with one another, the orientations of the metagenes were examined so that all of the metagenes were in the "correct" orientation (see Section 4.1).

Once the directions of the metagenes were determined, the obesity metagenes were clustered together in the heatmap to visualise the similarity of the obesity metagenes with the pathway metagenes. The heatmap revealed that all of the obesity metagenes that have originated from the CR data clustered together in a group with no other pathway metagenes correlating with these metagenes, which meant that the obesity metagenes were not similar to any of the pathway metagenes from the Gatza *et al.* (2010) study. The clustering of the obesity metagene was not surprising, as the results from Section 3.3.2 already showed that the metagenes were highly correlated with one another. This result showed that all of the obesity metagenes created from the CR data associated with an unknown genetic signature that was not related to any of the pathway associated genetic signatures from the Gatza *et al.* (2010) study.

The obesity metagene from the Fuentes-Mattei *et al.* (2014) study was a cluster of its own, where the metagene did not group with any of the obesity metagenes derived from the CR data set, nor with any of the GT pathway metagenes. Though Fuentes-Mattei *et al.* suggested a biological link with the Akt/mTOR pathway in their study, the result from the heatmap did not support that finding. This may have

been due to the inconsistency of the pathway metagene scores across different data sets as mentioned in Section 4.2. In fact, the Akt pathway metagene scores were variable across different cancer data sets (Appendix B), which suggested the Akt pathway genetic signature was of poor quality.

5.1.7 Linear models to predict the obesity metagenes

To further investigate whether there was any evidence of a pathway signature being related to the obesity associated genetic signatures, linear models were created in the NZBC data set with the pathway metagene scores. First, linear models were created with the patient BMI, BMI status and a selection of the most "consistent" pathway metagene scores (based on the consistency across different data sets). From these linear models, PR pathway metagene scores showed significant contribution to the model in predicting the obesity associated metagene scores in all of the linear models, with the exception of the FM obesity metagene (where Myc was the significant variable; Appendix B9).

This was not surprising since all but the FM obesity metagene were derived from the CR data set, and therefore the FM obesity metagene was likely to be different to those from the CR data set. Adding to this, differences between the CR and FM metagenes were also shown in the heatmap of the obesity and pathway metagenes mentioned earlier, where the FM metagene clustered with neither the CR obesity metagenes nor with any other pathway metagenes. This implied that the FM metagene differed fundamentally from the CR obesity metagenes. Nevertheless, to clarify the significance of the PR pathway with the obesity metagenes, linear models were also created with PR pathway metagene in combination with the patient BMI and/or BMI status.

The predictions of the obesity metagene scores were made using the linear models created in the NZBC data set. All of the predicted obesity metagene scores significantly correlated with the true obesity metagene scores in both the NZBC (training) and CR (testing) data sets. However, the R^2-values for all of these predictions were low (highest $R^2 \approx 0.42$; Appendix B11) with highly variable data points in the scatter plots. It was unclear from these predictions and plots whether the models explained the true underlying biological mechanisms of the obesity

associated genetic signatures. Nevertheless, the models did explain some variations in the obesity metagenes, but there may be some other unknown biological processes that provide more information about the signatures.

Since linear models were created only with the "consistent" pathway metagene scores, a step wise method was used to generate linear models that included the patient BMI, BMI status and all of the pathway metagenes that allowed the model to predict the true obesity metagene the best. As shown by the results in Section 4.3, all of the models predicted the true obesity metagene scores with high R^2-value in the CR data set (highest $R^2 \approx 0.89$) and were statistically significant. This meant that the variables that were significant in the linear models played a role in the obesity associated genetic signatures. There was no single pathway signature that was common to all of the linear models, but Akt, EGFR, TGFβ and Src pathway metagenes were included in many of the models, which suggested that these pathway signatures may have an important role in the obesity associated genetic signatures (Section 4.3.2).

5.1.8 Significance of the pathway associated signatures with the obesity associated signatures

The linear models that were created from the "consistent" pathway signatures in the NZBC data set showed significant association of the PR pathway with the obesity metagenes. But was this due to the true biological relationship between the PR pathway and the obesity metagenes, or only because the models were created in the NZBC data set? In fact, the proportion of the patients that were PR^+ was greater in the NZBC data set (62 PR^+ and 37 PR^- patients) compared with the CR data set (48 PR^+ and 50 PR^- patients), which implied that the patients in the NZBC data set was more likely to show association with the PR pathway metagene.

However, with this logic, the ER pathway signature would as likely to be associated, as the patient ER status was similar to the PR status in either data sets: 72 ER^+ and 27 ER^- patient in the NZBC data, and 58 ER^+ and 42 ER^- patients in the CR data set. The fact that the ER pathway signature was not significant in the models suggested that the obesity metagenes were related to the PR pathway

irrespective of the patient PR status. Furthermore, both the NZBC and the CR data sets showed strong association between the PR and ER pathway metagenes with the patient PR and ER status, respectively (Appendix A10). Since both ER and PR signatures were significantly associated with the patient ER and PR statuses, both signatures would have been as equally likely to show up significant in the models, supporting the idea that the ER and PR statuses were independent of the apparent significance of these signatures in the models.

In the step wise linear models, many variables were identified to be significantly associated with the obesity metagenes. This suggested that the pathway associated genetic signatures that were significant in these models may be biologically related to the obesity phenotype, or perhaps the obesity associated genetic signatures were detecting the signals from these pathways rather than the signals from the obesity phenotype. In fact, many of the variables included in the linear models were pathway metagenes that negatively correlated with the obesity metagenes in Figure 18 (Section 4.2.2). Since all of the obesity metagenes failed to show significant association with patient BMI, it may be possible that these pathways (Akt, ER, EGFR, PR, p53, Src and TGFβ) are actually what was being detected by the "obesity" metagenes.

With that said, even if these pathway signatures were related to the underlying biology of breast cancer, the fact that these results depended on the pathway associated genetic signatures from the Gatza *et al.* (2010) study makes the results questionable. This was due to the lack of consistency of the metagenes across different cancer data sets, where more than half of the pathway metagenes showed a correlation of less than 0.8 in at least one data set (Section 4.2). This meant that at least half of the signatures were specific to the GT data set and could be highly variable in the other data sets, reflecting the potential differences in the cohorts (at the molecular level). Thus, without further validation of the quality of the pathway associated genetic signatures and verification of the roles of these signatures in the biological relationship between obesity and cancer, these results should be interpreted with care.

5.2 Limitations

5.2.1 Definition of obesity

First and possibly the most important limitation to this project was the use of BMI as a measurement for obesity and its subsequent classification of the patients based on BMI. Obesity is a term given to those individuals with excessive amount of fat and can be defined in various ways. It was important to define the obese phenotype accurately, as all of the obesity associated genetic signatures, gene expression analyses and pathway enrichment analysis were dependent on the patient BMI status. BMI is perhaps the most popular measurement to estimate obesity owing to the relative ease of measurement in clinical settings, and hence used in various clinical trials and large epidemiological studies as an indication of the patients' metabolic state. However, there have been studies that showed other measurements, such as waist-to-height ratio and waist-to-hip ratio, to be a better representation of the patients' obesity status (Dalton *et al.*, 2003; Lee *et al.*, 2008).

Perhaps the reason why there were so many false positives when gene expression analyses were carried out in Section 5.1.5 was because BMI was used to classify the patients into their corresponding BMI statuses, instead of another measurement. It was likely that the patient BMI (and therefore BMI status) was not good enough a phenotypic characteristic to describe the underlying biology driven by obesity. Since the gene expression analyses were based on such "vague" parameter, it may have picked up many genes that were apparently related to obesity when in fact the genes were not related to obesity at all.

Ideally, multiple measurements that are indicative of obesity should be made in order to determine the obesity status of the patients to make a more accurate classification. Additional measurements such as blood lipid profiles and insulin levels may also help understand the underlying biological state in which the patients are in, and would greatly help focus on the genes that are truly affecting the patients. However, these measurements are difficult to carry out in a systematic fashion in a clinical setting, and to collate existing patient data with these measurements to make a data set that is large enough for the analyses done in this project may be impossible. Nevertheless, even the measurement of waist circum-

ference in addition to BMI may significantly help describe the obesity phenotype more accurately in future investigations.

5.2.2 Quality of the data

Another limitation of this work was the quality of the data sets. All of the data sets used in this project were from an online and publicly available source, and the types of data sets used were either microarray or RNA-seq data. The first potential problem with using open source data from other research groups is that the data are generated by multiple different groups with different technologies and experimental protocols. Since different laboratories have their own way to carry out certain experiments, the sequence data from each of the studies used in this project may have been produced and processed in a completely different manner, not to mention the use of different patient cohorts in each experiment.

For example, in the Creighton *et al.* (2012) study, the patient breast tumour biopsies were "trimmed such that all the samples had $\geq$ 70% tumour cells" in the samples. Fuentes-Mattei *et al.* (2014) only mentioned that they used breast tumour biopsies from patients with no comment about the quality or quantity of the biopsies. Without a doubt, the difference in how the samples were handled can affect the quality of the data, and make it challenging to compare different data sets (Irizarry *et al.*, 2005). In fact, observations were made in this project that suggested the data set from the Creighton *et al.* (2012) study was of poor quality (see Section 5.1.5).

Although the underlying quality of the data provided by these studies could not be improved in any way, best efforts were made to ensure the data sets were free of other factors that may affect the results. This leads to the second problem: the different technologies used in the studies. All of Creighton *et al.*, Fuentes-Mattei *et al.*, NZBC and Gatza *et al.* data sets were analysed using microarray technology, whereas all of the ICGC cancer data sets were generated with NGS technology. As mentioned in Section 2.3.1, the two technologies use fundamentally different principals to output the sequence information of the sample; microarray uses light intensity and NGS uses count data.

In order to make these two different types of data comparable to one another,

normalisation methods from the *limma* package were used (see Section 2.3.1). This enabled the standard *limma* analysis pipelines, normally used for the analysis of microarray data, to be used by the ICGC cancer RNA-seq data sets. In this way, the results from the two types of data sets were fairly reliable and consistent, as the same analysis methods were applied to the data sets. In addition to the difference in the sequencing technologies, the Gatza *et al.* data set comprised multiple different data sets, each from different sources. Since individual data sets had their own experimental differences (for example, "batch effects"), it was important to correct for these differences when multiple data sets were combined into one (see Section 2.3.4). In this project, batch corrections were made in the Gatza *et al.* data set and the combined ICGC cancer data sets (used in Section 3.5) to eliminate the batch effects.

5.2.3 Quality of the genetic signatures

The final limitation to consider was the quality of the genetic signatures used in this project. This limitation is partly related to Section 5.2.2, as the identification of the genetic signatures depended on the quality of the data as well as the definition of the signature in the first instance (for example, obesity or Akt over-expression).

Firstly, the obesity signatures identified by both Creighton *et al.* and Fuentes-Mattei *et al.* relied on the definition of obesity based on the patient BMI values. As discussed in Section 5.2.1, the use of BMI may not have been the best measurement for obesity. In fact, both CR and FM obesity signatures associated with different pathway signatures in Sections 4.2 and 4.3, which suggested that obesity as defined by BMI was not enough to reveal the underlying biological characteristics of obesity, and hence showed association with different pathway signatures. Furthermore, as mentioned in Section 5.2.2, it was likely that the CR data set (and perhaps other data sets as well) was of poor quality and may have affected the ability of the generated signatures as a marker for obesity in cancer.

Secondly, the method in which the pathway associated genetic signatures was defined in the Gatza *et al.* (2010) study may have been questionable. In their studies, Gatza *et al.* used model cell lines to identify the pathway associated genetic

signatures by altering the expression of the representative gene for the pathway and assigned all of the genes affected by the change in representative gene expression to be associated with that pathway (Bild *et al.*, 2006; Gatza *et al.*, 2010). However, some of these pathway associated genetic signatures were derived in different cell lines to one another, which meant that some pathway signatures may not reflect the true pathway activity *in vivo*. In fact, as seen in Sections 4.1 and 4.2, some pathway signatures produced more consistent metagenes from the data sets than the other signatures. These results were merely an indication that the pathway signatures may have reflected the pathway activities poorly, and it was not possible to determine whether these signatures were truly related to the corresponding pathways, as the original data was not available to reference the signatures back for validation. With that said, few of the pathway signatures correlated well with some clinical variables (such as ER and PR), so not all signatures may have been of poor quality (Appendix A10).

5.3 Conclusion

There were two main aims to this project: firstly to determine whether there were any obesity specific genetic signatures that could be transferred across multiple cancer types; and secondly to investigate whether there were any biological pathways dysregulated in the tumour samples from the patients that were obese compared to the samples from non-obese patients. These aims were addressed in Chapters 3 and 4, respectively.

As shown clearly by the results from Chapter 3, there seemed to be no genetic signatures that were able to differentiate between the tumour samples from the patients that were obese from those that were not, across different cancer types. Furthermore, there were no genes differentially expressed between obese and non-obese patients that were common across multiple cancer types. These results have suggested that there were no obesity specific genetic signatures that were common across different cancer types. The results from Chapter 4 showed that the obesity associated genetic signatures generated from the Creighton *et al.* data set were different from those generated from the Fuentes-Mattei *et al.* data set. None of the pathway associated genetic signatures showed positive correlation with any of the

obesity associated genetic signatures in the heatmap. However, all of the obesity metagenes derived from the Creighton *et al.* showed strong negative correlation with the Akt, ER, EGFR, PR, p53, Src and TGFβ pathways. This suggested that the obesity associated genetic signatures were picking up the signals from these pathways, rather than the signals from the obesity phenotype of the patients.

To summarise, there were no common genetic signatures that significantly associated with the obesity status of the patients. However, the Akt, EGFR, TGFβ and Src pathways may have a role in promoting the tumour progression in patients that are obese. Further investigations with good quality data sets and larger patient cohorts are required to clarify the biological relationship between obesity and cancer.

5.4 Future directions

This project has managed to show that there were no obesity specific genes that were common across multiple cancer types. Although this study failed to identify obesity specific genes that affected tumour biology, it is likely that there is some complex mechanism underlying the relationship between obesity and cancer; otherwise many of the clinical and epidemiological associations between obesity and cancer risk would not make any sense. Since Akt, EGFR, TGFβ and Src pathways were significant in the linear models that predicted the obesity metagenes, more thorough investigation of these pathways may be a good start to clarify the relationship. Further analyses of BLCA may help clarify the similarity between bladder and breast cancer types and provide insight into the association between the obesity associated genetic signatures and BLCA.

In this project, the lack of association of the obesity associated genetic signatures with any of the cancer types could have been due to the quality of both the raw data sets and the genetic signatures used in this project. Therefore, careful selection of the raw data sets with sufficient, and if possible, additional information (such as waist circumference) about the patients and well-validated genetic signatures will be essential for future investigations.

Appendix A

Additional results from Chapter 3

A1 Comparison of the Creighton *et al.* obesity metagene in standardised or non-standardised CR data

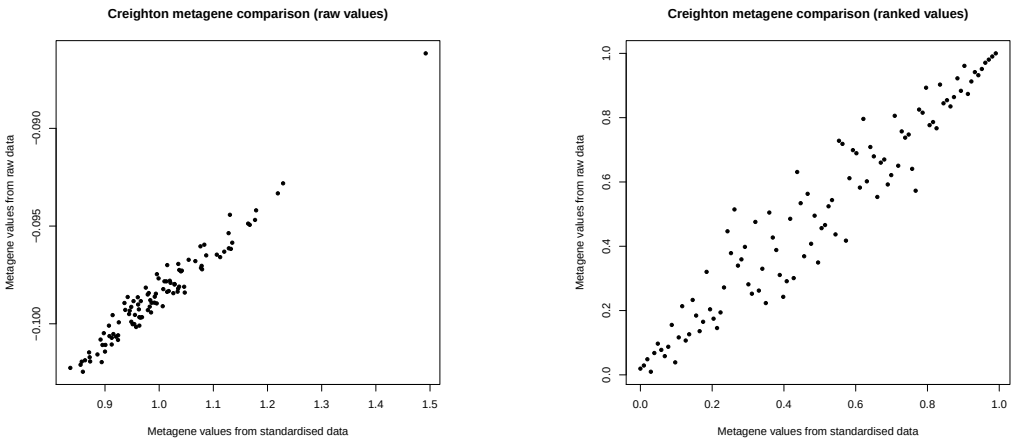


Figure S1: Scatter plots showing the raw and ranked Creighton *et al.* obesity metagene scores from the standardised or non-standardised CR data. The RMA-normalised CR data was standardised or used as is to generate the obesity metagene scores. The metagene scores were ranked based on the number of samples in the CR data set.

A2 Comparison of the Creighton *et al.* obesity metagene in standardised or non-standardised ICGC data

As shown in Figure S2, the Creighton *et al.* obesity metagene values generated from the transformation matrices (from standardised or non-standardised CR data) were more consistent (or less variable) in the standardised ICGC BLCA data, compared with the non-standardised data. Results from Figure S3 suggested that there was no obvious difference in the quality of the metagenes generated from standardised or non-standardised TM. Only the results from BLCA data set are shown, as the results from the other ICGC cancer types showed very similar results.

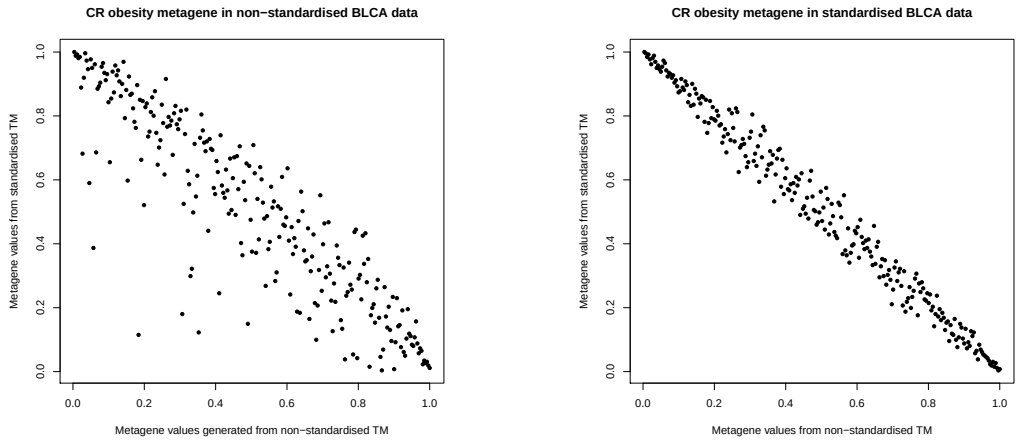


Figure S2: Scatter plots comparing the Creighton *et al.* obesity metagene generated from the transformation matrix (from standardised or non-standardised CR data) in the non-standardised (left) and standardised (right) ICGC BLCA data.

A3 Remainder of the results of FM obesity metagenes in the ICGC cancer data sets

Figure S4 shows the association of the FM obesity metagene with the gene expression pattern of the FM obesity associated genetic signature in the other ICGC cancer data sets. Figure S5 shows the association of the FM obesity metagene

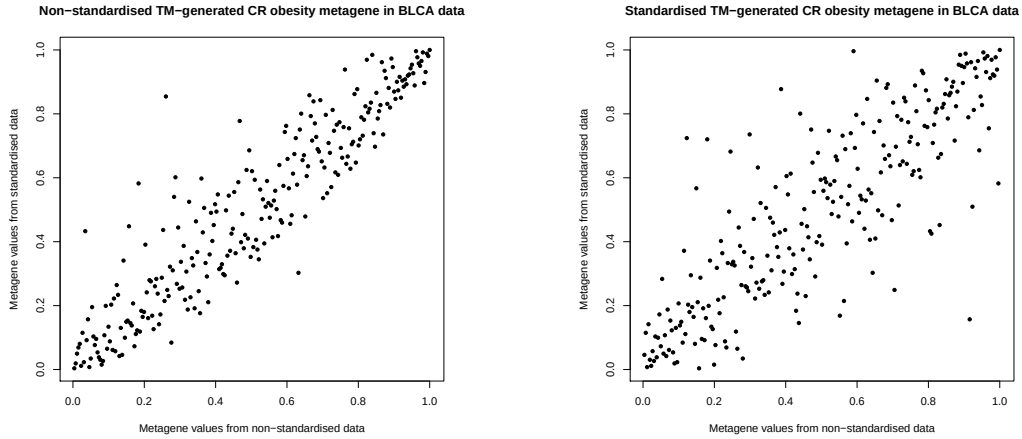


Figure S3: Scatter plots comparing the Creighton *et al.* obesity metagene from the non-standardised and standardised ICGC BLCA data, generated by the application of the transformation matrix from the non-standardised (left) or standardised (right) CR data.

scores with the patient BMI and BMI status. Clearly, the FM metagene was able to capture the overall gene expression patterns, but was not significantly associated with the patient BMI or BMI status.

A4 Direction of all the obesity metagenes from the CR data

The direction of the obesity metagenes identified from the CR data were examined and corrected for (Figure S6). Gene probes that were common to all obesity associated genetic signatures were used to examine whether the metagene was in the correct direction, relative to the other obesity metagenes

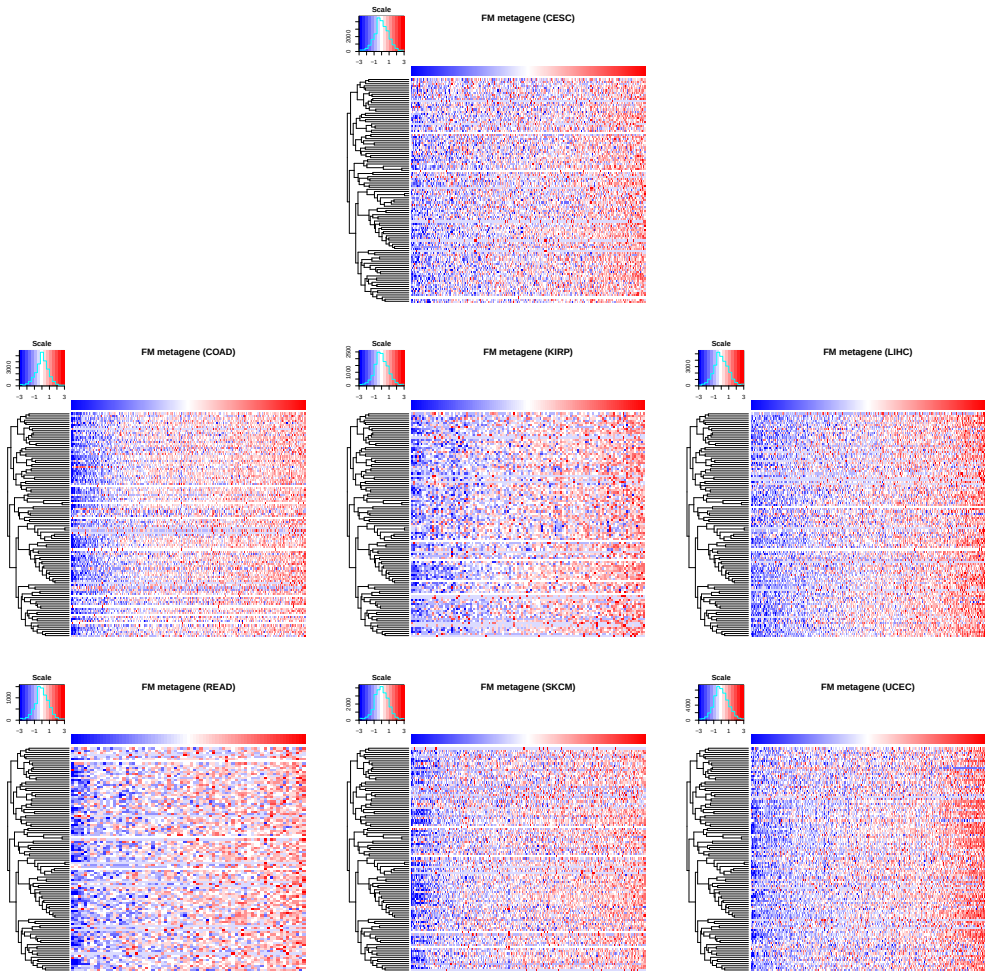


Figure S4: Heatmaps showing the association of the FM obesity metagenes with the sample gene expressions in the other ICGC data sets. Scales are as described in previous figures.

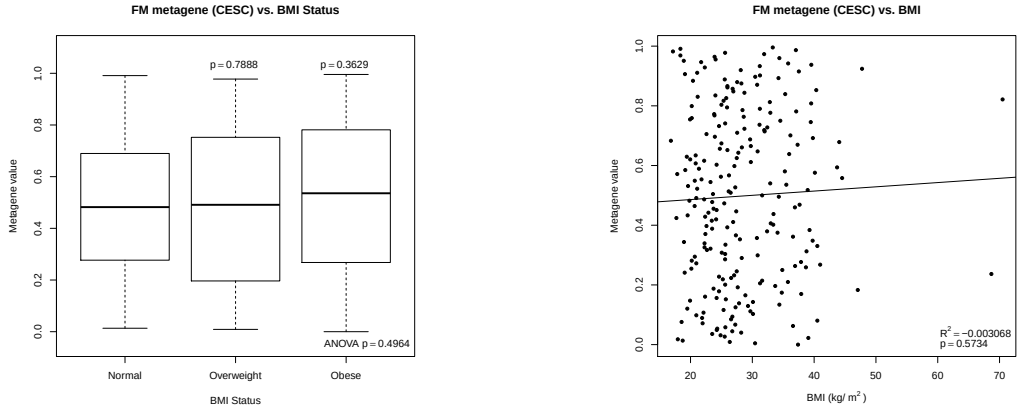


Figure S5: Box plots and scatter plots showing the association of the FM obesity metagene with the patient BMI status and BMI, respectively, in the ICGC data sets. P-values and R^2-value are as described in previous figures.

A5 Remainder of the results of the other CR obesity metagenes in the CR data set

All of the other obesity metagenes identified from the CR data set were able to capture the overall gene expressions of the samples in the CR data set (Figure S7). Furthermore, all of the CR obesity metagenes were significantly associated with the patient BMI/BMI status in the CR data set (Figure S8).

A6 Remainder of the results of the other CR obesity metagenes in NZBC data set

All of the other obesity metagenes from the CR data set were able to capture the overall gene expressions of the samples in the NZBC data set (Figure S9). However, none of the CR obesity metagenes were able to significantly associate with the sample BMI/BMI status in the NZBC data set (Figure S10).

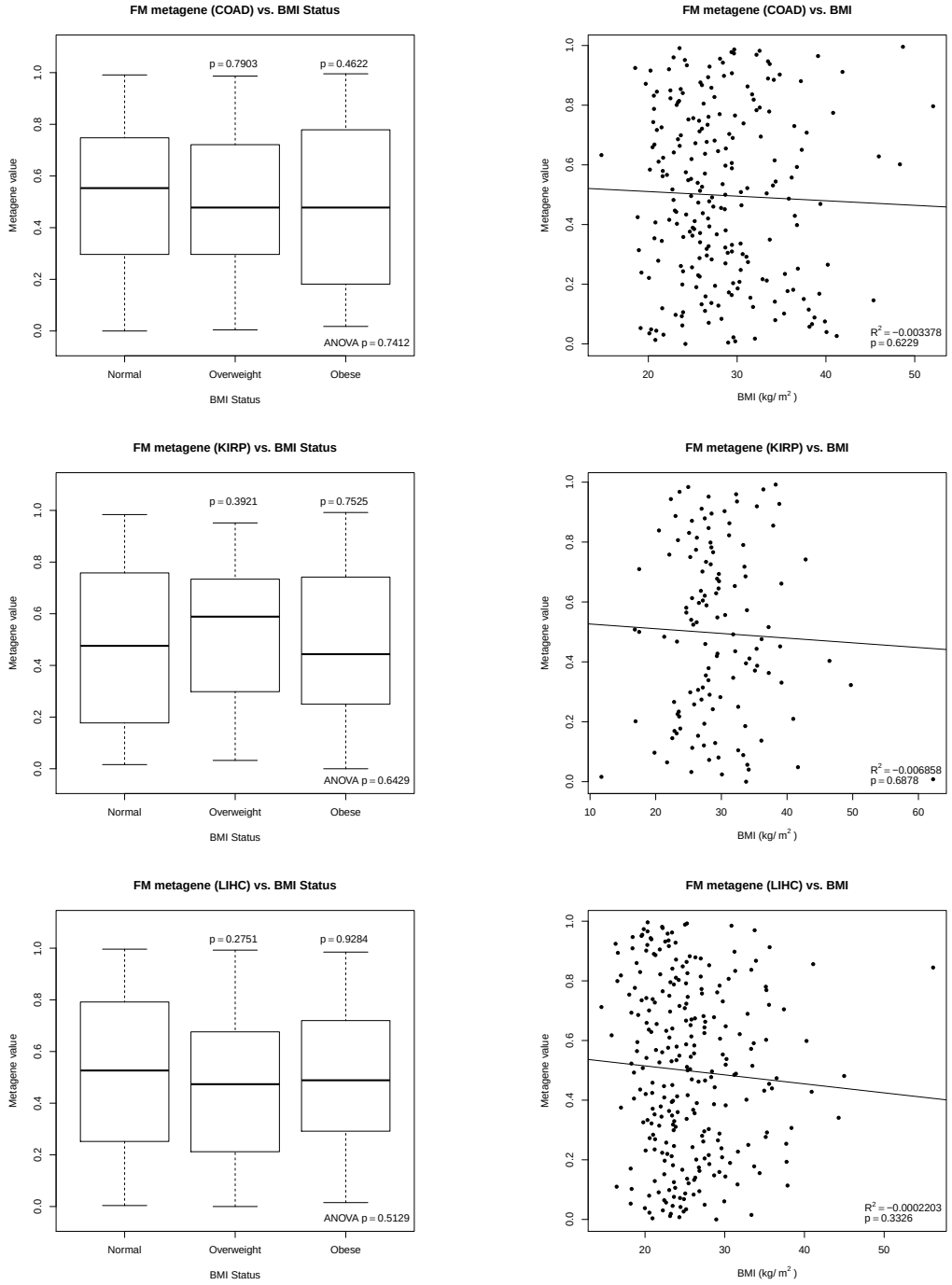


Figure S5 (cont.)

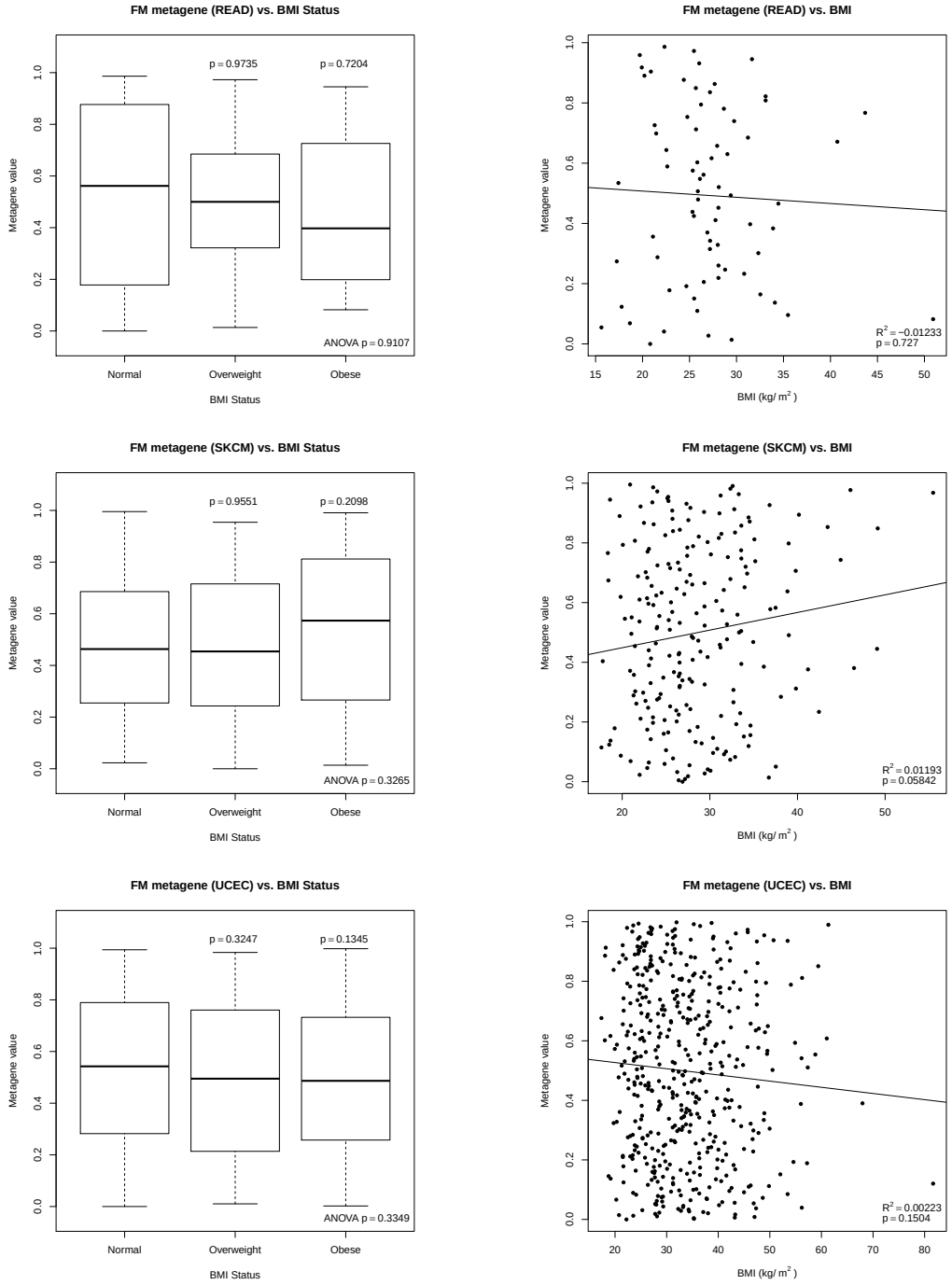


Figure S5 (cont.)

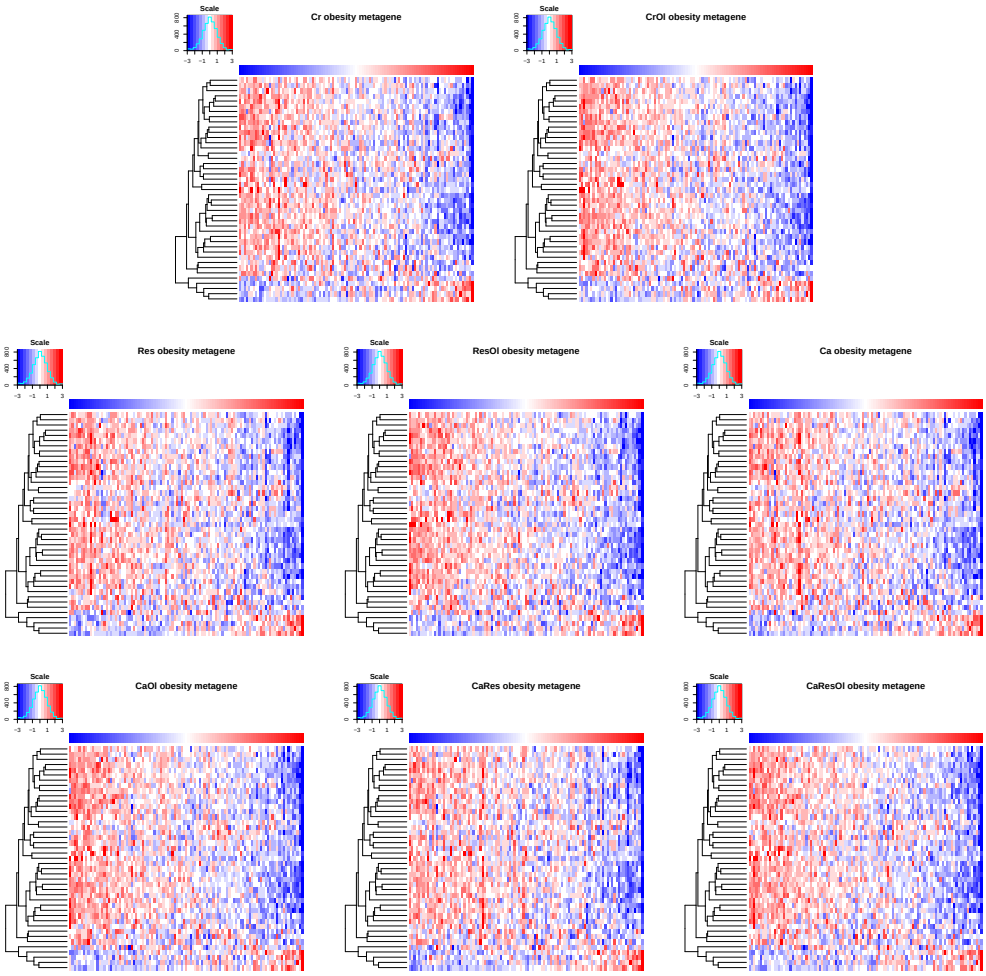


Figure S6: Heatmaps showing the gene expression of the obesity associated genetic signatures with the obesity metagene in RMA-normalised CR data set. The gene probes that were common to all of the obesity associated genetic signatures were used to determine the direction of the obesity metagenes. The scales and axes are the same as the previous heatmaps.

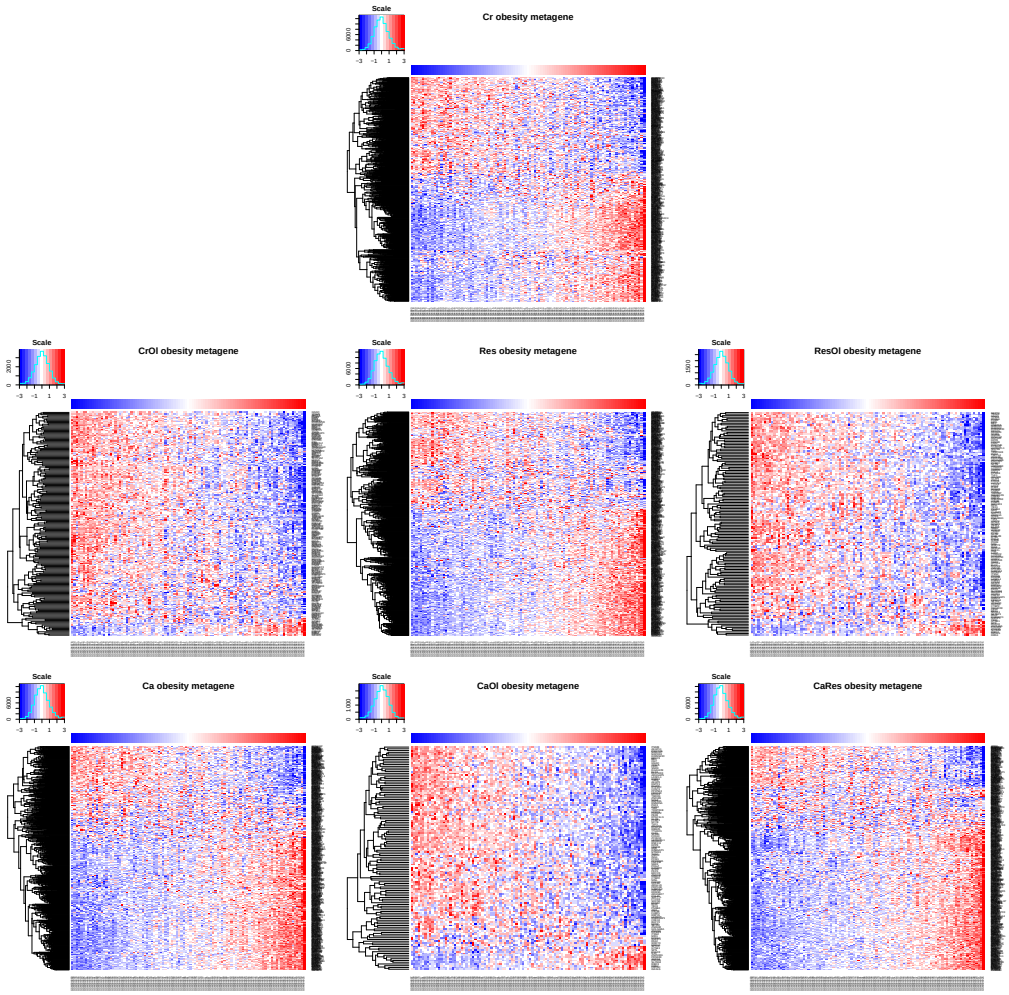


Figure S7: Heatmaps showing the association of the other obesity metagenes identified from the CR data with the sample gene expressions from the NZBC data set. Scales are as described in previous figures.

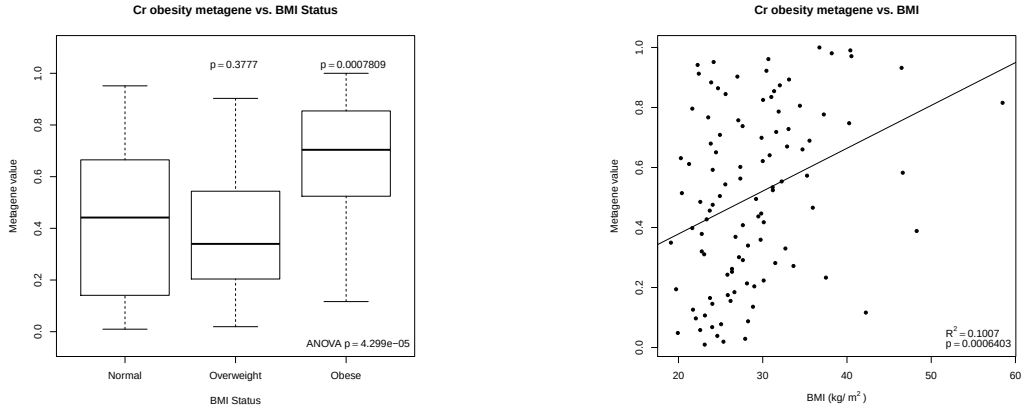


Figure S8: Box plots and scatter plots showing the association of the other CR obesity metagenes with the sample BMI status and BMI from the NZBC data set, respectively. P-values and R^2-value are as described in previous figures.

A7 Remainder of the results of the other CR obesity metagenes in the ICGC cancer data sets

As shown in Figure S11, the original obesity metagene from the Creighton *et al.* study was reflective of the sample gene expression levels of the obesity associated genetic signature in all of the other ICGC cancer data sets. However, none of the ICGC cancer data sets (other than the BLCA data set; Figure 5 in Section 3.1) showed any significant association of the sample BMI or BMI status with the Or obesity metagene (Figure S12). The heatmaps, box plots and scatter plots for the other CR obesity metagenes in other ICGC cancer types are not shown, as many of the plots were very similar to the plots from the BLCA data set. The statistics from the results of the box and scatter plots in the other ICGC cancer types were summarised in Tables S1 to S7.

A8 Obesity associated genetic signature using sample BMI values

Since all of these obesity associated genetic signatures were derived based on the discrete values of sample BMI staus, a genetic signature was identified using the

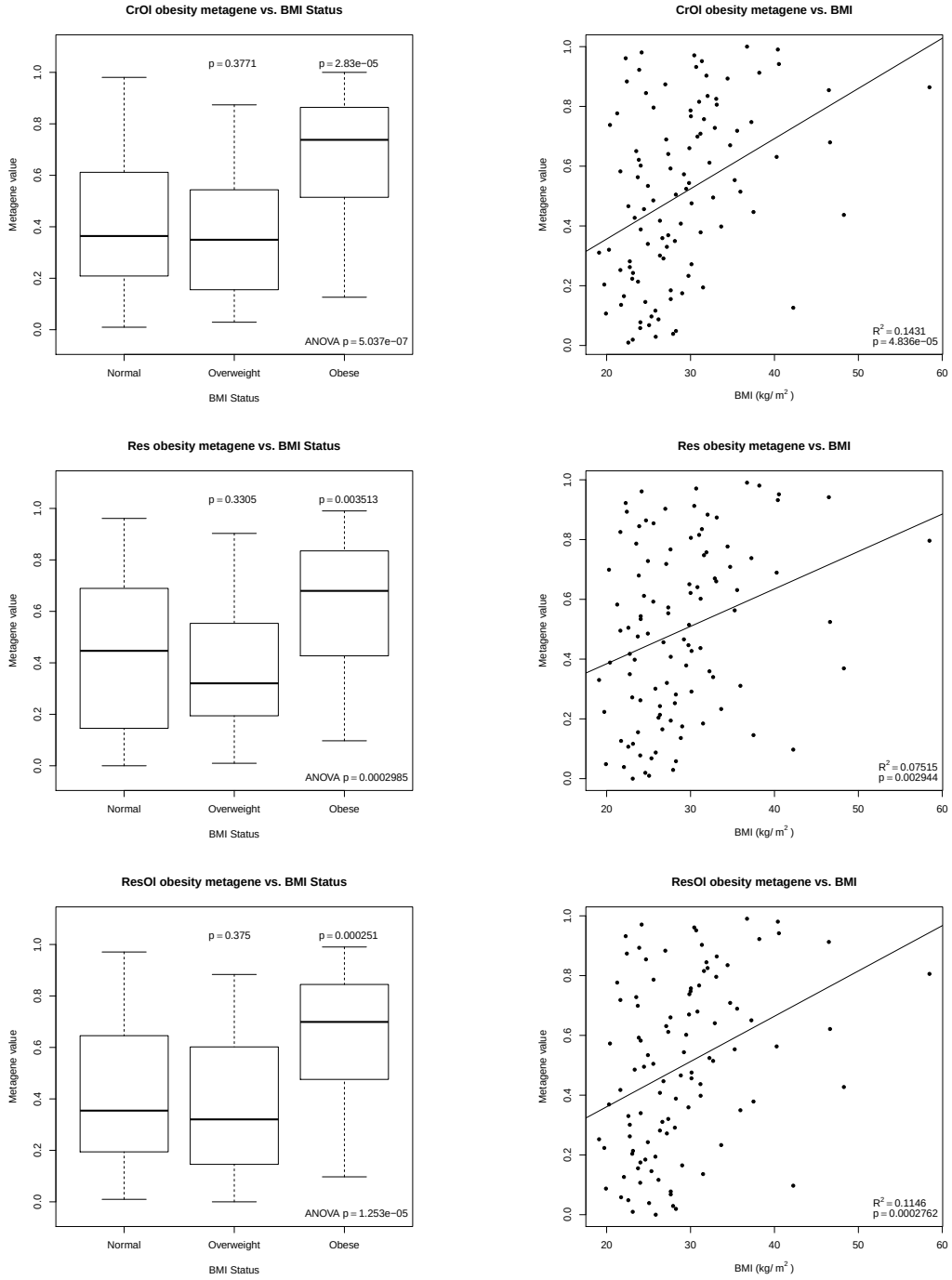


Figure S8 (cont.)

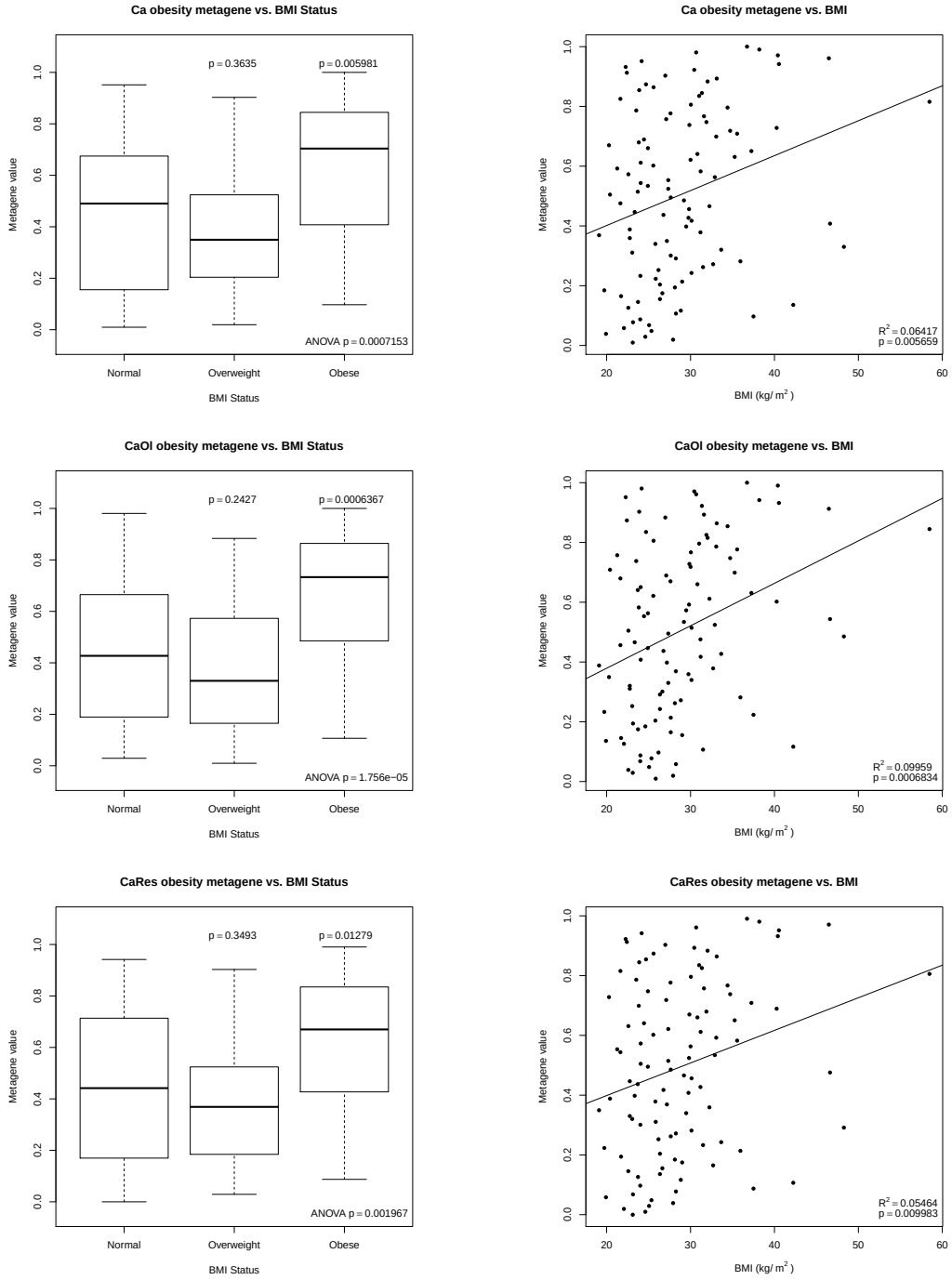


Figure S8 (cont.)

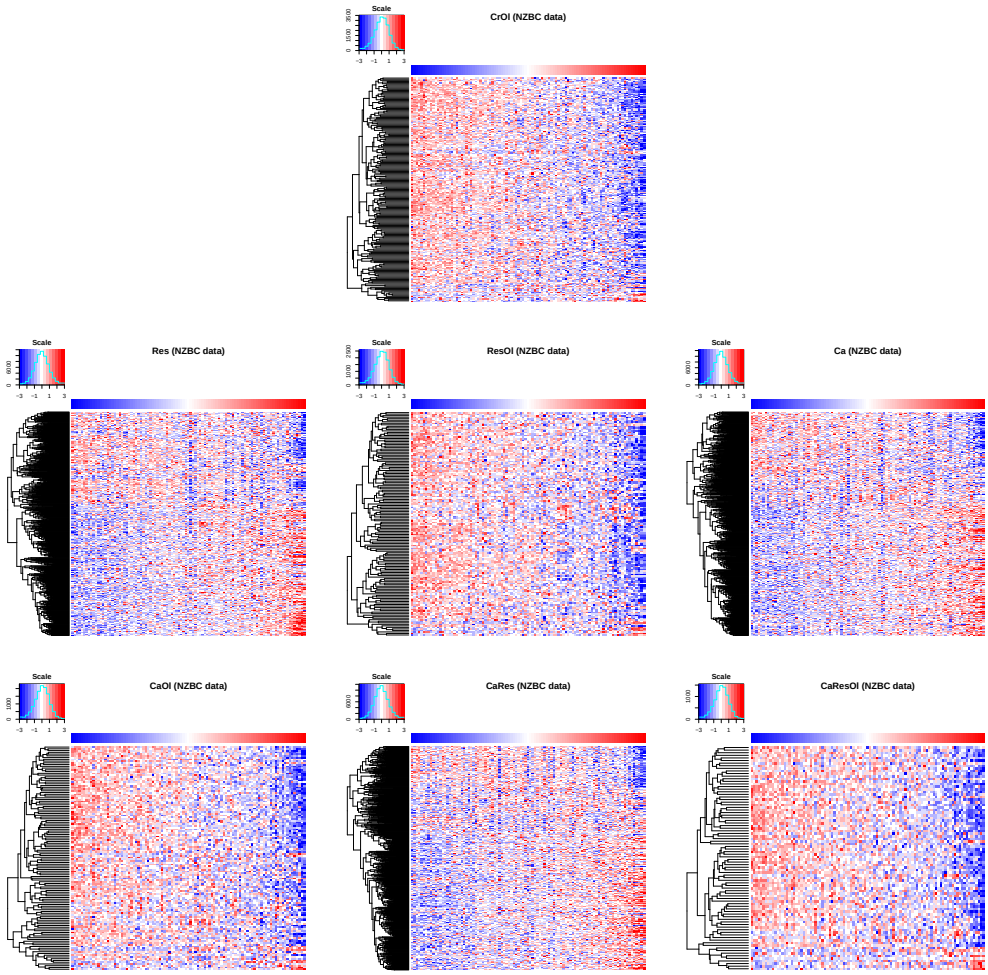


Figure S9: Heatmaps showing the association of the other obesity metagenes identified from the CR data with the sample gene expressions from the NZBC data set. Scales are as described in previous figures.

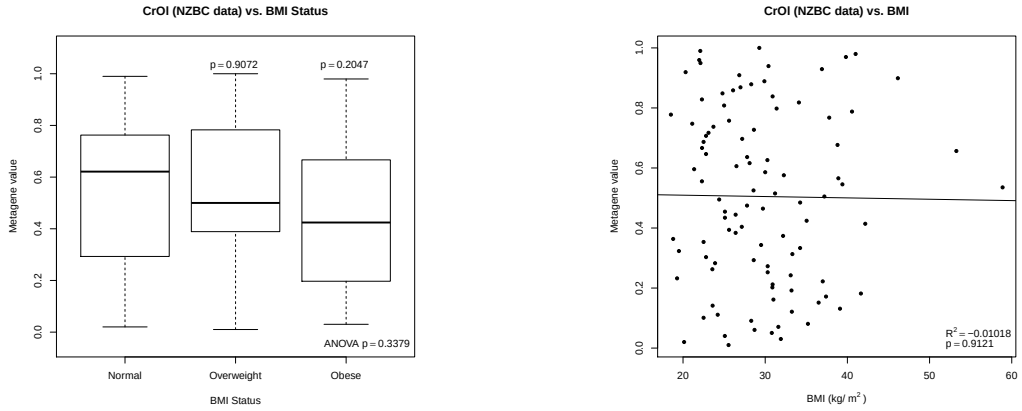


Figure S10: Box plots and scatter plots showing the association of the other CR obesity metagenes with the sample BMI status and BMI from the NZBC data set, respectively. P-values and R^2-value are as described in previous figures.

Table S1: Statistics of all the obesity metagenes in the ICGC CESC cancer data

	P-values			Regression line statistics	
Metagenes	Overweight	Obese	ANOVA	R^2	P
Cr	0.9332	0.5302	0.7622	-0.003950	0.7265
Res	0.9205	0.3369	0.6043	0.004249	0.8122
CrOl	0.8940	0.1430	0.2959	-0.0004927	0.3465
ResOl	0.9930	0.1705	0.3078	0.003263	0.1898
Ca	0.8861	0.3616	0.6426	-0.003307	0.6073
CaRes	0.8323	0.3330	0.6167	-0.003832	0.7001
CaOl	0.9124	0.1602	0.2600	-0.001612	0.4242
CaResOl	0.9646	0.1663	0.2863	0.003097	0.1946

Table S2: Statistics of all the obesity metagenes in the ICGC COAD cancer data

	P-values			Regression line statistics	
Metagenes	Overweight	Obese	ANOVA	R^2	P
Cr	0.2843	0.8176	0.5460	-0.004373	0.8865
Res	0.4091	0.8473	0.7064	-0.004243	0.8245
CrOl	0.4063	0.9295	0.6150	-0.00397	0.7400
ResOl	0.3748	0.8484	0.5295	-0.004020	0.7532
Ca	0.6507	0.7692	0.7579	-0.002493	0.5075
CaRes	0.4565	0.9808	0.7161	-0.004278	0.8385
CaOl	0.2931	0.9217	0.4621	-0.00425	0.8273
CaResOl	0.4203	0.9087	0.7017	-0.004461	0.9800

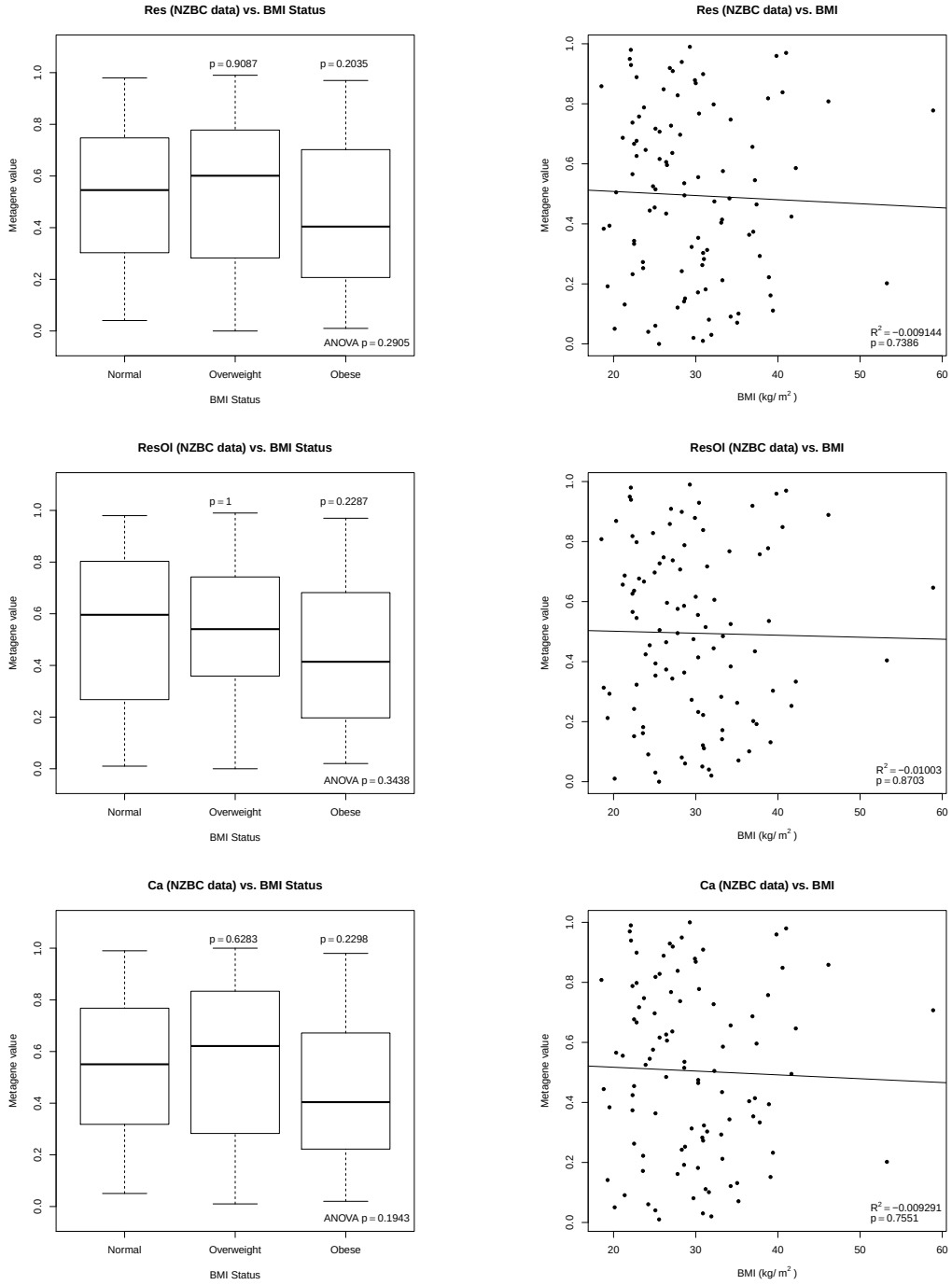


Figure S10 (cont.)

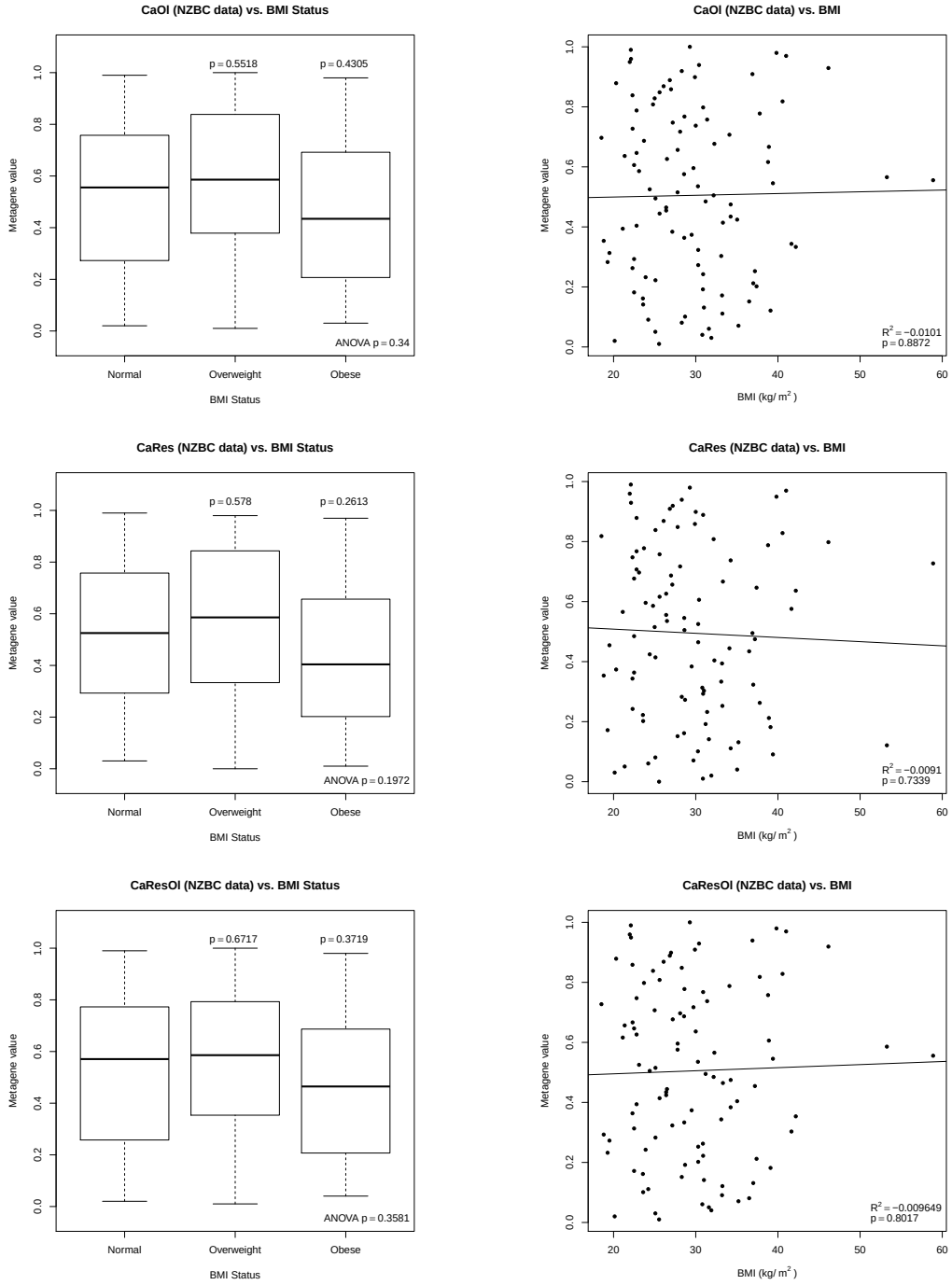


Figure S10 (cont.)

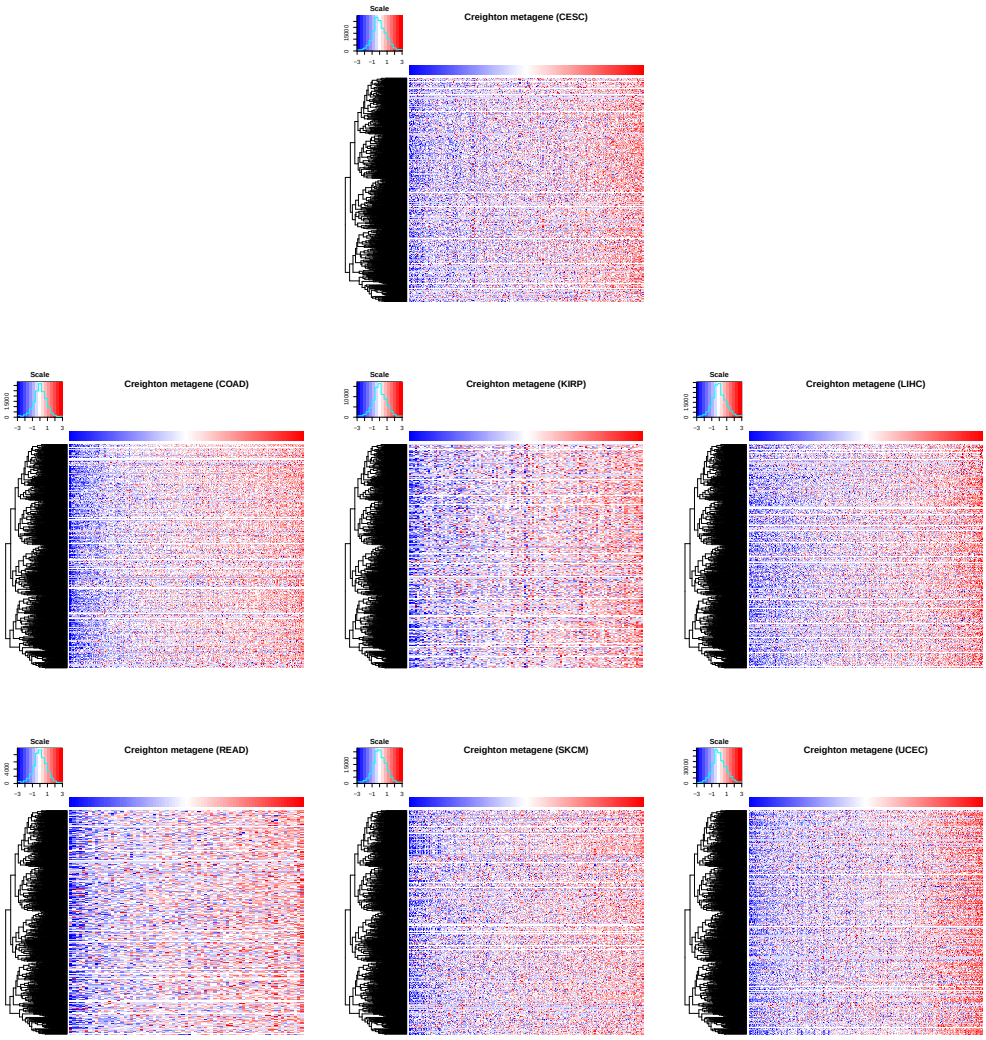


Figure S11: Heatmaps showing the association of obesity metagene from the Creighton *et al.* (2012) study with sample gene expression from the other ICGC data sets. Level of expression is represented in the top right histogram, where low and high gene expression were colour-coded with blue and red, respectively. Each row of the heatmap represents a gene from the obesity associated genetic signature, and each column of the heatmap represents a sample from the ICGC data. The obesity associated metagene scores of the samples are shown in a separate row at top of the heatmap, and the tree diagram of the heirarchical clustering of the genes is shown to the left of the heatmap.

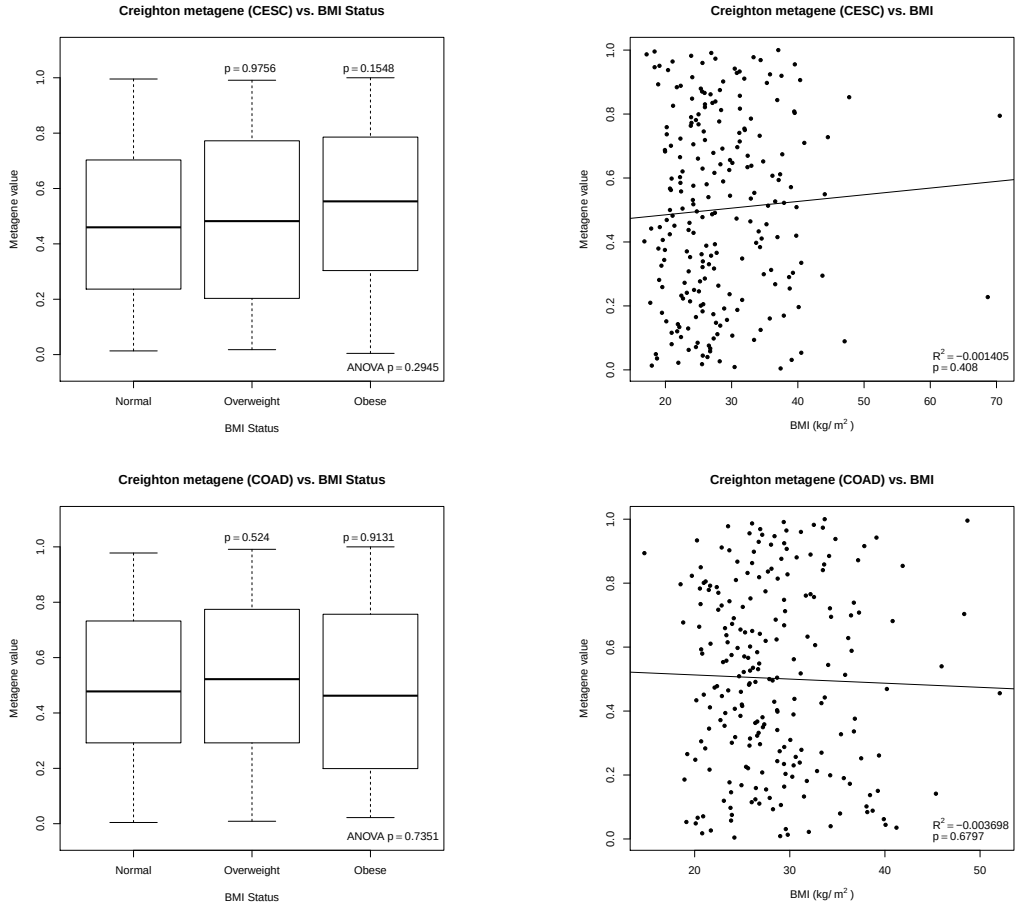


Figure S12: Box plots and scatter plots showing the association of obesity metagene from the Creighton *et al.* (2012) study with the sample BMI status and BMI from the other ICGC data sets, respectively. In the box plot, the p-values above the groups represent the statistical significance of the association of the metagene with the overweight or obese group compared with the normal weight group. The ANOVA p-value shows the statistical significance of the association of the metagene with the sample BMI groups. In the scatter plot, R^2- and p-values describe the adjusted coefficient of determination of the regression line and the statistical significance of the linear model used to draw the regression line, respectively.

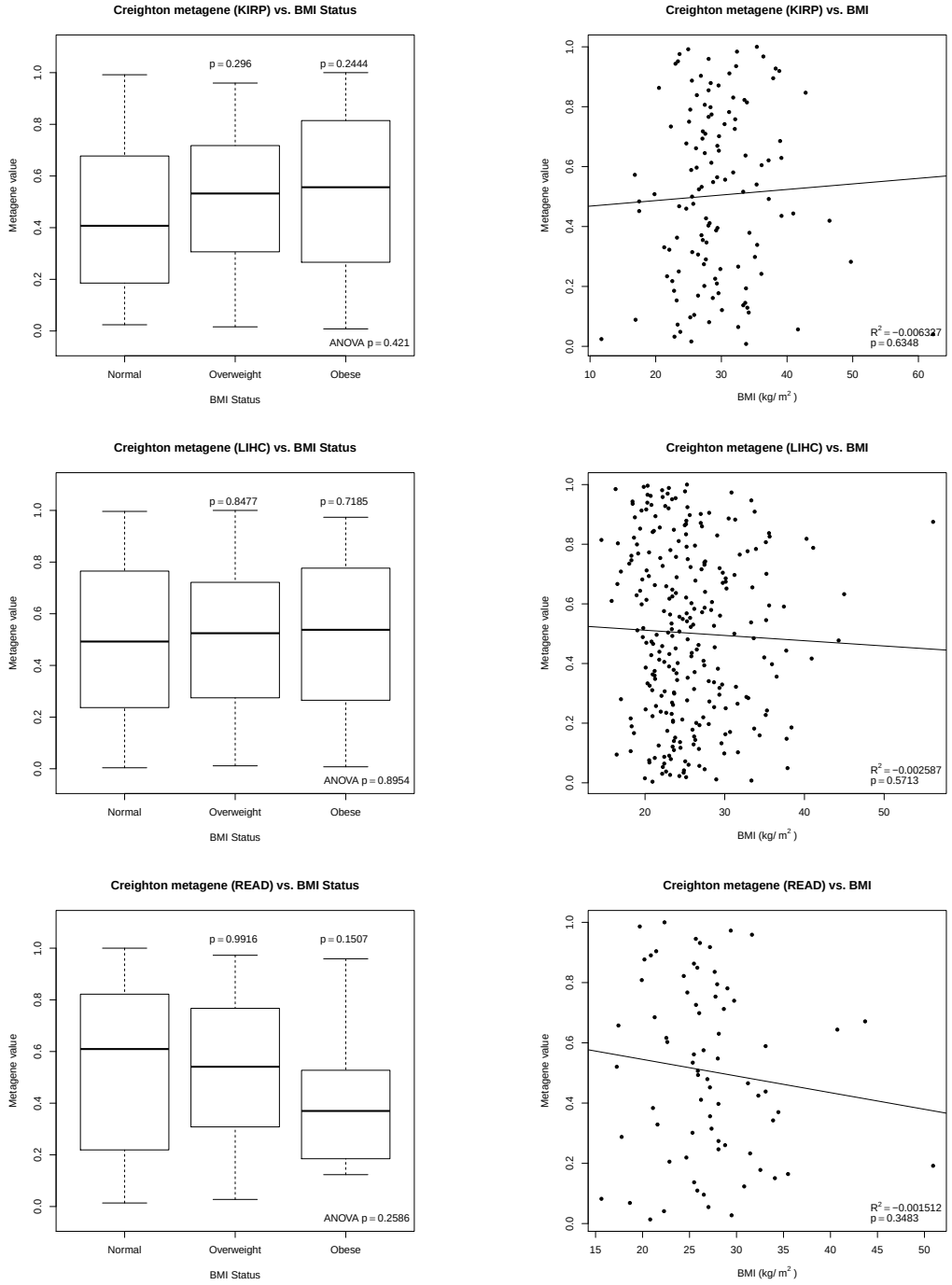


Figure S12 (cont.)

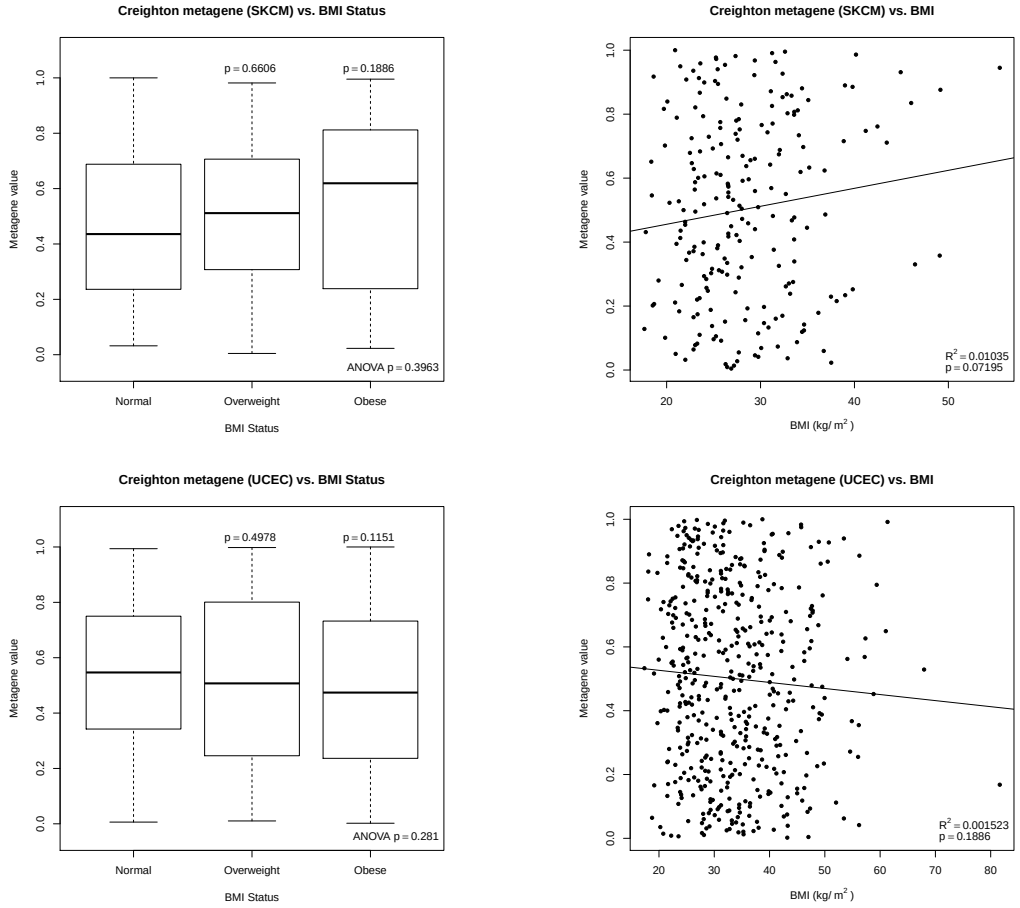


Figure S12 (cont.)

Table S3: Statistics of all the obesity metagenes in the ICGC KIRP cancer data

	P-values			Regression line statistics	
Metagenes	Overweight	Obese	ANOVA	R^2	P
Cr	0.1853	0.2590	0.3387	-0.005657	0.5798
Res	0.1063	0.1825	0.2128	-0.005037	0.5368
CrOl	0.2576	0.1757	0.3162	-0.005763	0.5879
ResOl	0.3120	0.1738	0.3198	-0.004513	0.5048
Ca	0.2431	0.2142	0.3591	-0.005582	0.5743
CaRes	0.2327	0.2344	0.3738	-0.006250	0.6279
CaOl	0.3426	0.2775	0.4693	-0.007333	0.7747
CaResOl	0.2643	0.1315	0.2410	-0.001408	0.3649

Table S4: Statistics of all the obesity metagenes in the ICGC LIHC cancer data

	P-values			Regression line statistics	
Metagenes	Overweight	Obese	ANOVA	R^2	P
Cr	0.4033	0.2900	0.5221	0.006348	0.1028
Res	0.5008	0.8743	0.8004	-0.0003709	0.3430
CrOl	0.7599	0.8562	0.9189	-0.001554	0.4424
ResOl	0.9238	0.9773	0.9938	-0.001283	0.4162
Ca	0.4898	0.8221	0.7917	0.0005899	0.2834
CaRes	0.5976	0.9929	0.8638	-0.001090	0.3990
CaOl	0.8717	0.8834	0.9679	-0.002010	0.4925
CaResOl	0.5369	0.7185	0.7073	-0.0004212	0.3465

Table S5: Statistics of all the obesity metagenes in the ICGC READ cancer data

	P-values			Regression line statistics	
Metagenes	Overweight	Obese	ANOVA	R^2	P
Cr	0.9129	0.4008	0.5942	-0.0132	0.8038
Res	0.7737	0.2244	0.4308	-0.008917	0.5484
CrOl	0.8439	0.1283	0.2643	0.001024	0.3036
ResOl	0.6953	0.06841	0.1766	0.01099	0.1840
Ca	0.6298	0.1326	0.2960	-0.001242	0.3432
CaRes	0.6302	0.1053	0.2453	0.001989	0.2885
CaOl	0.6871	0.1411	0.3358	-0.004847	0.4219
CaResOl	0.8821	0.08474	0.1817	0.001095	0.3025

Table S6: Statistics of all the obesity metagenes in the ICGC SKCM cancer data

	P-values			Regression line statistics	
Metagenes	Overweight	Obese	ANOVA	R^2	P
Cr	0.7115	0.3329	0.6264	0.002887	0.2033
Res	0.7604	0.5408	0.6396	-0.001725	0.4296
CrOl	0.4719	0.1224	0.2913	0.01261	0.05347
ResOl	0.8640	0.1668	0.3051	0.01541	**0.0372**[1]
Ca	0.8792	0.2084	0.3739	0.006674	0.1184
CaRes	0.8098	0.1950	0.3677	0.005746	0.1347
CaOl	0.9478	0.2726	0.4471	0.01009	0.07449
CaResOl	0.7309	0.2561	0.2830	0.01606	**0.03421**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S7: Statistics of all the obesity metagenes in the ICGC UCEC cancer data

	P-values			Regression line statistics	
Metagenes	Overweight	Obese	ANOVA	R^2	P
Cr	0.8357	0.08117	0.1039	0.006952	**0.03716**[1]
Res	0.8664	0.2023	0.3131	0.001998	0.1619
CrOl	0.9251	0.3800	0.5800	-0.000311	0.3569
ResOl	0.5826	0.09608	0.2156	0.003487	0.1021
Ca	0.7491	0.1103	0.1905	0.003811	0.09260
CaRes	0.8086	0.2002	0.3419	0.0005824	0.2584
CaOl	0.9068	0.2672	0.4153	0.00305	0.1166
CaResOl	0.8360	0.2440	0.4034	0.00022898	0.2863

[1] All values in bold are statistically significant ($p < 0.05$).

sample BMI value (continuous variable). Gene probe expressions were correlated with the sample BMI values and was corrected for multiple hypothesis testing with FDR. There were no significant gene that correlated with the sample BMI ($p < 0.05$).

With that said, the top 799 gene probes that correlated the most with the sample BMI were examined to see if there was any association. The results were similar to the other obesity associated genetic signatures: the continuous BMI signature significantly associated with sample BMI and BMI status in the CR data, but did not show any significance in any of the other cancer data.

A9 Pathways significant in each of the cancer types

The complete list of all the pathways significantly enriched using different pathway databases (FDR adjusted p-value < 0.05) are summarised in Tables S8 to S10.

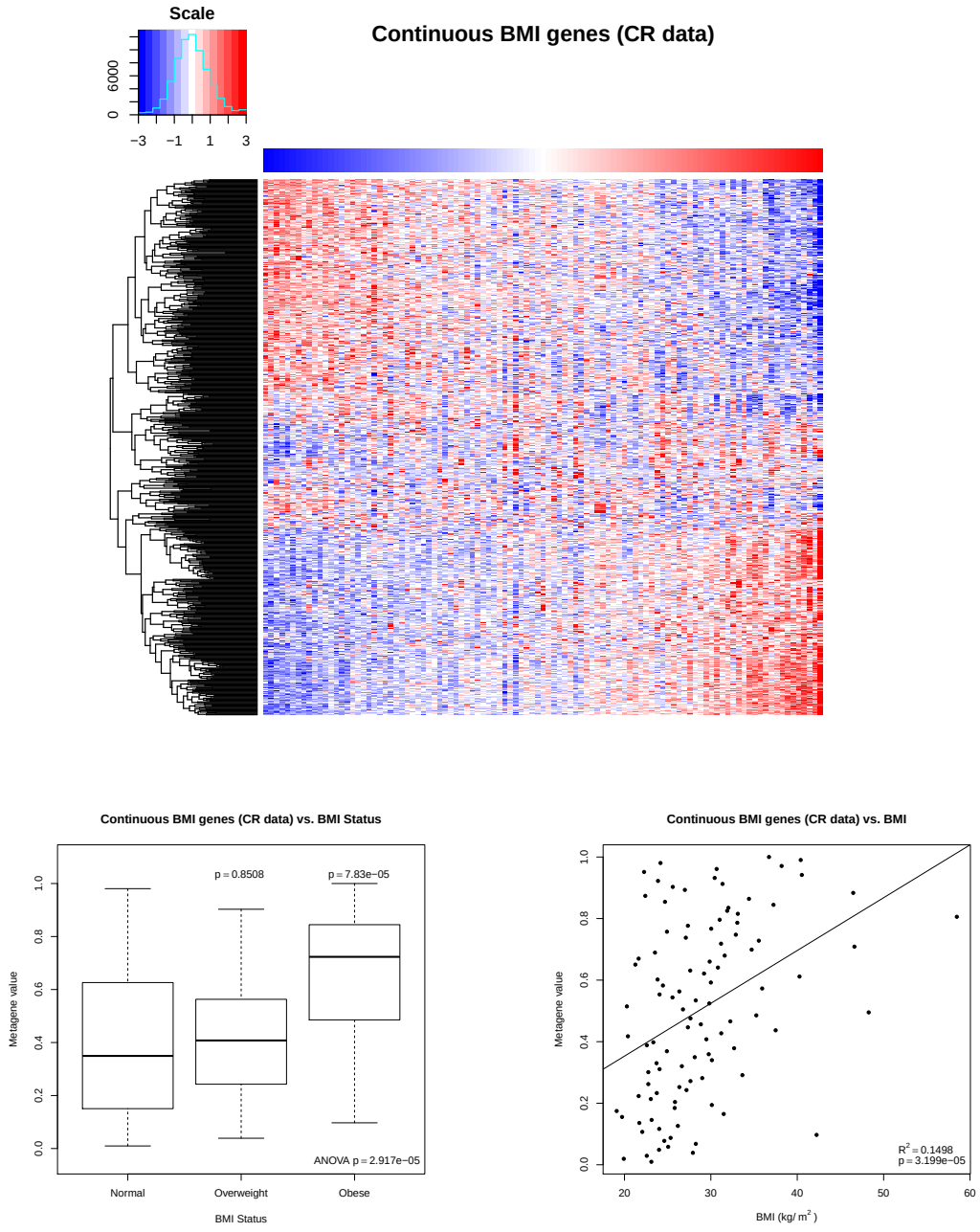


Figure S13: Heatmap, box plot and scatter plot showing the association of the continuous BMI metagene with the sample gene expression, BMI and BMI status, respectively, in various cancer data. Scales, p-values and R^2-value are as described in previous figures.

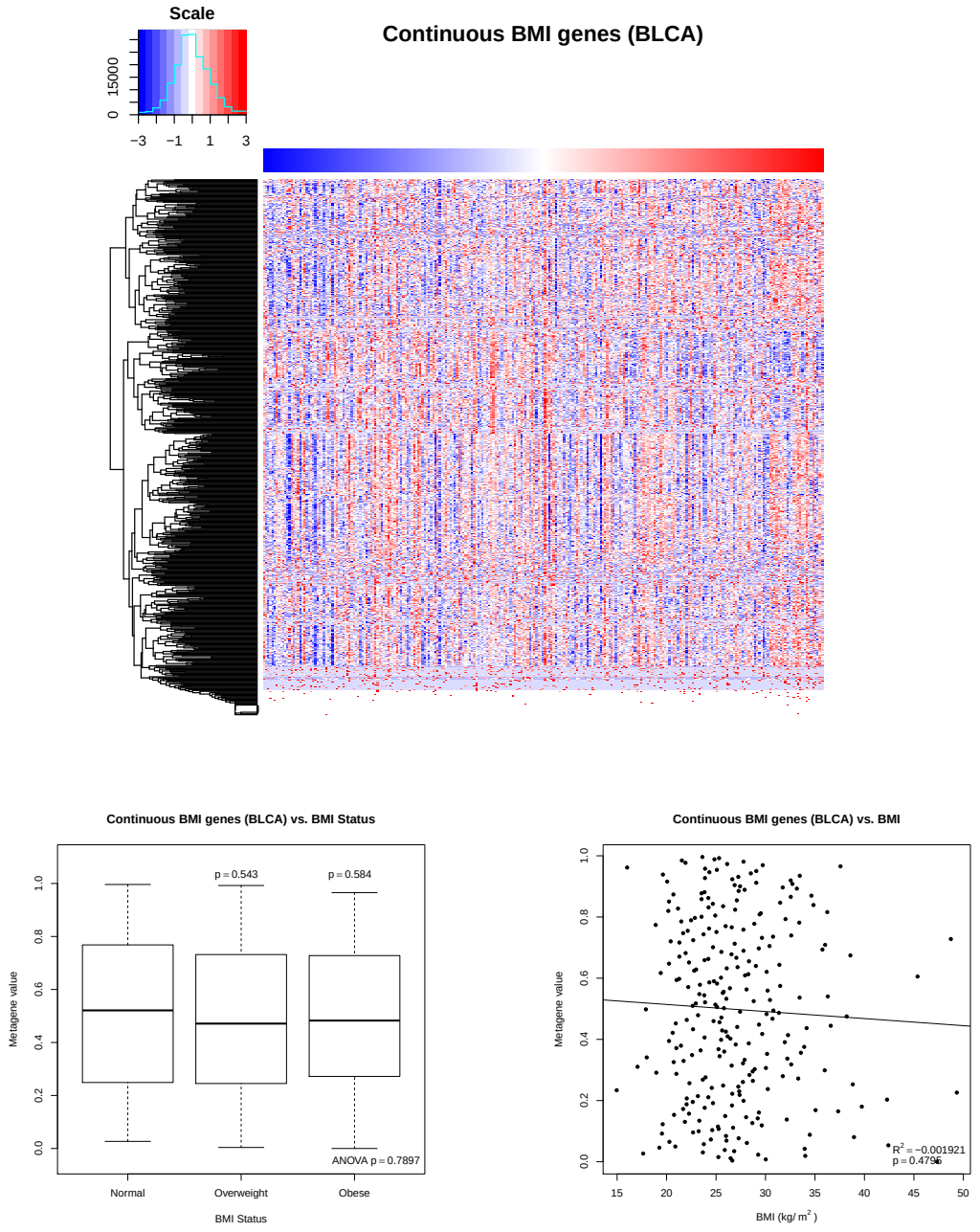
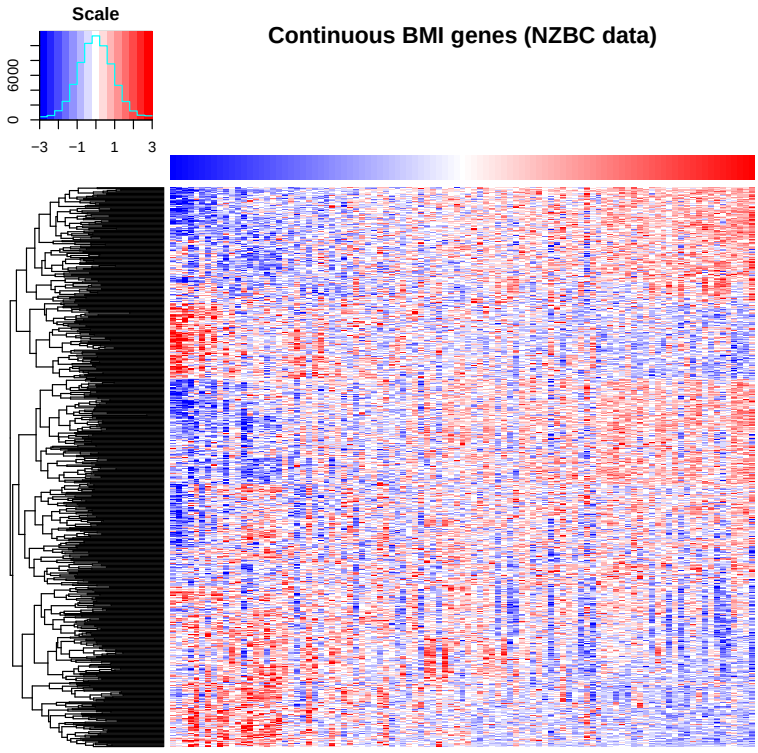
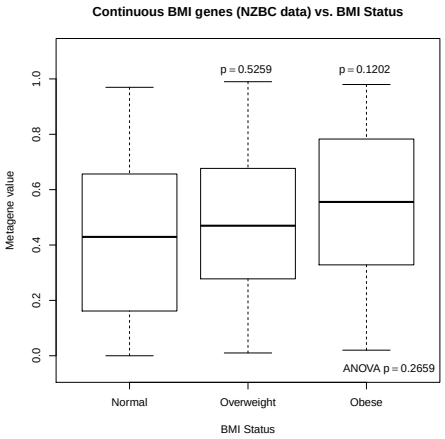


Figure S13 (cont.)





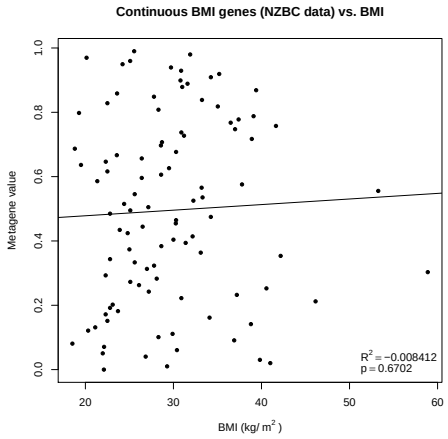


Figure S13 (cont.)

Table S8: Significantly enriched KEGG pathways in ICGC cancer data

Cancer type	Pathway
BLCA	NA[1]
CESC	ABC transporter
COAD	NA
KIRP	NA
LIHC	NA
READ	NA
SKCM	NA
UCEC	NA

[1] Not available (no pathway enriched)

Table S9: Significantly enriched Reactome pathways in ICGC cancer data

Cancer type	Pathway
BLCA	Phosphate bond hydrolysis by NUDT proteins
CESC	Mitochondrial ABC transporters
COAD	NA[1]
KIRP	NA
LIHC	NA
READ	NA
SKCM	NA
UCEC	NA

[1] Not available (no pathway enriched)

Table S10: Significantly enriched GO pathways in ICGC cancer data

Cancer type	Pathway
BLCA	Nucleobase-containing small molecule catabolic process
	SUMO-specific protease activity
	Diphosphoinositol-polyphosphate diphosphatase activity
	Inositol diphosphate tetrakisphosphate diphosphatase activity
	Inositol-1-diphosphate-2,3,4,5,6-pentakisphosphate diphosphatase activity
	Inositol-3-diphosphate-1,2,4,5,6-pentakisphosphate diphosphatase activity
	Inositol-5-diphosphate-1,2,3,4,6-pentakisphosphate diphosphatase activity
	Inositol-1,5-bisdiphosphate-2,3,4,6-tetrakisphosphate 1-diphosphatase activity
	Inositol-1,5-bisdiphosphate-2,3,4,6-tetrakisphosphate 5-diphosphatase activity
	Inositol-3,5-bisdiphosphate-2,3,4,6-tetrakisphosphate 5-diphosphatase activity
	Phosphatidylinositol 3-kinase complex, class IB
	Neurotrophin binding
	1-phosphatidylinositol-5-phosphate 4-kinase activity
	Coagulation
	Cholesterol O-acyltransferase activity
	Pyrimidine deoxyribonucleotide catabolic process
	Nucleotidase activity
	G-protein coupled neurotensin receptor activity
	8-oxo-7,8-dihydroguanosine triphosphate pyrophosphatase activity
	8-oxo-7,8-dihydrodeoxyguanosine triphosphate pyrophosphatase activity
	IDP catabolic process
	ADP-sugar diphosphatase activity

Table S10 (continued)

CESC	ATP-binding cassette (ABC) transporter complex
	Protein delipidation
	Atg8 ligase activity
	Cyclin-dependent protein kinase 5 holoenzyme complex
	Molybdopterin synthase complex
	Negative regulation of nodal signaling pathway
	Aspartate-tRNA ligase activity
	Peroxisomal long-chain fatty acid import
	Sterol-transporting ATPase activity
	Negative regulation of intestinal phytosterol absorption
	Negative regulation of intestinal cholesterol absorption
	Apolipoprotein A-I receptor activity
	Sulfonylurea receptor activity
	Atg12 transferase activity
	Superior olivary nucleus maturation
	Cyclin-dependent protein kinase 5 activator activity
	ATPase activity, coupled to transmembrane movement of substances
COAD	NA
KIRP	Trace-amine receptor activity
	Regulation of viral release from host cell
	Nuclear import
LIHC	NA
READ	Nuclear polyadenylation-dependent rRNA catabolic process
	Nuclear polyadenylation-dependent tRNA catabolic process
	Sodium channel complex
	U1 snRNA 3'-end processing
	U5 snRNA 3'-end processing
	Glucuronosyl-N-acetylglucosaminyl-proteoglycan 4-alpha-N-acetylglucosaminyltransferase activity
	Pyrroline-5-carboxylate reductase activity
	Hydroxymethylglutaryl-CoA reductase (NADPH) kinase activity

Table S10 (continued)

	Acetyl-CoA carboxylase kinase activity
	rRNA modification
	Rab GDP-dissociation inhibitor activity
	Nuclear retention of pre-mRNA with aberrant 3'-ends at the site of transcription
	Cellular polysaccharide biosynthetic process
	Heparan sulfate N-acetylglucosaminyltransferase activity
	N-acetylglucosaminyl-proteoglycan 4-beta-glucuronosyltransferase activity
	Glucuronyl-galactosyl-proteoglycan 4-alpha-N-acetylglucosaminyltransferase activity
	Regulation of RIG-I signaling pathway
	Regulation of cysteine-type endopeptidase activity
	Adenylate cyclase-inhibiting serotonin receptor signaling pathway
	Peptide-O-fucosyltransferase activity
	Interferon-gamma receptor activity
SKCM	Tissue morphogenesis
	Beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase activity
	N-acetyllactosaminide beta-1,6-N-acetylglucosaminyltransferase activity
	Posterior mesonephric tubule development
	Regulation of stem cell differentiation
	Negative regulation of nodal signaling pathway
	Aspartate-tRNA ligase activity
	Glutamate-cysteine ligase complex
	Glutamate-cysteine ligase activity
	Rab GDP-dissociation inhibitor activity
UCEC	SUMO-specific protease activity
	Regulation of dendritic cell cytokine production
	Detection of triacyl bacterial lipopeptide

Table S10 (continued)

	Cellular response to triacyl bacterial lipopeptide
	Toll-like receptor 1-Toll-like receptor 2 protein complex
	Detection of diacyl bacterial lipopeptide
	Cellular response to diacyl bacterial lipopeptide
	Toll-like receptor 2-Toll-like receptor 6 protein complex
	Diacyl lipopeptide binding
	Coagulation
	Polo kinase kinase activity
	Regulation of protein serine/threonine kinase activity
	DNA-dependent protein kinase complex
	Methylosome

[1] Not available (no pathway enriched).

A10 Results for ER and PR metagene analysis

To determine whether the results from all of the obesity metagene analyses were not showing any association with patient BMI because of the SVD/transformation matrix approach used, the same approach was used with ER and PR pathway associated genetic signatures from the Gatza *et al.* (2010) study. Firstly, both the ER and PR metagenes were generated in the RMA normalised CR data set. The generated metagenes were compared with the gene expressions of the corresponding pathway associated genetic signatures, and the association with the patient ER and PR statuses were also examined. As shown in Figures S14 and S15, both the ER and PR metagenes in the CR data showed assocation with the gene expression pattern of the pathway, and also showed significant association with the patient ER and PR statuses.

The transformation matrices for the ER and PR metagenes were created in the CR data, and was applied to the NZBC and FM data sets (RMA normalised). As clearly shown by Figures S16 to S19, both the ER and PR metagenes associated with the gene expressions of the corresponding pathway signature. More importantly, both metagenes were significantly associated with the patient ER and

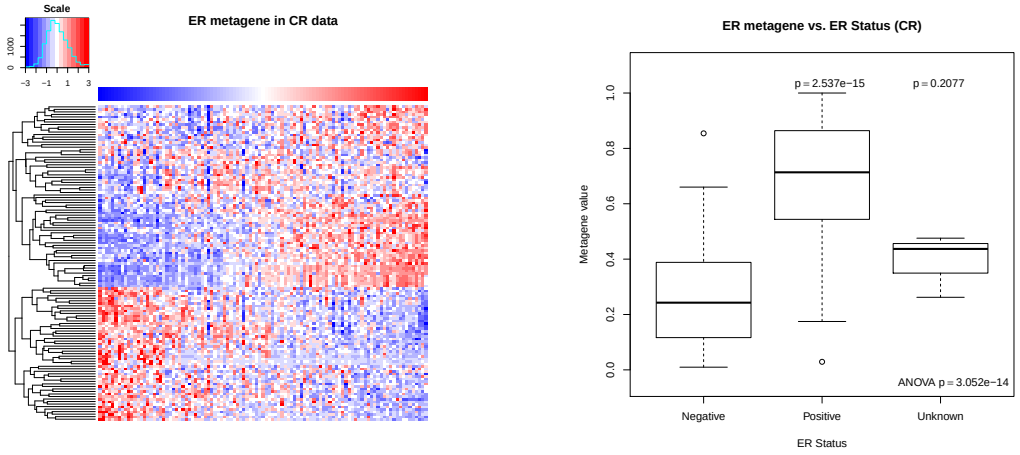


Figure S14: Heatmap and box plot showing the association of the ER metagene with the sample gene expression and patient ER status, respectively, in the CR data. Scales for the heatmap is as described in previou figures. P-values are calculated relative to the ER^- group.

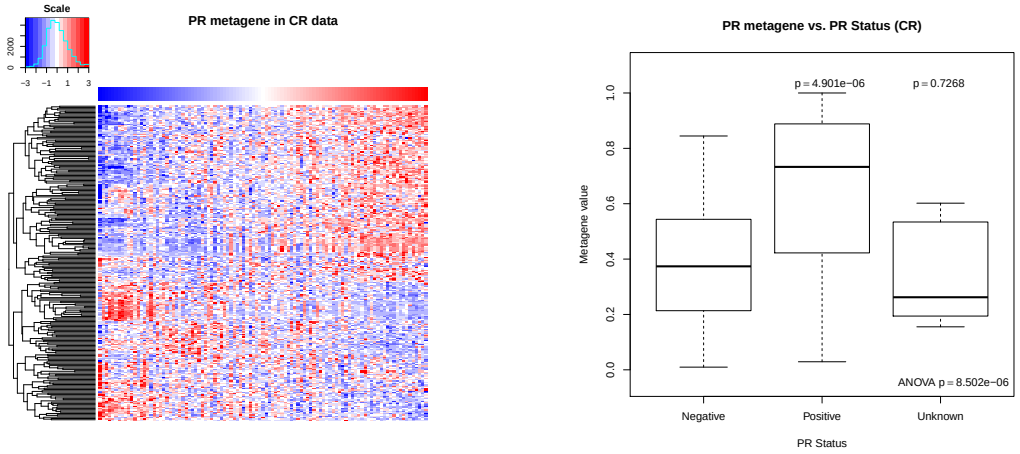


Figure S15: Heatmap and a box plot showing the association of the PR metagene with the sample gene expression and patient PR status, respectively, in the CR cancer data. Scales and p-values are as described in previous figures.

PR statuses, even though these metagenes were generated with the transformation matrices from the CR data. These results show that the approach taken in this project does not affect the link between the genetic signature and the patient phenotype, if such link exists between them.

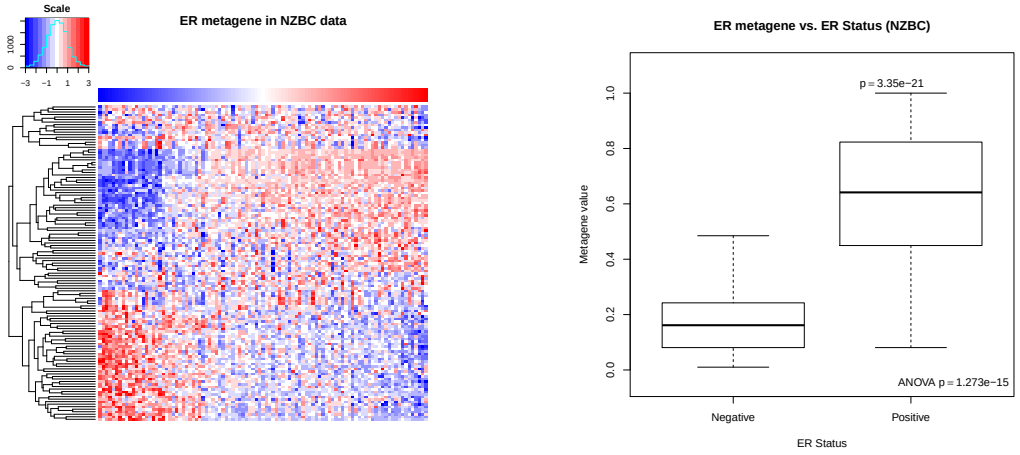


Figure S16: Heatmap and box plot showing the association of the ER metagene with the sample gene expression and patient ER status, respectively, in the NZBC data. Scales for the heatmap is as described in previou figures. P-values are calculated relative to the ER$^-$ group.

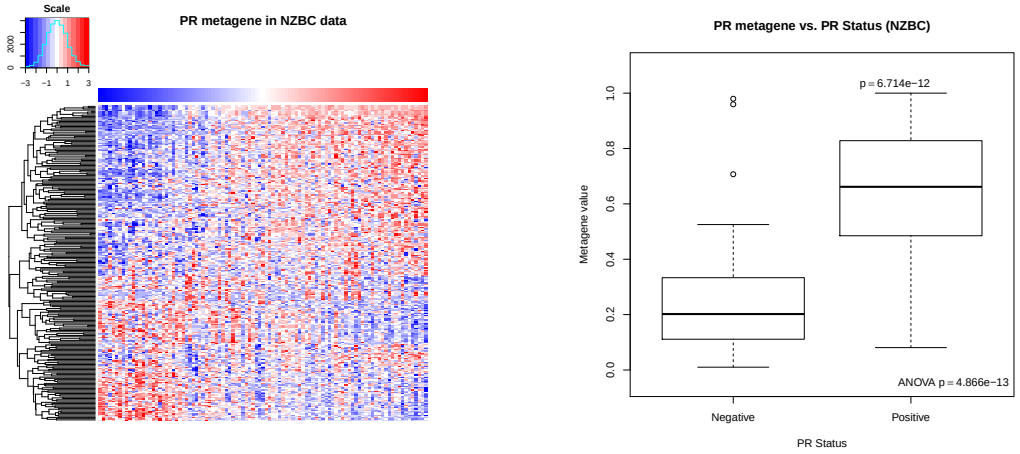


Figure S17: Heatmap and a box plot showing the association of the PR metagene with the sample gene expression and patient PR status, respectively, in the NZBC cancer data. Scales and p-values are as described in previous figures.

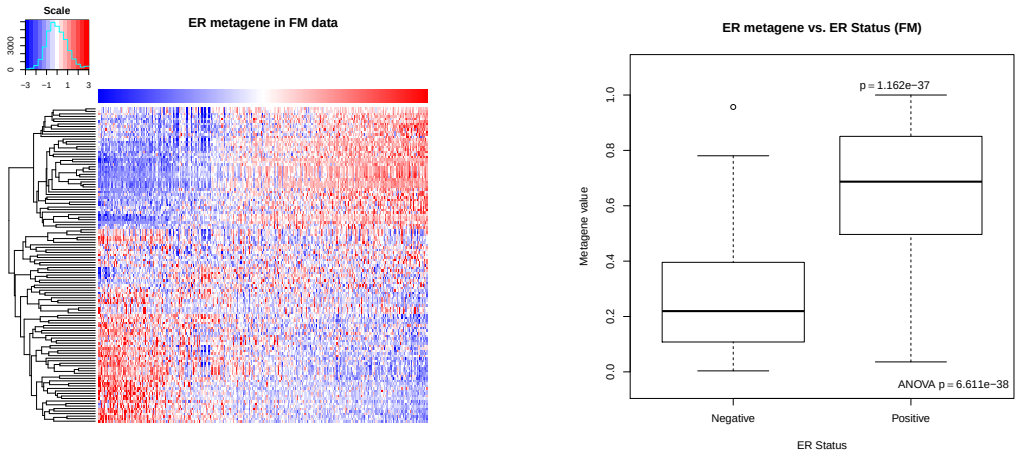


Figure S18: Heatmap and box plot showing the association of the ER metagene with the sample gene expression and patient ER status, respectively, in the FM data. Scales for the heatmap is as described in previou figures. P-values are calculated relative to the ER^- group.

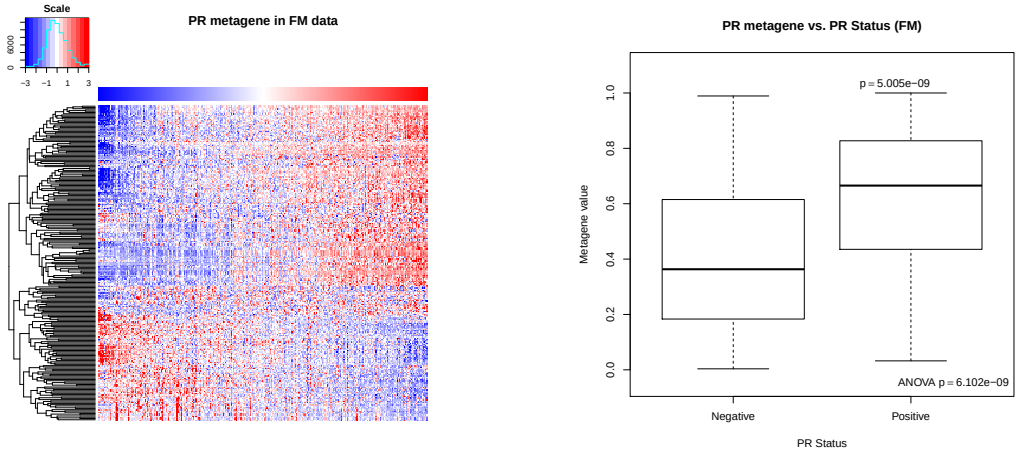


Figure S19: Heatmap and a box plot showing the association of the PR metagene with the sample gene expression and patient PR status, respectively, in the FM cancer data. Scales and p-values are as described in previous figures.

Appendix B

Additional results from Chapter 4

B1 Ranking method for the pathway associated genetic signatures

To determine whether it was the best to rank the metagene scores based on the number of samples or rank the scores with the probit method, the two ranking methods were compared in scatter plots with the GT pathway metagenes in RMA or MAS5 normalised GT data. Since the results from the RMA and MAS5 normalised GT data were so similar to one another, only the results from the RMA normalised data are presented (Figure S20). Figure S20 showed that the metagene scores for all of the GT pathway associated genetic signatures were approximately the same in either of the ranking methods used.

B2 Normalisation method for the pathway associated genetic signature transformation matrices

To investigate whether the normalisation methods were important in the creation of the pathway associated genetic signature transformation matrices, and whether the normalisation methods of the data affected the resulting metagenes, transformation matrices were generated in RMA or MAS5 normalised GT data set and

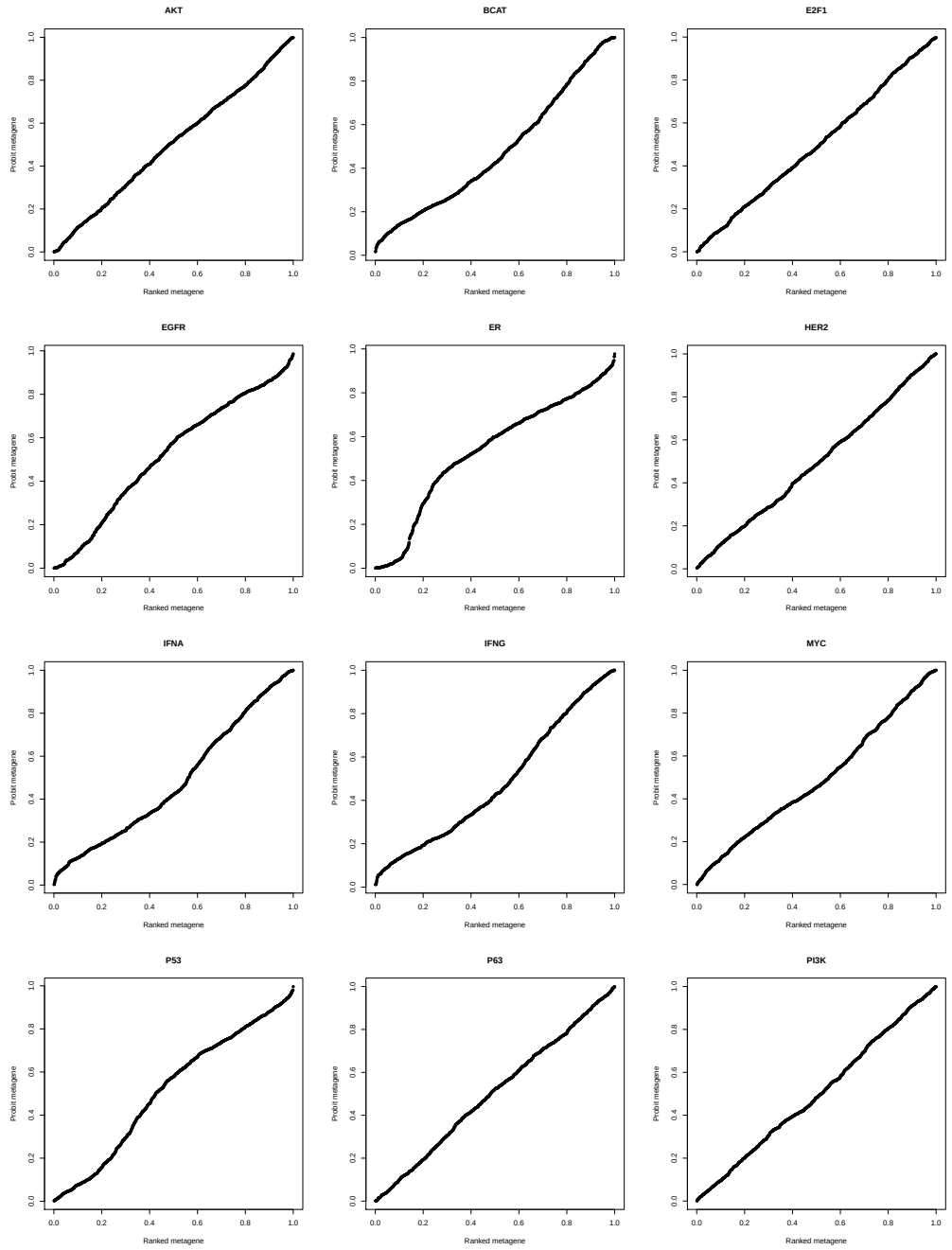


Figure S20: Scatter plots comparing the GT pathway metagenes ranked with probit and ranked based on number of samples in RMA normalised GT data.

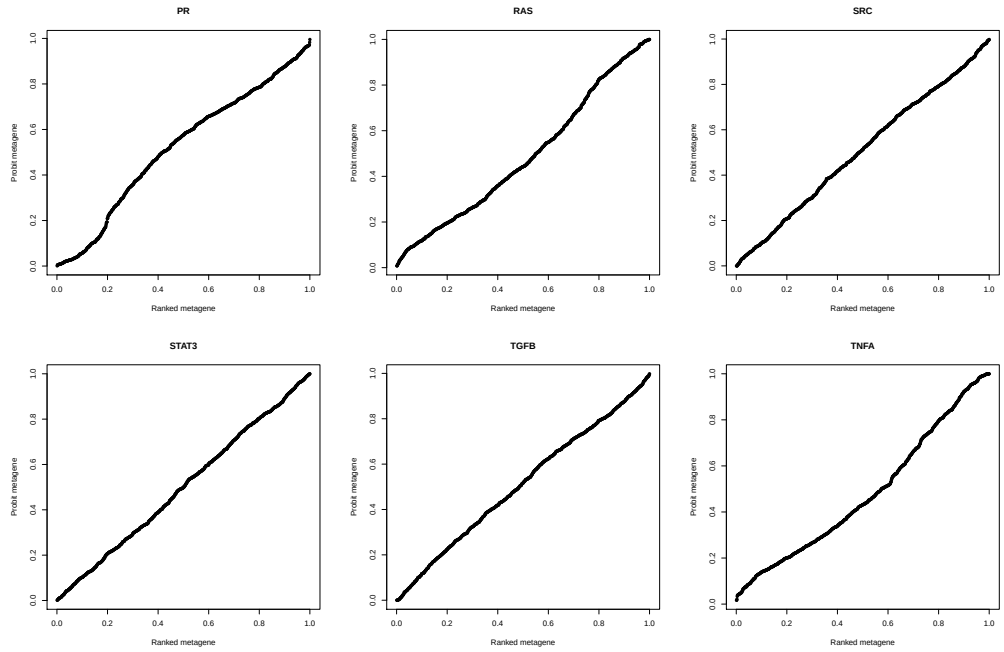


Figure S20 (cont.)

then applied to the RMA or MAS5 normalised GT data set. The results in Figure S21 showed less variability of the metagene scores when the metagenes were derived from the data with the same normalisation method, regardless of the transformation matrix used to produce the metagenes. This was most clearly demonstrated by the BCAT, ER, IFNα, IFNγ, p53 and PR pathway associated genetic signatures.

As mentioned in Section 4.1 some of the pathway metagenes showed greater variability than the other pathway metagene scores. This was also seen in other data sets as well (only the PR and TGFβ metagenes shown; Figure S22).

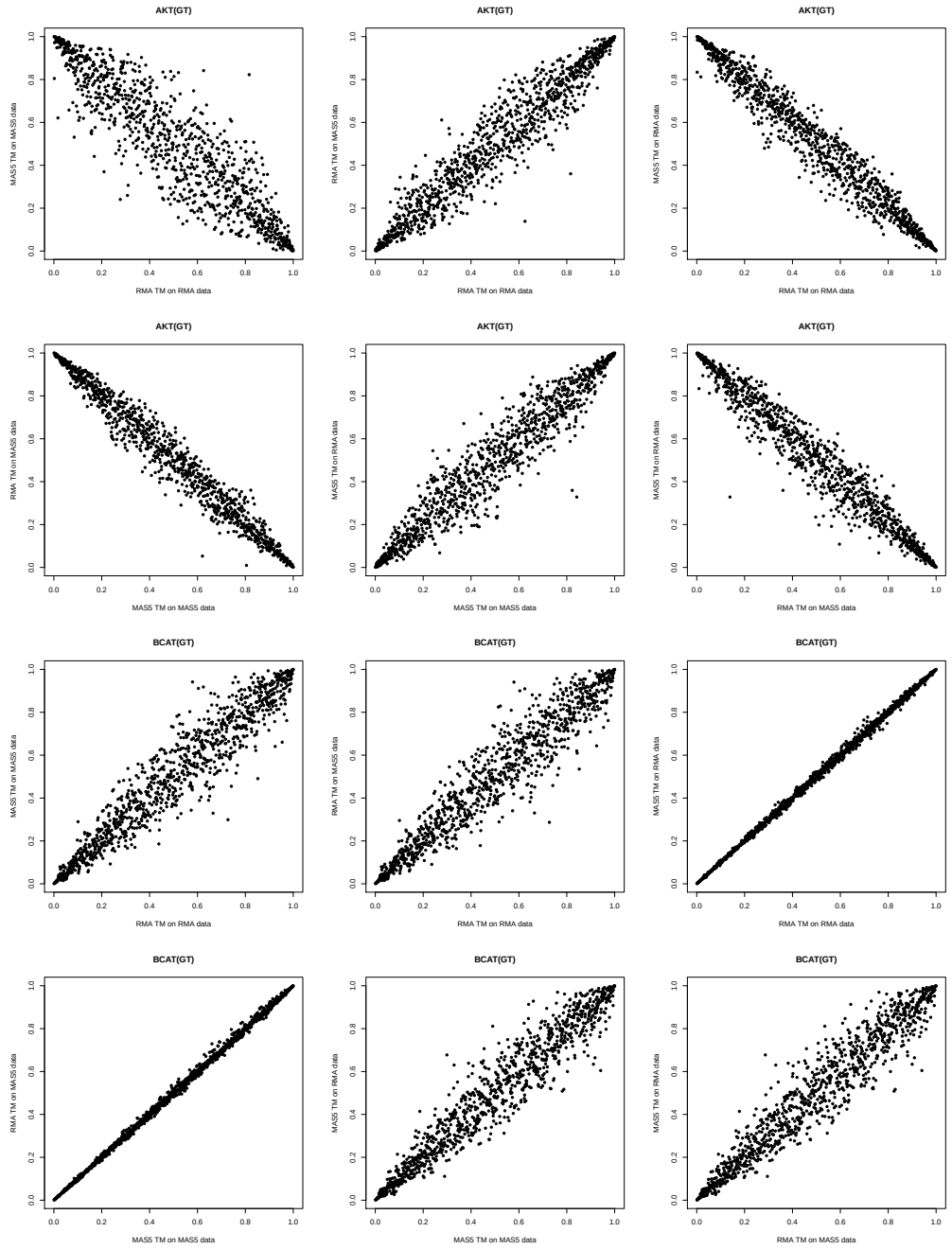


Figure S21: Scatter plots comparing the GT pathway metagenes derived from the transformation matrices in RMA- or MAS5-normalised GT data set. All of the transformation matrices were generated in either the RMA or MAS5 normalised GT data set.

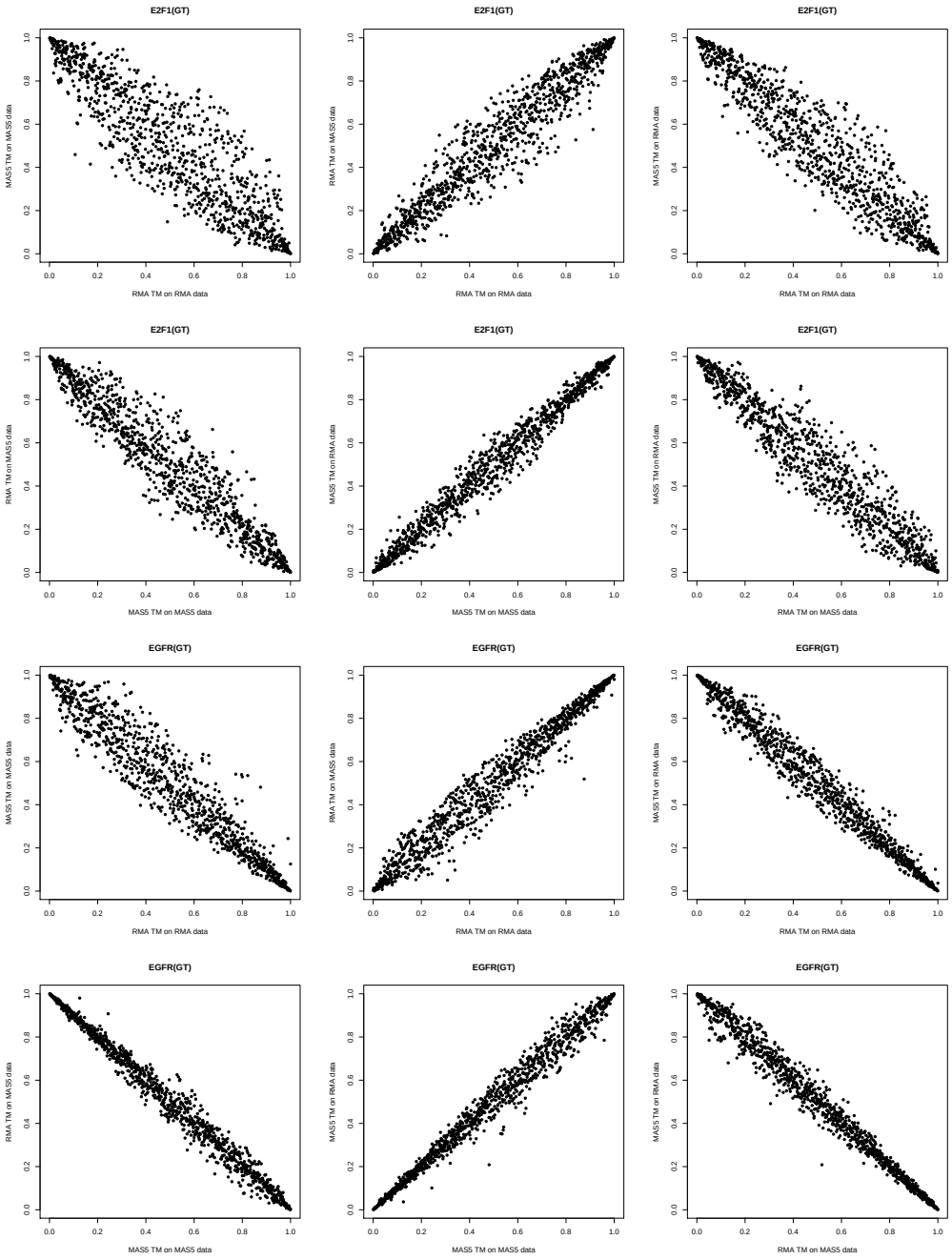


Figure S21 (cont.)

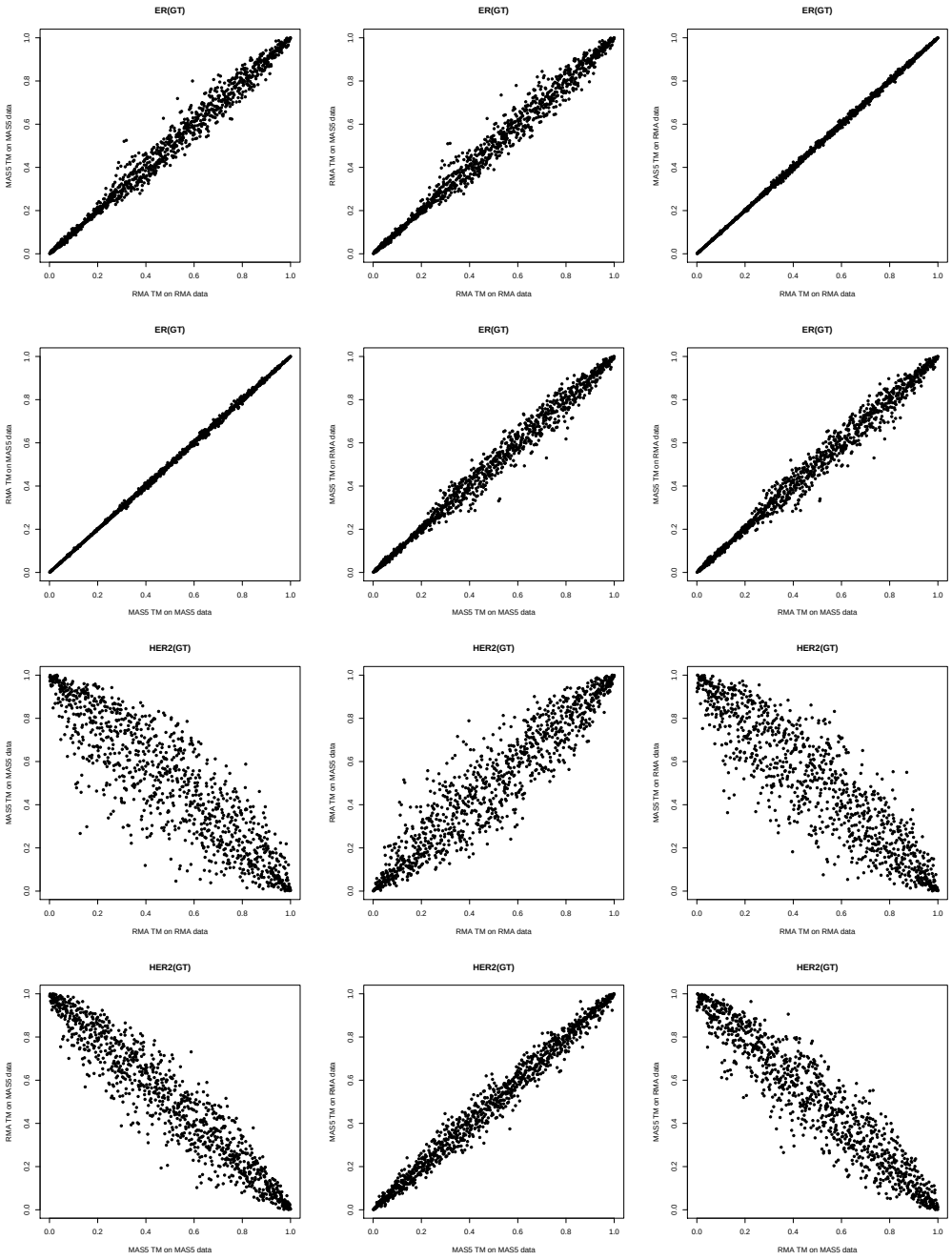


Figure S21 (cont.)

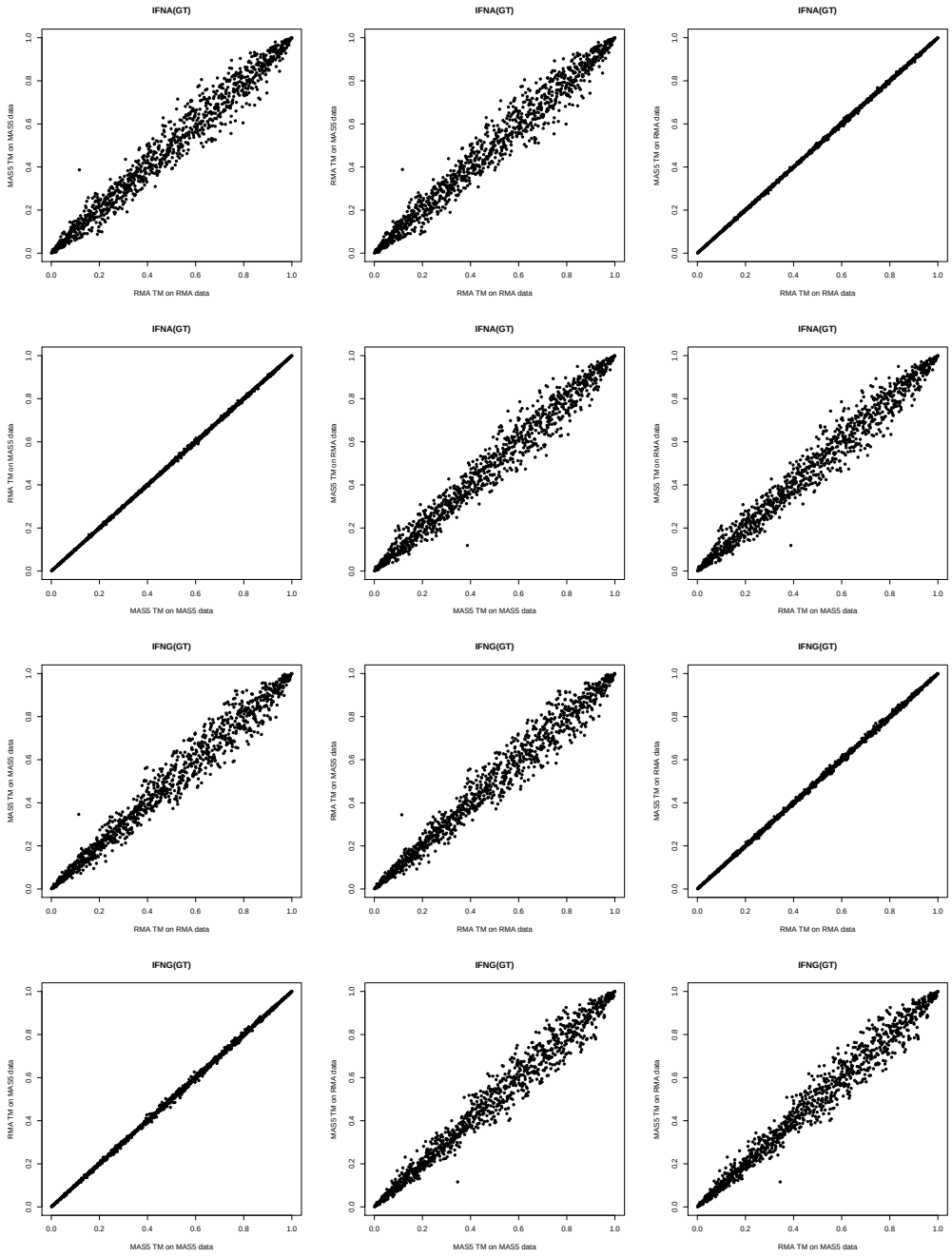


Figure S21 (cont.)

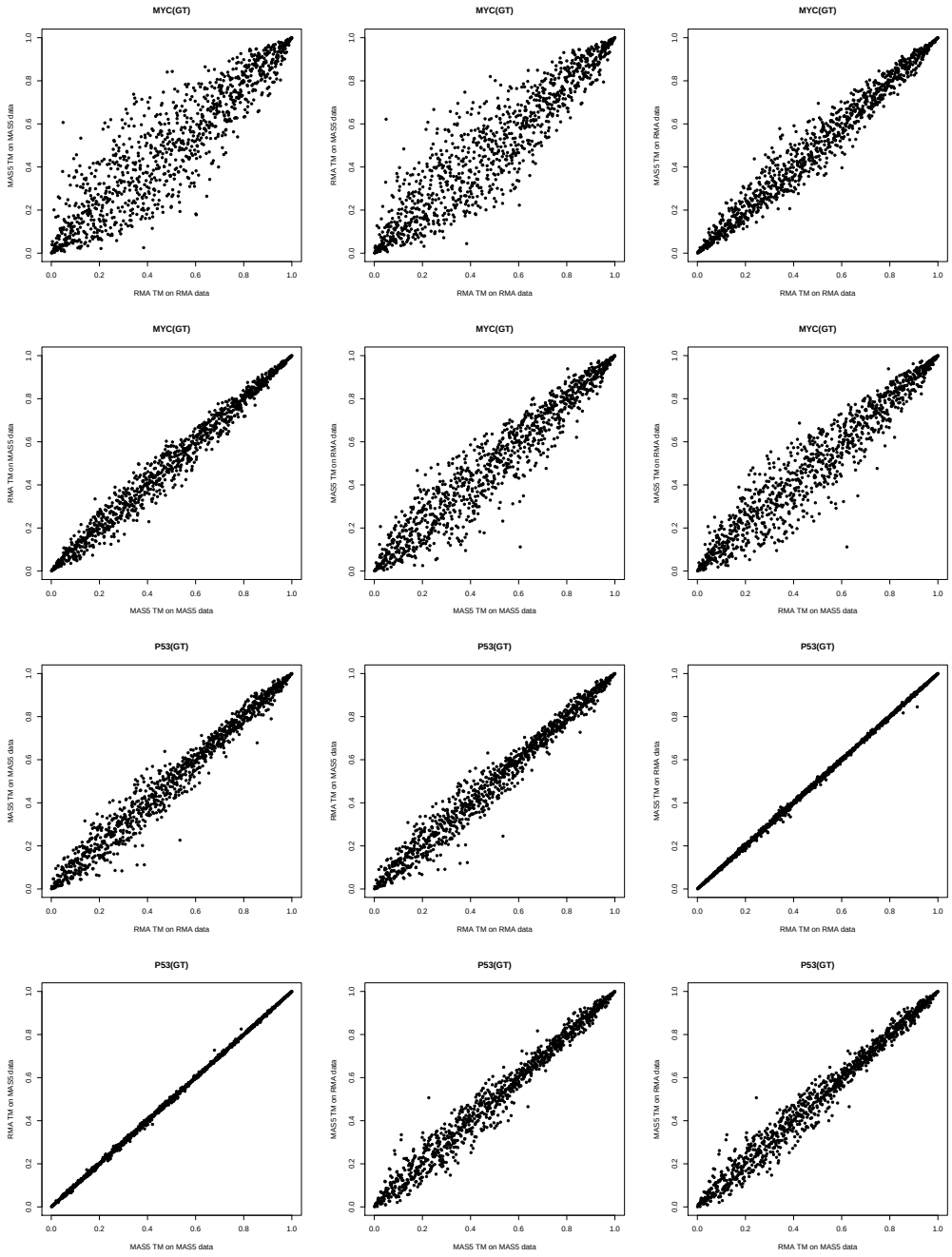


Figure S21 (cont.)

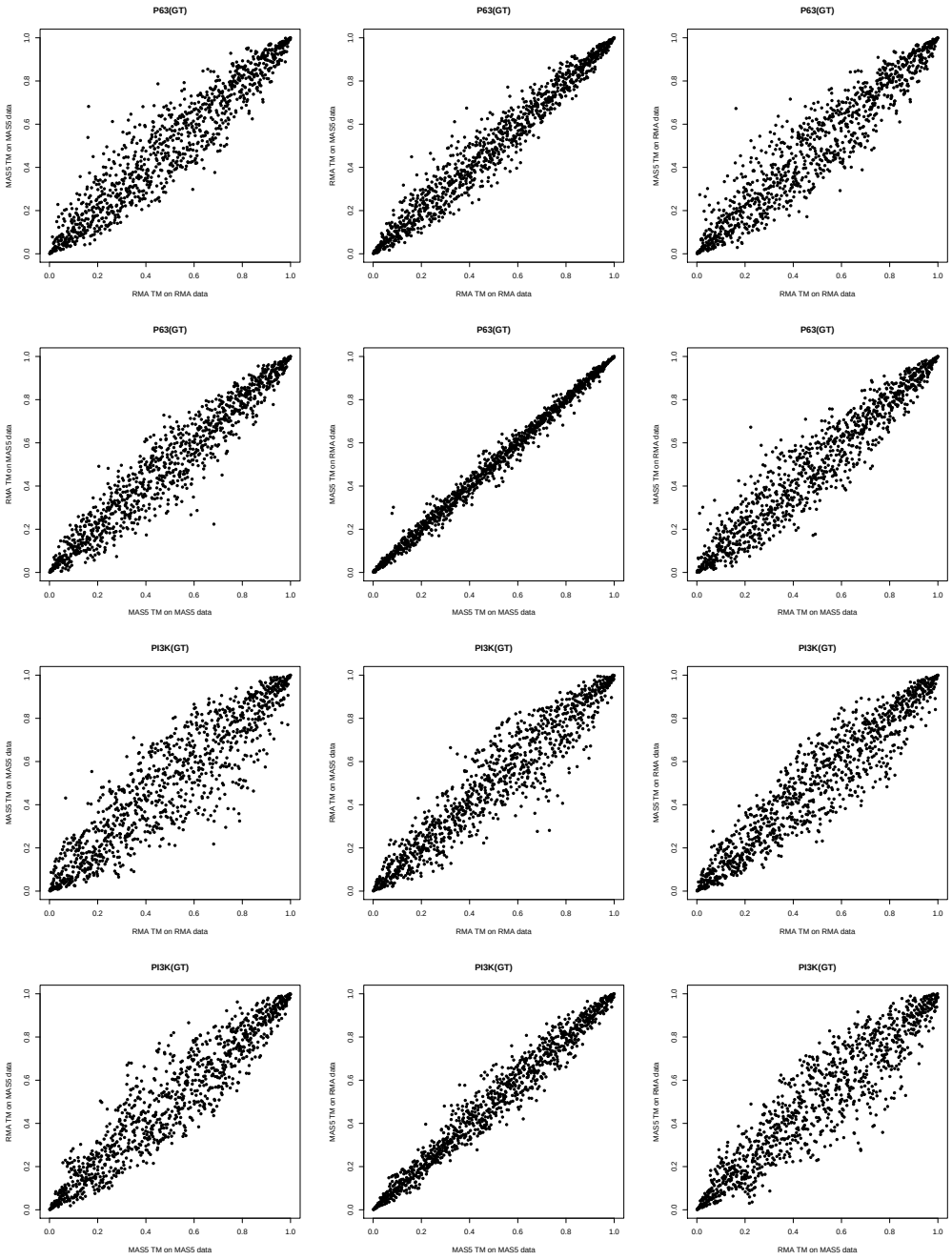


Figure S21 (cont.)

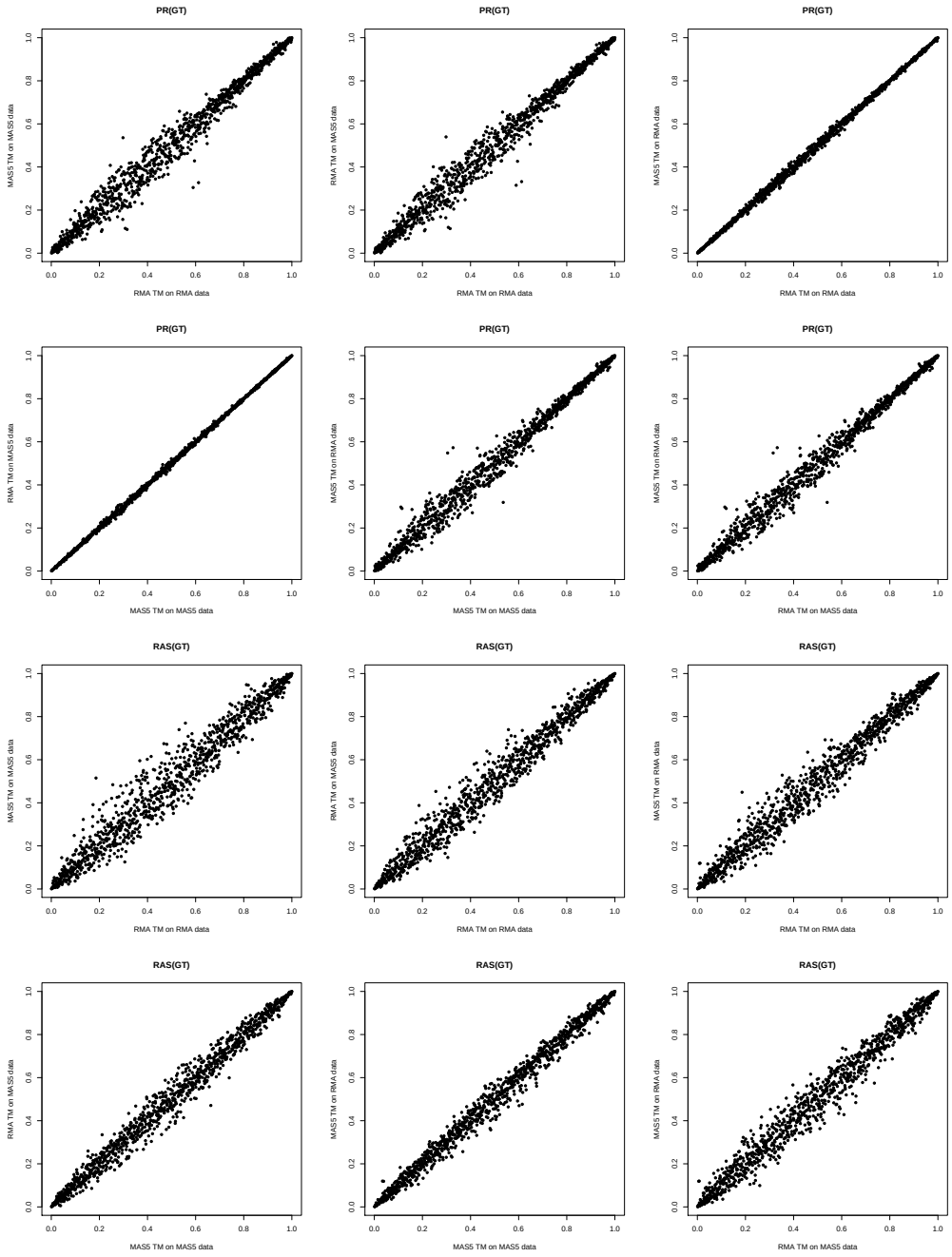


Figure S21 (cont.)

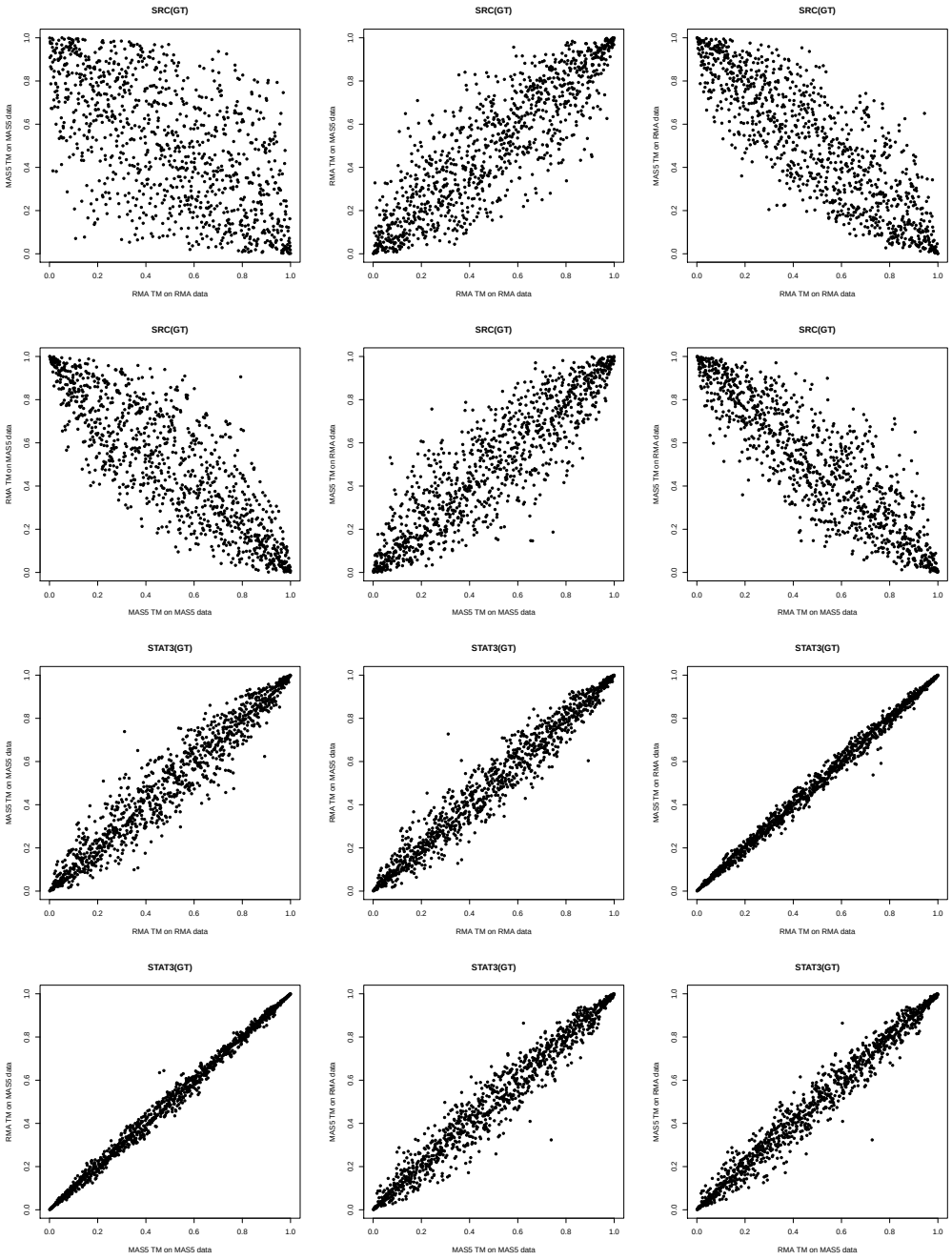


Figure S21 (cont.)

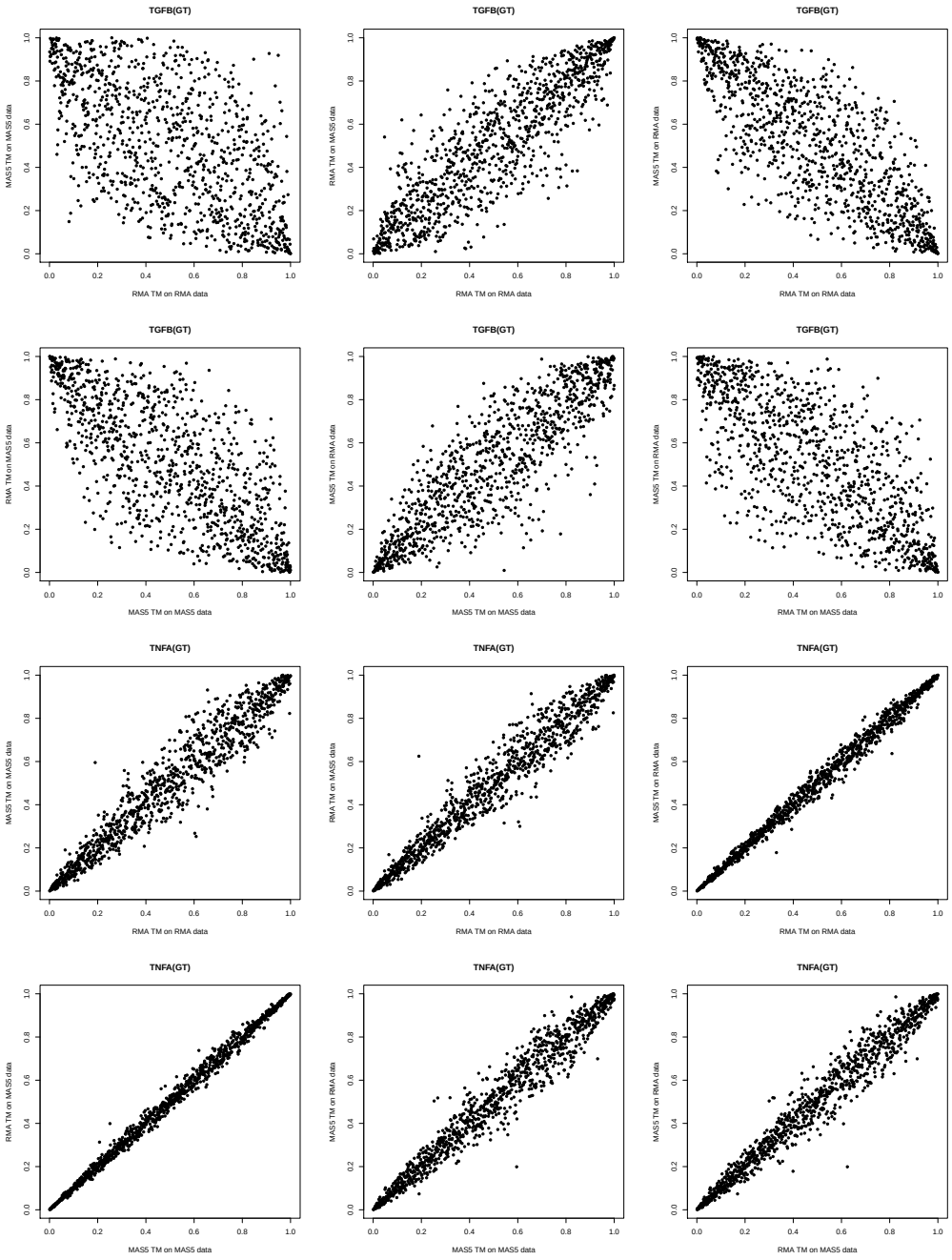


Figure S21 (cont.)

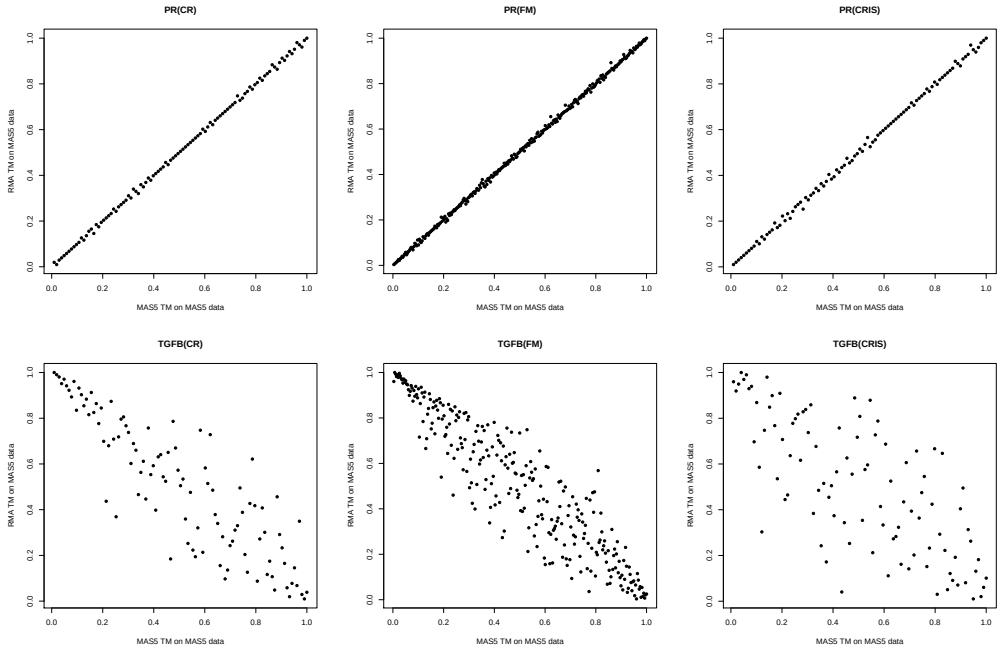


Figure S22: Comparison of the pathway metagenes generated in the MAS5 normalised CR, FM and NZBC data sets from the application of the transformation matrices derived from either the RMA or MAS5 normalised GT data. Only the PR (top) and the TGFβ (bottom) pathway metagenes are shown, due to the large number of scatter plots generated.

B3 Additional results used to determine the directionality of the pathway metagenes

The directionality of the pathway metagenes were determined by comparing the metagene values with the gene expression of the representative gene for the pathway associated genetic signature, as well as the gene expression pattern of the pathway signature (Figure S23). Together with the original results presented by Gatza *et al.* (2010) (Appendix B4), the final directions of the pathway metagenes were decided.

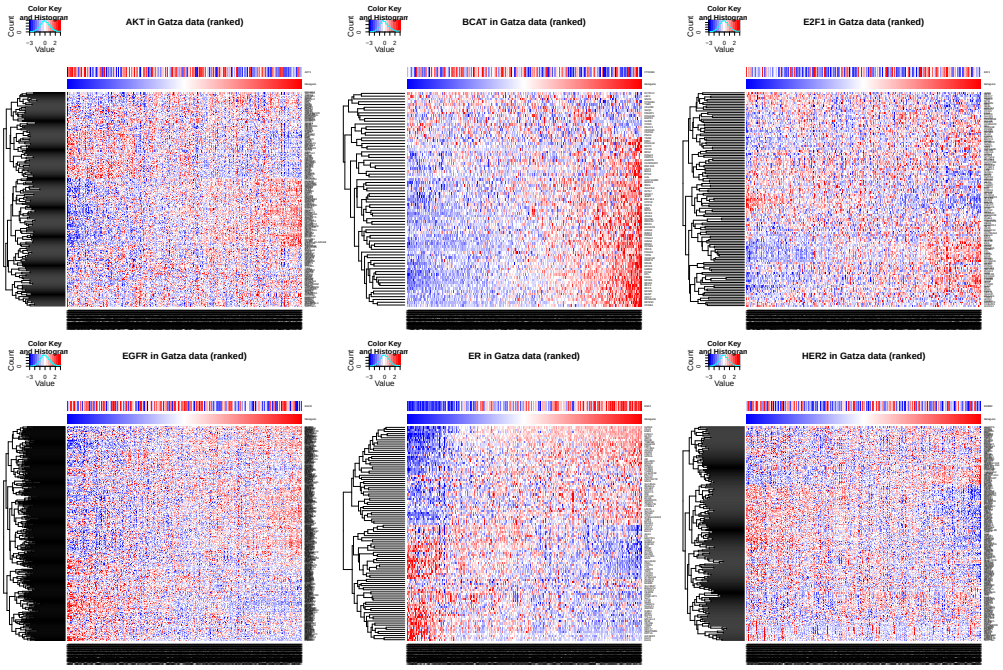


Figure S23: Heatmaps showing the gene expression of the pathway associated genetic signatures, pathway metagenes, and the expression of the representative pathway gene in RMA-normalised GT data set. For each pathway associated genetic signature, the gene expression pattern of the signature was plotted in the heatmap. The expression of the representative gene for the pathway signature (see Table 7 in Section 2.6.3) was plotted seaparately as a bar above the pathway metagene scores.

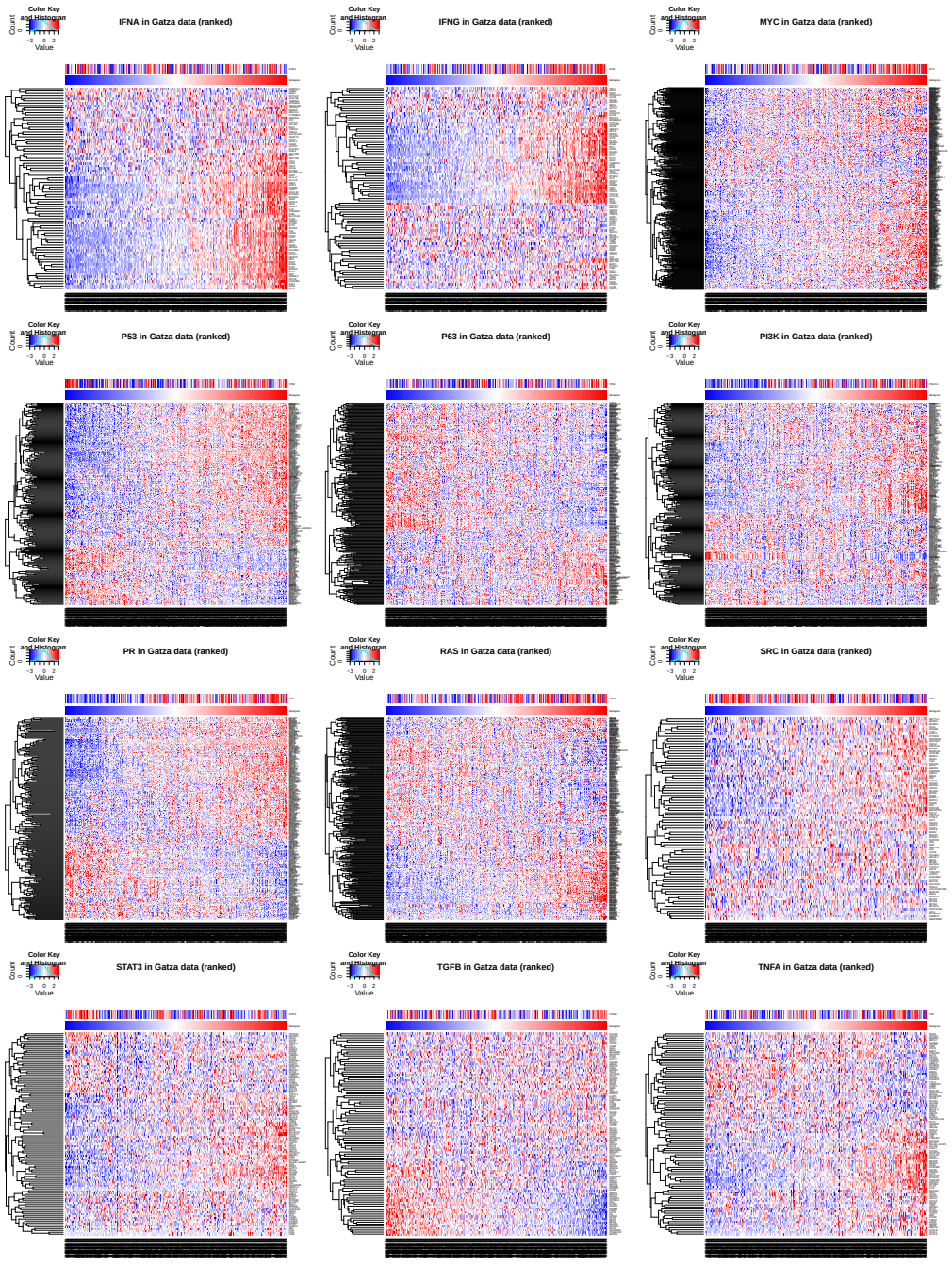


Figure S23 (cont.)

B4 Original result from the Gatza *et al.* (2010) study

The original result from the Gatza *et al.* (2010) study were used to determine the directions of the GT pathway associated genetic signatures in Section 4.1 (Figure S24).

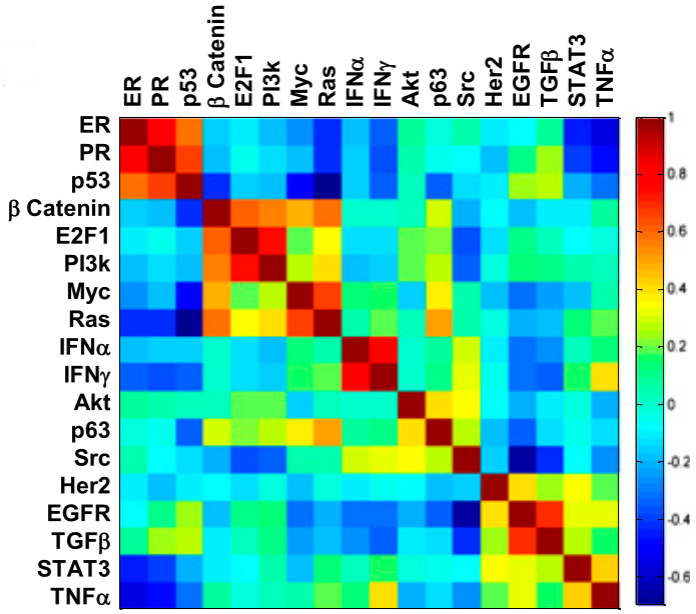


Figure S24: Heatmap showing the correlation of all of the GT pathway associated genetic signatures with one another. Figure adapted from the Gatza *et al.* (2010) study.

B5 Correlation of the GT pathway metagenes in other cancer data sets

The correlation of the pathway associated genetic signatures were plotted as heatmaps for the CR, FM and NZBC data sets to confirm that the pathway transformation matrices from the GT data set were generating metagenes that were similar to those in the original RMA normalised GT data. Most importantly, the pathway metagenes generated from the transformation matrices had to cluster in a similar pattern as the pathway metagenes in the GT data set. This was to confirm that

the directionality of the metagenes were conserved even after the transformation matrices were applied to the data set

B6 Directionality of the obesity metagenes in the GT data

The directionality of the CR obesity associated genetic signatures were determined in the RMA normalised GT data set to make sure these signatures were in the "correct" direction, and therefore able to be compared with the pathway associated genetic signatures appropriately. As clearly shown in Figure S26, the Res metagene was flipped as the Res metagene was in the oppsite direction compared with the other obesity metagenes.

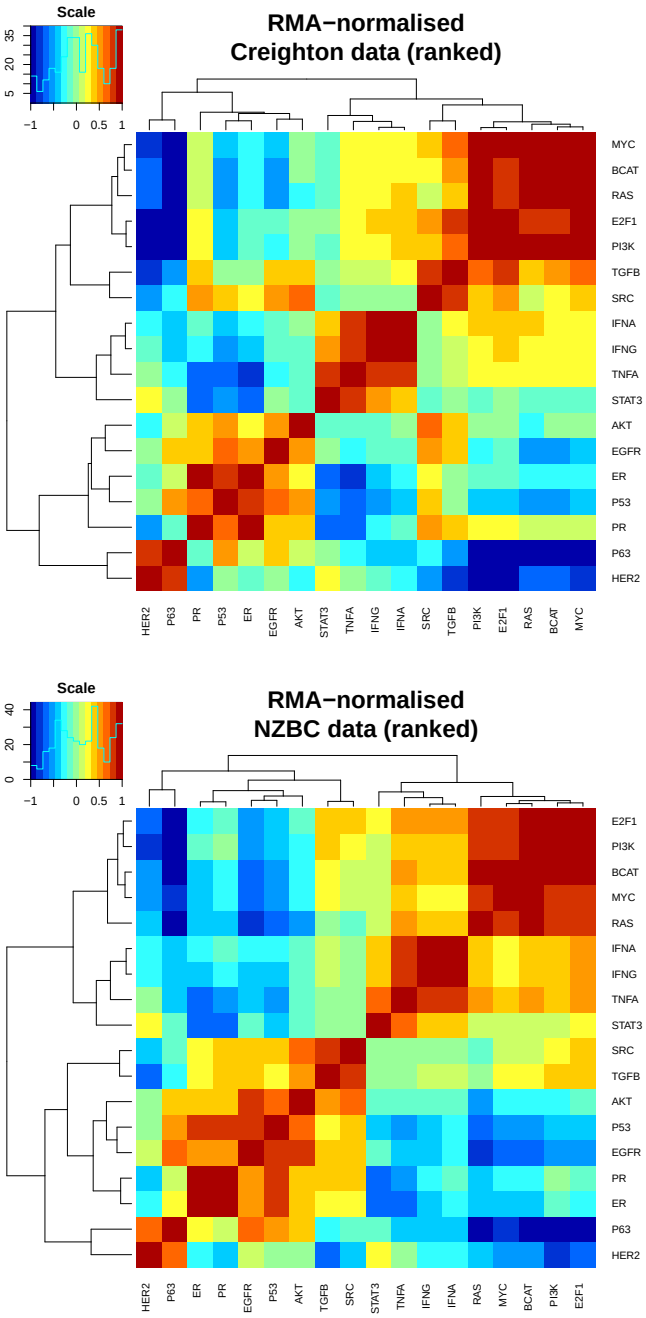


Figure S25: Heatmaps showing the correlation of all of the pathway metagenes with one another in RMA normalised CR, NZBC and FM data sets. High and low correlation were represented as red and blue, respectively, where the colours were matched with the values on the scale shown in the top left histogram.

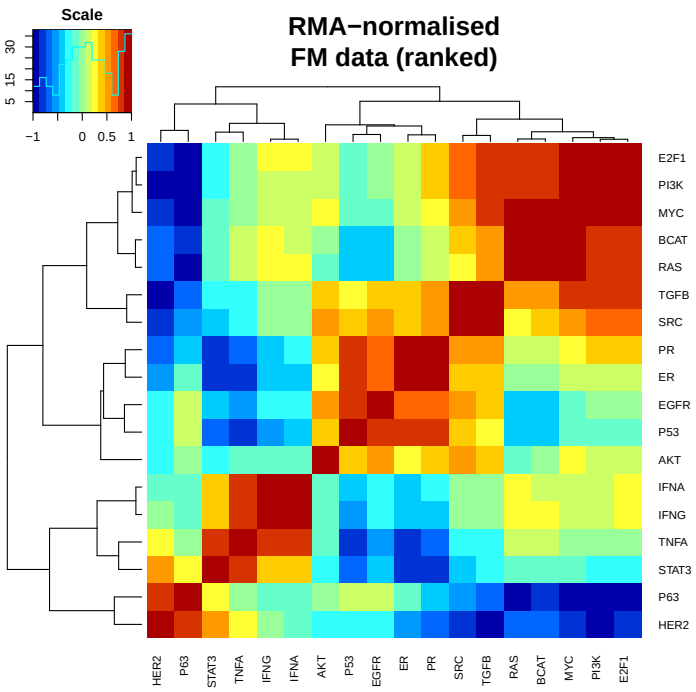


Figure S25 (cont.)

B7 Visual comparison of the correlation of SVD and TM generated pathway metagenes

This section provides visual comparison of the Spearman correlation of the SVD- and TM-generated pathway and obesity metagenes in the CR, FM, GT and NZBC data sets (Figure S27). The Akt, EGFR, HER2, p53, p63, PI3K, Ras, Src, STAT3, TGFβ and TNFα pathway signatures showed a correlation of 0.8 or less in at least one of the four data sets.

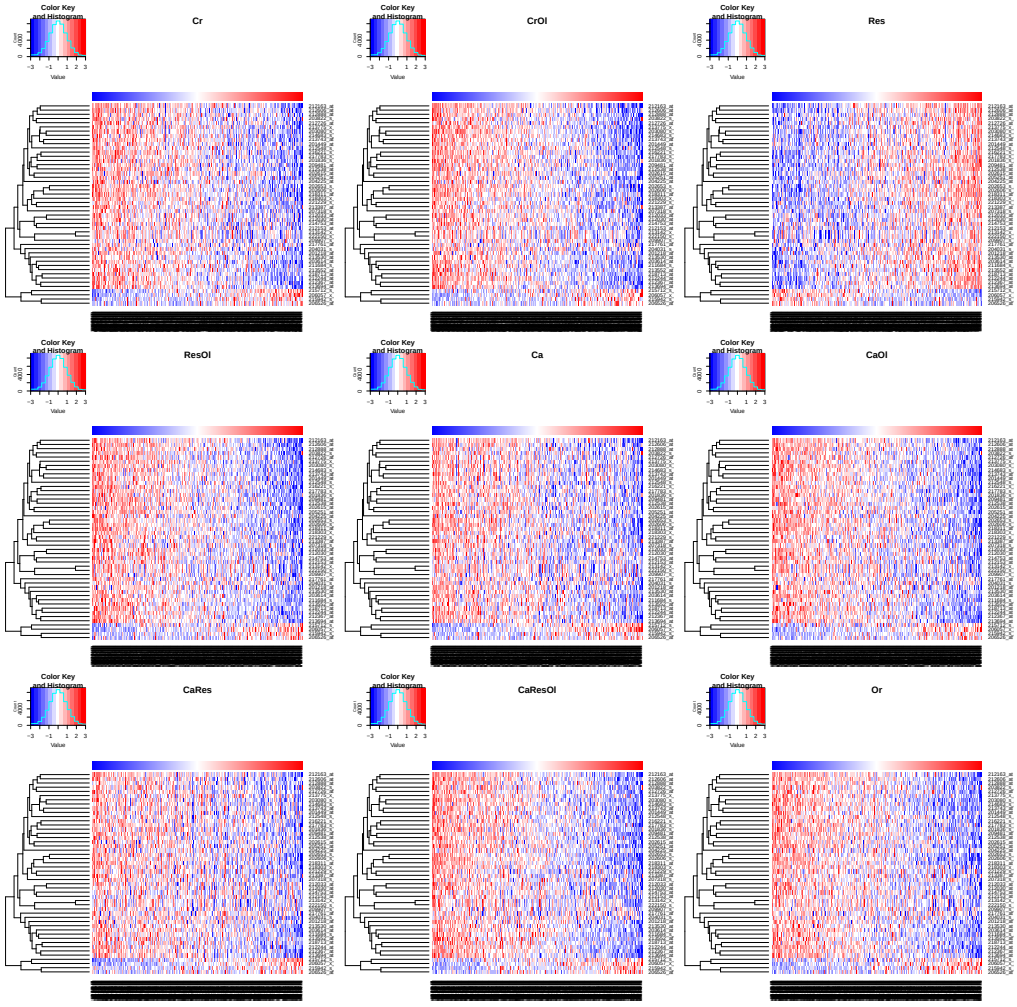


Figure S26: Heatmaps showing the obesity metagenes from the CR data set with the gene expressions of the obesity associated genetic signatures in the RMA normalised GT data set. High and low correlation were represented as red and blue, respectively, where the colours were matched with the values on the scale shown in the top left histogram.

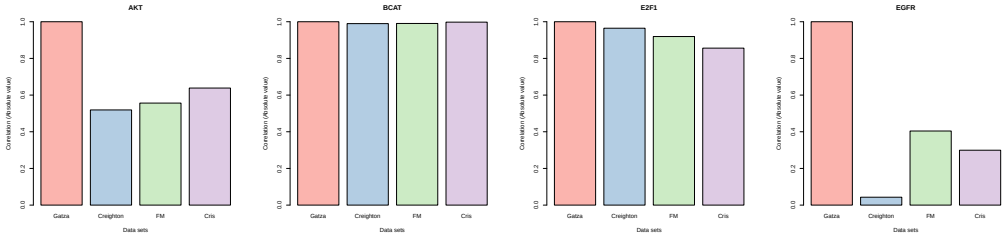


Figure S27: Visual comparison of the Spearman correlations of the SVD-derived and transformation matrix-derived pathway and obesity metagenes across different data sets

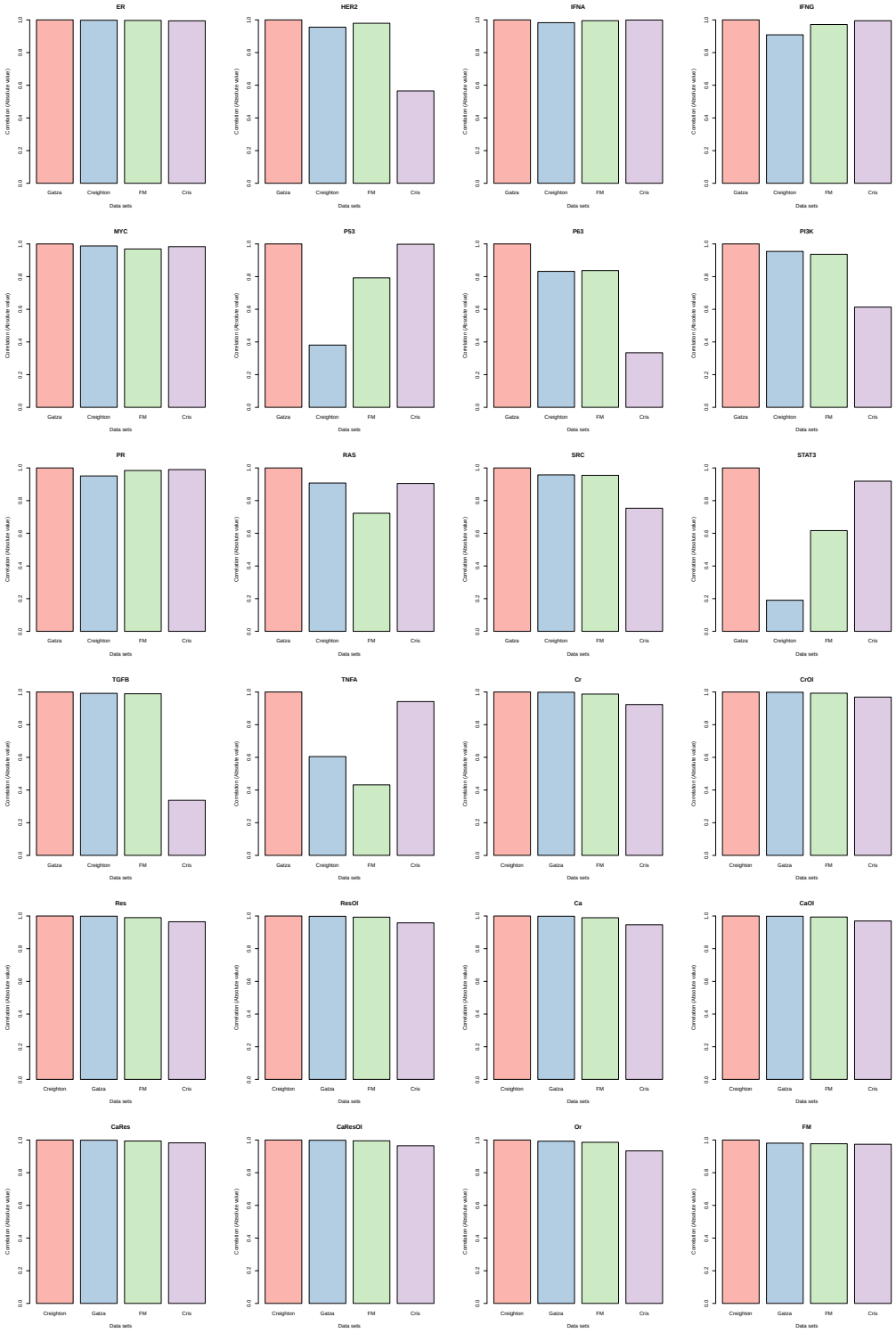


Figure S27 (cont.)

B8 Correlation of the pathway and obesity metagenes in other cancer data sets

Correlation of the pathway and obesity associated genetic signatures were plotted in a heatmap for the GT data in Section 4.2. In this section, the correlation of these genetic signatures were plotted as heatmaps for the CR, FM and NZBC data sets. Though each of the heatmaps show slightly different clustering patterns, the results were similar to that in the GT data set; the CR obesity metagenes clustered together in a single group, and FM metagene did not cluster with any other genetic signatures.

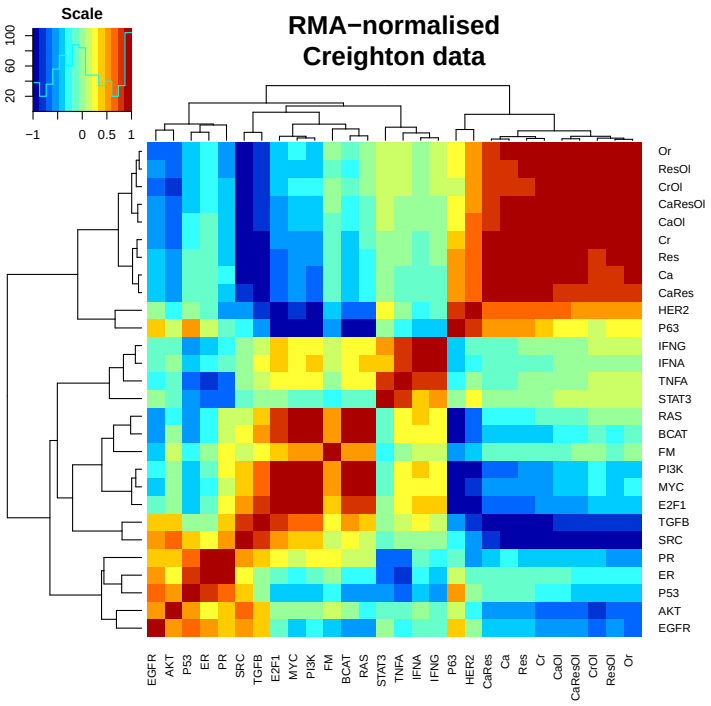


Figure S28: Heatmaps showing the correlation of all of the pathway and obesity metagenes with one another in the RMA-normalised CR, NZBC and FM data sets.

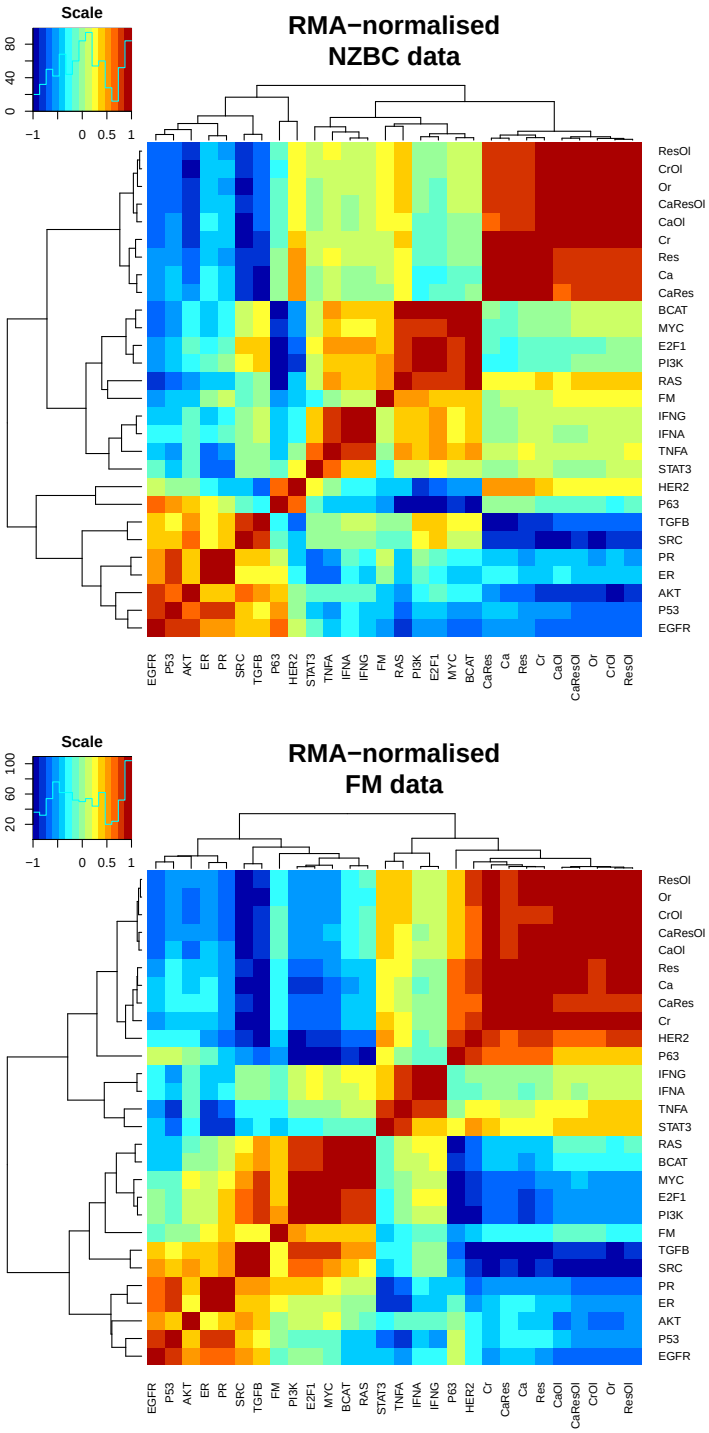


Figure S28 (cont.)

B9 Summary of the remainder of the linear models in NZBC data

All of the linear models for the other CR and FM obesity metagenes were summarised in this section. In all of the CR data set-derived obesity metagenes, the PR metagene showed up as a significant variable in the linear models (Tables S11 to S18). For the linear model that predicted the FM obesity metagene scores, the Myc pathway metagene was significant (Table S19).

Table S11: Description of the linear models constructed from the NZBC data to predict the CrOl obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	-0.001629	0.6900
BMI status only	Overweight	-0.02165	0.7800
	Obese	-0.1055	0.1360
BMI and BMI status	BMI	0.01128	0.1092
	Overweight	-0.07985	0.3480
	Obese	-0.2611	**0.0304**
Pathways only	BCAT	0.1561	0.4873
	ER	0.1234	0.6293
	IFNα	-0.4302	0.3892
	IFNγ	0.3784	0.4594
	Myc	0.1608	0.4846
	PR	0.5974	**0.0164**
BMI and Pathways	BMI	-0.002745	0.4736
	BCAT	0.1799	0.4297
	ER	0.1254	0.6247
	IFNα	-0.4754	0.3466
	IFNγ	0.4291	0.4074
	Myc	0.1846	0.4284
	PR	0.6080	**0.0151**
BMI status and Pathways	Overweight	-0.02348	0.7434
	Obese	-0.1120	**0.0933**
	BCAT	0.2204	0.3325
	ER	0.1045	0.6809
	IFNα	-0.6034	0.2341
	IFNγ	0.5467	0.2911
	Myc	0.2134	0.3559
	PR	0.5949	**0.0161**
BMI, BMI status and pathways	BMI	0.008834	0.1819
	Overweight	-0.06794	0.3890
	Obese	-0.2346	**0.0397**
	BCAT	0.2169	0.3380
	ER	0.08151	0.7478
	IFNα	-0.6259	0.2155
	IFNγ	0.5469	0.2889
	Myc	0.1962	0.3945
	PR	0.5607	**0.0232**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S12: Description of the linear models constructed from the NZBC data to predict the Res obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	-0.001422	0.7280
BMI status only	Overweight	-0.001804	0.9810
	Obese	-0.09683	0.1710
BMI and BMI status	BMI	0.01126	0.1099
	Overweight	-0.05990	0.4810
	Obese	-0.2522	**0.0364**
Pathways only	BCAT	0.09717	0.6541
	ER	0.07879	0.7496
	IFNα	-0.1758	0.7152
	IFNγ	0.07930	0.8723
	Myc	-0.04517	0.8387
	PR	0.5955	**0.0133**
BMI and Pathways	BMI	-0.001994	0.5902
	BCAT	0.1145	0.6029
	ER	0.08025	0.7461
	IFNα	-0.2086	0.6688
	IFNγ	0.1161	0.8163
	Myc	-0.02785	0.9015
	PR	0.6032	**0.0127**
BMI status and Pathways	Overweight	-0.004412	0.9493
	Obese	-0.09683	0.1328
	BCAT	0.1504	0.4935
	ER	0.05888	0.8106
	IFNα	-0.3441	0.4817
	IFNγ	0.2427	0.6271
	Myc	-0.001285	0.9954
	PR	0.5912	**0.0135**
BMI, BMI status and pathways	BMI	0.009426	0.1403
	Overweight	-0.05185	0.4953
	Obese	-0.2277	**0.0386**
	BCAT	0.1467	0.5013
	ER	0.03435	0.8883
	IFNα	-0.3681	0.4491
	IFNγ	0.2429	0.6246
	Myc	-0.01963	0.9295
	PR	0.5547	**0.0201**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S13: Description of the linear models constructed from the NZBC data to predict the ResOl obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	-0.0009922	0.8080
BMI status only	Overweight	-0.01515	0.8450
	Obese	-0.09063	0.2010
BMI and BMI status	BMI	0.01105	0.1178
	Overweight	-0.07219	0.3978
	Obese	-0.2432	**0.0442**
Pathways only	BCAT	0.09532	0.6683
	ER	0.03171	0.9004
	IFNα	-0.5797	0.2426
	IFNγ	0.5092	0.3157
	Myc	0.06345	0.7805
	PR	0.5420	**0.0276**
BMI and Pathways	BMI	-0.001950	0.6076
	BCAT	0.1123	0.6192
	ER	0.03313	0.8963
	IFNα	-0.6118	0.2231
	IFNγ	0.5452	0.2894
	Myc	0.08041	0.7279
	PR	0.5495	**0.0264**
BMI status and Pathways	Overweight	-0.01219	0.8640
	Obese	-0.09681	0.1440
	BCAT	0.1497	0.5070
	ER	0.01354	0.9570
	IFNα	-0.7389	0.1440
	IFNγ	0.6638	0.1980
	Myc	0.1081	0.6380
	PR	0.5387	**0.0280**
BMI, BMI status and pathways	BMI	0.009248	0.1596
	Overweight	-0.05873	0.4532
	Obese	-0.2252	**0.0466**
	BCAT	0.1461	0.5154
	ER	-0.01053	0.9667
	IFNα	-0.7624	0.1296
	IFNγ	0.6640	0.1955
	Myc	0.09012	0.6933
	PR	0.5029	**0.0399**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S14: Description of the linear models constructed from the NZBC data to predict the Ca obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	-0.001147	0.7790
BMI status only	Overweight	0.01804	0.8160
	Obese	-0.08982	0.2030
BMI and BMI status	BMI	0.01171	0.0957
	Overweight	-0.04241	0.6168
	Obese	-0.2515	**0.0365**
Pathways only	BCAT	0.1407	0.5367
	ER	0.1160	0.6545
	IFNα	-0.3332	0.5104
	IFNγ	0.3087	0.5515
	Myc	-0.04569	0.8445
	PR	0.5669	**0.0243**
BMI and Pathways	BMI	-0.002148	0.5804
	BCAT	0.1593	0.4906
	ER	0.1175	0.6514
	IFNα	-0.3686	0.4718
	IFNγ	0.3484	0.5072
	Myc	-0.02702	0.9090
	PR	0.5752	**0.0232**
BMI status and Pathways	Overweight	0.008987	0.9016
	Obese	-0.09970	0.1397
	BCAT	0.1934	0.4009
	ER	0.09245	0.7196
	IFNα	-0.5223	0.3089
	IFNγ	0.4921	0.3481
	Myc	-0.001892	0.9935
	PR	0.5606	**0.0249**
BMI, BMI status and pathways	BMI	0.009944	0.1376
	Overweight	-0.04106	0.6061
	Obese	-0.2378	**0.0392**
	BCAT	0.1895	0.4073
	ER	0.06657	0.7950
	IFNα	-0.5476	0.2832
	IFNγ	0.4923	0.3446
	Myc	-0.02125	0.9272
	PR	0.5221	**0.0362**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S15: Description of the linear models constructed from the NZBC data to predict the CaOl obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	-3.953×10^{-5}	0.9920
BMI status only	Overweight	0.02958	0.7040
	Obese	-0.06411	0.3650
BMI and BMI status	BMI	0.01136	0.1080
	Overweight	-0.02903	0.7334
	Obese	-0.2209	0.0671
Pathways only	BCAT	0.2366	0.3150
	ER	0.09521	0.7218
	IFNα	-0.4913	0.3476
	IFNγ	0.4447	0.4063
	Myc	0.1611	0.5033
	PR	0.4957	0.0555
BMI and Pathways	BMI	-0.001281	0.7495
	BCAT	0.2477	0.3004
	ER	0.09615	0.7205
	IFNα	-0.5124	0.3335
	IFNγ	0.4683	0.3887
	Myc	0.1722	0.4811
	PR	0.5006	0.0548
BMI status and Pathways	Overweight	0.02249	0.7655
	Obese	-0.07973	0.2536
	BCAT	0.2765	0.2479
	ER	0.07298	0.7846
	IFNα	-0.6604	0.2155
	IFNγ	0.6085	0.2638
	Myc	0.1946	0.4230
	PR	0.4885	0.0586
BMI, BMI status and pathways	BMI	0.009886	0.1549
	Overweight	-0.02727	0.7414
	Obese	-0.2170	0.0692
	BCAT	0.2726	0.2518
	ER	0.04725	0.8591
	IFNα	-0.6855	0.1964
	IFNγ	0.6087	0.2609
	Myc	0.1753	0.4683
	PR	0.4502	0.0808

[1] All values in bold are statistically significant ($p < 0.05$).

Table S16: Description of the linear models constructed from the NZBC data to predict the CaRes obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	0.001442	0.72408
BMI status only	Overweight	-0.01840	0.8120
	Obese	0.09338	0.1860
BMI and BMI status	BMI	-0.01141	0.1043
	Overweight	0.04047	0.6328
	Obese	0.2508	**0.0369**
Pathways only	BCAT	-0.2217	0.32284
	ER	-0.2308	0.36579
	IFNα	0.1853	0.70938
	IFNγ	-0.1528	0.76408
	Myc	-0.0737	0.74765
	PR	-0.6876	**0.00581**
BMI and Pathways	BMI	0.002580	0.49929
	BCAT	-0.2441	0.28322
	ER	-0.2327	0.36335
	IFNα	0.2278	0.65049
	IFNγ	-0.2005	0.69745
	Myc	-0.09612	0.67888
	PR	-0.6976	**0.00538**
BMI status and Pathways	Overweight	-0.005440	0.93926
	Obese	0.1015	0.12592
	BCAT	-0.2760	0.22351
	ER	-0.2077	0.41234
	IFNα	0.3736	0.45823
	IFNγ	-0.3355	0.51449
	Myc	-0.1187	0.60548
	PR	-0.6817	**0.00586**
BMI, BMI status and pathways	BMI	-0.008785	0.18237
	Overweight	0.0388	0.62094
	Obese	0.2235	**0.04881**
	BCAT	-0.2725	0.22730
	ER	-0.1848	0.46455
	IFNα	0.3959	0.43015
	IFNγ	-0.3356	0.51243
	Myc	-0.1016	0.65743
	PR	-0.6477	**0.00879**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S17: Description of the linear models constructed from the NZBC data to predict the CaResOl obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	0.0004131	0.9194
BMI status only	Overweight	0.02453	0.7520
	Obese	-0.06913	0.3290
BMI and BMI status	BMI	0.01335	0.0582
	Overweight	-0.04436	0.6008
	Obese	-0.2533	**0.0352**
Pathways only	BCAT	0.2439	0.2864
	ER	0.1383	0.5948
	IFNα	-0.4364	0.3903
	IFNγ	0.4301	0.4084
	Myc	0.2837	0.2264
	PR	0.5789	**0.0219**
BMI and Pathways	BMI	-0.001038	0.7900
	BCAT	0.2529	0.2767
	ER	0.1390	0.5947
	IFNα	-0.4535	0.3782
	IFNγ	0.4493	0.3947
	Myc	0.2928	0.2191
	PR	0.5829	**0.0219**
BMI status and Pathways	Overweight	0.01833	0.8024
	Obese	-0.08019	0.2373
	BCAT	0.2847	0.2209
	ER	0.1169	0.6525
	IFNα	-0.6015	0.2453
	IFNγ	0.5901	0.2646
	Myc	0.3178	0.1793
	PR	0.5723	**0.0231**
BMI, BMI status and pathways	BMI	0.01054	0.1181
	Overweight	-0.03471	0.6648
	Obese	-0.2265	0.0506
	BCAT	0.2806	0.2239
	ER	0.08944	0.7288
	IFNα	-0.6283	0.2216
	IFNγ	0.5903	0.2606
	Myc	0.2973	0.2058
	PR	0.5315	**0.0341**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S18: Description of the linear models constructed from the NZBC data to predict the Or obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	0.001654	0.68532
BMI status only	Overweight	0.005772	0.9410
	Obese	0.1009	0.1540
BMI and BMI status	BMI	-0.01106	0.1162
	Overweight	0.06284	0.4598
	Obese	0.2535	**0.0354**
Pathways only	BCAT	-0.2284	0.3074
	ER	-0.1481	0.5604
	IFNα	0.4504	0.3647
	IFNγ	-0.4013	0.4302
	Myc	-0.1402	0.5399
	PR	-0.6446	**0.009380**
BMI and Pathways	BMI	0.003004	0.4301
	BCAT	-0.2545	0.2618
	ER	-0.1502	0.5555
	IFNα	0.4999	0.3195
	IFNγ	-0.4568	0.3749
	Myc	-0.1663	0.4728
	PR	-0.6562	**0.0085**
BMI status and Pathways	Overweight	0.01412	0.8427
	Obese	0.1135	0.0867
	BCAT	-0.2921	0.1964
	ER	-0.1267	0.6153
	IFNα	0.6372	0.2056
	IFNγ	-0.5827	0.2570
	Myc	-0.1925	0.4010
	PR	-0.6407	**0.0092**
BMI, BMI status and pathways	BMI	-0.008654	0.1875
	Overweight	0.05768	0.4610
	Obese	0.2336	**0.03905**
	BCAT	-0.2887	0.1997
	ER	-0.1042	0.6788
	IFNα	0.6592	0.1890
	IFNγ	-0.5829	0.2549
	Myc	-0.1757	0.4421
	PR	-0.6072	**0.01353**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S19: Description of the linear models constructed from the NZBC data to predict the FM obesity metagene

Linear Model	Variables	Estimate	P-value
BMI only	BMI	0.002314	0.5708
BMI status only	Overweight	-0.1483	0.05670
	Obese	-0.08452	0.22910
BMI and BMI status	BMI	0.01503	**0.03091**
	Overweight	-0.2258	**0.007940**
	Obese	-0.2920	**0.01419**
Pathways only	BCAT	-0.1349	0.5334
	ER	0.1197	0.6271
	IFNα	-0.5624	0.2438
	IFNγ	0.5884	0.2340
	Myc	0.4177	0.06180
	PR	-0.08815	0.7090
BMI and Pathways	BMI	0.002392	0.5172
	BCAT	-0.1557	0.4786
	ER	0.1180	0.6333
	IFNα	-0.5231	0.2835
	IFNγ	0.5442	0.2767
	Myc	0.3969	0.07970
	PR	-0.09737	0.6817
BMI status and Pathways	Overweight	-0.1249	0.07330
	Obese	-0.08029	0.2104
	BCAT	-0.07254	0.7402
	ER	0.1304	0.5948
	IFNα	-0.5604	0.2517
	IFNγ	0.5876	0.2397
	Myc	0.4665	**0.03820**
	PR	-0.07506	0.7490
BMI, BMI status and pathways	BMI	0.01539	**0.01464**
	Overweight	-0.2024	**0.007460**
	Obese	-0.2940	**0.006660**
	BCAT	-0.0786	0.7119
	ER	0.09030	0.7053
	IFNα	-0.5994	0.2079
	IFNγ	0.5878	0.2265
	Myc	0.4365	**0.04623**
	PR	-0.1347	0.5575

[1] All values in bold are statistically significant ($p < 0.05$).

B10 Summary of the remainder of the PR linear models in NZBC data

The PR-only linear models for the other obesity associated genetic signatures that were generated in the NZBC data were summarised in Tables S20 to S28. The PR pathway metagene was significant in all of the linear models that predicted the obesity metagenes derived from the CR data set Tables S20 to S27. On the other hand, the sample BMI/BMI status were significant in the linear model that predicted the FM obesity metagene Table S28.

Table S20: Description of the linear models constructed from the NZBC data to predict the CrOl obesity, using only the sample BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	0.4484	$\mathbf{3.26 \times 10^{-6}}$ [1]
BMI and PR	BMI	-0.002084	0.5699
	PR	0.4497	$\mathbf{3.36 \times 10^{-6}}$
BMI status and PR	Obese	-0.09307	0.1440
	Overweight	-0.01989	0.7750
	PR	0.4418	$\mathbf{4.32 \times 10^{-6}}$
BMI, BMI status and PR	BMI	0.008072	0.2060
	Obese	-0.2048	0.06100
	Overweight	-0.06159	0.4230
	PR	0.4294	$\mathbf{7.93 \times 10^{-6}}$

[1] All values in bold are statistically significant ($p < 0.05$).

Table S21: Description of the linear models constructed from the NZBC data to predict the Res obesity, using only the sample BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	0.4926	$\mathbf{2.22 \times 10^{-7}}$ [1]
BMI and PR	BMI	-0.001923	0.5904
	PR	0.4938	$\mathbf{2.35 \times 10^{-7}}$
BMI status and PR	Obese	-0.08315	0.1790
	Overweight	0.0001257	0.9990
	PR	0.4862	$\mathbf{3.00 \times 10^{-7}}$
BMI, BMI status and PR	BMI	0.007716	0.2140
	Obese	-0.1899	0.0740
	Overweight	-0.03974	0.5950
	PR	0.4744	$\mathbf{5.78 \times 10^{-7}}$

[1] All values in bold are statistically significant ($p < 0.05$).

Table S22: Description of the linear models constructed from the NZBC data to predict the ResOl obesity, using only the sample BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	0.4581	$\mathbf{1.86 \times 10^{-6}}$ [1]
BMI and PR	BMI	-0.001457	0.6897
	PR	0.4590	$\mathbf{1.99 \times 10^{-6}}$
BMI status and PR	Obese	-0.07789	0.2200
	Overweight	-0.01336	0.8480
	PR	0.4526	$\mathbf{2.56 \times 10^{-6}}$
BMI, BMI status and PR	BMI	0.007763	0.2230
	Obese	-0.1853	0.0890
	Overweight	-0.05346	0.4860
	PR	0.4407	$\mathbf{4.71 \times 10^{-6}}$

[1] All values in bold are statistically significant ($p < 0.05$).

Table S23: Description of the linear models constructed from the NZBC data to predict the Ca obesity, using only the sample BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	0.3850	$\mathbf{8.34 \times 10^{-5}}$ [1]
BMI and PR	BMI	-0.001538	0.6848
	PR	0.3859	$\mathbf{8.67 \times 10^{-5}}$
BMI status and PR	Obese	-0.07917	0.2274
	Overweight	0.01954	0.7857
	PR	0.3782	**0.0001060**
BMI, BMI status and PR	BMI	0.008992	0.1718
	Obese	-0.2036	0.07034
	Overweight	-0.02692	0.7338
	PR	0.3644	**0.0001830**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S24: Description of the linear models constructed from the NZBC data to predict the CaOl obesity, using only the sample BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	0.3425	**0.0005200** [1]
BMI and PR	BMI	-0.0003869	0.9201
	PR	0.3428	**0.0005530**
BMI status and PR	Obese	-0.05462	0.4147
	Overweight	0.03092	0.6742
	PR	0.3373	**0.0006470**
BMI, BMI status and PR	BMI	0.008943	0.1842
	Obese	-0.1784	0.1206
	Overweight	-0.01528	0.8504
	PR	0.3236	**0.001060**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S25: Description of the linear models constructed from the NZBC data to predict the CaRes obesity, using only the sample BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	-0.4255	$\mathbf{1.13 \times 10^{-5}}$ [1]
BMI and PR	BMI	0.001874	0.6140
	PR	-0.4267	$\mathbf{1.17 \times 10^{-5}}$
BMI status and PR	Obese	0.08159	0.2040
	Overweight	-0.02006	0.7760
	PR	-0.4186	$\mathbf{1.44 \times 10^{-5}}$
BMI, BMI status and PR	BMI	-0.008379	0.1936
	Obese	0.1976	0.07310
	Overweight	0.02323	0.7645
	PR	-0.4057	$\mathbf{2.59 \times 10^{-5}}$

[1] All values in bold are statistically significant ($p < 0.05$).

Table S26: Description of the linear models constructed from the NZBC data to predict the CaResOl obesity, using only the sample BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	0.4094	$\mathbf{2.58 \times 10^{-5}}$ [1]
BMI and PR	BMI	-1.809d-06	0.9996
	PR	4.094d-01	$\mathbf{2.85 \times 10^{-5}}$
BMI status and PR	Obese	-0.05775	0.3740
	Overweight	0.02613	0.7140
	PR	0.4041	$\mathbf{3.36 \times 10^{-5}}$
BMI, BMI status and PR	BMI	0.01045	0.1091
	Obese	-0.2024	0.06910
	Overweight	-0.02786	0.7220
	PR	0.3881	$\mathbf{6.21 \times 10^{-5}}$

[1] All values in bold are statistically significant ($p < 0.05$).

Table S27: Description of the linear models constructed from the NZBC data to predict the Or obesity, using only the sample BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	-0.4463	$\mathbf{3.66 \times 10^{-6}}$ [1]
BMI and PR	BMI	0.002108	0.5660
	PR	-0.4476	$\mathbf{3.77 \times 10^{-6}}$
BMI status and PR	Obese	0.088520	0.1650
	Overweight	0.004028	0.9540
	PR	-0.4396	$\mathbf{4.82 \times 10^{-6}}$
BMI, BMI status and PR	BMI	-0.007866	0.2180
	Obese	0.1974	0.07100
	Overweight	0.04466	0.5620
	PR	-0.4275	$\mathbf{8.77 \times 10^{-6}}$

[1] All values in bold are statistically significant ($p < 0.05$).

Table S28: Description of the linear models constructed from the NZBC data to predict the FM obesity, using only the sample BMI, BMI status and the PR pathway metagene score

Linear Model	Variables	Estimate	P-value
PR only	PR	-0.1285	0.2050
BMI and PR	BMI	0.002446	0.5479
	PR	-0.1301	0.2012
BMI status and PR	Obese	-0.08819	0.2080
	Overweight	-0.1488	0.05500
	PR	-0.1306	0.1940
BMI, BMI status and PR	BMI	0.01619	**0.02008** [1]
	Obese	-0.3123	**0.008750**
	Overweight	-0.2324	**0.006030**
	PR	-0.1554	0.1165

[1] All values in bold are statistically significant ($p < 0.05$).

B11 Summary statistics of the predicted obesity metagenes with sample BMI/BMI status in the NZBC and CR data

Instead of presenting 11 scatter plots for the 10 obesity metagenes used in this project, all of the R^2- and p-values were summarised in Table S29.

Table S29: Table summarising the linear model statistics from the comparison of the various linear model-predicted obesity metagenes scores with the corresponding obesity metagene scores from the NZBC data set for all of the obesity metagenes

Obesity metagene	Model	R^2	P-value
Cr	BMI only	0.001852	0.2797
	BMI and BMI status	0.02834	0.05237
	Pathways only	0.2416	$\mathbf{1.431 \times 10^{-7}}$[1]
	BMI and pathways	0.2466	$\mathbf{1.031 \times 10^{-7}}$
	BMI status and pathways	0.2719	$\mathbf{1.883 \times 10^{-8}}$
	BMI, BMI status and pathways	0.2814	$\mathbf{9.811 \times 10^{-9}}$
	PR only	0.2041	$\mathbf{1.604 \times 10^{-6}}$
	BMI and PR	0.2047	$\mathbf{1.54 \times 10^{-6}}$
	BMI status and PR	0.2249	$\mathbf{4.237 \times 10^{-7}}$
	BMI, BMI status and PR	0.2372	$\mathbf{1.908 \times 10^{-7}}$
CrOl	BMI only	0.001295	0.2910
	BMI and BMI status	0.03509	**0.03517**
	Pathways only	0.2041	$\mathbf{1.604 \times 10^{-6}}$
	BMI and pathways	0.2173	$\mathbf{6.941 \times 10^{-7}}$
	BMI status and pathways	0.2411	$\mathbf{1.48 \times 10^{-7}}$
	BMI, BMI status and pathways	0.2597	$\mathbf{4.308 \times 10^{-8}}$
	PR only	0.2024	$\mathbf{1.786 \times 10^{-6}}$
	BMI and PR	0.2056	$\mathbf{1.458 \times 10^{-6}}$
	BMI status and PR	0.2212	$\mathbf{5.398 \times 10^{-7}}$
	BMI, BMI status and PR	0.2336	$\mathbf{2.417 \times 10^{-7}}$
Res	BMI only	0.0008245	0.3011
	BMI and BMI status	0.02452	0.06579

Table S29 (continued)

	Pathways only	0.2785	$\mathbf{1.196 \times 10^{-8}}$
	BMI and pathways	0.2780	$\mathbf{1.233 \times 10^{-8}}$
	BMI status and pathways	0.2914	$\mathbf{4.887 \times 10^{-9}}$
	BMI, BMI status and pathways	0.3195	$\mathbf{6.576 \times 10^{-10}}$
	PR only	0.2331	$\mathbf{2.492 \times 10^{-7}}$
	BMI and PR	0.2349	$\mathbf{2.215 \times 10^{-7}}$
	BMI status and PR	0.2538	$\mathbf{6.38 \times 10^{-8}}$
	BMI, BMI status and PR	0.2631	$\mathbf{3.421 \times 10^{-8}}$
ResOl	BMI only	-0.002442	0.3851
	BMI and BMI status	0.03878	**0.02835**
	Pathways only	0.2384	$\mathbf{1.762 \times 10^{-7}}$
	BMI and pathways	0.2371	$\mathbf{1.927 \times 10^{-7}}$
	BMI status and pathways	0.2670	$\mathbf{2.616 \times 10^{-8}}$
	BMI, BMI status and pathways	0.2878	$\mathbf{6.263 \times 10^{-9}}$
	PR only	0.2251	$\mathbf{4.198 \times 10^{-7}}$
	BMI and PR	0.2257	$\mathbf{4.044 \times 10^{-7}}$
	BMI status and PR	0.2392	$\mathbf{1.675 \times 10^{-7}}$
	BMI, BMI status and PR	0.2568	$\mathbf{5.207 \times 10^{-8}}$
Ca	BMI only	-0.0004067	0.3296
	BMI and BMI status	0.03576	**0.03383**
	Pathways only	0.2043	$\mathbf{1.585 \times 10^{-6}}$
	BMI and pathways	0.2077	$\mathbf{1.28 \times 10^{-6}}$
	BMI status and pathways	0.2342	$\mathbf{2.321 \times 10^{-7}}$
	BMI, BMI status and pathways	0.2487	$\mathbf{8.927 \times 10^{-8}}$
	PR only	0.1362	**0.0001009**
	BMI and PR	0.1364	$\mathbf{9.944 \times 10^{-5}}$
	BMI status and PR	0.1643	$\mathbf{1.876 \times 10^{-5}}$
	BMI, BMI status and PR	0.1744	$\mathbf{1.012 \times 10^{-5}}$
CaOl	BMI only	-0.006593	0.5509
	BMI and BMI status	0.0301	**0.04718**
	Pathways only	0.1323	**0.0001273**
	BMI and pathways	0.1347	**0.0001106**

Table S29 (continued)

	BMI status and pathways	0.1487	$\mathbf{4.802 \times 10^{-5}}$
	BMI, BMI status and pathways	0.1657	$\mathbf{1.725 \times 10^{-5}}$
	PR only	0.1103	**0.0004599**
	BMI and PR	0.1106	**0.0004539**
	BMI status and PR	0.1235	**0.0002134**
	BMI, BMI status and PR	0.1383	$\mathbf{8.92 \times 10^{-5}}$
CaRes	BMI only	0.001142	0.2943
	BMI and BMI status	0.03048	**0.04612**
	Pathways only	0.2181	$\mathbf{6.586 \times 10^{-7}}$
	BMI and pathways	0.2373	$\mathbf{1.902 \times 10^{-7}}$
	BMI status and pathways	0.2431	$\mathbf{1.298 \times 10^{-7}}$
	BMI, BMI status and pathways	0.2501	$\mathbf{8.152 \times 10^{-8}}$
	PR only	0.1566	$\mathbf{2.979 \times 10^{-5}}$
	BMI and PR	0.1586	$\mathbf{2.654 \times 10^{-5}}$
	BMI status and PR	0.1844	$\mathbf{5.485 \times 10^{-6}}$
	BMI, BMI status and PR	0.1957	$\mathbf{2.707 \times 10^{-6}}$
CaResOl	BMI only	-0.006037	0.5225
	BMI and BMI status	0.0413	**0.02449**
	Pathways only	0.1719	$\mathbf{1.183 \times 10^{-5}}$
	BMI and pathways	0.1721	$\mathbf{1.167 \times 10^{-5}}$
	BMI status and pathways	0.1961	$\mathbf{2.642 \times 10^{-6}}$
	BMI, BMI status and pathways	0.2095	$\mathbf{1.141 \times 10^{-6}}$
	PR only	0.1571	$\mathbf{2.894 \times 10^{-5}}$
	BMI and PR	0.1574	$\mathbf{2.843 \times 10^{-5}}$
	BMI status and PR	0.17	$\mathbf{1.327 \times 10^{-5}}$
	BMI, BMI status and PR	0.1939	$\mathbf{3.044 \times 10^{-6}}$
Or	BMI only	0.0008035	0.3015
	BMI and BMI status	0.03699	**0.03147**
	Pathways only	0.2554	$\mathbf{5.738 \times 10^{-8}}$
	BMI and pathways	0.2568	$\mathbf{5.234 \times 10^{-8}}$
	BMI status and pathways	0.28	$\mathbf{1.075 \times 10^{-8}}$
	BMI, BMI status and pathways	0.2974	$\mathbf{3.194 \times 10^{-9}}$

Table S29 (continued)

	PR only	0.2388	$\mathbf{1.717 \times 10^{-7}}$
	BMI and PR	0.2382	$\mathbf{1.79 \times 10^{-7}}$
	BMI status and PR	0.2531	$\mathbf{6.663 \times 10^{-8}}$
	BMI, BMI status and PR	0.2684	$\mathbf{2.382 \times 10^{-8}}$
FM	BMI only	-0.008238	0.6563
	BMI and BMI status	0.07438	**0.003652**
	Pathways only	0.268	$\mathbf{2.449 \times 10^{-8}}$
	BMI and pathways	0.2645	$\mathbf{3.106 \times 10^{-8}}$
	BMI status and pathways	0.3067	$\mathbf{1.654 \times 10^{-9}}$
	BMI, BMI status and pathways	0.3653	$\mathbf{2.112 \times 10^{-11}}$
	PR only	-0.004177	0.4434
	BMI and PR	0.007149	0.1946
	BMI status and PR	0.0394	**0.02734**
	BMI, BMI status and PR	0.1166	**0.0003183**

[1] All values in bold are statistically significant ($p < 0.05$).

Table S30: Table summarising the linear model statistics from the comparison of the predicted obesity metagenes scores from the BMI status-only model with the corresponding obesity metagene scores from the NZBC data set for all of the obesity metagenes

	P-values		
Obesity metagene	Overweight	Obese	ANOVA
Cr	0.2002	0.1501	0.2591
CrOl	0.3126	0.1532	0.3134
Res	0.9708	0.1764	0.2576
ResOl	0.4001	0.1919	0.3950
Ca	0.1847	0.1099	0.2000
CaOl	0.3466	0.2325	0.4273
CaRes	0.1713	0.1154	0.2017
CaResOl	0.2976	0.1742	0.3439
Or	0.2669	0.1621	0.3074
FM	0.4191	0.05207	0.1206

Table S31: Table summarising the linear model statistics from the comparison of the various linear model-predicted obesity metagenes scores with the corresponding obesity metagene scores from the CR data set for all of the obesity metagenes

Obesity metagene	Model	R^2	P-value
Cr	BMI only	0.1239	**0.000157**[1]
	BMI and BMI status	0.0471	**0.01568**
	Pathways only	0.1756	$\mathbf{6.307 \times 10^{-6}}$
	BMI and pathways	0.1407	$\mathbf{5.596 \times 10^{-5}}$
	BMI status and pathways	0.05509	**0.009721**
	BMI, BMI status and pathways	0.04186	**0.02147**
	PR only	0.1327	$\mathbf{9.191 \times 10^{-5}}$
	BMI and PR	0.1116	**0.0003317**
	BMI status and PR	0.03887	**0.02571**
	BMI, BMI status and PR	0.04176	**0.02161**
CrOl	BMI only	0.18	$\mathbf{4.738 \times 10^{-6}}$
	BMI and BMI status	0.06523	**0.005312**
	Pathways only	0.1296	**0.000111**
	BMI and pathways	0.09006	**0.00121**
	BMI status and pathways	0.02455	**0.06179**
	BMI, BMI status and pathways	0.01941	**0.08532**
	PR only	0.1817	$\mathbf{4.249 \times 10^{-6}}$
	BMI and PR	0.1558	$\mathbf{2.188 \times 10^{-5}}$
	BMI status and PR	0.07293	**0.003359**
	BMI, BMI status and PR	0.07116	**0.003732**
Res	BMI only	0.09072	**0.001163**
	BMI and BMI status	0.04082	**0.02287**
	Pathways only	0.2263	$\mathbf{2.275 \times 10^{-7}}$
	BMI and pathways	0.2012	$\mathbf{1.206 \times 10^{-6}}$
	BMI status and pathways	0.1184	**0.0002191**
	BMI, BMI status and pathways	0.1005	**0.0006487**
	PR only	0.1419	$\mathbf{5.195 \times 10^{-5}}$
	BMI and PR	0.1224	**0.0001718**
	BMI status and PR	0.05342	**0.01074**

Table S31 (continued)

	BMI, BMI status and PR	0.05419	**0.01026**
ResOl	BMI only	0.1535	$\mathbf{2.536 \times 10^{-5}}$
	BMI and BMI status	0.0466	**0.01615**
	Pathways only	0.1581	$\mathbf{1.899 \times 10^{-5}}$
	BMI and pathways	0.1329	$\mathbf{9.083 \times 10^{-5}}$
	BMI status and pathways	0.05085	**0.01252**
	BMI, BMI status and pathways	0.04523	**0.01754**
	PR only	0.2067	$\mathbf{8.421 \times 10^{-7}}$
	BMI and PR	0.1936	$\mathbf{1.979 \times 10^{-6}}$
	BMI status and PR	0.1149	**0.0002723**
	BMI, BMI status and PR	0.1097	**0.0003713**
Ca	BMI only	0.07979	**0.002233**
	BMI and BMI status	0.05026	**0.01297**
	Pathways only	0.2379	$\mathbf{1.042 \times 10^{-7}}$
	BMI and pathways	0.2081	$\mathbf{7.653 \times 10^{-7}}$
	BMI status and pathways	0.09383	**0.0009655**
	BMI, BMI status and pathways	0.07756	**0.00255**
	PR only	0.09938	**0.000692**
	BMI and PR	0.08114	**0.00206**
	BMI status and PR	0.01059	0.1512
	BMI, BMI status and PR	0.008212	0.1774
CaOl	BMI only	0.1242	**0.0001542**
	BMI and BMI status	0.05609	**0.009156**
	Pathways only	0.08453	**0.001683**
	BMI and pathways	0.07397	**0.003157**
	BMI status and pathways	0.003647	0.244
	BMI, BMI status and pathways	0.0003557	0.3111
	PR only	0.1253	**0.0001444**
	BMI and PR	0.1243	**0.0001534**
	BMI status and PR	0.02838	**0.04876**
	BMI, BMI status and PR	0.02625	0.0556
CaRes	BMI only	0.05877	**0.007805**

Table S31 (continued)

	BMI and BMI status	0.0415	**0.02194**
	Pathways only	0.2438	$\mathbf{6.952 \times 10^{-8}}$
	BMI and pathways	0.2037	$\mathbf{1.023 \times 10^{-6}}$
	BMI status and pathways	0.1132	**0.0003018**
	BMI, BMI status and pathways	0.0995	**0.0006872**
	PR only	0.1096	**0.0003748**
	BMI and PR	0.08751	**0.001409**
	BMI status and PR	0.02671	0.05406
	BMI, BMI status and PR	0.02392	0.06429
CaResOl	BMI only	0.1167	**0.0002439**
	BMI and BMI status	0.04905	**0.01395**
	Pathways only	0.07821	**0.002453**
	BMI and pathways	0.06736	**0.004681**
	BMI status and pathways	-0.0005575	0.3338
	BMI, BMI status and pathways	-0.003248	0.4151
	PR only	0.1632	$\mathbf{1.377 \times 10^{-5}}$
	BMI and PR	0.1628	$\mathbf{1.412 \times 10^{-5}}$
	BMI status and PR	0.05905	**0.007678**
	BMI, BMI status and PR	0.05391	**0.01043**
Or	BMI only	0.1246	**0.0001508**
	BMI and BMI status	0.03912	**0.02533**
	Pathways only	0.1649	$\mathbf{1.237 \times 10^{-5}}$
	BMI and pathways	0.1311	**0.0001014**
	BMI status and pathways	0.05614	**0.009131**
	BMI, BMI status and pathways	0.04897	**0.01402**
	PR only	0.1805	$\mathbf{4.604 \times 10^{-6}}$
	BMI and PR	0.1626	$\mathbf{1.431 \times 10^{-5}}$
	BMI status and PR	0.08503	**0.001634**
	BMI, BMI status and PR	0.08819	**0.001353**
FM	BMI only	-0.004507	0.4632
	BMI and BMI status	0.01947	0.08502
	Pathways only	0.4202	$\mathbf{8.085 \times 10^{-14}}$

Table S31 (continued)

	BMI and pathways	0.4213	$\mathbf{7.317 \times 10^{-14}}$
	BMI status and pathways	0.3386	$\mathbf{6.885 \times 10^{-11}}$
	BMI, BMI status and pathways	0.2734	$\mathbf{8.759 \times 10^{-9}}$
	PR only	0.09798	**0.0007529**
	BMI and PR	0.09642	**0.0008267**
	BMI status and PR	-0.00988	0.9634
	BMI, BMI status and PR	-0.008862	0.7478

[1] All values in bold are statistically significant ($p < 0.05$).

Table S32: Table summarising the linear model statistics from the comparison of the predicted obesity metagenes scores from the BMI status-only model with the corresponding obesity metagene scores from the CR data set for all of the obesity metagenes

	P-values		
Obesity metagene	Overweight	Obese	ANOVA
Cr	$\mathbf{3.005 \times 10^{-6}}$ [1]	**0.0002419**	$\mathbf{6.931 \times 10^{-6}}$
CrOl	$\mathbf{2.924 \times 10^{-7}}$	$\mathbf{8.516 \times 10^{-6}}$	$\mathbf{1.467 \times 10^{-7}}$
Res	0.2783	$\mathbf{2.146 \times 10^{-5}}$	$\mathbf{5.715 \times 10^{-5}}$
ResOl	$\mathbf{2.023 \times 10^{-6}}$	$\mathbf{3.31 \times 10^{-5}}$	$\mathbf{1.052 \times 10^{-6}}$
Ca	**0.00283**	$\mathbf{8.769 \times 10^{-5}}$	**0.0002354**
CaOl	**0.0001869**	$\mathbf{3.328 \times 10^{-6}}$	$\mathbf{4.371 \times 10^{-6}}$
CaRes	**0.008582**	**0.0004552**	**0.001334**
CaResOl	**0.0004192**	$\mathbf{5.338 \times 10^{-6}}$	$\mathbf{8.806 \times 10^{-6}}$
Or	$\mathbf{5.806 \times 10^{-6}}$	**0.0002258**	$\mathbf{8.092 \times 10^{-6}}$
FM	0.8573	0.185	0.3568

[1] All values in bold are statistically significant ($p < 0.05$).

B12 Code used in this project

All of the R source code used in this project is publicly available at: https://github.com/rikutakei/bmi-cancer-code.git. This thesis was written with LaTeX, and the source code is publicly available at https://github.com/rikutakei/mastersDoc.git.

References

Alter, O., Brown, P.O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.* **97**, 10101–10106

Ames, B.N., Swirsky, L. and Willettt, W.C. (1995) The causes and prevention of cancer. *Proc. Natl. Acad. Sci. USA* **92**, 5258–5265

Armstrong, B.K. and Kricker, A. (2001) The epidemiology of UV induced skin cancer. *J. Photochem. Photobiol. B Biol.* **63**, 8–18

Barsh, G.S. and Schwartz, M.W. (2002) Genetic approaches to studying energy balance: perception and integration. *Nat. Rev. Genet.* **3**, 589–600

Basen-Engquist, K. and Chang, M. (2011) Obesity and cancer risk: Recent review and evidence. *Curr. Oncol. Rep.* **13**, 71–76

Bell, C.G., Walley, A.J. and Froguel, P. (2005) The genetics of human obesity. *Nat. Rev. Genet.* **6**, 221–234

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300

Bhowmick, N.a., Neilson, E.G. and Moses, H.L. (2004) Stromal fibroblasts in cancer initiation and progression. *Nature* **432**, 332–337

Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.B., Harpole, D., Lancaster, J.M., Berchuck, A., Olson, J.a., Marks, J.R., Dress-

man, H.K., West, M. and Nevins, J.R. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357

Bild, A.H., Parker, J.S., Gustafson, A.M., Acharya, C.R., Hoadley, K.A., Anders, C., Marcom, P.K., Carey, L.A., Potti, A., Nevins, J.R. and Perou, C.M. (2009) An integration of complementary strategies for gene-expression analysis to reveal novel therapeutic opportunities for breast cancer. *Breast cancer Res.* **11**, R55

Bray, F., Ren, J.S., Masuyer, E. and Ferlay, J. (2013) Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int. J. Cancer* **132**, 1133–1145

Brüning, J.C., Gautam, D., Burks, D.J., Gillette, J., Schubert, M., Orban, P.C., Klein, R., Krone, W., Müller-Wieland, D. and Kahn, C.R. (2000) Role of Brain Insulin Receptor in Control of Body Weight and Reproduction. *Science* **289**, 2122–2125

Cairns, R., Harris, I. and Mak, T. (2011) Regulation of cancer cell metabolism. *Nat. Rev. Cancer* **11**, 85–95

Calle, E.E. and Kaaks, R. (2004) Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. *Nat. Rev. Cancer* **4**, 579–591

Calle, E.E., Rodriguez, C., Walker-Thurmond, K. and Thun, M.J. (2003) Overweight, obesity, and mortality from cancer in a prospectively studied cofort of U.S. adults. *N. Engl. J. Med.* **348**, 1625–1638

Carlson, M. (2016a) GO.db: A set of annotation maps describing the entire Gene Ontology. R package version 3.4.0

Carlson, M. (2016b) hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a). R package version 3.2.3

Carlson, M. (2016c) KEGG.db: A set of annotation maps for KEGG

Coller, H.A., Grandori, C., Tamayo, P., Colbert, T., Lander, E.S., Eisenman, R.N. and Golub, T.R. (2000) Expression analysis with oligonucleotide microarrays

reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 3260–3265

Creighton, C.J., Sada, Y.H., Zhang, Y., Tsimelzon, A., Wong, H., Dave, B., Landis, M.D., Bear, H.D., Rodriguez, A. and Chang, J.C. (2012) A gene transcription signature of obesity in breast cancer. *Breast Cancer Res. Treat.* **132**, 993–1000

Dalton, M., Cameron, A., Zimmet, P., Shaw, J., Jolley, D., Dunstan, D. and Welborn, T. (2003) Waist circumference, waist–hip ratio and body mass indexand their correlation with cardiovascular disease risk factorsin Australian adults. *J. Intern. Med.* **254**, 555–563

Damrauer, J.S., Hoadley, K.A., Chism, D.D., Fan, C., Tiganelli, C.J., Wobker, S.E., Yeh, J.J., Milowsky, M.I., Iyer, G., Parker, J.S. and Kim, W.Y. (2014) Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 3110–3115

Dar, A.C., Das, T.K., Shokat, K.M. and Cagan, R.L. (2012) Chemical genetic discovery of targets and anti-targets for cancer polypharmacology. *Nature* **486**, 80–84

Dina, C., Meyre, D., Gallina, S., Durand, E., Körner, A., Jacobson, P., Carlsson, L.M.S. *et al.* (2007) Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat. Genet.* **39**, 724–726

Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210

El-Sayed Moustafa, J.S. and Froguel, P. (2013) From obesity genetics to the future of personalized obesity therapy. *Nat. Rev. Endocrinol.* **9**, 402–413

Elinav, E., Nowarski, R., Thaiss, C.A., Hu, B., Jin, C. and Flavell, R.A. (2013) Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. *Nat. Rev. Cancer* **13**, 759–771

Frayling, T.M., Timpson, N.J., Weedon, M.N., Freathy, R.M., Lindgren, C.M., Perry, J.R.B., Katherine, S. *et al.* (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894

Fridman, J.S. and Lowe, S.W. (2003) Control of apoptosis by p53. *Oncogene* **22**, 9030–9040

Friedberg, E.C. (2003) DNA damage and repair. *Nature* **421**, 436–440

Fuentes-Mattei, E., Velazquez-Torres, G., Phan, L., Zhang, F., Chou, P.C., Shin, J.H., Choi, H.H. *et al.* (2014) Effects of Obesity on Transcriptomic Changes and Cancer Hallmarks in Estrogen Receptor-Positive Breast Cancer. *J. Natl. Cancer Inst.* **106**, dju158

Fuqua, S., Fitzgerald, S. and Chamness, G. (1991) Variant human breast tumor estrogen receptor with constitutive transcriptional activity. *Cancer Res.* **8**, 105–109

Gallagher, R.P. and Lee, T.K. (2006) Adverse effects of ultraviolet radiation: A brief review. *Prog. Biophys. Mol. Biol.* **92**, 119–131

Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A.B., Maisonneuve, P. and Boyle, P. (2008) Tobacco smoking and cancer: A meta-analysis. *Int. J. Cancer* **122**, 155–164

Garofalo, C. and Surmacz, E. (2006) Leptin and Cancer. *J. Cell. Physiol.* **207**, 12–22

Garrow, J.S. (1988) *Obesity and related diseases*. Churchill Livingstone, London

Gatza, M.L., Lucas, J.E., Barry, W.T., Kim, J.W., Wang, Q., Crawford, M.D., Datto, M.B., Kelley, M., Mathey-Prevot, B., Potti, A. and Nevins, J.R. (2010) A pathway-based classification of human breast cancer. *Proc. Natl. Acad. Sci. USA* **107**, 6994–6999

Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315

Gelber, R.P., Gaziano, J.M., Orav, E.J., Manson, J.E., Buring, J.E. and Kurth, T. (2008) Measures of Obesity and Cardiovascular Risk Among Men and Women. *J. Am. Coll. Cardiol.* **52**, 605–615

Gene Ontology Consortium (2000) Gene Ontology: Tool for The Unification of Biology. *Nat. Genet.* **25**, 25–29

Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, 258D–261

Gerken, T., Girard, C.A., Tung, Y.C.L., Webby, C.J., Saudek, V., Hewitson, K.S., Yeo, G.S.H. *et al.* (2007) The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* **318**, 1469–1472

Ghilardi, N., Ziegler, S., Wiestner, A., Stoffel, R., Heim, M.H. and Skoda, R.C. (1996) Defective STAT signaling by the leptin receptor in diabetic mice. *Proc. Natl. Acad. Sci. USA* **93**, 6231–6235

Gilbert, C.a. and Slingerland, J.M. (2013) Cytokines, obesity, and cancer: new insights on mechanisms linking obesity to cancer risk and progression. *Annu. Rev. Med.* **64**, 45–57

Giovannucci, E. (1995) Insulin and colon cancer. *Cancer Causes Control* **6**, 164–179

Girke, T. (2016) overLapper.R

Goeman, J.J. and Bühlmann, P. (2007) Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics* **23**, 980–987

Golub, G.H. and Reinsch, C. (1970) Singular value decomposition and least squares solutions. *Numer. Math.* **14**, 403–420

Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell* **100**, 57–70

Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: The next generation. *Cell* **144**, 646–674

Hecht, S. (1999) Tobacco smoke carcinogen and lung cancer. *J. Natl. Cancer Inst.* **91**, 1194–1210

Hochberg, Y. and Tamhane, A.C. (1987) *Multiple comparison procedures*. John Wiley and Sons, New York

Hoeijmakers, J.H.J. (2001) Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366–374

Hurd, P.J. and Nelson, C.J. (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings Funct. Genomics Proteomics* **8**, 174–183

International Cancer Genome Consortium (2016) International Cancer Genome Consortium.
Available: http://icgc.org/ [7/9/2015]

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264

Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E. *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–349

Johnson, W.E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E. and Stein, L. (2005) Reactome: A knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432

Kalluri, R. and Weinberg, R.a. (2009) Review series The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, 480–484

Kazerounian, S., Yee, K.O. and Lawler, J. (2008) Thrombospondins: From structure to therapeutics - Thrombospondins in cancer. *Cell. Mol. Life Sci.* **65**, 700–712

Kearney, J. (2010) Food consumption trends and drivers. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2793–2807

Keitt, T. (2012) colorRamps: Builds color tables

Kelesidis, I., Kelesidis, T. and Mantzoros, C.S. (2006) Adiponectin and cancer : a systematic review. *Br. J. Cancer* **94**, 1221–1225

Khatri, P., Sirota, M. and Butte, A.J. (2012) Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**

Langfelder, P. and Horvath, S. (2008) WGCNA : an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559

Law, C.W., Chen, Y., Shi, W. and Smyth, G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29

Lee, C.M.Y., Huxley, R.R., Wildman, R.P. and Woodward, M. (2008) Indices of abdominal obesity are better discriminators of cardiovascular risk factors than BMI: a meta-analysis. *J. Clin. Epidemiol.* **61**, 646–653

Lee, G.H., Proenca, R., Montez, J.M., Carroll, K.M., Darvishzadeh, J.G., Lee, J.I. and Friedman, J.M. (1996) Abnormal splicing of the leptin receptor in diabetic mice. *Nature* **379**, 632–635

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. and Storey, J.D. (2012) The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883

Leibowitz, M.L., Zhang, C.Z. and Pellman, D. (2015) Chromothripsis: A New Mechanism for Rapid Karyotype Evolution. *Annu. Rev. Genet.* **49**, 183–211

Levine, A.J. (1997) P53, the Cellular Gatekeeper for Growth and Division. *Cell* **88**, 323–331

Lightenberg, W. (2016) reactome.db: A set of annotation maps for reactome

Lim, S.S., Vos, T., Flaxman, A.D., Danaei, G., Shibuya, K., Adair-Rohani, H., Amann, M. *et al.* (2012) A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2224–2260

Liotta, L.A. and Kohn, E.C. (2001) The microenvironment of the tumour-host interface. *Nature* **411**, 375–379

Lumeng, C.N. and Saltiel, A.R. (2011) Inflammatory links between obesity and metabolic disease. *J. Clin. Invest.* **121**, 2111–2117

Malik, V.S., Willett, W.C. and Hu, F.B. (2013) Global obesity: trends, risk factors and policy implications. *Nat. Rev. Endocrinol.* **9**, 13–27

Mantovani, A., Allavena, P., Sica, A. and Balkwill, F. (2008) Cancer-related inflammation. *Nature* **454**, 436–444

McKeown-Eyssen, G. (1994) Epidemiology Triglycerides of Colorectal and / or Plasma Cancer Glucose Revisited : Associated Are Serum with Risk ? *Cancer Epidemiol. Biomarkers Prev.* **3**, 687– 695

Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46

Moelling, K., Schad, K., Bosse, M., Zimmermann, S. and Schweneker, M. (2002) Regulation of Raf-Akt cross-talk. *J. Biol. Chem.* **277**, 31099–31106

Montague, C.T., Farooqi, I.S., Whitehead, J.P., Soos, M.a., Rau, H., Wareham, N.J., Sewter, C.P., Digby, J.E., Mohammed, S.N., Hurst, J.a., Cheetham, C.H., Earley, a.R., Barnett, a.H., Prins, J.B. and O'Rahilly, S. (1997) Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387**, 903–908

Muthukaruppan, A., Lasham, A., Woad, K.J., Black, M.A., Blenkiron, C., Miller, L.D., Harris, G., McCarthy, N., Findlay, M.P., Shelling, A.N. and Print, C.G. (2016) Multimodal assessment of oestrogen receptor mRNA profiles to quantify oestrogen pathway activity in breast tumours. *Clin. Breast Cancer* pp. 1–15

National Cancer Institute and National Human Genome Research Institute (2016) The Cancer Genome Atlas.
Available: https://cancergenome.nih.gov/ [2015-04-01]

National Center for Biotechnology Information (2016) Gene Expression Omnibus.
Available: https://www.ncbi.nlm.nih.gov/geo/ [2015-03-31]

Neuwirth, E. (2014) RColorBrewer: ColorBrewer Palettes

New Zealand Ministry of Health (2016a) *Annual Update of Key Results 2015/16: New Zealand Health Survey.* Ministry of Health, Wellington

New Zealand Ministry of Health (2016b) *Cancer : New registrations and deaths 2013*. Ministry of Health, Wellington

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D. *et al.* (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993

Pikarsky, E., Porat, R.M., Stein, I., Abramovitch, R., Amit, S., Kasem, S., Gutkovich-Pyest, E., Urieli-Shoval, S., Galun, E. and Ben-Neriah, Y. (2004) NF-kappaB functions as a tumour promoter in inflammation-associated cancer. *Nature* **431**, 461–466

R Development Core Team (2016) R: A language and environment for statictical computing.
Available: https://www.r-project.org/ [2016-01-09]

Renehan, A.G., Frystyk, J. and Flyvbjerg, A. (2006) Obesity and cancer risk: the role of the insulin-IGF axis. *Trends Endocrinol. Metab.* **17**, 328–336

Renehan, A.G., Tyson, M., Egger, M., Heller, R.F. and Zwahlen, M. (2008) Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet* **371**, 569–78

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies **43**, 1–13

Roberts, D.L., Dive, C. and Renehan, A.G. (2010) Biological mechanisms linking obesity and cancer risk: New perspectives. *Annu. Rev. Med.* **61**, 301–316

Rodríguez-Hernández, H., Simental-Mendía, L.E., Rodríguez-Ramírez, G. and Reyes-Romero, M.A. (2013) Obesity and inflammation: epidemiology, risk factors, and markers of inflammation. *Int. J. Endocrinol.* pp. 1–11

Saltiel, A.R. and Pessin, J.E. (2002) Insulin signaling pathways in time and space. *Trends Cell Biol.* **12**, 65–71

Samatar, A.a. and Poulikakos, P.I. (2014) Targeting RAS–ERK signalling in cancer: promises and challenges. *Nat. Rev. Drug Discov.* **13**, 928–942

Satyamoorthy, K., Li, G., Gerrero, M.R., Brose, M.S., Volpe, P., Weber, B.L., Belle, P.V., Elder, D.E. and Herlyn, M. (2003) Constitutive Mitogen-activated Protein Kinase Activation in Melanoma Is Mediated by Both BRAF Mutations and Autocrine Growth Factor Stimulation Advances in Brief Constitutive Mitogen-activated Protein Kinase Activation in Melanoma Is Mediated by Both BRAF. *Cancer Res.* **63**, 756–759

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470

Schulze, A. and Downward, J. (2001) Navigating gene expression using microarrays–a technology review. *Nat. Cell Biol.* **3**, E190–E195

Scuteri, A., Sanna, S., Chen, W.M., Uda, M., Albai, G., Strait, J., Najjar, S. *et al.* (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* **3**, 1200–1210

Shaffer, J.P. (1995) Multiple hypothesis testing. *Annu. Rev. Psychol.* **46**, 561–584

Smyth, G.K. (2004) Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 1–26

Smyth, G.K. (2005) Limma: linear models for microarray data. In *Bioinforma. Comput. Biol. Solut. Using R Bioconductor*, pp. 397–420

Spanswick, D., Smith, M.a., Groppi, V.E., Logan, S.D. and Ashford, M.L. (1997) Leptin inhibits hypothalamic neurons by activation of ATP-sensitive potassium channels. *Nature* **390**, 521–525

Spanswick, D., Smith, M.a., Mirshamsi, S., Routh, V.H. and Ashford, M.L. (2000) Insulin activates ATP-sensitive K+ channels in hypothalamic neurons of lean, but not obese rats. *Nat. Neurosci.* **3**, 757–758

Spiegelman, B.M. and Flier, J.S. (2001) Obesity and the regulation of energy balance. *Cell* **104**, 531–543

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40

Stratton, M.R., Campbell, P.J. and Futreal, P.A. (2009) The cancer genome. *Nature* **458**, 719–724

Su Huang, H.J., Nagane, M., Klingbeil, C.K., Lin, H., Nishikawa, R., Ji, X.D., Huang, C.M., Gill, G.N., Wiley, H.S. and Cavenee, W.K. (1997) The enhanced

tumorigenic activity of a mutant epidermal growth factor receptor common in human cancers is mediated by threshold levels of constitutive tyrosine phosphorylation and unattenuated signaling. *J. Biol. Chem.* **272**, 2927–2935

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.a., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550

Talmadge, J.E. and Fidler, I.J. (2010) AACR centennial series: The biology of cancer metastasis: Historical perspective. *Cancer Res.* **70**, 5649–5669

Teng, M.W., Swann, J.B., Koebel, C.M., Schreiber, R.D. and Smyth, M.J. (2008) Immune-mediated dormancy: an equilibrium with cancer. *J. Leukoc. Biol.* **84**, 988–993

Vander Heiden, M.G., Cantley, L.C., Thompson, C.B., Mammalian, P., Exhibit, C. and Metabolism, A. (2009) Understanding the Warburg Effect : The metabolic requirements of cell proliferation. *Science* **324**, 1029–1034

Vogelstein, B. and Kinzler, K. (2004) Cancer genes and the pathways they control. *Nat. Med.* **10**, 789–799

Wardburg, O. (1956) On the origin of cancer cells. *Science* **123**, 309–314

Warnes, G., Bolker, B., Lodewijk, B., Gentleman, R., Andy Liaw, W., Lumley, T., Maechler, M., Magnusson, A., Steffen, M., Schwartz, M. and Venables, B. (2016) gplots: Various R Programming Tools for Plotting Data

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A., Marks, J.R. and Nevins, J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA* **98**, 11462–11467

Wilhelm, S., Carter, C., Lynch, M., Lowinger, T., Dumas, J., Smith, R.a., Schwartz, B., Simantov, R. and Kelley, S. (2006) Discovery and development

of sorafenib: a multikinase inhibitor for treating cancer. *Nat. Rev. Drug Discov.* **5**, 835–844

Wiseman, B.S. and Werb, Z. (2002) Stromal Effects on Mammary Gland Development and Breast Cancer. *Science* **296**, 1046–1049

Woods, S.C., Lotter, E.C., McKay, L.D. and Porte, D. (1979) Chronic intracerebroventricular infusion of insulin reduces food intake and body weight of baboons. *Nature* **282**, 503–505

World Health Organisation (2000) Obesity : preventing and managing the global epidemic **894**, 6–8

World Health Organisation (2014) Global status report on noncommunicable diseases 2014. *Geneva: World Health Organization*

World Health Organisation (2016) Cancer factsheet.
Available: http://www.who.int/mediacentre/factsheets/fs297/en/ [2016-11-16]

Wright, W.E., Pereira-Smith, O.M. and Shay, J.W. (1989) Reversible cellular senescence: implications for immortalization of normal human diploid fibroblasts. *Mol. Cell. Biol.* **9**, 3088–3092

Wu, D. and Smyth, G.K. (2012) Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40**, e133

Yancopoulos, G.D., Gale, N.W., Davis, S., Gale, N.W., Rudge, J.S., Wiegand, S.J. and Holash, J. (2000) Vascular-Specific Growth Factors and Blood Vessel Formation. *Nature* **407**, 242–248

Yu, H., Kortylewski, M. and Pardoll, D. (2007) Crosstalk between cancer and immune cells: role of STAT3 in the tumour microenvironment. *Nat. Rev. Immunol.* **7**, 41–51

Yu, H., Lee, H., Herrmann, A., Buettner, R. and Jove, R. (2014) Revisiting STAT3 signalling in cancer : new and unexpected biological functions. *Nat. Rev. Cancer* **14**, 736–746

Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., Wong-Erasmus, M., Yao, L. and Kasprzyk, A. (2011) International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database* **2011**, 1–10

Zhang, Y., Proenca, R., Maffei, M., Barone, M., Leopold, L. and Friedman, J.M. (1994) Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–432

Zimmermann, S. and Moelling, K. (1999) Phosphorylation and regulation of Raf by Akt (protein kinase B). *Science* **286**, 1741–1744