Using data from The Cancer Genome Atlas to analyse DNA methylation and copy number changes in the Y chromosome in male cancers

by

Robert John Hollows

A thesis submitted to the University of Birmingham for the degree of Doctor of Philosophy

The Institute of Cancer and Genomic Sciences

College of Medical and Dental Sciences

University of Birmingham

February 2016

Abstract

Many human cancers are more prevalent in men than women. This disparity is not fully explained by differences in key risk factor exposures, which suggests a possible genetic cause. Recent research has reported a link between loss of the Y chromosome (LoY) and increased incidence of non-haematological cancers.

Using data from The Cancer Genome Atlas, I conducted an integrated, multi-‘omic analysis of Y chromosome methylation and copy number aberrations in three different cancers – colon, head and neck and kidney.

My results indicate that aberrant methylation of the Y chromosome is common in all three cancer types. Hyper-methylation occurs in short, discrete regions, interspersed among wider regions of more general hypo-methylation. I also show that LoY is the most common aneuploidy in all three cancers, affecting between one third and one half of patients. Furthermore, both aberrant methylation and LoY are associated with reduced expression of potentially important genes.

Most interestingly, for HPV negative head and neck cancer patients, I show a statistically significant association between LoY and worse survival, and that LoY may be linked to smoking. Subject to further validation, this suggests that LoY could be important in the pathogenesis of head and neck cancer for HPV negative patients.

Acknowledgements

I would like to thank Professor Paul Murray and Dr Wenbin Wei for their valuable assistance and guidance with this project, and the Medical Research Council for their provision of funding.

Table of contents

List of figures

List of tables

List of abbreviations

1stExon - first exon

27k - HumanMethylation27 BeadChip

3’UTR - untranslated region at 3’ end

450k - HumanMethylation450 BeadChip

5’UTR - untranslated region at 5’ end

AMELY - amelogenin Y-linked gene

ATM - ataxia telangiectasia mutated gene

BCORL2 - BCL6 corepressor-like 2 gene

BCR - Biospecimen Core Resource

BLAST - Basic Local Alignment Search Tool

CDKN2A - cyclin-dependent kinase inhibitor 2A gene

chrY - chromosome Y

CIMP - CpG island methylator phenotype

CNV - copy number variation

COAD – TCGA’s colon adenocarcinoma dataset

COSMIC - catalogue of somatic mutations in cancer

CpG - cytosine and guanine dinucleotide

CYorf15A - taxilin gamma pseudogene, Y-linked

DCC - Data Coordinating Center

DDX3Y - DEAD-box helicase3, Y-linked gene

DNA - deoxyribonucleic acid

DNMT1 - DNA methyltransferase 1 gene

EIF1AY - eukaryotic translation initiation factor 1A, Y-linked gene

FISH - fluorescence in-situ hybridisation

GCC - Genome Characterization Center

GDAC - Genome Data Analysis Center

GSC - Genome Sequencing Center

H3K27 - lysine 27 on histone 3

hg18 - human genome reference sequence 18 (UCSC)

hg19 - human genome reference sequence 19 (UCSC)

HNSC – TCGA’s head and neck squamous cell carcinoma dataset

HPV - human papillomavirus

HR - hazard ratio

KDM5D - lysine demethylase 5D gene

KIRC - kidney renal clear-cell carcinoma dataset

K-W - Kruskal-Wallis test

LoY - loss of Y chromosome

MDM2 - mouse double minute 2 homolog gene

MiRNA - microRNA

mRNA - messenger ribonucleic acid

NCBI - National Center for Biotechnology Information

NCRNA00185 - non-coding RNA 185

NLGN4Y - neuroligin 4, Y-linked gene

NOTCH1 - notch homolog 1 gene

PAR - pseudoautosomal region

PCA - principal components analysis

PCR - polymerase chain reaction

PD-1 - programmed cell death protein 1

PD-L1 - programmed death ligand 1

PRKY - protein kinase, Y-linked gene

P29 - 29th percentile

P72 - 72nd percentile

R - R programming language

Rb - retinoblastoma gene

RefSeq - NCBI Reference Sequence Database

RNA - ribonucleic acid

RNA-seq - RNA sequencing

RPS4Y1 - ribosomal protein S4, Y-linked 1 gene

Sd - standard deviation

SNP - single nucleotide polymorphism

SRY - sex determining region Y gene

TAP1/2 - transporter associated with antigen processing 1 / 2 gene

TBL1Y - transducin beat-like 1, Y-linked gene

TCGA - The Cancer Genome Atlas

TFI - total female intensity

TMSB4Y - thymosin beta 4, Y-linked gene

TP53 - tumour protein 53 gene

TP53+ve - with TP53 mutation

TP53-ve - without TP53 mutation

TSG - tumour suppressor gene

TSS - transcriptional start site

TSS1500 - region between 200 and 1500 bases upstream of transcriptional start site

TSS200 - region within 200 bases upstream of transcriptional start site

TTTY14/15 - testis-specific transcript, Y-linked 14 / 15 gene

UCSC - University of California, Santa Cruz

USP9Y - ubiquitin-specific peptidase 9, Y-linked gene

UTY - ubiquitously transcribed tetratricopeptide repeat containing, Y-linked gene

Yp - short arm of chromosome Y

Yq - short arm of chromosome Y

ZFY - zinc finger protein, Y-linked gene

1. Introduction

Many human cancers are more prevalent in men than in women[1]. This discrepancy is not fully explained by men having greater exposure to key risk factors[2], which raises the possibility that there may be underlying genetic differences between the sexes which result in men being more susceptible to cancer.

The most fundamental genetic difference between men and women is that the former have one maternally inherited X chromosome and one paternally inherited Y chromosome, whereas the latter have two X chromosomes. So it is reasonable to speculate that aberrations in either or both of these chromosomes could be, at least partially, responsible for the difference in cancer prevalence.

1.1 Project outline

My project is focussed on the Y chromosome. Recent research[3] has suggested that loss of the Y chromosome in peripheral blood may be linked to increased risk of cancer in men and also shorter survival times. Additional research by the same group[4] has linked Y chromosome loss to smoking. Furthermore, there is recent evidence that the Y chromosome contains a small number of genes whose functions are fundamentally important to male survival[5], including two potential tumour suppressor genes[6].

The initial focus of my project was on methylation of the Y chromosome and how this is altered in male tumours. I chose methylation as this is a biological process which is known to be altered in tumours[7], and because there is very little research into aberrant methylation of the Y chromosome.

I have studied three different cancer types using data provided on-line by The Cancer Genome Atlas ("TCGA")[8]. All of these cancer types are more prevalent in men than in women[1].

My initial analyses were based on TCGA's kidney renal clear-cell carcinoma (KIRC) dataset[9]. I chose this dataset because, at the time, it was relatively under-studied compared with other datasets, it has a large amount of methylation data, and, in particular, it has a large number of normal tissue samples against which the tumour samples can be compared.

I then performed similar analyses on TCGA's colon adenocarcinoma dataset (COAD)[10]. I chose this dataset because aberrant methylation is a well-known feature in a subset of colon cancers[11].

Finally, I also studied TCGA's head and neck squamous cell carcinoma dataset (HNSC)[12], which has a very large number of samples processed on TCGA's chosen methylation analysis platform. A further advantage of this dataset is that useable data on smoking history is also available (which is not the case for the KIRC and COAD datasets). In addition, one of the risk factors for head and neck cancer is infection with the human papillomavirus (HPV)[13], so by analysing this dataset I was able to compare tumours with and without viral infection.

An advantage of the way in which TCGA's methylation data are produced is that they can also be used to estimate copy number variations[14]. So the next part of my project used the methylation data for this purpose. TCGA also provide separate copy number data, so I used these additional data to corroborate my results based on the methylation data.

For each of the three cancer types, I then used TCGA's gene expression data to look for Ychromosome genes whose expression appeared to be affected by either aberrant methylation or copy number variation (in particular loss).

I was then interested to discover whether the Y chromosome changes I had found were associated with any differences in survival. TCGA also provide clinical information for the samples they have processed, and this contains data on overall survival.

Finally, for the HNSC dataset only, I performed more detailed analyses of the associations between Y chromosome aberrations and the key risk factors of smoking and HPV infection. For this purpose I also made use of TCGA’s somatic mutation data.

To my knowledge this is the first time that a multi-‘omic study of changes in the Y chromosome in human cancer has been undertaken. Figure 1 below summarises the order in which my results are presented:

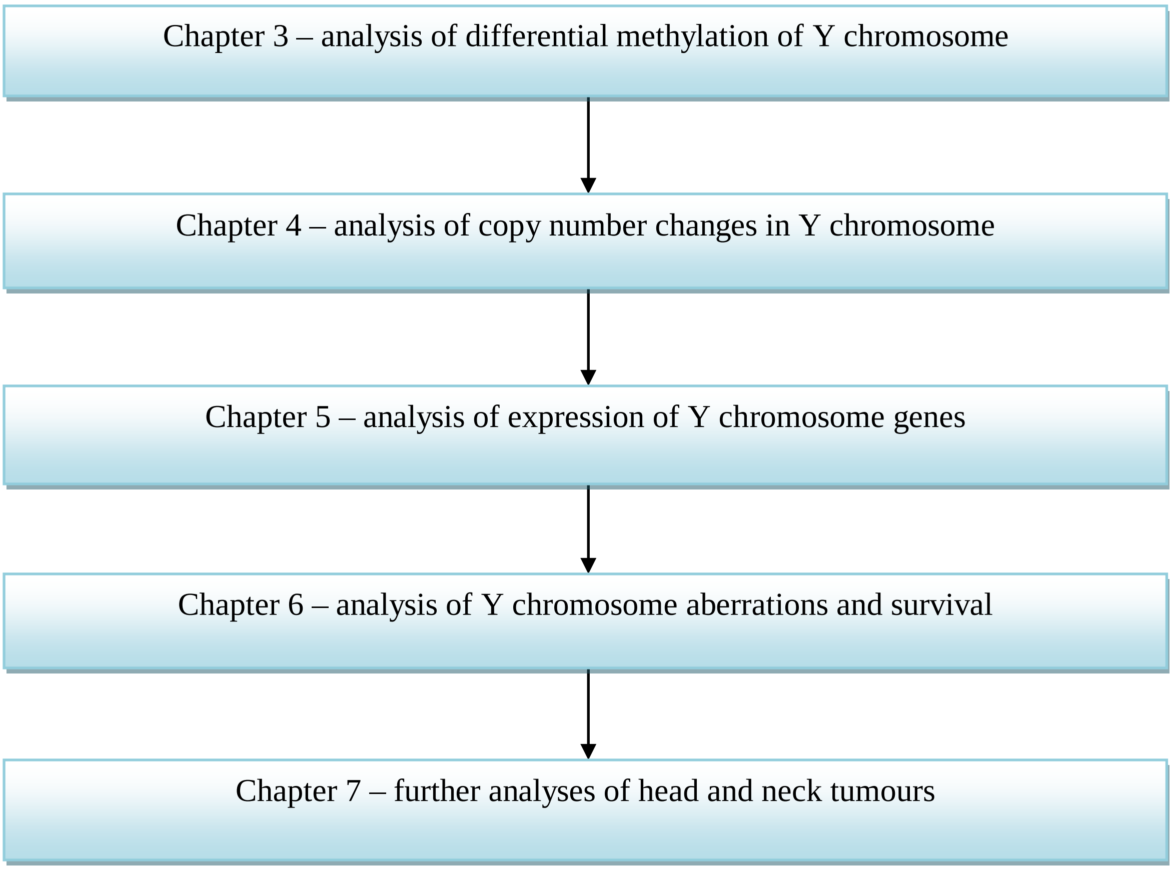

Chapter 3 – analysis of differential methylation of Y chromosome
Chapter 4 – analysis of copy number changes in Y chromosome
Chapter 5 – analysis of expression of Y chromosome genes
Chapter 6 – analysis of Y chromosome aberrations and survival
Chapter 7 – further analyses of head and neck tumours

Figure 1 - overview of results chapters

1.2 The Y chromosome

The Y chromosome is the third shortest of all the human chromosomes at around 60 million base pairs (Mbp). In contrast to its sex-chromosomal partner, the X chromosome, it contains the genetic material which uniquely defines the male gender, and this is its primary, and most researched, function.

The Y chromosome is fundamentally different to the 22 human autosomes in that it only recombines (with the X chromosome) over a very small part of its overall length. Furthermore, its transmission from generation to generation is wholly via the male gender. These two factors have had important implications for Y chromosome evolution. Over recent years, research has provided insights into this evolutionary process, along with details of the Y chromosome's genetic make-up and functionality beyond sex-determination[15,16].

1.2.1 Evolution of the Y chromosome

Separate sex-defining chromosomes have evolved independently in many different species. The human X and Y chromosomes are believed to have originated over 200 million years ago from an ancestral pair of autosomes in eutherian mammals[15].

Whilst the precise mechanism of how the process of sex-differentiation started is unclear, it is thought that an important first step was the acquisition of a key sex-determining gene on the prototype Y chromosome (the SRY gene in humans). The process was then accelerated by suppression of recombination between the two erstwhile autosomes, possibly by means of chromosomal inversions[15]. Comparison of the DNA sequences of homologous genes on the X and Y chromosomes has revealed heterogeneous levels of sequence divergence, suggesting that suppression of recombination was a multi-step process, leading to distinct evolutionary strata on the Y chromosome[17]. The afore-mentioned SRY gene is located within the oldest identified strata, consistent with the idea that it was a very early acquisition to the Y chromosome.

Suppression of recombination means that natural selection acts over the whole length of the chromosome (except for the two pseudoautosomal regions – see below). This leads to greater accumulation of deleterious mutations than would otherwise occur in a normally recombining pair of chromosomes, which in turn results in substantial levels of gene decay. Research suggests that gene decay on the Y chromosome occurred at a rapid rate initially, but slowed down over time, until a steady cohort of genes was established[15]. This and the fact that transmission of the Y chromosome occurs wholly through males, implies that those genes which survive are likely to be important for male viability.

1.2.2 Genetic content of the Y chromosome

The Y chromosome is acrocentric with its centromere located at around 12.5Mbp. It can be broken down into a number of different regions, with different properties, as shown in figure 2 (reproduced from Bachtrog[15]):

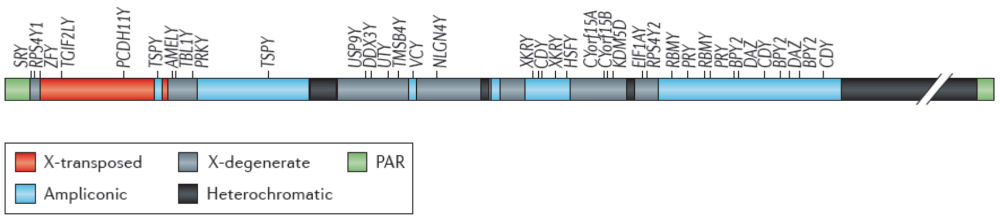


Figure 2 - overview of the human Y chromosome
Schematic diagram of the Y chromosome (taken from Bachtrog[15]) showing gene locations and the different regions (colour-coded as per legend). The diagram runs from the end of the short arm (left) to the end of the long arm (right).

At either end of the chromosome are the two pseudoautosomal regions (“PARs”). These short regions still undergo recombination during meiosis with homologous regions at the ends of the X chromosome, a process which is essential for correct segregation of the chromosomes[18].

PAR1 is located at the tip of the short arm and is around 2.7Mbp long. PAR2 is located at the opposite end of the chromosome, and is much shorter at around 0.3Mbp[19]. The two PARs contain 24 and five genes respectively, all of which are also found on the X chromosome[20]. The possible existence of a third PAR on the short arm of the Y chromosome has also been suggested[21], although this is not generally accepted yet.

In between the two PARs is the so-called “male-specific region”, which does not recombine during meiosis, and which can be sub-divided into four sub-regions. A significant part (around 70%) is hetereochromatic and contains no known functioning genes. The remaining, euchromatic part contains around 60 known genes, and can be sub-divided into three distinct regions[15,16].

1. X-degenerate – contains 16 functional genes which have survived chromosomal decay from the original ancestral autosome, and which all have homologs on the X chromosome. These genes are expressed in many areas of the body, and have a broad range of house-keeping functionality.
2. X-transposed – a short 3.4Mbp region which was transposed from the X chromosome around 3-5 million years ago and which contains only two genes.
3. Ampliconic – contains highly repetitive DNA sequences which are home to nine distinct gene families. These genes have male-specific functions, and are mainly expressed only in the testes.

My project focuses on the male-specific region, and in particular, though not exclusively, on 12 genes in this region which have recently been proposed as potentially essential to male viability[5]. The genes concerned, all of which have homologs on the X chromosome which escape X inactivation, are: RPS4Y1, ZFY, TBL1Y, PRKY, USP9Y, DDX3Y, UTY, TMSB4Y, NLGN4Y, CYorf15A, KDM5D and EIF1AY. All of these genes are located within the X-degenerate sub-region.

1.2.3 Y chromosome and cancer

The Y chromosome is often overlooked in cancer research. Indeed, TCGA's own studies into the datasets I have used do not feature it[9,10,12]. This is likely to be because much research is performed on a unisex basis, and hence any results involving the Y chromosome are diluted by a proportion of the samples being female. However, there is evidence of Y chromosome aberrations in numerous cancer types.

Aberrant methylation of the Y chromosome in cancer has had very little attention. However, some research has indicated that hyper-methylation of certain Y-linked genes, including the key sex-determining gene SRY, is observed in prostate cancer[22,23].

There is substantially more evidence that loss of the Y chromosome ("LoY") is a common feature in numerous cancer types[24], including the three types that I have investigated[25,26,27]. LoY in peripheral blood cells has also been linked to increased risk of solid tumours[3].

There has been some debate as to whether LoY is simply a by-product of general chromosomal instability, possibly linked to ageing, or whether it is an important causal event in tumour formation. However, there is evidence for the latter. For example, research into the effect of smoking on LoY in peripheral blood cells has indicated that different types of haematopoietic cell suffer differing degrees of loss, suggesting that it is not a totally random process across all cell-types[4]. Secondly, research into head and neck cancer has indicated that LoY is not age-related and may impact on survival[27,28]. Finally, a pan-cancer study which looked at the mutational profiles of over 8,200 human tumour samples has suggested that two genes on the Y chromosome (UTY and ZFY) may be tumour suppressor genes[6].

There is also some evidence that aberrant expression of the TSPY gene may have a role in the formation of germ cell tumours[29].

1.3 The Cancer Genome Atlas

Following on from the initial complete sequencing of the human genome in 2003 by the Human Genome Project, attention was given to how this knowledge could be applied to the cure and prevention of human diseases, including cancer. The research community already appreciated that the molecular causes of cancer were many and varied, and, therefore, that a coordinated approach to cancer research was vital.

The need for greater collaboration between research teams was heightened by the explosive growth of biological "omic" data which was becoming available as a result of technological advances such as microarrays and next generation sequencing. These technologies enabled researchers to investigate the molecular characteristics of cancer on a genome-wide scale using actual clinical samples, with a view to extracting biomarker data which could be useful for diagnosis and prognosis[30]. However, they generate vast amounts of data, giving rise to significant issues of data handling and analysis, which are difficult for small research teams to deal with. Furthermore, piece-meal research by small teams has led to many novel findings which have been difficult to validate independently because they are based on small data sets and hence lack statistical credibility[31]. Hence a more coordinated approach is required so that analyses can be made more credible by being based on larger data sets.

In recognition of the need to foster greater collaboration between research teams, The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov/) was formed in the USA in 2006 with funding from the National Cancer Institute (NCI: http://www.cancer.gov/) and National Human Genome Research Institute (NHGRI: http://www.genome.gov/)[8]. The key purpose of TCGA is to provide, on a standardised basis, a central database of molecular and clinical data relating to cancer, which can be accessed freely by the international research community. The intention is that this will foster greater sharing of knowledge and ideas amongst researchers generally.

A vast amount of biological and clinical data is freely available on-line via TCGA's data-portal (https://tcga-data.nci.nih.gov/tcga/). The database contains information on over twenty of the most commonly occurring human cancers, and, to date, has records in respect of over 10,000 actual human samples. These include both cancerous and matched normal tissues, enabling case / control comparisons to be made.

A number of different types of biological data can be downloaded:

- Gene expression
- DNA sequencing
- SNPs
- MiRNA expression
- Protein expression
- Copy number
- DNA methylation

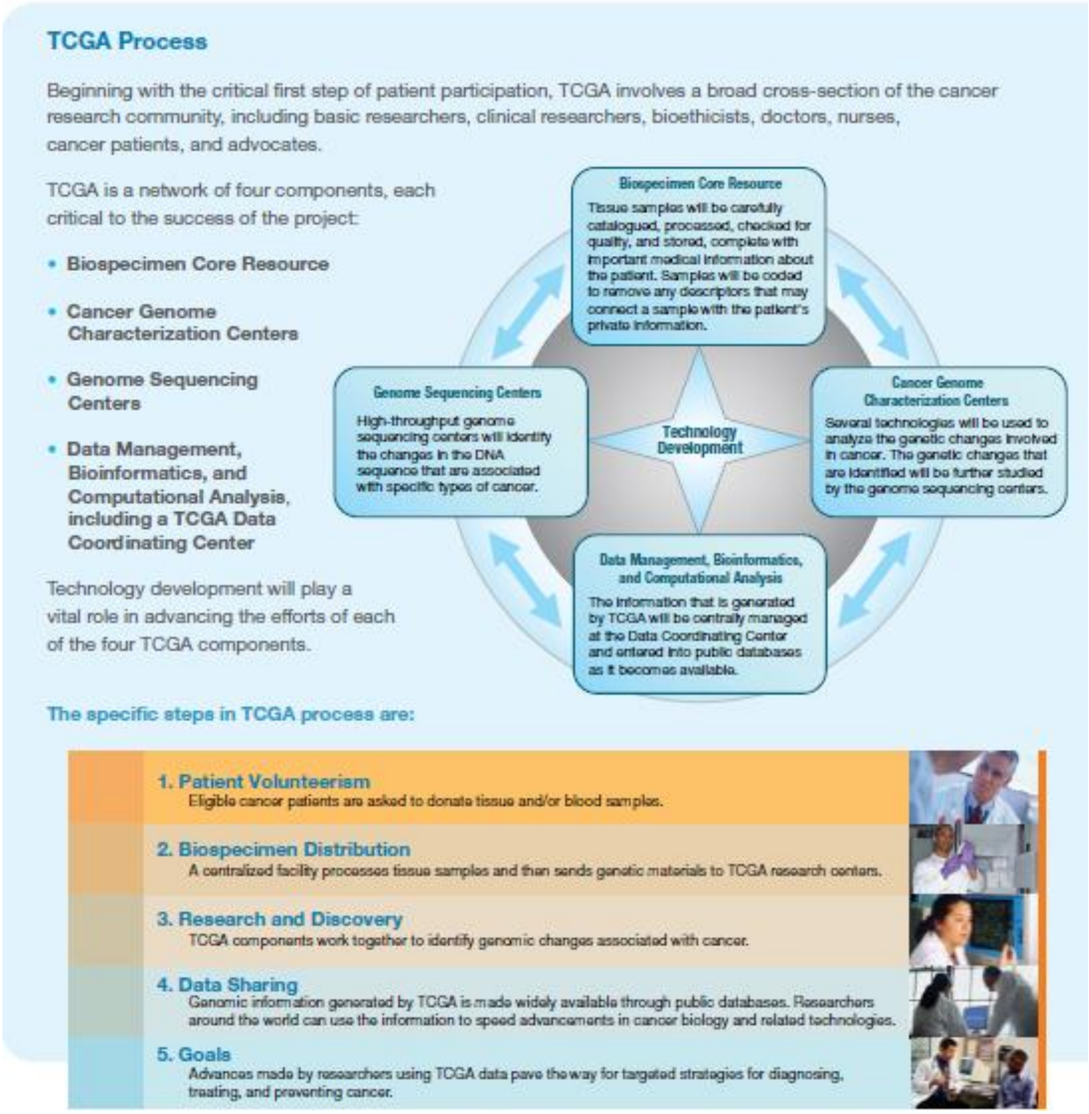
Within each type of data, samples have been processed in the same way using the same technologies, thereby facilitating comparison between samples (NB for some data types, for example DNA methylation, more than one different technology has been used with samples grouped by each technology). The data are presented with various (typically 3) levels of pre-processing having already been performed. Level 1 data are the "raw" data taken straight from the relevant microarray / sequencing technology etc with no pre-processing, with subsequent levels having increasing levels of processing. For example, the level 2 DNA methylation data have had some background correction (see later) applied to the level 1 intensity values, and the level 3 data are the calculated methylation levels (called beta values) for probes which have passed certain filtering criteria.

TCGA has developed a systematic pipeline in order to standardise and streamline the production of data. There are four key components of this pipeline:

1. The Biospecimen Core Resource (BCR), located at The Research Institute at Nationwide Children’s Hospital, Columbus, Ohio - this institution is responsible for reviewing and handling all tissue samples so that data-handling is performed in a consistent manner for all samples and to the standards required by TCGA.
2. Genome Characterization Centers (GCCs) - using the relevant technologies (e.g. microarray), these centres process the tissue samples which have been passed on by the BCR. Different centres specialize in different types of analysis (for example, University of Southern California and John Hopkins University specialize in DNA methylation analysis).
3. Genome Sequencing Centers (GSCs) - these centres perform genome-wide DNA sequencing on data passed on by the BCR.

4. The Data Coordinating Center (DCC) at SRA International, Arlington, Virginia - this centre manages the collection, storage and distribution to the wider research community of the data produced by the GCCs and GSCs.

Figure 3, taken from the TCGA website, is a schematic of how the various parts of TCGA fit together.



TCGA Process

Beginning with the critical first step of patient participation, TCGA involves a broad cross-section of the cancer research community, including basic researchers, clinical researchers, bioethicists, doctors, nurses, cancer patients, and advocates.

TCGA is a network of four components, each critical to the success of the project:

- Biospecimen Core Resource
- Cancer Genome Characterization Centers
- Genome Sequencing Centers
- Data Management, Bioinformatics, and Computational Analysis, including a TCGA Data Coordinating Center

Technology development will play a vital role in advancing the efforts of each of the four TCGA components.

Biospecimen Core Resource
Tissue samples will be carefully catalogued, processed, checked for quality, and stored, complete with important medical information about the patient. Samples will be coded to remove any descriptors that may connect a sample with the patient's private information.

Genome Sequencing Centers
High-throughput genome sequencing centers will identify the changes in the DNA sequence that are associated with specific types of cancer.

Technology Development

Cancer Genome Characterization Centers
Several technologies will be used to analyze the genetic changes involved in cancer. The genetic changes that are identified will be further studied by the genome sequencing centers.

Data Management, Bioinformatics, and Computational Analysis
The information that is generated by TCGA will be centrally managed at the Data Coordinating Center and entered into public databases as it becomes available.

The specific steps in TCGA process are:

1. Patient Volunteerism
   Eligible cancer patients are asked to donate tissue and/or blood samples.
2. Biospecimen Distribution
   A centralized facility processes tissue samples and then sends genetic materials to TCGA research centers.
3. Research and Discovery
   TCGA components work together to identify genomic changes associated with cancer.
4. Data Sharing
   Genomic information generated by TCGA is made widely available through public databases. Researchers around the world can use the information to speed advancements in cancer biology and related technologies.
5. Goals
   Advances made by researchers using TCGA data pave the way for targeted strategies for diagnosing, treating, and preventing cancer.

Figure 3 - overview of TCGA process

In addition to the facilities mentioned above, TCGA has also established Genome Data Analysis Centers (GDACs) whose role is to develop new bioinformatic tools for processing and analysing TCGA data.

1.4 DNA methylation

DNA methylation is a form of epigenetic modification of genomic DNA, which means that it does not actually change the sequence of DNA bases themselves, but that it can still be passed on to future generations of cells via mitosis (and also potentially future generations of the host organism).

The extent of methylation and mechanism by which it occurs vary significantly between different organisms[32]. In mammals, methylation most commonly, though not exclusively, affects cytosine/guanine di-nucleotides (referred to as "CpG sites"), whereby the cytosine half of the pair is "tagged" with a methyl group which replaces one of the hydrogen atoms.

The distribution of CpG sites across the human genome is not uniform. The genome is generally depleted of CpG sites, there being approximately 28 million CpG sites out of a total of around 3 billion bases. This depletion is believed to be caused by the fact that methylated cytosines can be converted naturally to thymines by deamination. There are, however, concentrated areas (typically around 1,000 bases in length) of relatively high CpG density which are known as CpG islands, and these are present in over half of human genes[33].

Across all cell types in the human body, the overall profile of methylation throughout the whole genome is bimodal, with well over half of CpG sites being highly methylated whereas CpG islands have typically low levels of methylation[34]. It should be pointed out, however, that any such measurement of methylation level is an average across cells, since for each individual CpG site on each individual piece of DNA, methylation is essentially a binary state - either the CpG site is unmethylated or it is methylated.

Methylation of genomic DNA is typically carried out by a family of enzymes known as methyltransferases. Two of these, known as DNMT3A and DNMT3B, are *de novo* methyltransferases which can tag previously unmethylated DNA, whilst DNMT1 is a maintenance methyltransferase which facilitates the copying of methylation status during cell division by recreating the methyl group on the replicated strand.

Unlike genomic DNA, which is mostly the same in all cells within the human body, the methylation profile of a cell can vary considerably, both across different cell types and also over time. This means that the influence of methylation also varies between cell types and over time. In particular, variation in methylation profiles between different cell types reflects, in part, different gene expression profiles between cell types (see below).

Recent research[34] suggests that in early embryonic development there is wide-scale demethylation across the entire human genome, with the exception of certain DNA sequences such as those involved with genomic imprinting. There then follows, at around the time of implantation, a reestablishment of global methylation patterns, and these are then maintained throughout future cell division. Again there are some exceptions, such as CpG islands, which are believed to be protected from this genome-wide methylation. Furthermore, methylation states after implantation are not entirely static, and can be influenced by environmental factors (such as diet and alcohol consumption) and / or sporadic genetic mutations[35]. Increasing methylation levels are also believed to occur as a result of the aging process.

Methylation is thought to have a number of functions[33]. For example, it is important for chromosomal stability when it occurs in repeat DNA regions such as centromeres[36], and aids genomic stability by repressing other repeat regions such as transposable elements. Its main function, however, concerns the pattern of tissue-specific gene expression patterns.

There has been much research in recent years investigating the impact of methylation on gene expression patterns. It is currently believed that the influence which methylation exerts on transcription of any particular gene depends on both its location relative to the gene and also on whether it occurs within a CpG island or not[33].

As mentioned above, CpG islands have typically low levels of methylation. Hence when a gene which contains a CpG island near its transcriptional start site (TSS) is repressed, this is generally by means of some cause other than methylation - for example by interaction of the gene's promoter with the so-called Polycomb complex of proteins. This form of repression allows some flexibility so that genes can be expressed in a tissue-specific manner as and when necessary[33].

There are, however, some CpG islands associated with TSSs which are highly methylated. These areas tend to occur in genes which are subject to long-term, stable repression, such as imprinted genes and inactivated X-chromosome genes in females. In this scenario, research has linked the methylation of CpG islands to the inability of transcription to be initiated[37]. However, the current line of thinking is that for these genes, methylation does not directly cause repression of transcription, but rather that it is a mechanism for stabilising repression which has already been brought about through another mechanism[33]. In this way, the flexibility for the affected genes to be expressed at certain times / places is lost and instead permanent silencing of the genes is achieved.

The situation is different for TSSs which do not contain CpG islands. In this scenario, the level of CpG methylation is much more variable and tissue-specific, and there appears to be an inverse correlation between methylation and gene expression levels[38]. Again it is believed that methylation is not the initial cause of transcription repression, although there is no definitive evidence for this yet.

The considerations for methylation in gene bodies are different again. Most gene bodies have low CpG density and are highly methylated. Seemingly paradoxical to the inverse correlation noted above for TSSs, methylation levels in gene bodies have been found to be positively correlated to gene expression[39]. One theory for this is that methylation in TSSs blocks the initiation of transcription, but when it occurs in gene bodies it does not prevent the propagation of transcription that has already started upstream of the gene body.

A further potential complication to the above analysis is that most genes have more than one TSS, which means that at least one downstream TSS will effectively be in the body of the transcriptional region of another upstream TSS. Hence, if the measurement of gene expression level is not calculated specifically to each TSS (as is often the case), then direct linkage of the methylation level of the downstream TSS to the gene expression level could lead to incorrect conclusions about the association between methylation and expression being drawn[33].

Finally, genome-wide methylation profiling research has observed that exons tend to have higher levels of methylation than introns. Furthermore, the switch from high to low methylation appears to occur around exon / intron boundaries. One potential implication of this is that methylation may influence the locations where gene splicing takes place[40].

1.5 DNA methylation and cancer

In recent years there has been a significant amount of scientific research into links between DNA methylation and cancer, and this has shown that abnormal levels of methylation, both low and high, play an important part in cancer progression[7]. This involvement is not surprising given methylation's role in the control of transcription repression, as outlined in the previous section. However, research has also suggested that abnormal methylation does not necessarily act alone in tumorigenesis, but rather that it is coordinated with both genetic alterations and also potentially with other epigenetic changes such as histone modifications and nucleosome remodelling[7].

Two key types of aberrant patterns of DNA methylation have been consistently reported in various types of cancer[34]. Firstly, increased (hyper) methylation of large numbers of CpG islands positioned in gene promoters has been identified in cancerous tissue relative to normal tissue. Secondly, associations have also been made to genome-wide reduced (hypo) methylation in non-CpG islands relative to non-cancerous tissue. Each of these phenomena is considered in turn below.

1.5.1 Cancer-related hyper-methylation

The first of these effects has been the most extensively researched to date, given that methylation of CpG island promoters has been shown to repress gene transcription permanently[41].

Normally, permanent repression of this nature only occurs in imprinted genes, inactivated X-chromosome genes and germ-cell specific genes[41]. So when abnormal hyper-methylation of other genes occurs, this could potentially lead to the permanent repression of genes whose controlled regulation via other mechanisms is essential for the normal functioning of cells. Tumour-suppressor genes are an obvious potential example of this.

Hyper-methylation has been observed to occur in hundreds (typically between 5% and 10%) of CpG island promoters within the same tumour[41], and this has given rise to the term CpG island methylator phenotype (CIMP), to define a subset of CpG island promoters which are all simultaneously hyper-methylated in cancer[11]. Whilst this is merely a hypothesis at the present time, there is some evidence suggesting that it may be a real effect[42,43].

How and when such hyper-methylation is involved in cancer progression is not entirely clear yet[34]. With regards to timing, there is evidence that hyper-methylation may be an early event in cancer progression. For example, in studies of colon cancer, hyper-methylation has been observed to already occur in polyps[44].

One model for how hyper-methylation may play an early part in tumorigenesis[45] hypothesises that increased methylation (for example caused by aging), may affect critical "epigenetic gatekeeper" genes which, under normal conditions, are responsible for the regulation and differentiation of stem cells. Permanent silencing of such genes causes abnormal growth of these cells and stops their differentiation. This in turn provides time for genetic mutations to occur in genes which are involved in critical developmental pathways (such as the Wnt pathway in colon cancer), leading to increased tumour progression.

A justification for this hypothesis is based on the idea of cancer stem cells being key initiators of cancer formation[7]. In normal stem cells, long-term repression of genes is controlled by the Polycomb complex of proteins. Many genes involved in the CIMP have also been found to be regulated by this complex[46]. There is evidence that genes regulated by Polycomb proteins are targeted by DNA methylation[47]. So increased methylation (e.g. as a result of aging) may be targeted to these stem cell genes, transforming the cells into cancer stem cells which then drive tumour formation.

On the other hand, hyper-methylation may also be a downstream consequence of other genetic mutations. For example, the TET group of proteins are known to remove methylation marks[48]. Hence a mutation which causes loss of function in one of these enzymes could lead to aberrantly high methylation, though this is yet to be proved.

Another way in which hyper-methylation could promote tumour progression is by repressing microRNAs[41]. Such repression could cause increased transcription of the miRNA's target gene. If this gene were an oncogene (e.g. BCL-6)[49] then this could promote tumour growth.

1.5.2 Cancer-related hypo-methylation

As mentioned above, at the same time as the hyper-methylation of CpG island promoters is occurring in cancer, the opposite effect of hypo-methylation has also been observed to occur in non-CpG islands. Whilst research has generally concentrated on the former, characteristics of these hypo-methylated regions are now also being elucidated[41].

Although this cancer-related hypo-methylation occurs on a genome-wide basis, recent research suggests that, rather than being randomly distributed, it tends to be concentrated in large, megabase stretches of DNA which in normal conditions are highly methylated[41]. These cancer-related regions of hypo-methylation are broken up by much smaller regions of hyper-methylation, typically affecting CpG islands[50]. A key feature of these parts of the genome is that they appear to be associated with the nuclear envelope lamina[51], and this has two potential implications.

Firstly, DNA associated with the nuclear envelope lamina is subject to late replication. This potentially provides an explanation for the regions of hypo-methylation, in that the DNA maintenance methyltransferase DNMT1 may be overwhelmed by the extra methylation that has occurred in the hyper-methylated CpG islands. Hence there may not be sufficient DNMT1 to fully maintain the methylation of the other, normally highly methylated areas, which therefore experience passive demethylation via cell division and thereby become hypo-methylated in cancer[34].

Secondly, in stem cells the nuclear envelope lamina region contains the majority of genes which are potentially subject to hyper-methylation in cancer[41]. Hence this gives further weight to the cancer stem cell theory mentioned above.

It is not clear yet whether this wide-scale hypo-methylation has any causal effect on cancer, or whether it is simply a passive side-effect. However, there is some evidence that demethylation may be an important cause in some cancers (for example in some leukemias)[52].

1.6 Difficulties with genome-wide methylation analyses

There are a number of technical difficulties in connection with the profiling of genome-wide methylation levels[53]. Firstly, DNA samples usually contain a mixture of cell types, each of which may have their own different methylation profiles. Hence, any measurement of methylation levels will be an average over all the different cells present in the sample.

Of course the cell-type mixture is a problem for many other analytical techniques (eg gene expression analysis using microarrays). However, there are a couple of other issues which are specific to methylation profiling. Firstly, methylated cytosines are frequently mutated to thymine naturally, but this depletion does not occur at a uniform rate across the whole genome. Hence, there is a non-uniform distribution of methylated cytosines, which can have implications for the way in which experimental assays for detecting methylation are designed, especially if a genome-wide analysis is sought.

Secondly, methylated and unmethylated cytosines cannot be distinguished by methods which rely on DNA hybridisation (such as microarrays). Furthermore, methylation tags are erased when DNA is amplified by PCR. Hence experimental techniques are required to pre-process the DNA so that there is no subsequent erasure of information and / or so that hybridisation methods will work. Three key techniques, endonuclease digestion, methylated DNA immunoprecipitation, and bisulphite conversion have been devised for this purpose[53,54].

TCGA's methylation data used in this project are based on bisulphite conversion. This method makes use of the fact that when denatured DNA is treated with sodium bisulphite, unmethylated cytosines are converted to uracil much more rapidly than methylated cytosines[55]. Hence, a difference in methylation status can effectively be converted to a genetic difference, which is then amenable to detection by techniques such as microarray hybridisation or next generation sequencing. Figure 4 is a schematic of the bisulphite conversion process taken from Illumina's "Epigenetics" data-sheet.

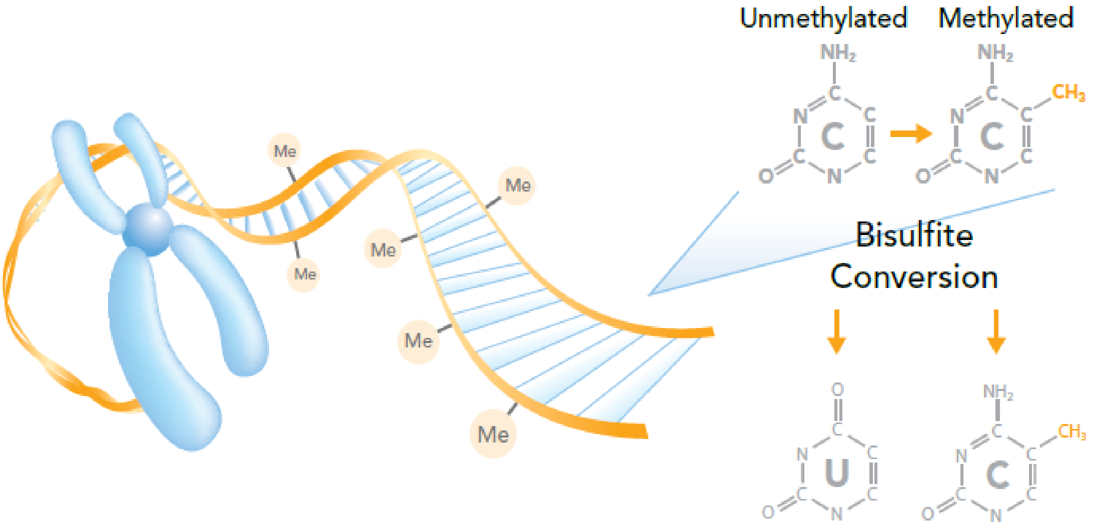


Figure 4 - schematic of bisulphite conversion process
Schematic diagram of the bisulphite conversion process (taken from Illumina's "Epigenetics" data-sheet). Unmethylated cytosines are converted to uracils, whereas methylated cytosines are not converted.

However, the reduced complexity of the genome after conversion (effectively from four bases to three, except for methylated cytosines) means that careful consideration has to be given to the design of hybridisation probes or methods for sequence alignment to allow for the reduced specificity[53].

The most widely used microarray technology for methylation profiling after bisulphite conversion is Illumina's HumanMethylation Beadchip[54]. This provides genome-wide methylation profiling at individual CpG site resolution. However, as with all microarray technologies, it suffers from the fact that the genome coverage is restricted to the actual CpG sites queried by the array - the more modern 450k version queries around 485,000 sites, whereas there are estimated to be around 28 million CpG sites in total throughout the human genome.

1.7 TCGA methylation data

TCGA use two different, but related, technology platforms to analyse the methylation levels of their samples. Both are produced by Illumina Inc. The older technology, which was in use until 2012, is the HumanMethylation27 BeadChip ("27k platform")[56]. This was superseded by the HumanMethylation450 BeadChip ("450k platform")[57], which is the technology now being used. Both use microarray technology applied to DNA which has been bisulphite converted, as described in the previous section. After bisulphite conversion, the DNA is amplified using PCR, fragmented and then hybridised to specially designed oligonucleotides on the platform[53].

1.7.1 27k platform

The 27k platform is the older technology which was introduced by Illumina in 2009[56]. It is capable of measuring genome-wide methylation levels of 27,578 different CpG sites for up to 12 samples simultaneously. The CpG sites which are interrogated were chosen, primarily, because they fall within the proximal promoter regions of genes. 14,475 genes are covered (by an average of 2 sites per gene), including over 200 cancer-related or imprinted genes. 254 sites within the promoters of 110 miRNAs are also included. Where possible, CpG sites which fall within so-called CpG islands[58] were selected.

The 27k platform uses Illumina's proprietary Infinium I technology to measure the methylation states of individual CpG sites. The array is designed such that for each targeted CpG site there are two sets of "beads" which are used to hybridise with the sample DNA. One set is designed to hybridise with unmethylated cytosines, whereas the other is designed for methylated cytosines. Both types of hybridisation are recorded by the incorporation of a fluorescently labelled nucleotide of the same colour (which can be red or green depending on the based immediately before the interrogated cytosine). The methylation level for any particular CpG site is calculated as the ratio of the total number of hybridisations with methylated beads to the total number of hybridisations with both types of bead.

1.7.2 450k platform

The 450k platform[57] is an updated version of the 27k platform, and therefore has many features in common with its predecessor. Again, up to 12 samples can be processed on one array. However, the 450k platform has much greater coverage of the genome, as it is capable of simultaneously measuring the methylation levels of 485,577 different CpG sites. There is a much more varied distribution of CpG sites across promoters / non-promoters and CpG islands / non-CpG islands. 99% of RefSeq genes are covered, at an average of 17 sites per gene, as are 96% of known CpG islands. Regions adjacent to CpG islands are also extensively covered.

The other key difference relative to the 27k platform is that the 450k platform uses a combination of the Infinium I technology and Illumina's more recently developed Infinium II technology. Whereas the former uses two beads for each CpG site, the latter uses only one bead. It works by single-base extension, whereby the colour of the incorporated, fluorescently labelled nucleotide is dependent on the methylation state of the interrogated CpG site. For an unmethylated site, a red-labelled adenine base, complementary to the bisulphite-converted uracil, will be incorporated, whereas for a methylated site a green-labelled guanine will be incorporated. The ratio of green fluorescence to total fluorescence detected then determines the methylation level.

72% of the CpG sites on the 450k platform are interrogated using the Infinium II technology, with the other 28% being interrogated using the Infinium I technology. A diagram of how the Infinium techniques work, taken from the second Bibikova paper[57], is included in figure 5.

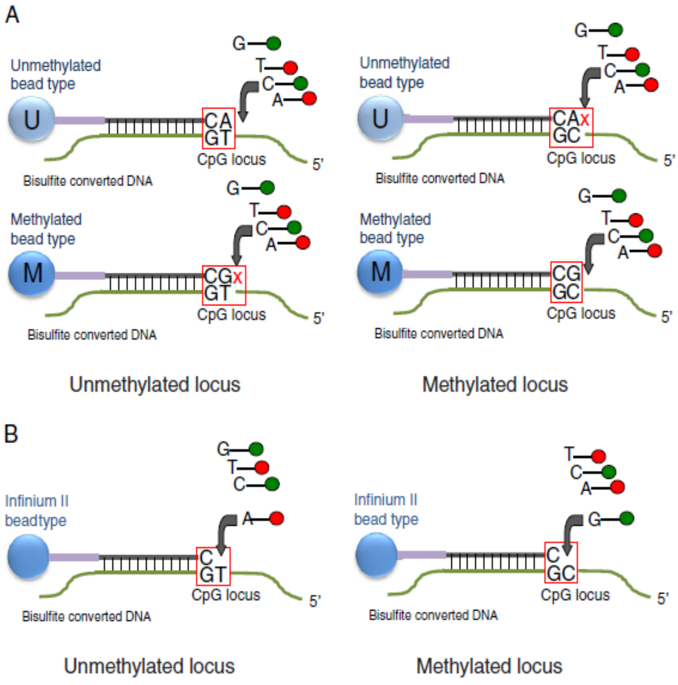


Figure 5 - diagram of (A) Infinium I and (B) Infinium II technology

Schematic diagrams of (A) Infinium I and (B) Infinium II technologies (taken from Bibikova[57]). In (A) there are two separate probes, one for unmethylated cytosines and the other for methylated cytosines. Bisulphite-converted DNA will hybridise to only one of these depending on the methylation status of the target CpG site. Both types of hybridisation are recorded by the incorporation of a fluorescently labelled nucleotide of the same colour (which can be red or green depending on the base immediately before the interrogated cytosine). In (B) there is only one probe for each target CpG site. The colour of the incorporated, fluorescently labelled nucleotide is dependent on the methylation state of the interrogated CpG site. For an unmethylated site, a red-labelled adenine base, complementary to the bisulphite-converted uracil, will be incorporated, whereas for a methylated site a green-labelled guanine will be incorporated.

Out of the 27,578 CpG sites interrogated by the 27k platform, 25,978 are also interrogated by the 450k platform. Of these 25,978, 23,663 are measured using the Infinium II technology. Illumina claim that the correlation between platforms for the 25,978 common sites is greater than 95%[57].

1.7.3 Annotation of probes

Illumina provide, for both platforms, annotation details of the interrogated CpG sites, which can be used for subsequent analysis of the methylation measurements. These details include items such as chromosome, starting position on chromosome and UCSC reference gene to which a site is associated.

Also included are two separate labels, one of which indicates each site's position relative to the nearest gene ("gene location"), and the other its position relative to the nearest CpG island[36] ("CpG island region").

The classifications used to annotate the gene locations are as follows:

- "Body" - the region between 3'UTR and the 1st exon
- 1st exon - the first transcribed exon
- 3'UTR - the untranslated region at the 3' end
- 5'UTR - the untranslated region at the 5' end
- TSS1500 - the region between 200 and 1,500 bases upstream of the transcriptional start site
- TSS200 - the region between the transcriptional start site and 200 bases upstream of this
- other sites which do not fall into any of the above categories are unlabelled.

The classifications used to annotate the CpG island regions are as follows:

- Island[58] - a region of at least 500bp with more than 50% GC composition and observed / expected ratio of CpG dinucleotides of at least 60%
- North shore[59] - 2kb region immediately upstream of an island
- South shore - 2kb region immediately downstream of an island
- North shelf - 2kb region immediately upstream of a north shore
- South shelf - 2kb region immediately downstream of a south shore
- other sites which do not fall into any of the above categories are unlabelled.

A schematic representation of both of these sets of labels, taken from the second Bibikova paper[57], is shown in figure 6:

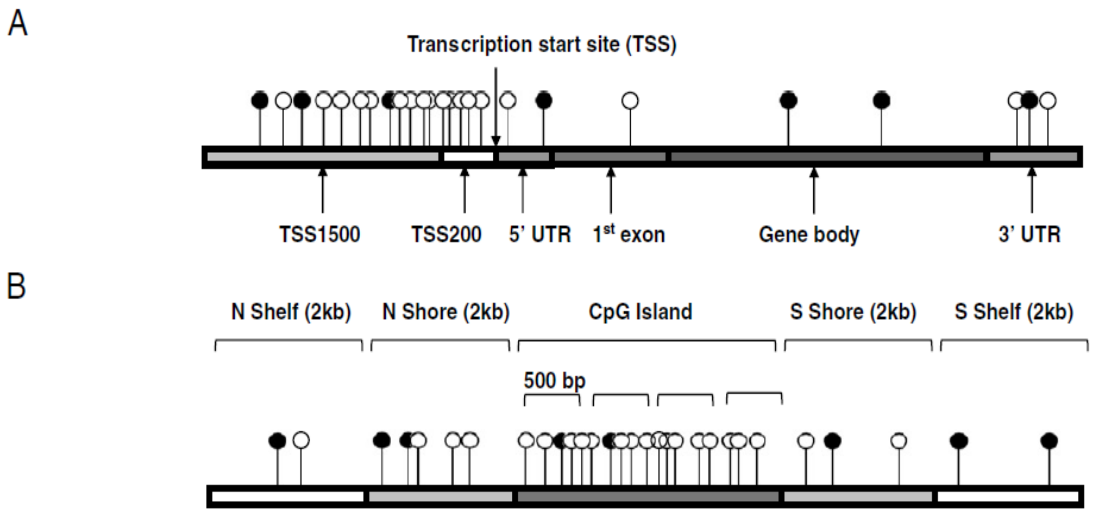


Figure 6 - schematic of (A) gene locations and (B) CpG island regions
Schematic diagrams of (A) gene location and (B) CpG island region annotation details provided by Illumina for each targeted CpG site (taken from Bibikova[57]). (A) shows the different types of gene location dependent on where the target CpG site is located relative to the nearest gene. (B) shows the different island regions dependent on where the target CpG site is located relative to the nearest CpG island.

1.7.4 TCGA level 1 data

The primary outputs from the Illumina platforms are the fluorescent signals emitted by the incorporated nucleotides and detected by the imaging hardware. The initial processing of the images to produce raw intensity values is usually performed using the BeadScan software provided by Illumina themselves. The resultant intensity values are stored in special .idat files. These can then be input into Illumina's own GenomeStudio software for further processing, or other packages can be used for this purposes – Minfi[60] and Methylumi[61], both available in R[62], being two common examples. The .idat files produced from TCGA's application of the Illumina platforms to their samples are freely available for download in their level 1 DNA methylation data.

1.7.5 TCGA level 2 data

The raw intensity values stored in the .idat files potentially need some degree of pre-processing before they can be used to calculate methylation ("beta") values[63]. In particular, intensity values may be inflated, to varying degrees, by "background" levels of fluorescence which are in no way related to differences in the biological samples being analysed. Also, due to imperfections in the imaging hardware, the way and extent to which intensity values have been detected may vary between samples for reasons other than genuine biological differences. Hence there may be a need to "normalise" measurements between samples.

Neither of these "correction" processes is straightforward, and several approaches have been proposed for each [64-68]. The key issue is to try and remove as much unwanted variation between samples as possible, whilst at the same time retaining as much genuine biological variation as possible. In their level 2 methylation data, TCGA provide separate intensity values for methylated and unmethylated probes which have been adjusted for both background correction and normalisation using TCGA’s preferred methods.

For background correction, TCGA use a method called normal exponential convolution[65], which makes use of negative control probes on the Illumina platform. These probes are not intended to hybridise to the DNA samples, and so their purpose is to capture any background signal "noise" which is independent of the samples used. The method assumes that the signal for each probe can be considered to be the sum of the actual signal and the background noise, where the actual signal is assumed to follow an exponential distribution and the background signal is assumed to be normally distributed. The parameters for these distributions are then determined empirically from the data.

TCGA also undertake a basic level of normalisation whereby samples in different “batches” have their intensities in each of the two channels (green or red) multiplicatively scaled to match a reference sample (the sample with red / green ratio closest to 1.0.)

1.7.6 TCGA level 3 data

TCGA’s level 3 methylation data contain, for each sample and probe, the estimated beta value calculated as the ratio of the level 2 methylated intensity to the sum of the level 2 methylated and unmethylated intensities. However, one further adjustment is made, because there may be some probes which have failed to work properly or which may need to be filtered out because they are prone to cross-hybridisation. Again the extent to which this adjustment is made is a matter of judgement, rather than there being any hard and fast rule.

Probes which have failed to work properly, for whatever reason, can be identified through the so-called "detection p-values" which are produced by the Illumina software for each probe for every sample. These values are calculated using the negative control probes on the Illumina platforms, and represent the probability for each given sample that the probe in question has registered a genuine signal rather than simply recording some background noise. The lower the detection p-value, the less likely it is that the probe has simply made a background reading (i.e. it is more likely that the probe has worked properly).

The probabilities are calculated based on the distribution of actual measurements recorded by the negative control probes. For any given probability threshold, the user can decide to exclude any probes which have detection p-values greater than the threshold for a specified proportion of samples.

The TCGA level 2 data include the detection p-values as a separate item for each probe on each sample, and the level 3 beta value data are replaced by "NA" where the p-value has exceeded 5%. The end user can then decide what proportion of samples with "NA" needs to be reached before any individual probe is filtered out across all samples for this reason.

Incorrect hybridisation is a problem for certain subsets of probes. Probes which have common genetic variants in their sequences or which overlap with repetitive stretches of DNA are particularly prone to this[63]. So users need to be aware of this and potentially filter out affected probes. In their level 3 beta value data, TCGA have replaced the calculated beta values with "NA" for any probes whose sequences either:

1. contain a Single Nucleotide Polymorphism within 10 base pairs of the interrogated CpG site (the dbSNP database is used for this purpose); or
2. overlap with a repetitive DNA element within 15 base pairs of the interrogated site (based on UCSC hg19 human reference genome).

1.7.7 Data used for project

The Y chromosome is the least represented chromosome on both Illumina platforms. The 27k platform contains only seven probes which interrogate CpG sites on the Y chromosome. However, the 450k platform contains 416 probes covering the vast majority of known Y chromosome genes in the male-specific region. Given the much greater coverage, my project has, therefore, been based on data obtained using the more modern platform.

As indicated above, TCGA's level 3 data already excludes beta values for certain subsets of probes. I preferred to work with the full probe-set, and make my own decisions on which probes to exclude. I, therefore, decided to use the background-corrected and normalised intensity values in TCGA's level 2 data, from which I could derive beta values. One further advantage of this approach is that the total (methylated plus unmethylated) intensity values can also be used to estimate copy number variations[14].

1.7.8 Other processing issues

For the 450k platform there is a further complication not allowed for by TCGA, namely that two different technologies (Infinium I and Infinium II) are used for different subsets of probes as described previously. Research has shown that the distribution of methylation values for the two subsets are different, in particular that the Infinium II probes have a smaller range of values than the Infinium I probes even when differences in the characteristics of the targeted probes are taken into account[69]. Hence this difference could lead to further distortions in any subsequent analyses, as Infinium I probes may appear to be have higher levels of differential methylation relative to Infinium II probes simply because of their greater dynamic range.

A number of techniques have been proposed to adjust for this imbalance between the two Infinium technologies[69-74]. However, it is not clear which , if any, is the most suitable, nor indeed whether further normalisation (on top of the background and normalisation correction already undertaken by TCGA) is required at all. A large-scale benchmarking study would be required to answer this, and would need an objective, independent method (for example pyrosequencing) to compare results and determine the most appropriate technique. In the absence of such a study, the final answer is a judgement, which can be assisted, for example, by investigating any major inconsistencies in results for probes produced using the different technologies[63].

Even after corrections for background effects and normalisation have been made, and unreliable probes have been filtered out, there is yet one more potential source of inaccuracy in the data. These are generally known as "batch" effects, and relate to the fact that different groups (or batches) of samples which have been processed at different times and / or by different researchers may have been subject to slightly different experimental conditions[63]. It is possible that such batch effects may introduce systematic bias in any downstream analysis if they are not corrected for.

The existence of batch effects can often be identified by using unsupervised clustering of samples (for example using Principal Components Analysis) to discover whether samples group together according to their methylation profiles. If such clustering is observed, then some form of correction may be necessary. Batch effects have been shown to be a potential problem for methylation data generated on the Illumina platforms[75]. Furthermore, batch effects are potentially an even greater problem for total intensity values.

There are two ways, in particular, in which TCGA samples have been grouped which could cause batch effects. Firstly, different groups of samples have been collected by different institutions (known as "tissue source sites"). Secondly, TCGA groups samples into "batches" when handling them. Both of these groupings could potentially produce unwanted batch effects due to any unintentional differences in the way different groups of samples are handled (even though the overall handling procedures are in theory the same). However, both of these pieces of information are available to download for all TCGA samples, so the existence of any batch effects can be investigated and corrected if necessary.

1.8 Other data types used

My project has been based primarily on TCGA's methylation data, both for methylation analyses and also copy number estimation. However, I have also used three other biological datasets from TCGA to supplement my analyses:

1.8.1 Copy number data

I used TCGA's level 3 copy number data to act as a check on my estimates based on the methylation data and for some subsequent analyses. The copy number data had been produced using the Affymetrix Genome-Wide Human SNP Array 6.0[76]. The raw values had then been tangent-normalized[77] and segmented using Circular Binary Segmentation[78].

1.8.2 Gene expression data

To analyse the transcriptomic effects of aberrant methylation and copy number variations, I then used TCGA's level 3 RNA sequencing data, which provides a measure of gene expression. Raw read data had been produced using the Illumina HiSeq 2000 RNA Sequencing platform[79] and aligned to the reference human genome using MapSplice[80]. The RSEM algorithm[81] had then been used to estimate expression levels from the raw read data, and these estimates had been normalised between samples at the gene-level by setting the upper quartile value to 1,000.

1.8.3 Somatic mutation data

Finally, for the HNSC dataset only, I also made use of TCGA's level 2 somatic mutation data. In particular I used these data to identify those tumour samples which harboured a mutation of the TP53 gene. Raw sequencing data had been produced using the Illumina HiSeq200 platform and aligned to the reference genome using the Burrows-Wheeler alignment algorithm[82]. Mutations had then been identified using the MuTect algorithm[83].

1.9 Cancer types investigated

As mentioned previously, my analyses have been based on three different cancer types. Brief overviews of key characteristics of each of these are set out below:

1.9.1 Colon adenocarcinoma

Colon cancer is the third most common cancer worldwide, with over one million new cases reported every year. It is more prevalent in men than in women, and incidence is strongly age-related. Around 5% of cases are believed to be inherited, but the vast majority are sporadic[84].

As well as a family history of cancer, there are several other risk factors, including inflammatory bowel disease, smoking, alcohol consumption, eating red meat, obesity and diabetes. Treatment is usually via a mixture of surgery, radiotherapy and chemotherapy. Survival rates have increased gradually over time, with five year survival now averaging around 65%[84].

Colon cancer is known for its molecular heterogeneity. Three main molecular features have been observed, involving micro-satellite instability (MIN), chromosomal instability (CIN) and CpG island methylator phenotype (CIMP) [85]. These features are not mutually exclusive. Around 70% of cases involve an early mutation in the APC tumour suppressor gene. Other frequently mutated genes include the KRAS and BRAF oncogenes and the TP53 tumour suppressor gene.

Aberrant methylation has been extensively studied in colon cancer, which was the first cancer type in which the CIMP was observed[11]. CIMP tumours typically also include BRAF mutations and high micro-satellite instability[86]. Promoter methylation of the MLH1 tumour suppressor gene is also a common characteristic.

Very little research has been reported on the effects of Y chromosome aberration in colon cancer. However, there is evidence that LoY is a common feature in male colon cancers[26].

1.9.2 Head and neck squamous cell carcinoma

Head and neck cancer, of which squamous cell carcinoma is by far the most common subtype, is the sixth most common cancer worldwide[1] with around 600,000 new cases each year[87]. It is more prevalent in men than in women. It is sub-categorised by location into five sub-regions: oral cavity, larynx, hypopharynx, oropharynx and nasopharynx[88]. It is treated with a mixture of surgery, radiotherapy and chemotherapy, but its high level of molecular heterogeneity makes prediction of outcome very difficult, with five year survival rates of around only 50%[13].

Mutational inactivation of the TP53 tumour suppressor gene is a common feature of head and neck cancer, and this often results in general chromosomal instability[13]. The two most common risk factors are smoking and alcohol consumption, and research has indicated that these may act synergistically in the carcinogenic process[89]. A further, important risk factor is infection with the human papillomavirus (HPV), which is believed to inactivate both TP53 and Rb tumour suppressor genes. HPV induced cancers represent around 25% of new cases, and are particularly prevalent in the orophayrnx[13]. It is believed that their aetiology is different to other head and neck cancers[90], and that they may have a better prognosis[91].

Again there has been very little research into the effects of Y chromosome aberrations. However, LoY has been observed in high proportions of male head and neck cancers, and has been reported as possibly being linked to impaired survival[27,28].

1.9.3 Kidney renal clear cell carcinoma

There are around 300,000 new cases of kidney cancer diagnosed worldwide each year, and around 75% of these are clear-cell carcinomas[92]. The prevalence in males is roughly 1.5-2 times higher than in females, and key risk factors are smoking, hypertension and obesity[93]. Surgery is the main form of treatment, as kidney tumours are notoriously resistant to chemotherapy and radiotherapy. However, some do respond to angiogenesis inhibitors[94].

A small proportion of cases are hereditary and associated with von Hippel Lindau disease, involving mutation of the VHL gene on the short arm of chromosome 3. VHL is also inactivated in around 90% of sporadic tumours, either by mutation or hyper-methylation[92].

There is considerable mutational heterogeneity within tumours[95]. Loss of chromosome 3p is also commonly observed. Other genes on 3p are also inactivated, either by mutation or hyper-methylation, including PBRM1 and RASSFIA[95,96].

Once again there has been very little research into Y chromosome aberrations and kidney cancer, but LoY has been observed to occur in around 55% of male cases[25].

2. Materials and methods

My project considers the following changes in the male-specific region of the Y chromosome, and their impacts in male cancers:

1. Aberrant DNA methylation
2. Loss of chromosome
3. Correlation of 1 and 2 with changes in gene expression
4. Impact of 1 and 2 on survival
5. Association of 2 with key risk factors for head and neck tumours

This chapter sets out details of the data, software and methods used for the analyses described in subsequent chapters.

2.1 Data used

My project is based entirely on TCGA data which can be freely downloaded via their data portal (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). I downloaded all data using the "Data matrix" option within TCGA's data portal.

2.2 Datasets used

As the initial focus of my project was on DNA methylation, I reviewed the available data to find some large methylation datasets which could be used for my analyses. I concentrated on data produced using Illumina's 450k platform, as the 27k platform contains only 7 probes for the Y chromosome. The 450k platform contains 416 Y-linked probes.

As previously mentioned I initially chose TCGA's kidney renal clear cell carcinoma (KIRC) dataset. I downloaded the KIRC methylation and clinical data on 30 January 2013.

I subsequently selected a second dataset so that I could compare results with those obtained from my analyses of the KIRC dataset. For this purpose, I chose the colon adenocarcinoma (COAD) dataset. I initially downloaded the COAD methylation and clinical data on 23 July 2013.

In order to create a consolidated set of analyses for both case studies based on the most up to date information available at the time, I downloaded the methylation and clinical data for both datasets again on 2 December 2013.

Finally, on 4 February 2015 I selected and downloaded methylation and clinical data for a third dataset. This time I used TCGA's head & neck squamous cell carcinoma (HNSC) dataset.

I used the methylation data for each of the three cancer types to:

1. analyse changes in DNA methylation in respect of probes targeted at the Y chromosome; and
2. perform initial analyses of copy number changes for the Y chromosome (in particular chromosomal loss).

Subsequent to the above analyses, I downloaded the following additional data for each cancer type to perform further analyses:

1. Level 3 copy number data to verify the results of 2 above and for further downstream analyses; and
2. Level 3 RNA-sequencing data to investigate the effects of aberrant methylation and loss of Y chromosome on expression of Y-linked genes

I also downloaded level 2 somatic mutation data for the HNSC dataset only, to investigate correlations between loss of Y chromosome and frequently occurring mutations (in particular in TP53).

I set out further details of the data which I downloaded in the following sections.

2.3 Methylation data

I downloaded all available level 2 and level 3 methylation data generated on the Illumina 450k platform for each of the three datasets.

Table 1 provides a summary of the numbers of samples for which I downloaded methylation data:

	COAD	**HNSC**	**KIRC**
Tumour samples	291	528	301
Normal samples	38	52	160
Total	329	580	461

Table 1 - summary of downloaded methylation data

All data are stored in individual text files for each sample, and were downloaded in compressed format. For each cancer type I merged the individual files into one large data file using R[62] for subsequent analyses.

On inspection of the data, I noticed that a small number of samples contained duplicate records. I therefore removed the extra copies – this resulted in 5 KIRC duplicates and 17 COAD duplicates being removed.

My analyses are based on the level 2 data which contain, for each sample, measurements of methylated intensity, unmethylated intensity and detection probability values for each probe. I derived beta values for each sample / probe as the ratio of methylated intensity to total intensity, the latter being the sum of methylated and unmethylated intensities.

I also downloaded annotation data for the 450k platform directly from Illumina's website.

2.4 Clinical data

TCGA also make freely available anonymised clinical information for patients. I downloaded all available clinical data in the "Biotab" format for each dataset at the same time as the corresponding methylation data were downloaded.

The clinical data are provided across several text files, and can be sub-divided into three broad categories:

1. General information – e.g. gender, age, year of diagnosis
2. General cancer-related information – e.g. stage, grade, histological type, treatment details
3. Cancer-type-specific information – e.g. anatomic subdivision (COAD), HPV status (HNSC), laterality (KIRC)

Unfortunately, no clinical data were available for 3 COAD, 17 HNSC and 11 KIRC samples, and so I have excluded these samples from all my analyses.

Table 2 sets out the final numbers of methylation data samples which I have used in my analyses, after removal of samples which were either duplicated or had no corresponding clinical information:

	COAD		HNSC		KIRC	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Tumour samples	148	122	375	138	189	96
Normal samples	21	18	38	12	106	54
Total	169	140	413	150	295	150

Table 2 - summary of final methylation data numbers

2.5 Copy number data

In order to check the Y chromosome copy number estimates derived from the methylation data, and also for further use in downstream analyses, I downloaded TCGA's copy number data for each dataset. For this purpose I used the level 3 data produced using the Affymetrix Genome-Wide Human SNP Array 6.0.

I downloaded the level 3 copy number data for COAD, HNSC and KIRC on 21 January 2015, 5 February 2015 and 3 December 2014 respectively. These data contain segmented copy number data for each chromosome (including Y).

For each dataset there were four separate text files for each sample which had been processed on the Affymetrix platform. Two of the four files contain results which had been produced using version hg18 of the human reference genome, with the other two containing results based on version hg19. The latter is consistent with the reference genome used by Illumina's 450k platform, and I, therefore, used these data for my analyses.

Within the hg19 data there were two separate files for each sample – one of these was labelled "nocnv" and contains only somatic copy number variations. The other file potentially also contains germline as well as somatic copy number variations. Ideally I would have used the “nocnv” data, but unfortunately I discovered, by inspection, that these files do not contain any results for the Y chromosome. I, therefore, decided to use the other data files. I performed comparisons of the results from the two types of file for other chromosomes, and found that they were highly consistent (data not shown). I was, therefore, confident that I could use my chosen files for my analyses.

In order to organise the data into a more helpful format, for each cancer type I merged all the results for all chromosomes and samples into one large array using R. I restricted myself to those samples for which 450k methylation data were also available, and also removed duplicate samples. My final merged data files contained 308 COAD samples (out of 309 on the 450k platform), 555 HNSC samples (out of 563) and 433 KIRC samples (out of 445).

2.6 Gene expression data

To analyse the association between aberrant methylation / copy number loss of the Y chromosome with gene expression, I downloaded TCGA’s RNA-sequencing data for each dataset. For this purpose I used the level 3 data produced using the Illumina HiSeq 2000 RNA Sequencing platform and TCGA’s version 2 pipeline.

I downloaded the level 3 RNA-sequencing data for COAD, HNSC and KIRC on 21 January 2015, 14 February 2015 and 28 November 2014, respectively.

For each dataset there were six separate files for each sample which had been processed on the Illumina HiSeq platform. For my analyses, I used the files labelled “rsem.genes.normalized_results”, which contain normalised expression levels at gene-level for over 20,000 genes, including those on the Y chromosome.

Once again, I merged all the results for all genes and samples into one large file using R, and restricted myself to those samples for which 450k data were also available. My final merged data files contained 271 COAD samples (out of 309 on the 450k platform), 542 HNSC samples (out of 563) and 308 KIRC samples (out of 445).

2.7 Somatic mutation data

The final data I downloaded from TCGA contained somatic mutations, for the HNSC dataset only. For this purpose I used TCGA's level 2 data produced using the IlluminaGA DNASeq platform. I downloaded these data on 29 March 2015. The output consisted of a database containing a list of all somatic mutations identified for each sample. The data included gene name and type of mutation.

2.8 Software used for analyses

All analyses were performed using a combination of Microsoft Excel and (mainly) programs written in R. During my analyses, I used a number of special R packages which can be freely downloaded. Details of these packages are set out in table 3 below:

Package	**Use**
BSgenome.Hsapiens.UCSC.hg19	Calculating CpG densities in chapter 3
DNAcopy	Segmentation of copy number data derived from methylation data in chapter 4
Scatterplot3d	Producing 3D scatterplots in chapter 5
Survival	Survival analyses in chapter 6

Table 3 - R packages used in analyses

2.9 Statistical analyses

I used a number of standard statistical methods in my analyses. In particular:

- Kruskal-Wallis test for comparing extent of loss of chromosomal arms between different sub-groups of male patients
- Kaplan-Meier (with log-rank test) and Cox Proportional Hazards methods for survival analyses

I used a significance level of 5% in all statistical tests.

3. Differential methylation of the Y chromosome

In this chapter I set out details of my analyses of differential methylation, between normal and tumour samples, of CpG sites on the Y chromosome.

An overview of the process I have followed is provided in figure 7:

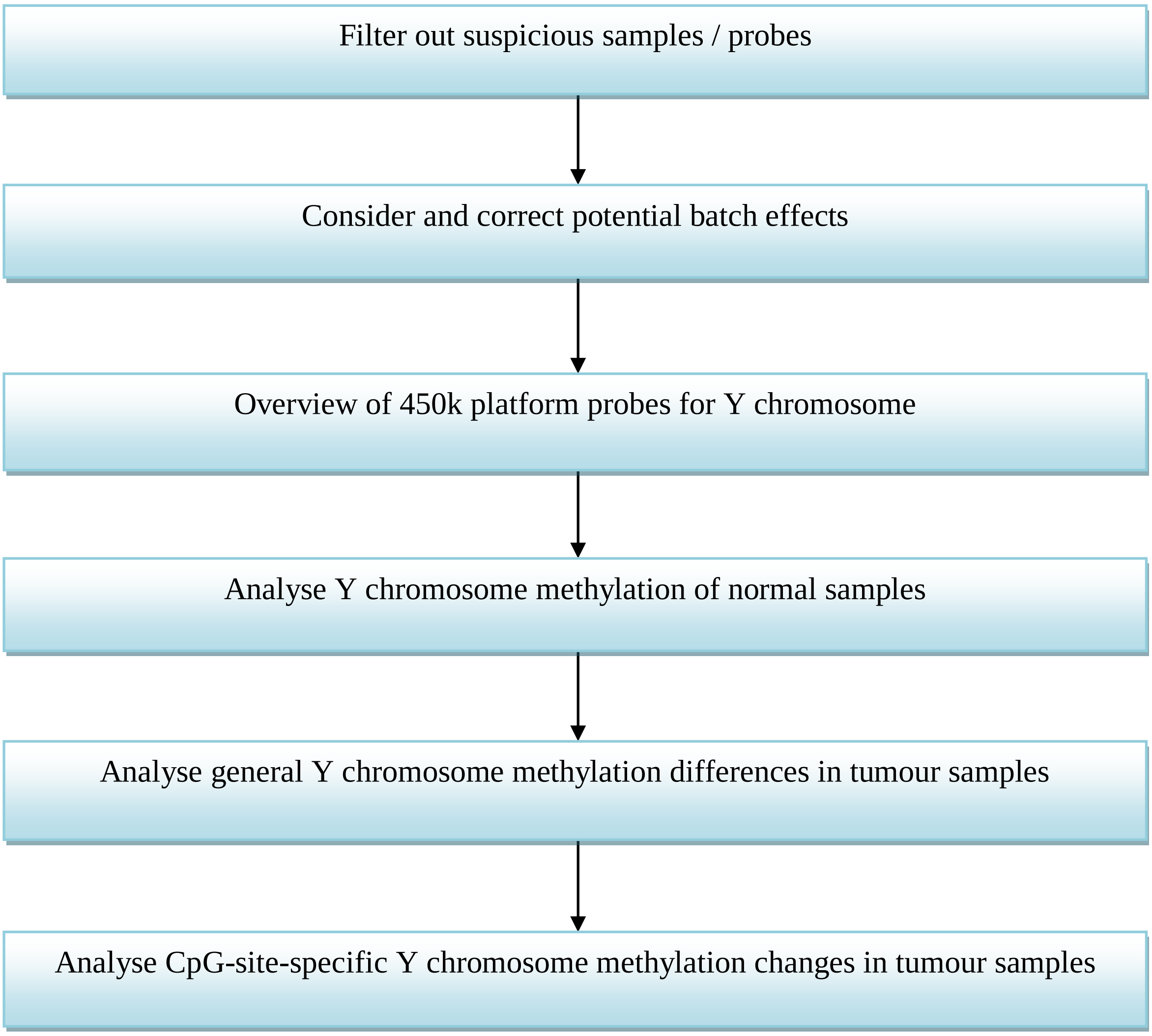

Filter out suspicious samples / probes
Consider and correct potential batch effects
Overview of 450k platform probes for Y chromosome
Analyse Y chromosome methylation of normal samples
Analyse general Y chromosome methylation differences in tumour samples
Analyse CpG-site-specific Y chromosome methylation changes in tumour samples

Figure 7 - overview of experimental process for differential methylation analyses

I set out details of each of the above stages in the following sections.

3.1 Filtering of samples and probes

My first consideration was to check whether there are any Y chromosome probes on the 450k platform which may be prone to cross-hybridisation with other parts of the genome and which, therefore, could provide misleading measurements. My plan was to identify any such probes and remove them from my analyses.

In order to perform this check, I decided to investigate the total (methylated plus unmethylated) probe intensities for the female samples – my reasoning being that since females do not possess a Y chromosome, their total intensities should be at a very low background level, and much lower than the corresponding values for male samples. Any probes for which female samples have high total intensities, comparable with male values, are likely to have cross-hybridised with other parts of the genome.

3.1.1 Filtering of samples

Whilst carrying out my analysis, I also observed that there were a number of samples, both male and female, whose total intensity measurements were consistent with someone of the opposite gender – i.e. it is likely that their gender had been misrecorded.

In total I identified nine female samples (six COAD, one HNSC and two KIRC – the latter being one patient for whom there was both a normal and tumour sample, the others all being tumour samples) whose Y chromosome probe intensities suggested that they were more likely to be male.

Although these samples were not important for the main parts of my analyses, I decided to remove them at this stage before checking the total intensities of individual probes.

I also observed six COAD, one HNSC and two KIRC male samples whose Y chromosome intensities were consistent with those of female samples (i.e. very low). It is possible that these males had suffered a complete loss of Y chromosome. However, their intensity measurements were so significantly lower than those of any other males, including patients who appeared to have suffered significant Y chromosome loss (see chapter 4), that I considered this possibility unlikely.

It is also possible that the gender labels for the nine females and nine males had in some way been interchanged, but I considered this unlikely (I double-checked my merging of the data to ensure that I had not inadvertently caused an error). Hence, as I could not be certain of the reason for the apparent discrepancies, I decided to also remove the nine male samples from my analyses.

I set out in table 4 the final data numbers allowing for the removal of the 18 samples described above:

	COAD		**HNSC**		**KIRC**	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Tumour samples	142	116	374	137	188	95
Normal samples	21	18	38	12	105	53
Total	163	134	412	149	293	148

Table 4 - summary of final numbers of methylation samples used

3.1.2 Filtering of probes

I then looked at the total female intensities (tfi) for all 416 Y chromosome probes. Figure 8 shows a plot of the tfis for each probe summed across all female KIRC samples, in increasing order from left to right:

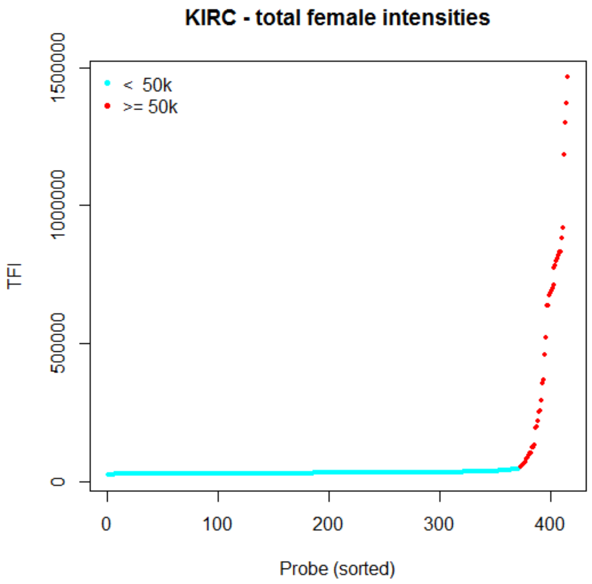


Figure 8 - plot of KIRC total female intensities
The total intensity measurements across all female samples were calculated for each of the 416 Y chromosome probes. The 416 total intensity values are shown in increasing order from left to right.

The plot shows that there is a small proportion of probes which have very high tfis – there were 43 probes with a tfi of at least 50,000. These probes are potentially cross-hybridising with other parts of the genome. I chose a threshold of 50,000 because there was a significant increase in the differences between consecutive, ordered tfis at this point – from 46,817 to 53,607, which was more than four times higher than the previous highest difference.

I observed similar tfi patterns for both the COAD and HNSC datasets. For COAD there were 42 probes with a tfi of at least 50,000, and for HNSC there were 40 probes with a tfi of at least 100,000.

Figure 9 compares, for each probe, the KIRC tfis against COAD (left) and HNSC (right):

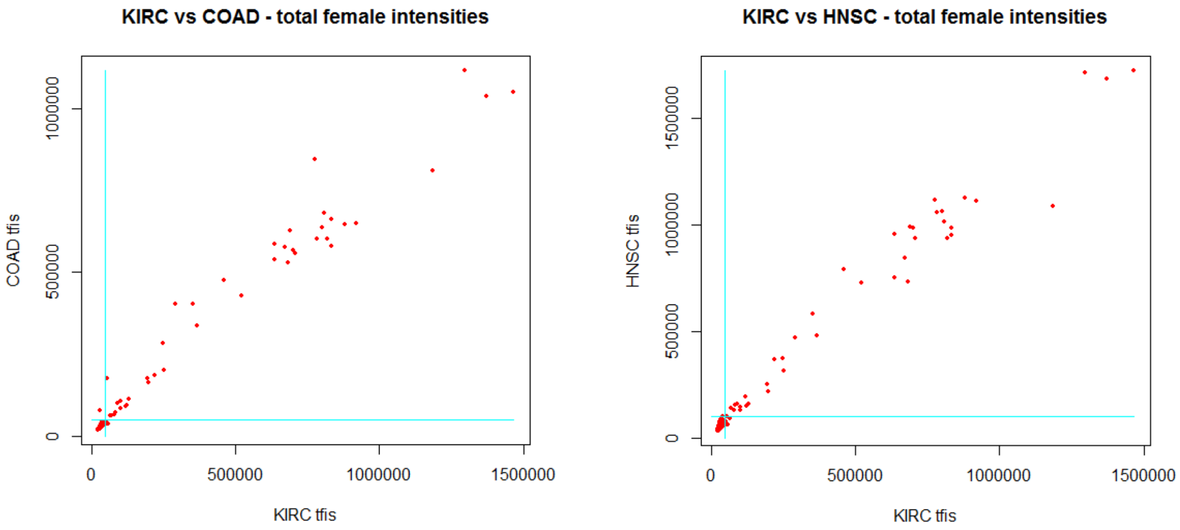


Figure 9 - comparison of KIRC total female intensities vs COAD & HNSC
The chart on the left shows KIRC total female intensities plotted against COAD total female intensities for each of the 416 probes. The blue lines represent the thresholds of 50,000 in both cases. The chart on the right shows KIRC versus HNSC total female intensities, with the blue lines representing the thresholds of 50,000 and 100,000 respectively.

Figure 9 shows that there is excellent consistency between the different datasets (Pearson correlation of 0.99, p-value less than 0.0001 in both cases).

There were 39 probes for which the tfi exceeded 50,000 for both COAD and KIRC and 100,000 for HNSC, and these were my prime candidates for cross-hybridising probes. I then used NCBI's nucleotide BLAST facility to check whether there is any evidence that these probes actually map to other parts of the genome. I discovered that 32 of the 39 probes have high (89% or more) sequence homology with regions on the X-chromosome. I, therefore, decided to exclude these 32 probes from my subsequent analyses. A list of the 32 probes is included in Appendix 1. I retained the other 384 of the original 416 probes for further analyses.

3.1.3 Summary

18 samples were removed across the three datasets on account of their gender information looking suspicious. 32 Y chromosome probes were removed because of potential cross-hybridisation issues, leaving 384 probes for further analysis.

3.2 Consideration of potential batch effects

I then considered whether any batch effects were present in the male methylation data. I investigated two data items of potential interest – "batch" and "tissue source site" – and performed separate analyses of beta values and total intensities. In all cases I performed principal components analysis (PCA) to determine whether there were any batch effects.

3.2.1 Impact of low total intensity measurements

Figure 10 shows a plot of the first two principal components of my PCA for the KIRC male, Y chromosome beta values.

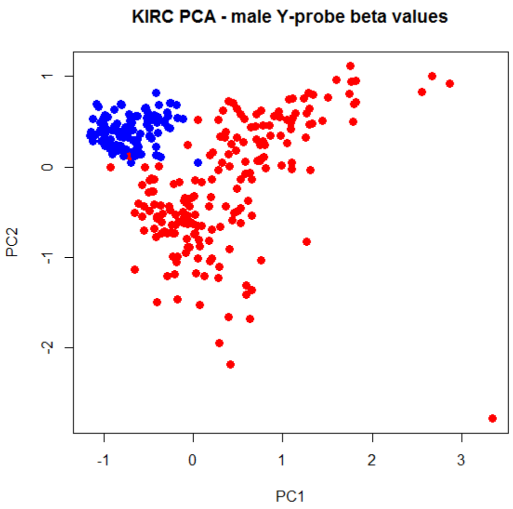


Figure 10 - PCA of KIRC beta values
Principal components analysis was performed using the KIRC male Y chromosome beta values. The first two principal components for each sample are shown on the x- and y- axes respectively. Tumour samples are shown in red and normal samples in blue.

The normal samples (in blue) are grouped together at the top left of the chart, whereas there appear to be two distinct groups of tumour samples (in red). I, therefore, next considered whether the grouping of tumour samples may have been caused by batch issues. Figure 11 shows the same plot as in figure 10 but with samples now colour-coded according to "batch" (in the left hand plot) and "tissue source site" (in the right hand plot):

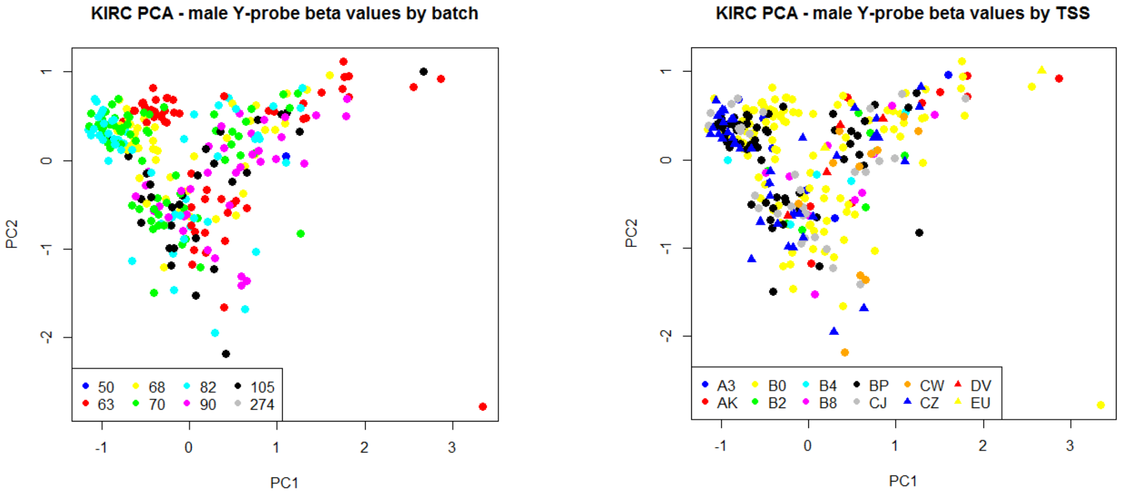


Figure 11 - PCA of KIRC beta values using "batch" and "tissue source site"
Principal components analysis was performed using the KIRC male Y chromosome beta values. The first two principal components for each sample are shown on the x- and y- axes respectively. Samples are colour-coded according to "batch" in the left hand chart and "tissue source site" in the right hand chart.

These plots show that the observed grouping of tumour samples was not caused by the "batch" or "tissue source site" data items. Similar results were obtained for the COAD and HNSC datasets (data not shown).

Whilst I was reviewing the data, I had noticed that there were some male samples (all tumours) in each dataset which had low total intensity values for most of the Y chromosome probes. I believe these samples to be cases which have suffered some loss of chromosomal DNA (see next chapter). I wondered whether copy number loss of Y chromosome may impact on beta value measurements and may be the cause of the observed grouping.

As a preliminary step, I therefore plotted, for each of the 384 probes, total intensity measurement against beta value for all male samples. By reviewing these plots, I noticed a couple of anomalies which were common to all three datasets:

- For CpG sites which were generally methylated at low levels, those tumour samples which had low total intensity measurements tended to have higher beta values than the other samples; and
- Conversely, the same samples tended to have lower beta values than other samples for CpG sites which were generally highly methylated.

Figure 12 shows a couple of plots using the KIRC dataset to illustrate these issues:

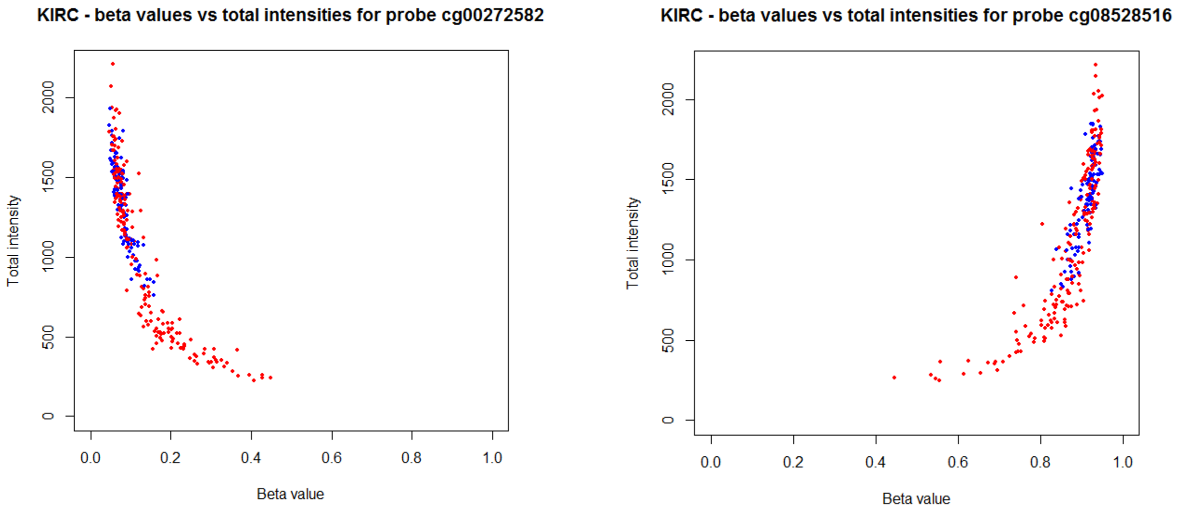


Figure 12 - examples of beta value anomalies caused by low intensity measurements
The charts show plots of beta value (x-axis) versus total intensity (y-axis) for all male KIRC samples. The left hand chart is for probe cg00272582, and the right hand chart is for probe cg08528516. Tumour samples are shown in red and normal samples in blue.

In both plots, the beta values for tumour samples with low total intensity measurements tend towards the centre of the plot (i.e. beta value of 0.5). Including these samples in my differential methylation analyses could distort the results.

In order to assess the scale of the problem, I plotted the total intensities for all samples (normal and tumour) in ascending order, as shown for the KIRC dataset in figure 13:

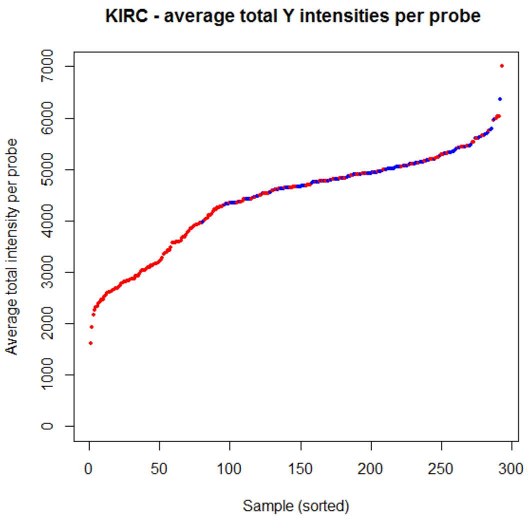


Figure 13 - plot of total intensities for KIRC samples in ascending order
For each KIRC sample, the average total intensity measurement was calculated across all 384 probes, and plotted in ascending order from left to right. Tumour samples are shown in red and normal samples in blue.

Figure 13 shows that there is a sizeable subset of tumour samples whose total intensities are lower than the corresponding measurements for all normal samples, and that the latter are grouped together in a relatively narrow band. I observed the same pattern for both the COAD and HNSC datasets.

I decided that, for the purposes of my differential methylation analyses, I would remove all tumour samples for which the total intensity measurements (across all 384 probes) were lower than the lowest corresponding measurement across all the normal samples within the same dataset. As a result, I removed 70 (out of 142) COAD tumour samples, 153 (out of 374) HNSC samples, and 80 (out of 188) KIRC samples.

Figure 14 shows the same plots as shown in figure 12 after the low intensity tumour samples had been removed:

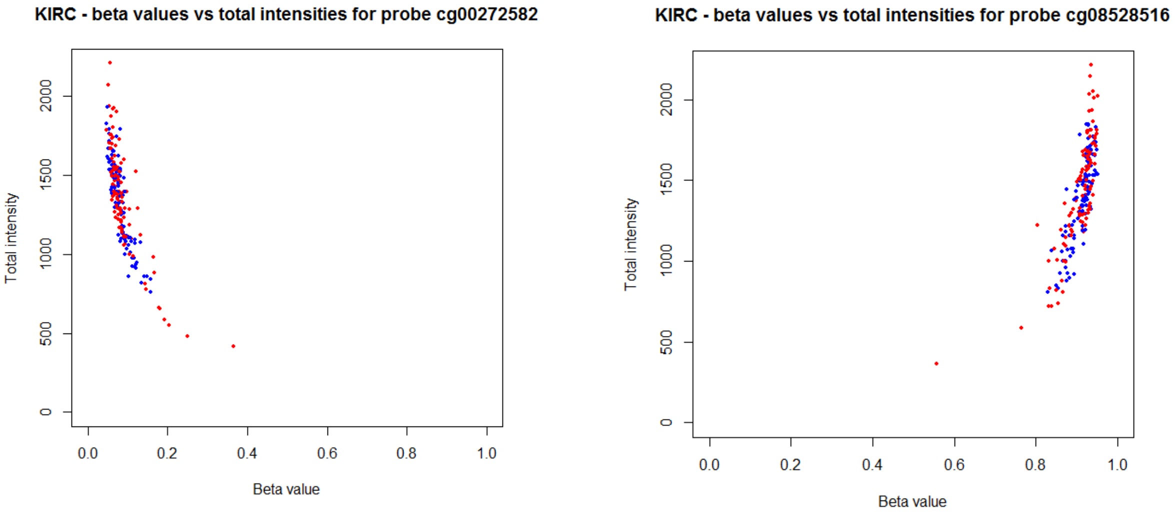


Figure 14 - examples of removal of beta value anomalies
The charts show plots of beta value (x-axis) versus total intensity (y-axis) for male KIRC samples after the low intensity cases had been removed. The left hand chart is for probe cg00272582, and the right hand chart is for probe cg08528516. Tumour samples are shown in red and normal samples in blue.

Figure 15 is a repeat of figure 10 with the low intensity tumour samples now coloured green:

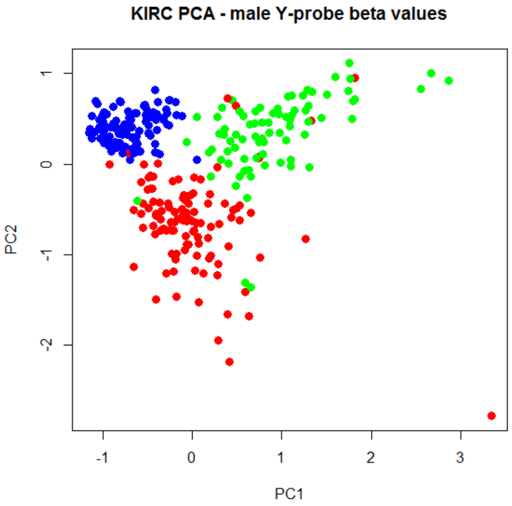


Figure 15 - PCA of KIRC beta values showing low intensity cases separately
Principal components analysis was performed using the KIRC male Y chromosome beta values. The first two principal components for each sample are shown on the x- and y- axes respectively. Low intensity tumour samples are shown in green, other tumour samples are shown in red and normal samples in blue.

The plot clearly shows that the low intensity cases form a distinct group from the other tumour samples. Similar results were obtained for the COAD and HNSC datasets (data not shown).

I then repeated my initial PCA after the low intensity tumour samples had been removed. A plot of the first two principal components is shown in figure 16:

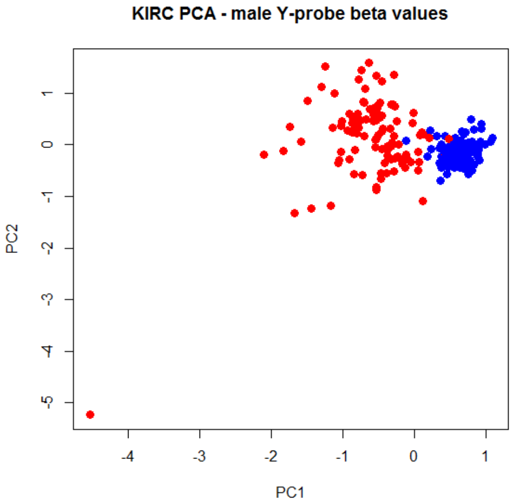


Figure 16 - PCA of KIRC beta values excluding low intensity cases
Principal components analysis was performed using the KIRC male Y chromosome beta values, with the low intensity cases removed. The first two principal components for each sample are shown on the x- and y- axes respectively. Tumour samples are shown in red and normal samples in blue.

After the low intensity samples had been removed, normal and tumour samples clearly formed two separate clusters based on the first principal component (shown on the x-axis), and there were no longer two distinct groups of tumour samples. Again, similar results were obtained for the COAD and HNSC datasets (data not shown).

3.2.2 Checking of batch effects – beta values

Although the low intensity issue had been the main confounding factor in the beta value data, I also wanted to check that there were no serious batch effects for the samples which I had retained for my methylation analyses. Figure 17 shows the results of my PCA for the KIRC dataset, with samples colour-coded according to “batch” and “tissue source site”:

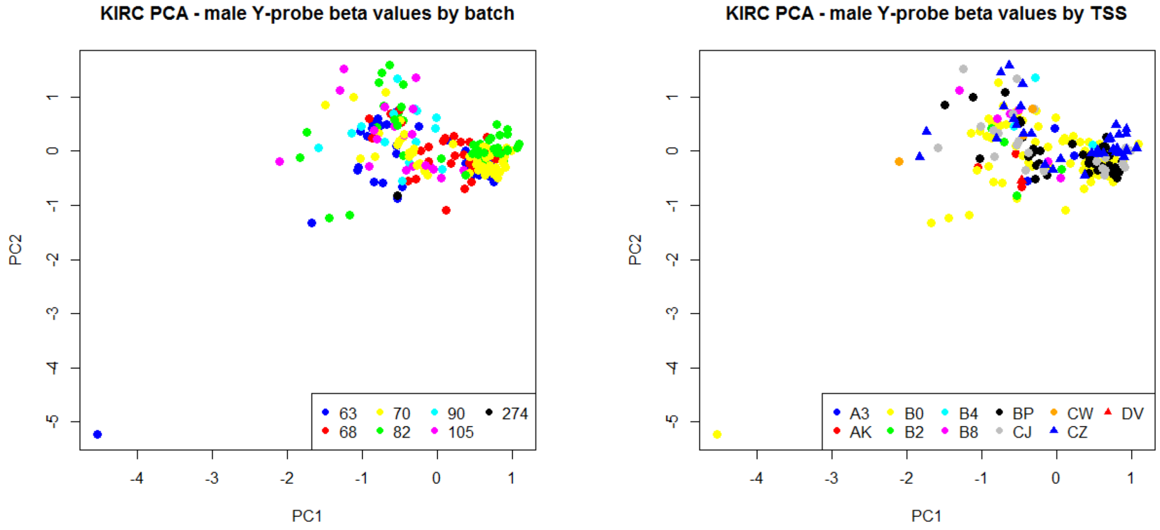


Figure 17 - PCA of KIRC beta values using "batch" and “tissue source site”
Principal components analysis was performed using the KIRC male Y chromosome beta values, after the low intensity cases had been removed. The first two principal components for each sample are shown on the x- and y- axes respectively. Samples are colour-coded according to “batch” in the left hand chart and “tissue source site” in the right hand chart.

There are no obvious batch effects present in the above plots as there is no obvious grouping of samples by either variable. Given this and the fact that sample type (tumour or normal) was clearly the most significant discriminator between samples (confirmed by multiple linear regression analysis – p-value < 0.0001 for association between the first principal component and sample type), I decided that no further remedial action was required in respect of the beta value measurements. Similar results were obtained for the COAD and HNSC datasets.

3.2.3 Checking of batch effects – total intensity values

In the next chapter I look at Y-chromosome copy number variations, using, in the first instance, the total methylation intensity data to estimate copy number. I, therefore, also wanted to check whether there were any serious batch effects which may have distorted the total intensity measurements.

Figure 18 shows the results of my initial PCA of the total intensity measurements using all the male samples:

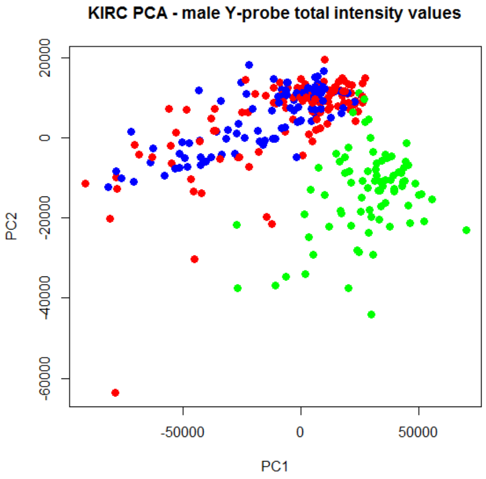


Figure 18 - PCA of KIRC total intensity measurements
Principal components analysis was performed using the KIRC male Y chromosome total intensity values. The first two principal components for each sample are shown on the x- and y- axes respectively. Low intensity tumour samples, as previously defined, are shown in green, other tumour samples are shown in red and normal samples in blue.

The previously identified low intensity cases (in green) form a distinct subgroup of samples, as I had expected. The remaining tumour and normal samples form another group. There is, however, some variability within both groups, suggesting that there might be some batch effects. Similar results were also obtained for the COAD and HNSC datasets (data not shown).

I then performed the same analysis again, but this time colour-coding the samples according to either "batch" or "tissue source site". Figure 19 shows the first two principal components of my PCA for the KIRC total intensity measurements using "batch" and "tissue source site" as the items of interest:

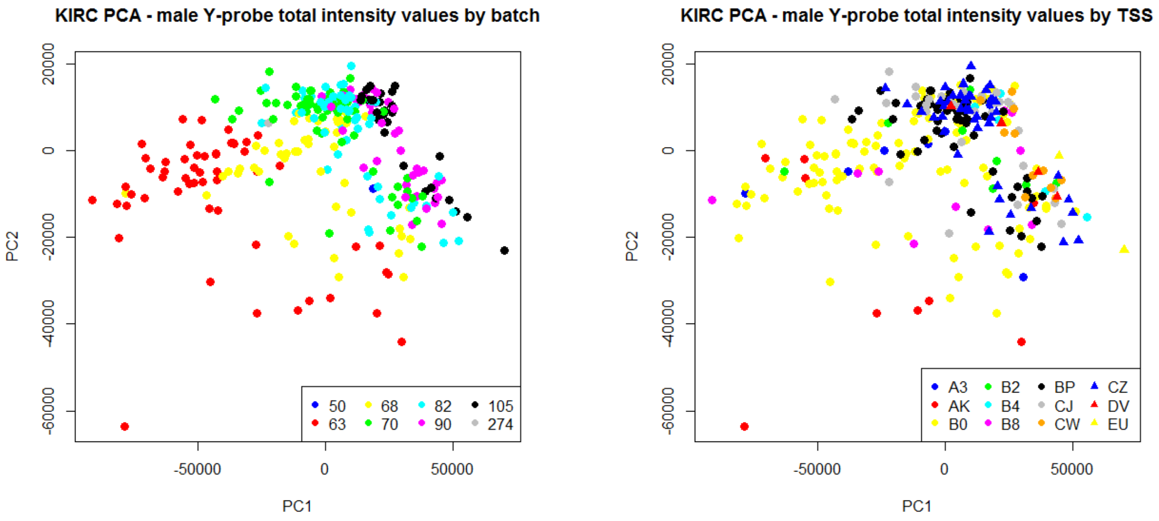


Figure 19 - PCA of KIRC total intensities using "batch" & "tissue source site"
Principal components analysis was performed using the KIRC male Y chromosome total intensity values. The first two principal components for each sample are shown on the x- and y- axes respectively. Samples are colour-coded according to "batch" in the left hand chart and "tissue source site" in the right hand chart.

The plots confirm that there is a potential batch effect present, especially in relation to the "batch" variable. A similar issue was observed for both the COAD and HNSC datasets. To illustrate this point further, figure 20 shows a plot of all male total intensity values (averaged across all 384 probes) for the KIRC dataset, broken down by "batch":

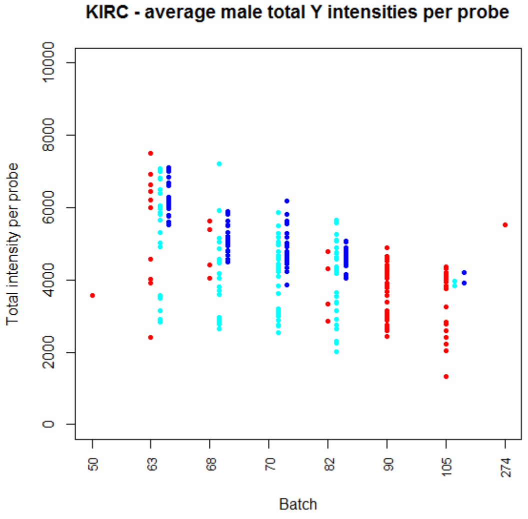


Figure 20 - plot of KIRC average male total intensities split by "batch"
For each KIRC sample, the average total intensity measurement was calculated across all 384 probes, and plotted against the y-axis. Normal samples are shown in blue, corresponding matched tumour samples in light blue, and other unmatched tumour samples in red. Samples are subdivided into "batch" along the x-axis.

The above plot illustrates that there are some discrepancies between batches. In particular, looking at the values for normal samples (in dark blue), batch 63 values are higher than for other batches. Once again, a similar issue was observed for the COAD and HNSC datasets.

I, therefore, decided to make some adjustment to the total intensity values, so that the distributions of values across different batches were more consistent. I decided to base my adjustment on the measurements for normal samples, as there is much lower variability between normal samples compared with tumour samples. I calculated the median average intensity across all normal samples within each batch, and then adjusted the total intensity values for all samples within a batch by the ratio:

median normal intensity across all batches / batch-specific median normal intensity

This method better aligned total intensity values between batches, as shown in figure 21 for the KIRC dataset:

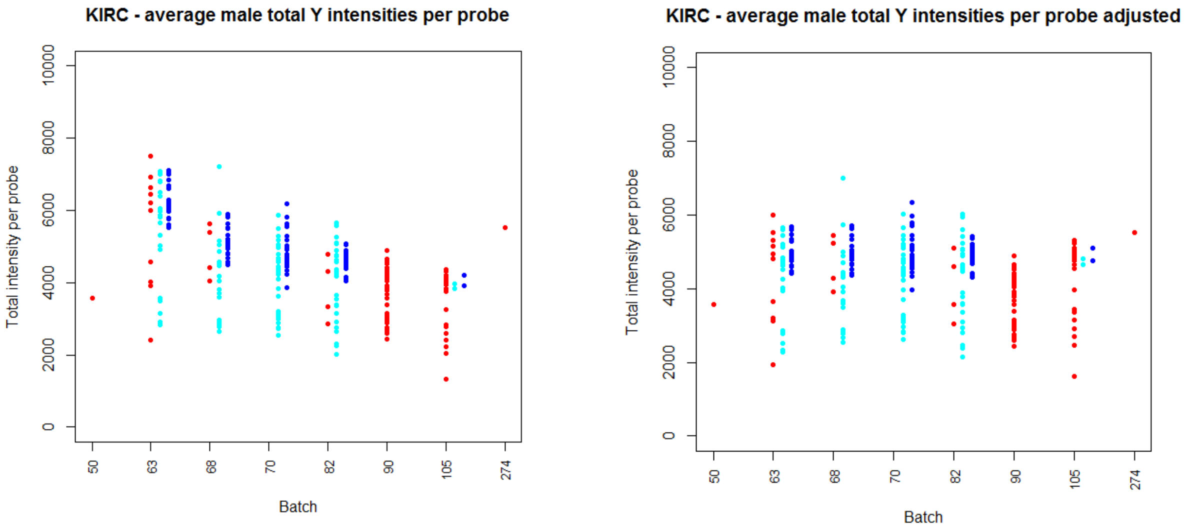


Figure 21 - KIRC average male total intensities split by "batch" before and after adjustment
For each KIRC sample, the average total intensity measurement was calculated across all 384 probes, and plotted against the y-axis. Normal samples are shown in blue, corresponding matched tumour samples in light blue, and other unmatched tumour samples in red. Samples are subdivided into “batch” along the x-axis. The left hand plot shows values before “batch” correction, and the right hand plot show values after correction.

I employed the same process for both the COAD and HNSC datasets. Note that beta values were left unadjusted – only total intensity values were adjusted.

One issue with this method is that it did not make any adjustment for batches which contained no normal samples – this was a particular issue for the COAD and HNSC datasets. However, I took the view that there was too much variability in measurements for tumour samples for me to be able to make a sensible adjustment for such batches and, consequently, I accepted that there might be a small amount of error in total intensity measurements for samples in these batches. I also tried another method for making these adjustments, known as trimmed mean quantile normalisation[67] (applied to all 450k platform probes, not just those targeting the Y chromosome), and obtained very similar results (data not shown).

3.2.4 Summary

Beta values appeared to be distorted by low overall intensity measurements. Samples with consistently low intensity measurements were removed from subsequent analyses which used the methylation data. No residual batch effects were detected for beta values. However, batch effects were detected for total intensity values, and were partially corrected.

3.3 Overview of Y chromosome probes on 450k platform

Following the filtering out of potentially cross-hybridising probes, 384 Y chromosome probes on the 450k platform remained for further analysis. These probes span the male-specific region on the Y chromosome between 2.65 million and 28.55 million bases, which represents less than one half of the length of the entire chromosome (approx 59 million bases). However, they cover 55 genes (out of an estimated total of around 60).

I set out in the following sections details of key characteristics of the 384 probes, and how they compare to the whole set of probes on the 450k platform.

3.3.1 CpG density of Y chromosome probes

As described in chapter 1, some regions of the human genome are sparsely populated with CpG sites, whereas others are densely populated. It is possible that CpG density may influence changes in methylation patterns in cancer[34]. I, therefore, investigated the CpG density of those regions to which the 384 Y chromosome probes are targeted.

For each CpG site, I counted the number of CpG dinucleotides located in the 500bp window centred on the site. For this purpose I used the “BSgenome.Hsapiens.UCSC.hg19” package in R.

Figure 22 shows the distribution of CpG densities for the 384 probes, and the corresponding distribution for all the probes on the 450k platform:

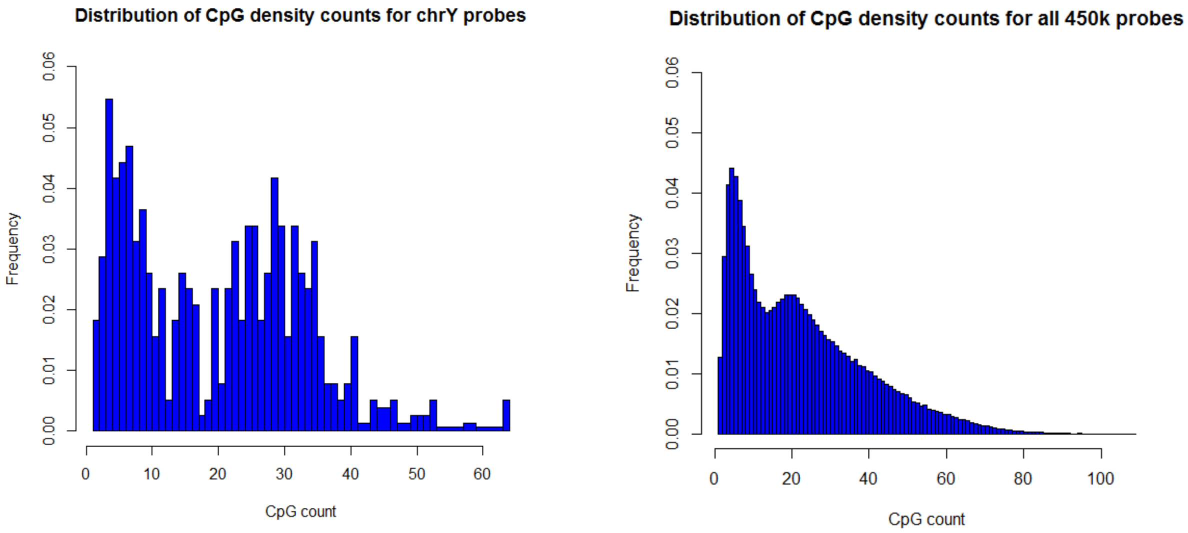


Figure 22 - distribution of CpG densities of Y probes and all 450k probes
The number of CpG dinucleotides in the 500bp window centred on each targeted CpG site was calculated. The left hand histogram shows the distribution of counts for the 384 Y chromosome probes. The right hand histogram shows the distribution for all probes on the 450k platform.

Figure 22 shows that there appears to be a bi-modal distribution of CpG density, with a peak at very low values, and a second slightly lower peak at a count of around 30 CpG sites. This pattern is similar to the pattern for all 450k probes, although for the latter there is a smoother distribution (as there are a lot more probes), and the second peak occurs earlier at a count of around 20 CpG sites.

3.3.2 Breakdown of Y chromosome probes by island regions

As described in Chapter 1, Illumina provide for the 450k platform an annotation item which describes the relationship between each interrogated CpG site and the nearest CpG island. I will use the term “island region” for this item.

I set out in figure 23 a pie-chart showing the distribution of island regions for the 384 Y probes, along with a corresponding pie-chart for all 450k probes:

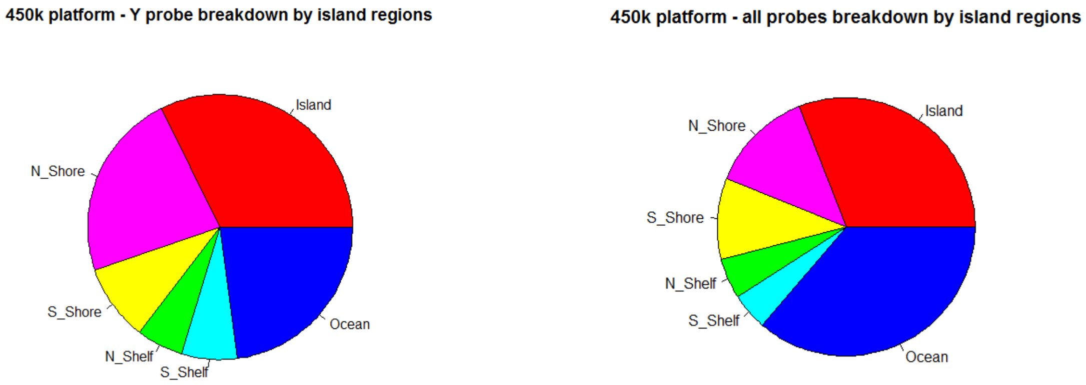


Figure 23 - distribution of island regions of Y probes and all 450k probes
Pie-charts showing the distribution of island regions in which probes on the 450k platform are located. The left hand chart shows the distribution for the 384 Y chromosome probes. The right hand chart shows the distribution for all probes on the 450k platform.

Figure 23 shows some differences between the distribution of island regions for the 384 Y probes and the distribution for the whole 450k platform. In particular, the Y probes contain a greater proportion of sites located in "north shores" (23% vs 13%), and a lower proportion (23% vs 36%) in areas more than 4,000 base pairs away from the nearest island (which I have called "oceans").

In figure 24 I compare the distribution of CpG densities across the different island regions:

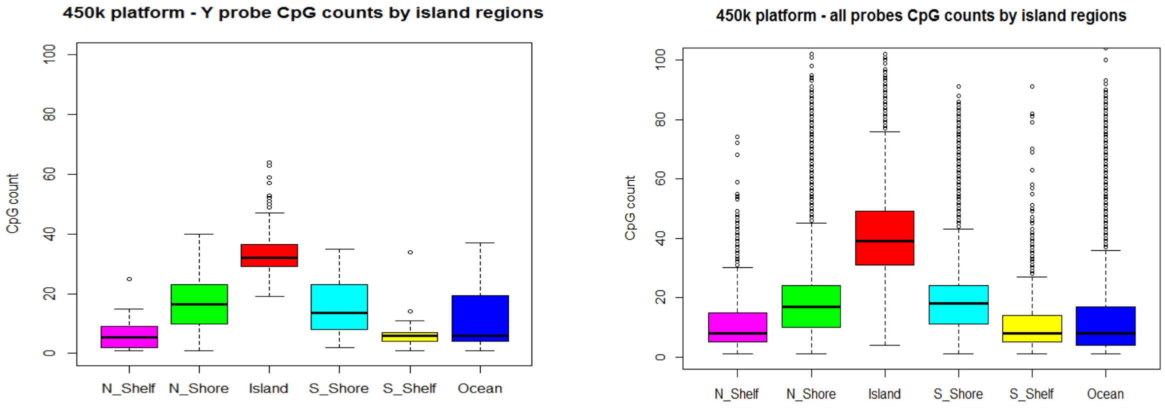


Figure 24 - distribution of CpG counts across island regions for Y probes and all 450k probes
Boxplots showing the distribution of CpG counts for probes on the 450k platform split by island region. The left hand chart shows the distribution for the 384 Y chromosome probes. The right hand chart shows the distribution for all probes on the 450k platform.

The above plots show, as expected, that islands typically have higher CpG densities than the other regions, and that CpG density generally declines as the distance from the nearest island increases.

3.3.3 Genes covered by Y chromosome probes

265 out of the 384 probes are linked to genes on the Y chromosome. 55 separate genes are covered by these 265 probes – a list of the genes included is set out in table 5:

AMELY	BCORL2	CD24	CYorf15A	DAZ1
DAZ2	DDX3Y	EIF1AY	FAM197Y2	FAM41AY2
HSFY1	KDM5D	LOC100101115	LOC100101121	LOC401629
LOC401630	NCRNA00185	NLGN4Y	PCDH11Y	PRKY
RPS4Y1	RPS4Y2	SRY	TBL1Y	TMSB4Y
USP9Y	UTY	ZFY		
RBMY1A3P / 1D / 1F / 1J/ 2EP / 3AP				
TSPY1 / 2 / 3/ 4				
TTTY1 / 4 / 4B / 5 / 8 / 8B / 10 / 11 / 12 / 13 / 14 / 15 / 16 / 18 / 19 / 20 / 22				

Table 5 - Y chromosome genes covered by probes on the 450k platform

Genes shown in red are of particular interest as they have previously been proposed as potential tumour suppressor genes[6]. Furthermore, there are another 10 genes, shown in orange, which have been proposed as being essential for male viability[5]. 125 probes are targeted at these 12 genes, though the coverage is very uneven – ranging from just one probe for USP9Y to 19 for NLGN4Y.

3.3.4 Breakdown of Y chromosome probes by gene locations

Illumina also provide for the 450k platform an annotation item which describes the relationship between each interrogated CpG site and the nearest gene. I will use the term “gene location” for this item.

I set out in figure 25 a pie-chart showing the distribution of gene locations for the 384 probes, along with a corresponding pie-chart for all 450k probes:

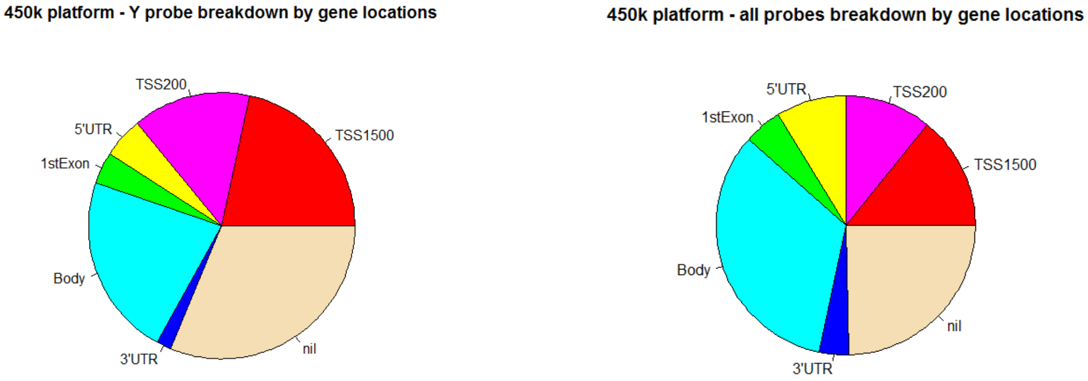


Figure 25 - distribution of gene locations of Y probes and all 450k probes
Pie-charts showing the distribution of gene locations in which probes on the 450k platform are located. The left hand chart shows the distribution for the 384 Y chromosome probes. The right hand chart shows the distribution for all probes on the 450k platform.

Figure 25 shows some differences between the distribution of gene locations for the 384 Y chromosome probes and the distribution for the whole 450k platform. The Y chromosome probes contain higher proportions of sites within 200 and 1,500 base pairs of a gene's transcriptional start site (14% vs 11% and 22% vs 14% respectively), and also sites not linked to any gene (labelled "nil" – 31% vs 25%). Conversely, there are lower proportions of Y chromosome probes within gene bodies, 3'UTRs and 5'UTRs (22% vs 33%, 2% vs 4% and 5% vs 9% respectively).

In figure 26 I compare the distribution of CpG densities across the different types of gene location:

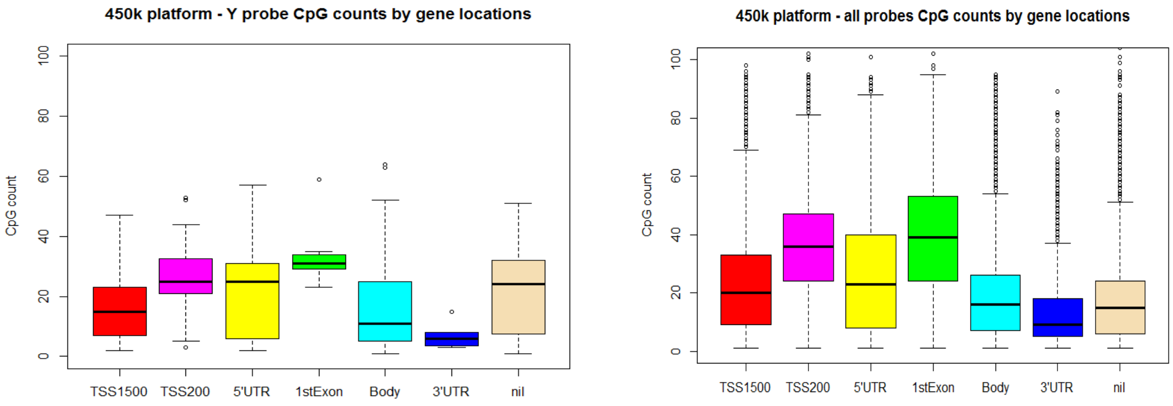


Figure 26 - distribution of CpG counts across gene locations for Y probes and all 450k probes
Boxplots showing the distribution of CpG counts for probes on the 450k platform split by gene location. The left hand chart shows the distribution for the 384 Y chromosome probes. The right hand chart shows the distribution for all probes on the 450k platform.

The above plot shows that CpG density varies across different gene locations, with sites located in TSS200s and first exons typically being in regions of higher density (interquartile ranges above 20 counts), and sites in 3’UTRs in lower density regions (median counts = 6 and 9 respectively).

3.3.5 Summary

The 384 probes retained cover just under one half of the total length of the male-specific region of the Y chromosome, but cover the vast majority of known genes in this region, including 12 which could be of particular interest to cancer development. The characteristics of the retained probes are broadly consistent with the 450k platform as a whole, but with some differences.

3.4 Methylation of normal male samples

In this section I outline key characteristics of the Y chromosome methylation profiles of normal samples. I will then use these characteristics as a base for investigating differential methylation of tumour samples in subsequent sections.

The KIRC dataset has many more normal male samples (105) than either the COAD (21) or HNSC (38) datasets. My analysis of normal samples is, therefore, based on the KIRC dataset, but includes comparison with the other datasets.

3.4.1 Overall methylation profile of KIRC normal samples

Figure 27 shows the distribution of all beta values across all normal samples and all 384 Y chromosome probes for the KIRC dataset:

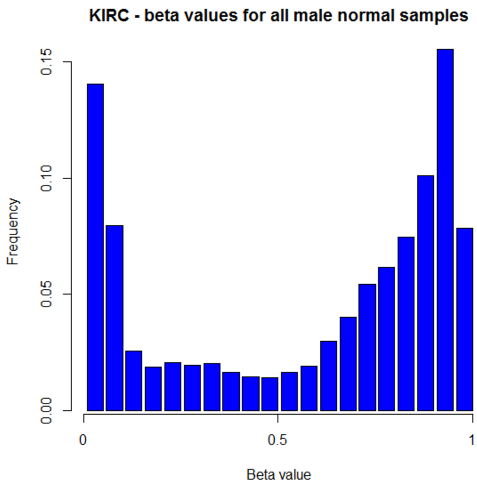


Figure 27 - distribution of beta values for all KIRC normal samples and probes
Histogram of all beta values for all KIRC normal, male samples across all 384 Y chromosome probes

The above plot shows that there are two clear peaks in the distribution of beta values, one at very low (i.e. unmethylated) values and the other at very high (i.e. fully methylated) values, with much smaller frequencies for intermediate values.

Figure 28 shows how the methylation profile for the KIRC normal samples varies according to island region:

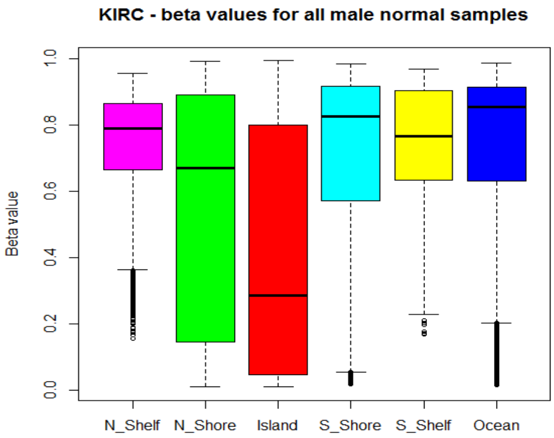


Figure 28 - distribution of KIRC normal beta values across island regions
Boxplots showing the distribution of beta values across all 384 Y chromosome probes for all KIRC normal male samples split by island region.

The plot shows that, whilst there is some variability within all regions, CpG sites located in islands tend to be less methylated than sites in other regions (median beta value = 0.29, compared with greater than 0.60 for all other regions).

Figure 29 shows, for each of the 384 probes, a plot of mean beta value for the KIRC normal samples against CpG density:

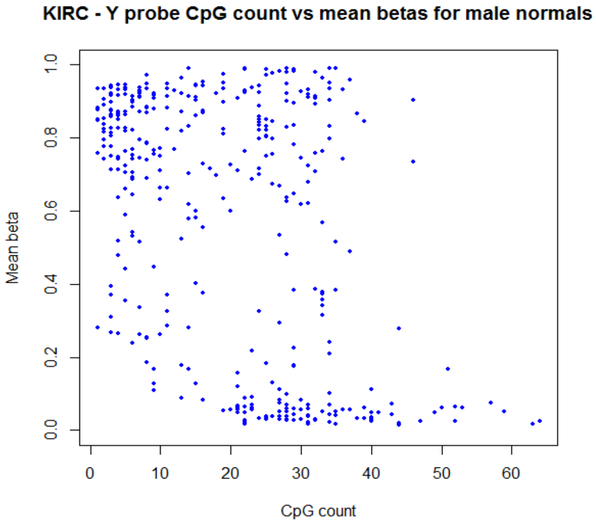


Figure 29 - plot of mean normal KIRC beta values vs CpG count
For each of the 384 Y chromosome probes, the CpG count based on the 500bp window centred on the targeted site was plotted against the mean beta value across all KIRC normal male samples.

Again there is some variability. However, figure 29 shows that CpG sites in areas of low CpG density tend to be highly methylated, whereas sites in regions of high CpG density tend to be unmethylated (Pearson correlation = -0.46, p-value <0.0001).

Finally, figure 30 shows how the methylation profile for the KIRC normal samples varies according to gene location:

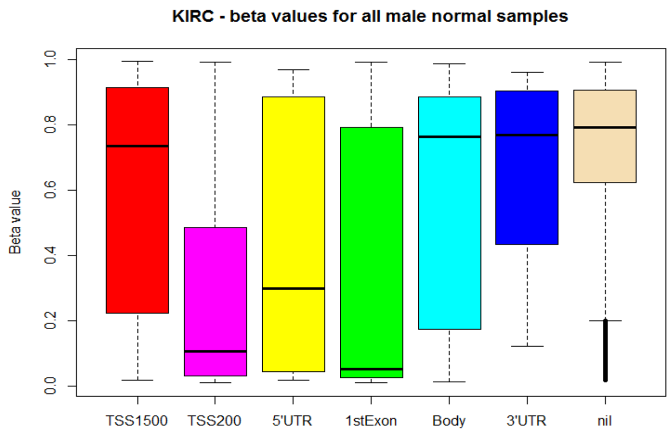


Figure 30 - distribution of KIRC normal beta values across gene locations
Boxplots showing the distribution of beta values across all 384 Y chromosome probes for all KIRC normal male samples split by gene location.

Once again there is a large amount of variability within each location type. However, CpG sites located in TSS200s and 1stExons tend to be lowly methylated (median values = 0.105 and 0.051 respectively), whereas sites located in TSS1500s, gene bodies and 3’UTRs tend to be highly methylated (median values = 0.735, 0.763 and 0.768 respectively), as do sites not linked to any gene (median value = 0.792). Sites located within 5’UTRs tend to have intermediate methylation levels (median value = 0.299).

3.4.2 Comparison of normal methylation profiles between datasets

I also looked at distributions of overall Y chromosome methylation for both the COAD and HNSC datasets, as shown in figure 31:

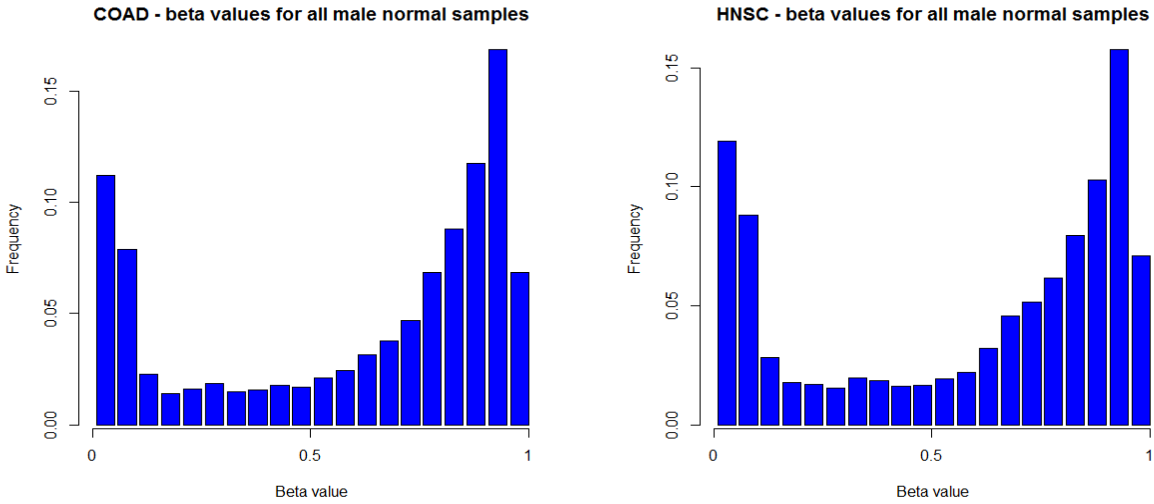


Figure 31 - distribution of beta values for all COAD and HNSC normal samples and probes
Histograms of all beta values for all COAD (left) and HNSC (right) normal, male samples across all 384 Y chromosome probes

When compared with the KIRC methylation distribution shown in figure 27, both the COAD and HNSC datasets, but in particular COAD, have slightly lower peaks at low beta values and slightly higher peaks at high beta values.

I also performed a probe-by-probe comparison between the datasets, using the mean beta values for each probe. Figure 32 shows the comparison of mean beta values between KIRC and COAD datasets on the left and between KIRC and HNSC datasets on the right:

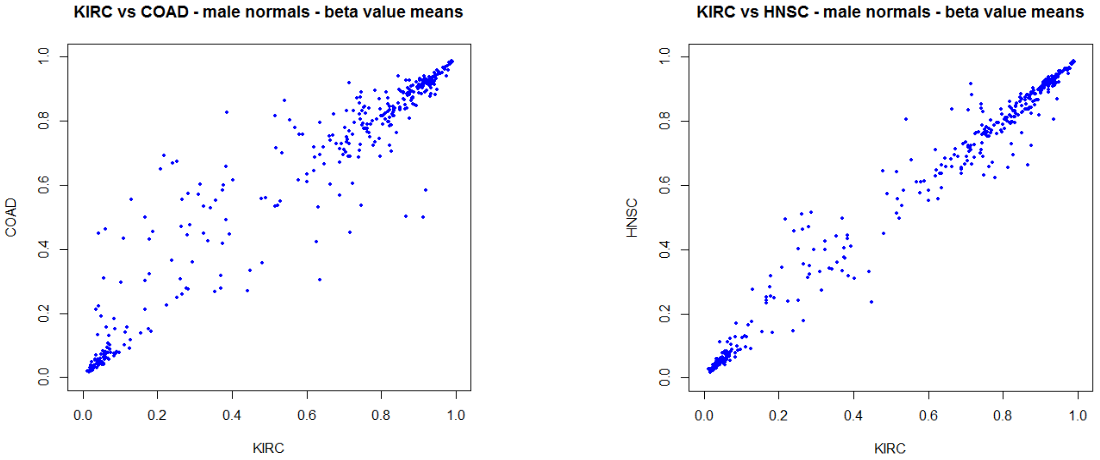


Figure 32 - comparison of mean Y chromosome beta values between datasets
For each dataset, the mean beta value for each of the 384 Y chromosome probes was calculated across all normal, male samples. The left hand plot shows KIRC vs COAD values, and the right hand plot shows KIRC vs HNSC values.

The above plots show that there is generally very good correlation in mean beta values between datasets (Pearson correlation = 0.95 and 0.99 respectively, p-values < 0.0001). However, there are some inconsistencies, in particular with the COAD dataset.

3.4.3 Investigation of probes with inconsistent methylation profiles

I was then interested to investigate those CpG sites where there was a large discrepancy in mean beta values between datasets. These sites may be important in distinguishing between different tissue types, and may also be a source of aberrant methylation in cancer. Table 6 shows those sites for which there was a difference in mean beta values of at least 0.3 between at least two of the datasets (this threshold yielded the top 5% of all 384 sites):

CpG site ID	Position (Mbp)	Gene	Gene location	Island region	CpG count	Mean beta values		
						COAD	HNSC	KIRC
cg13654344	2.65	SRY	3'UTR	N_Shelf	15	0.615	0.310	0.402
cg11898347	2.66	SRY	TSS200	N_Shore	9	0.433	0.131	0.111
cg27636129	2.66	SRY	TSS200	N_Shore	9	0.500	0.251	0.168
cg09595415	2.66	SRY	TSS200	N_Shore	8	0.673	0.403	0.252
cg18058072	2.81	ZFY	5'UTR	S_Shelf	6	0.864	0.806	0.541
cg09728865	6.78	TBL1Y	5'UTR	Island	23	0.693	0.495	0.217
cg08921682	6.89	TBL1Y	5'UTR	Ocean	4	0.305	0.592	0.636
cg14671357	7.68	TTTY12	Body	Ocean	3	0.583	0.933	0.919
cg14210405	9.93	nil	nil	Island	34	0.668	0.459	0.241
cg14720093	10.03	nil	nil	Island	35	0.816	0.643	0.515
cg01463110	14.65	nil	nil	N_Shore	4	0.501	0.663	0.865
cg17430262	14.65	nil	nil	Island	34	0.650	0.344	0.209
cg18188392	14.65	nil	nil	Island	43	0.449	0.111	0.044
cg05608794	14.65	nil	nil	Island	26	0.556	0.277	0.131
cg02730008	15.81	TMSB4Y	TSS1500	N_Shore	11	0.500	0.926	0.913
cg26198148	15.81	TMSB4Y	TSS1500	N_Shore	24	0.452	0.883	0.715
cg10363397	15.86	nil	nil	Island	32	0.826	0.317	0.386
cg09748856	16.64	NLGN4Y	Body	Island	39	0.462	0.074	0.062
cg16894943	20.49	LOC401630	TSS200	Island	33	0.602	0.272	0.315

Table 6 - CpG sites differentially methylated in normal samples
The table shows CpG sites for which there was a difference of at least 0.3 in mean beta values between any pair of normal sample datasets.

Nine out of the 19 sites are located in CpG islands, with a further six found within 2,000 base pairs of an island in north shores. Seven genes are represented in the table, including four out of the 12 genes of interest previously mentioned in section 3.3.3 (NLGN4Y, TBL1Y, TMSB4Y and ZFY), and the SRY gene, which has four sites in the list. However, the differences between the datasets are not consistent across the different genes / CpG sites.

For each of the four CpG sites targeted at the SRY gene, the COAD normal samples are more highly methylated than both the HNSC and KIRC normal samples. Figure 33 compares the beta values across all three datasets for one of these sites (cg09599415):

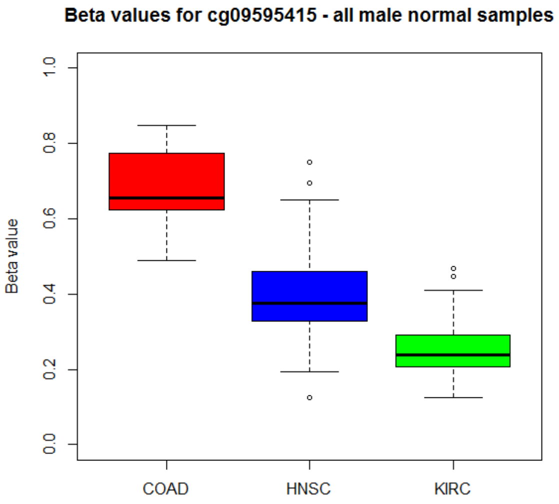


Figure 33 - comparison of normal sample beta values for cg09595415
Boxplots of all normal, male beta values for CpG site cg09595415 within the SRY gene, split by dataset. The plots for COAD, HNSC and KIRC samples are shown in red, blue and green respectively.

The KIRC normal samples are the least methylated (median = 0.239), the COAD samples are the most highly methylated (median = 0.655), and the HNSC samples have intermediate methylation levels (median = 0.377).

One of the TBL1Y sites (cg09728865) has similar methylation patterns to the SRY sites. However, the other TBL1Y site (cg08921682) has the opposite pattern, with the COAD samples being the least methylated (median = 0.288 vs 0.581 and 0.649 respectively for HNSC and KIRC), as shown in figure 34:

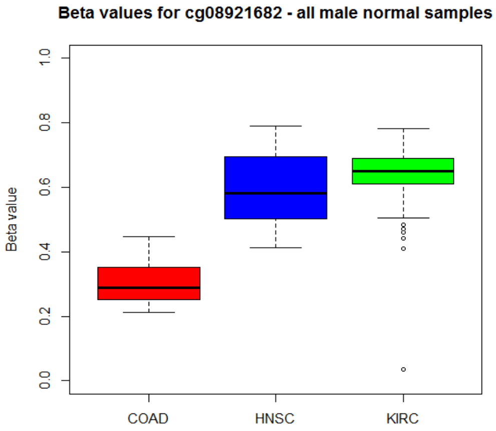


Figure 34 - comparison of normal sample beta values for cg08921682
Boxplots of all normal, male beta values for CpG site cg08921682 within the TBL1Y gene, split by dataset. The plots for COAD, HNSC and KIRC samples are shown in red, blue and green respectively.

CpG site cg18058072 targeted at the ZFY gene has a different pattern again, with both COAD and HNSC samples being highly methylated (median = 0.877 and 0.828 respectively), but KIRC samples having lower methylation levels (median = 0.541), as shown in figure 35:

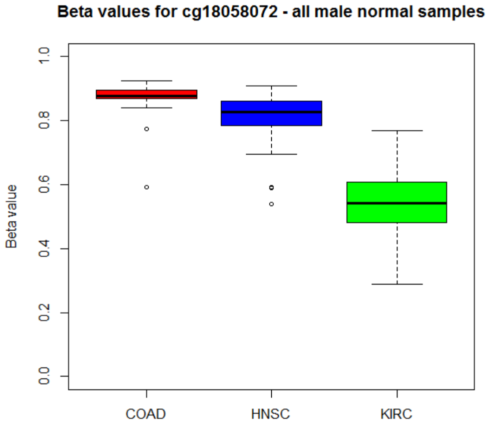


Figure 35 - comparison of normal sample beta values for cg18058072
Boxplots of all normal, male beta values for CpG site cg18058072 within the ZFY gene, split by dataset. The plots for COAD, HNSC and KIRC samples are shown in red, blue and green respectively.

The two CpG sites within the TMSB4Y gene have similar methylation patterns, which are again different from the other genes. The comparison across the datasets for site cg02730008 is shown in figure 36. This time the HNSC and KIRC samples are highly methylated (median = 0.930 and 0.917 respectively), whereas the COAD samples have methylation levels around 50% (median = 0.510).

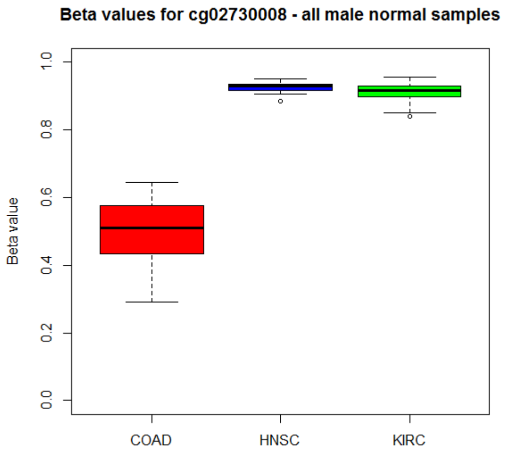


Figure 36 - comparison of normal sample beta values for cg02730008
Boxplots of all normal, male beta values for CpG site cg02730008 within the TMSB4Y gene, split by dataset. The plots for COAD, HNSC and KIRC samples are shown in red, blue and green respectively.

Finally, the pattern for site cg09748856 within the NLGN4Y gene is shown in figure 37:

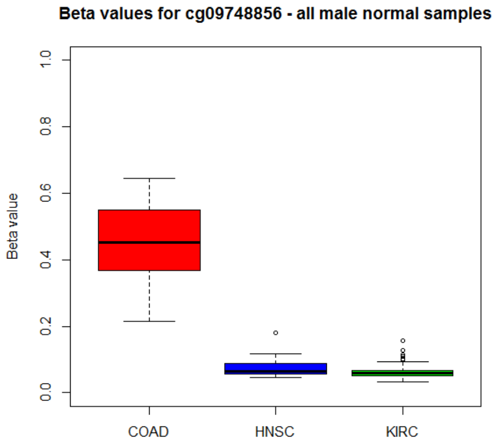


Figure 37 - comparison of normal sample beta values for cg09748856
Boxplots of all normal, male beta values for CpG site cg09748856 within the NLGN4Y gene, split by dataset. The plots for COAD, HNSC and KIRC samples are shown in red, blue and green respectively.

This time the HNSC and KIRC samples have low methylation levels (median = 0.065 and 0.058 respectively), whereas the COAD samples have higher levels (median = 0.453).

In all five cases, the Kruskal-Wallis test comparing datasets gave a significant result (all p-values less than 0.0001).

There is also evidence of some clustering of methylation differences, with four of the 19 sites very close together on the Y chromosome at around 14.65 million base pairs, in addition to the four sites within the SRY gene and the two pairs of sites within the TBL1Y and TMSB4Y genes.

3.4.4 Summary

The normal samples within all three datasets had similar overall Y chromosome methylation profiles, with peaks at both low and high methylation values. However, COAD samples tended to be slightly more methylated on average than both HNSC and KIRC samples. Furthermore, for a small number of individual CpG sites, there were some significant differences in methylation profiles between the tissue types.

3.5 Overview of differential methylation in tumours

Next I considered how Y chromosome methylation differed between normal and tumour samples in my three chosen datasets. It is possible that aberrant methylation of the Y chromosome could be an important factor in tumour development.

I initially decided to split my analysis into two parts. Firstly I considered, for each dataset, those tumour samples for which there were also matching normal samples. I then investigated all the other tumours for which there were no matching normal samples.

The key advantage of comparing matched normal / tumour pairs is that I could make a direct comparison of changes in methylation for each CpG site within each matched pair of samples. However, this comes at the cost of not utilising all the available data, which is a particular issue for the COAD and HNSC datasets for which there are only relatively small numbers of normal samples compared with the numbers of tumour samples.

There is also the possibility that unmatched tumour samples may exhibit different methylation profiles from the matched samples. Therefore, by initially splitting my analyses, I could check whether there were any such differences.

3.5.1 Overall methylation changes in matched tumour samples

I first investigated general differences in methylation between normal and matched tumour samples. Initially, I compared mean normal and tumour beta values for each probe, as shown in figure 38 for the KIRC dataset:

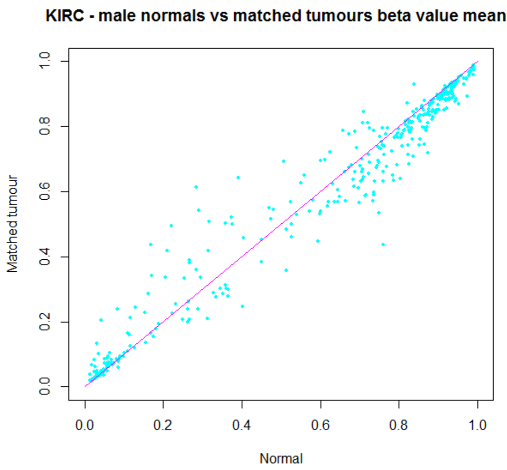


Figure 38 - comparison of mean beta values for matched samples (KIRC)
Plot of mean beta values for all 384 Y chromosome probes for KIRC normal vs matched tumour samples.

There is high correlation between the mean normal and matched tumour methylation values for the KIRC dataset (r=0.98, p-value < 0.0001), although that there are some differences. In figure 39 I show the corresponding plots for the COAD and HNSC datasets:

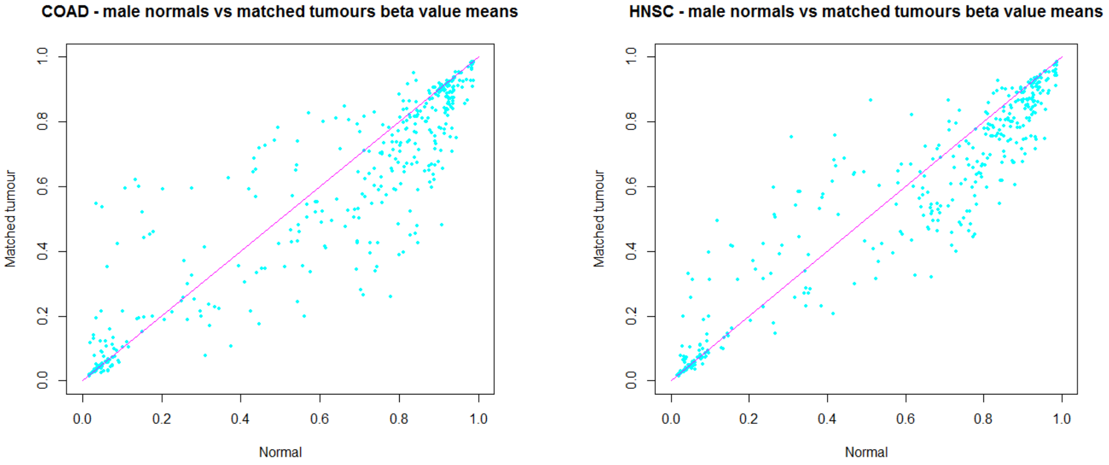


Figure 39 - comparison of mean beta values for matched samples (COAD & HNSC)
Plot of mean beta values for all 384 Y chromosome probes for COAD (left) and HNSC (right) normal (x-axis) vs matched tumour samples (y-axis).

For both there is again high correlation (r=0.90 and 0.94 respectively, p-values < 0.0001), although the correlations are lower than for the KIRC dataset, in particular for COAD.

I then compared the variation in methylation between normal and tumour samples. For this purpose I calculated the standard deviation of beta values for each probe. Figure 40 shows that methylation is clearly more variable within KIRC matched tumour samples compared with normal samples (Wilcoxon p-value < 0.0001):

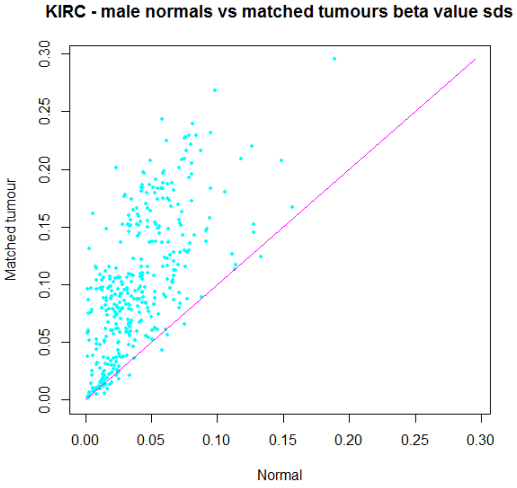


Figure 40 - comparison of beta value variability for matched samples (KIRC)
Plot of standard deviation of beta values for all 384 Y chromosome probes for KIRC normal (x-axis) vs matched tumour samples (y-axis).

Figure 41 shows a similar picture for the COAD and HNSC datasets:

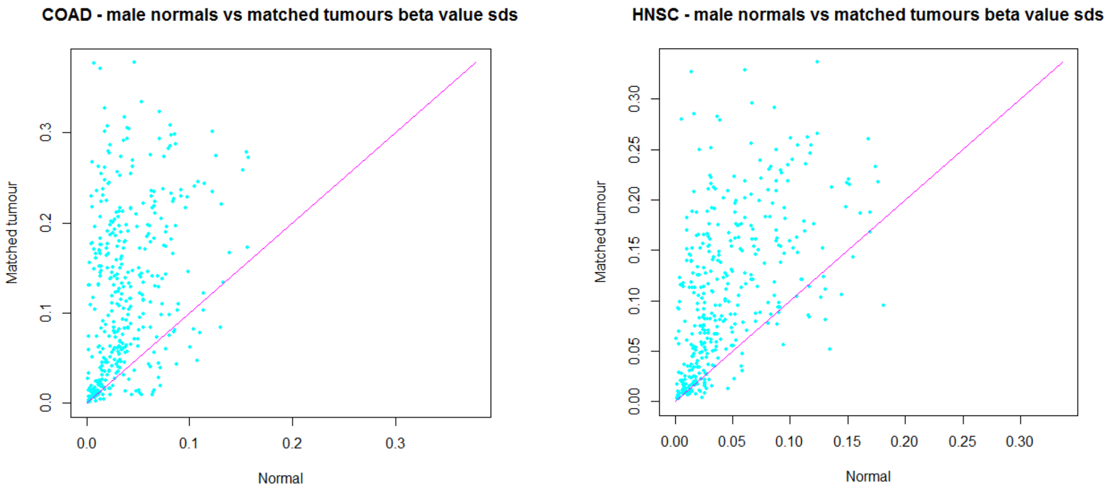


Figure 41 - comparison of beta value variability for matched samples (COAD & HNSC)
Plot of standard deviation of beta values for all 384 Y chromosome probes for COAD (left) and HNSC (right) normal (x-axis) vs matched tumour samples (y-axis).

3.5.2 Overall methylation changes in unmatched tumour samples

Next I separately considered tumour samples for which there were no matched normal samples. For the COAD and HNSC datasets there were many more unmatched tumour samples than there were matched samples.

As for the matched cases, I observed:

- some differences in mean beta values between tumour and normal samples for all datasets;
- greater variability in methylation amongst the tumour samples for all datasets;
- larger differences within the COAD dataset.

Figure 42 shows comparisons of the means and standard deviations of beta values between normal and unmatched tumour samples for the COAD dataset:

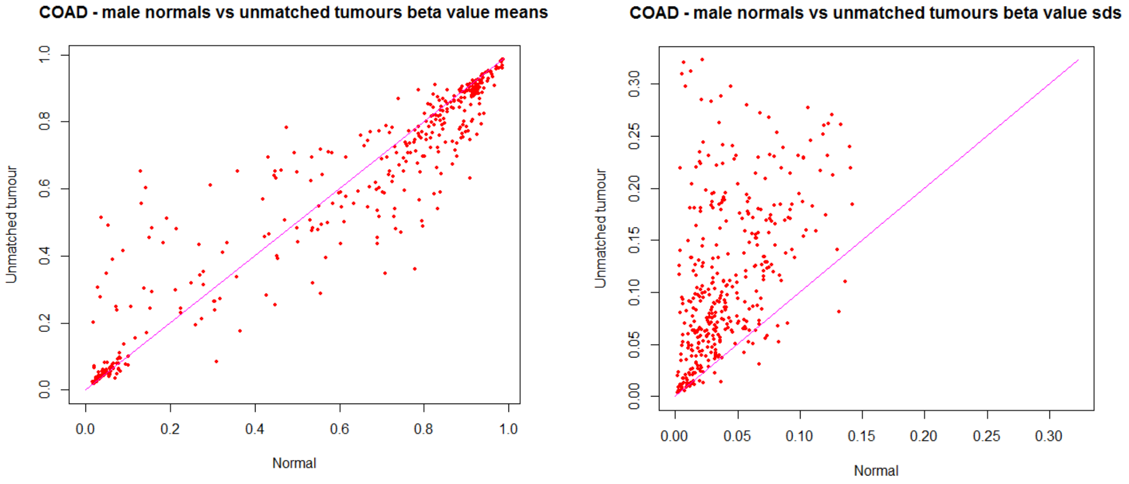


Figure 42 - comparison of unmatched tumour and normal samples (COAD)
Left hand side - plot of mean beta values for all 384 Y chromosome probes for COAD normal (x-axis) vs unmatched tumour samples (y-axis). Right hand side – plot of standard deviation of beta values for all 384 probes for COAD normal (x-axis) vs unmatched tumour samples (y-axis).

3.5.3 Comparison of methylation changes in matched and unmatched tumour samples

I next wanted to check whether the methylation profiles of unmatched and matched tumour samples were similar. For this purpose I calculated, for both groups of tumour samples, the differences in mean beta values between tumour and normal samples for each of the 384 probes. Figure 43 shows that there is very high correlation (r=0.95) for the KIRC dataset:

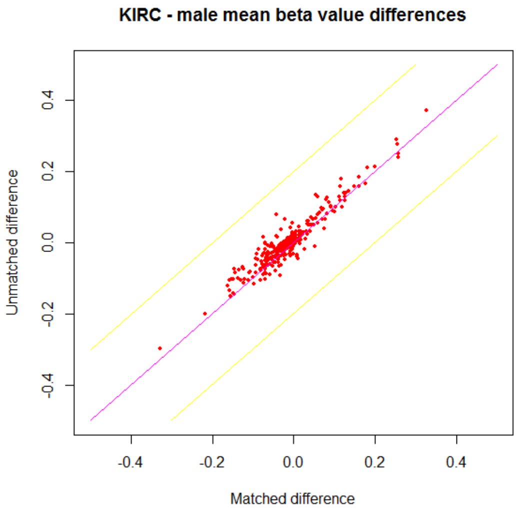


Figure 43 - comparison of mean beta value differences for matched/unmatched tumour samples (KIRC)
Plot of differences in mean beta values for all 384 Y chromosome probes for KIRC tumour samples. X-axis shows difference between mean values for matched tumours and normal samples, y-axis shows corresponding values for unmatched tumours.

There is also high correlation for the COAD and HNSC datasets (r=0.94 and 0.96 respectively), as shown in figure 44, although the correlation is slightly lower for COAD:

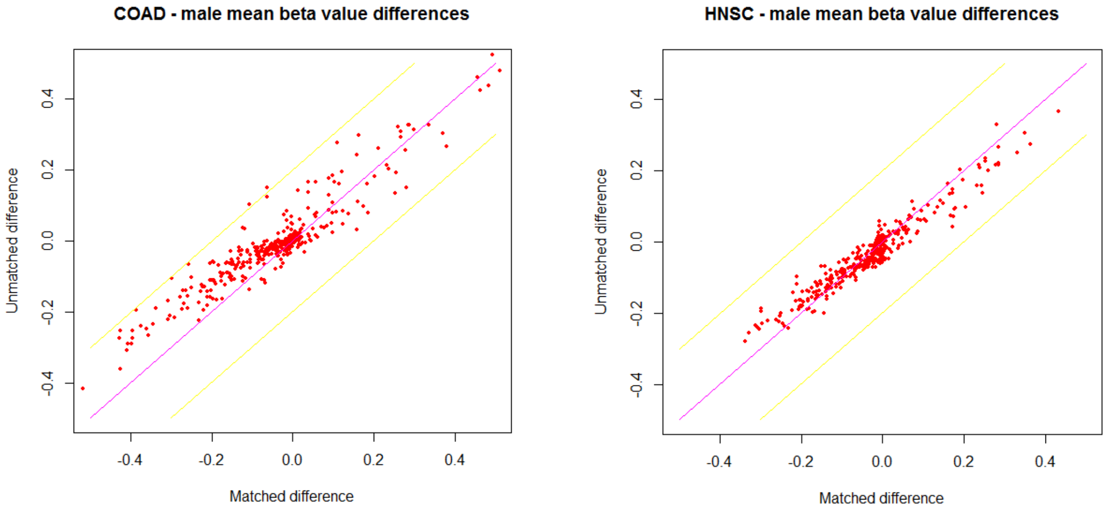


Figure 44 - comparison of mean beta value differences for matched/unmatched tumour samples (COAD & HNSC)
Plot of differences in mean beta values for all 384 Y chromosome probes for COAD (left) and HNSC (right) tumour samples. X-axis shows difference between mean value for matched tumours and normal samples, y-axis shows corresponding values for unmatched tumours.

These plots led me to believe that the methylation changes, relative to normal samples, were similar between unmatched and matched tumour samples.

As a further check, I performed principal components analyses using all Y chromosome beta values and all male samples. Figure 45 shows that the matched and unmatched tumours are intermingled for the KIRC dataset:

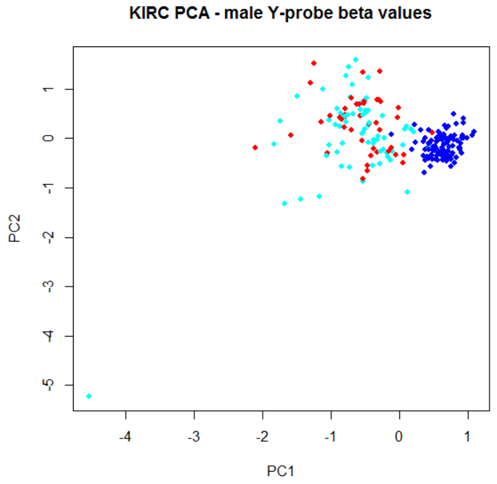


Figure 45 - PCA of all male Y chromosome beta values (KIRC)
Principal components analysis was performed using the KIRC male Y chromosome beta values. The first two principal components for each sample are shown on the x- and y- axes respectively. Matched tumour samples are shown in light blue, unmatched tumour samples are shown in red and normal samples in blue.

Figure 46 shows the corresponding PCA plots for the COAD and HNSC datasets:

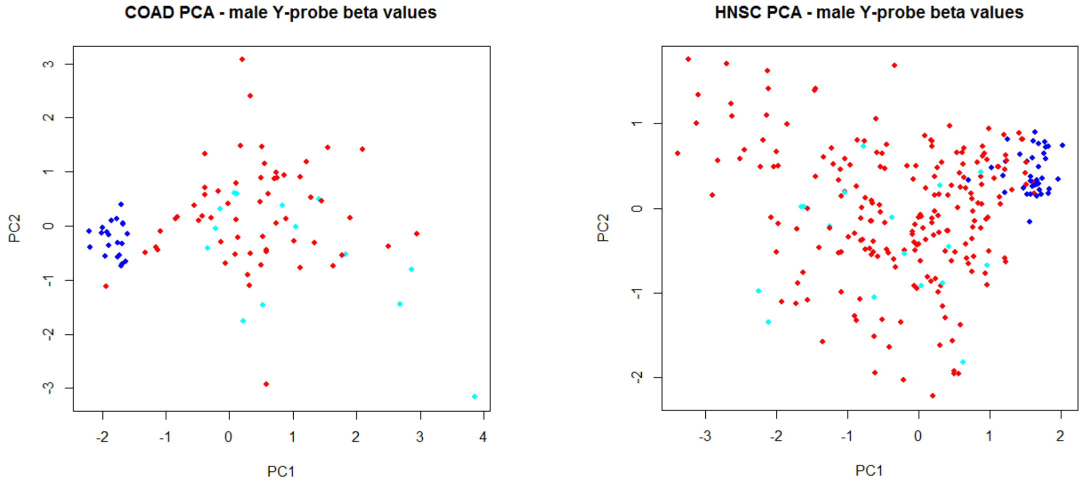


Figure 46 - PCA of all male Y chromosome beta values (COAD & HNSC)
Principal components analysis was performed using the COAD (left) and HNSC (right) male Y chromosome beta values. The first two principal components for each sample are shown on the x- and y- axes respectively. Matched tumour samples are shown in light blue, unmatched tumour samples are shown in red and normal samples in blue.

In both cases, there is again no clear distinction between matched and unmatched tumour samples.

On the basis of my comparisons of mean changes in beta value and my principal components analyses, I concluded that the methylation profiles of matched and unmatched tumours were similar and that I could, therefore, merge the two groups together for my subsequent analyses.

3.5.4 Summary

There was clear evidence of differences in methylation in tumour samples compared with normal samples, both in terms of mean methylation level and variability. I observed similar patterns for both matched and unmatched tumour samples, and decided to group these samples together for my subsequent analyses.

3.6 Site specific methylation changes in tumours

Having considered general methylation changes between normal and tumour samples, I turned my attention to changes at an individual CpG site level.

3.6.1 Locations of largest mean beta value changes

First, for each of the 384 CpG sites, I plotted the difference in mean beta value between tumour and normal samples against location on the Y chromosome. The resultant plots, one for each of the three datasets, are shown in figure 47 overleaf. A hyper-methylated site is indicated by a red line above the x-axis, and a hypo-methylated site by a blue line below the x-axis.

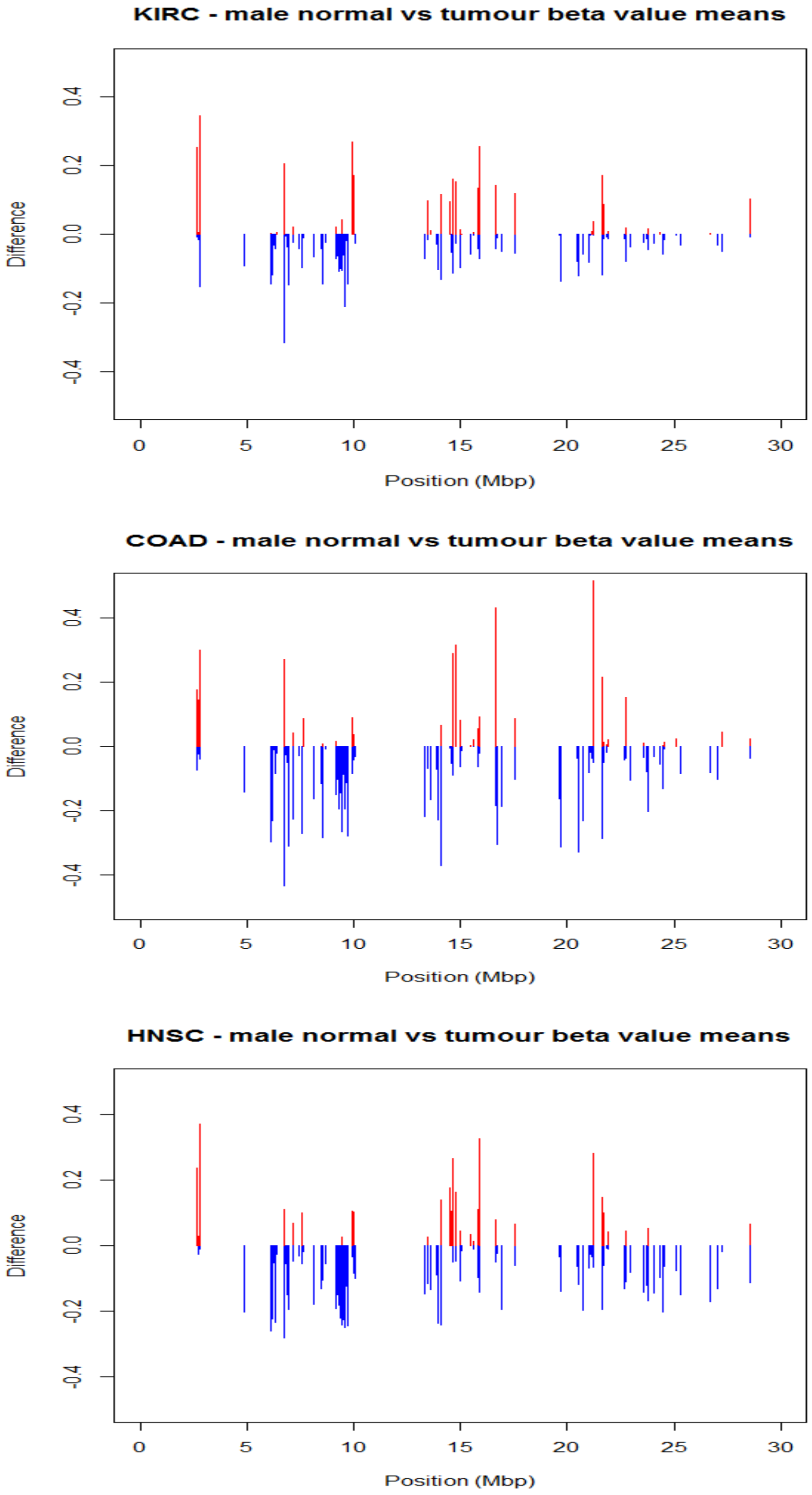


Figure 47 - locations of mean beta value changes in tumours (all datasets)
For each of the 384 Y chromosome probes, the difference in mean tumour beta value and mean normal beta value was calculated. These were then plotted (y-axis) against the position of each probe on the Y chromosome (x-axis). Results for the KIRC, COAD and HNSC datasets are shown separately in the top, middle and bottom plots respectively.

The plots show that hypo-methylation is more common than hyper-methylation in all three datasets, but that hyper-methylation when it occurs tends to be at a higher level. For example, in the KIRC dataset there were 242 probes for which the difference between tumour and normal mean beta values was negative and 142 for which it was positive. However, only two of the 242 negative values were lower than -0.2, whereas six of the positive values were greater than 0.2. COAD samples have the greatest amount of aberrant methylation on average (average hypo-methylation of -0.076 and average hyper-methylation of +0.084), with KIRC samples having the least amount (averages of -0.039 and + 0.047 respectively).

The patterns of hyper-methylation are similar across all datasets. In particular, CpG sites at the following locations have high levels of hyper-methylation:

- 2.656Mbp, covering the SRY gene;
- 2.802Mbp, covering the ZFY gene;
- 6.779Mbp, covering the TBL1Y gene;
- 9.993Mbp, no genes in the vicinity;
- 14.650Mbp, no gene in the vicinity;
- 14.774Mbp, covering the TTTY15 gene;
- 21.239Mbp, covering the TTTY14 gene; and
- 21.665Mbp, covering the BCORL2 gene.

There are a couple of other areas of high hyper-methylation specific to the COAD dataset, at 2.710Mbp (RPS4Y1 gene) and 16.635Mbp (NLGN4Y gene).

There are also locations with a high density of hypo-methylation in all datasets at:

- 6.137Mbp, no genes in the vicinity;
- 6.743Mbp, covering the AMELY gene;
- 6.954Mbp, covering the TBL1Y gene;
- Between 9.306 and 9.747Mbp, covering the TSPY1, TSPY3, TSPY4, RBMY3AP, TTTY1, TTTY8 and TTTY22 genes;
- Between 13.914 and 14.104Mbp, no genes in the vicinity;
- 19.679Mbp, no genes in the vicinity;
- 20.736Mbp, covering the HSFY gene;
- 21.666Mbp, covering the BCORL2 gene;

Additionally, there is evidence of hypo-methylation near the NLGN4Y gene in the COAD and HNSC datasets.

3.6.2 Analysis of large individual beta value differences

I next decided to refine my analysis by looking at the magnitude of methylation changes for each tumour sample and for each probe, and to compare them against a suitable threshold. For each CpG site I compared each tumour sample's beta value against the mean of all normal sample values.

For this purpose I needed to determine an appropriate threshold of change, above which I would determine a CpG site to be either hypo- or hyper- methylated in the tumour sample relative to normal samples. To help me decide on the threshold, I looked at the distribution of differences in beta values (tumour minus normal mean) across all probes and tumour samples for the three datasets. A plot of these distributions is set out in figure 48:

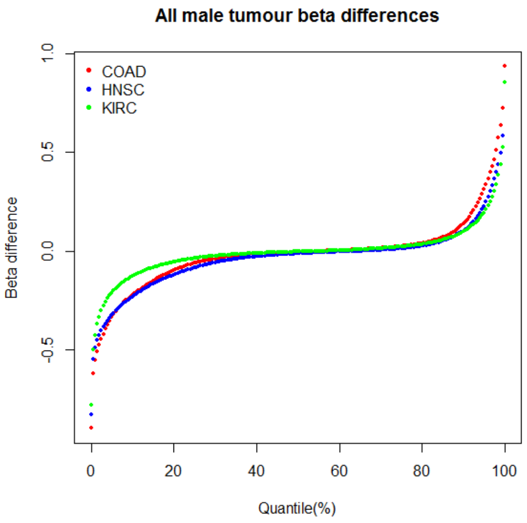


Figure 48 - distribution of methylation changes in all tumour samples and probes
For each of the 384 Y chromosome probes, the difference between tumour beta value and mean normal beta value was calculated for each tumour sample. The plot shows the cumulative percentiles of all of these differences. Results for COAD (red), HNSC (blue) and KIRC (green) are shown separately.

By inspection of the distributions, I calculated that just over 5% of all beta value changes (hypo- or hyper-methylated) across all probes / samples exceeded a threshold of 0.4. I, therefore, decided to use a change in beta value of 0.4 as my chosen threshold. I also used 0.2 as a secondary threshold (this value is consistent with Bibikova's analysis of the precision of the 450k platform[57]).

Using these thresholds I produced heatmaps for each dataset indicating those tumour samples which were either hypo- or hyper-methylated. Figure 49 shows the heatmap for the KIRC tumour samples. Each column represents one of the 384 probes, which are ordered from left to right according to position on the Y chromosome. Each row represents a single tumour sample, and samples have been ordered from bottom to top in increasing order of the overall change in methylation levels across all 384 probes relative to normal samples (i.e. most hyper-methylated overall is at the top). Dark blue indicates hypo-methylation and red indicates hyper-methylation.

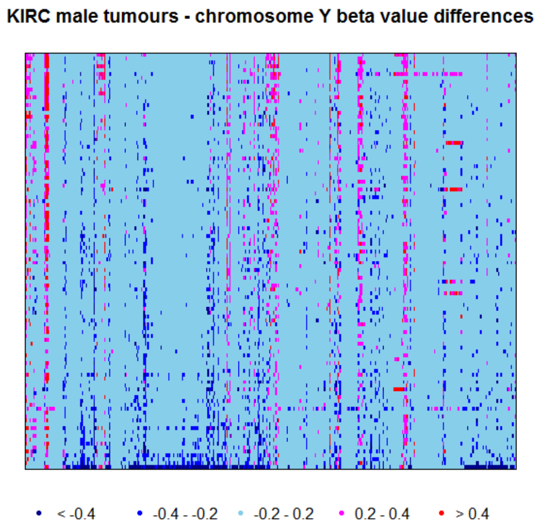


Figure 49 - heatmap of hypo- and hyper-methylated tumour samples (KIRC)
Each column represents one of the 384 probes, which are ordered from left to right according to position on the Y chromosome. Each row represents a single KIRC tumour sample, and samples have been ordered from bottom to top in increasing order of the overall change in methylation levels across all 384 probes relative to normal samples (i.e. most hyper-methylated overall is at the top). Dark blue indicates that a sample was hypo-methylated at that probe and red indicates hyper-methylation.

Figure 50 shows corresponding heatmaps for the COAD and HNSC datasets:

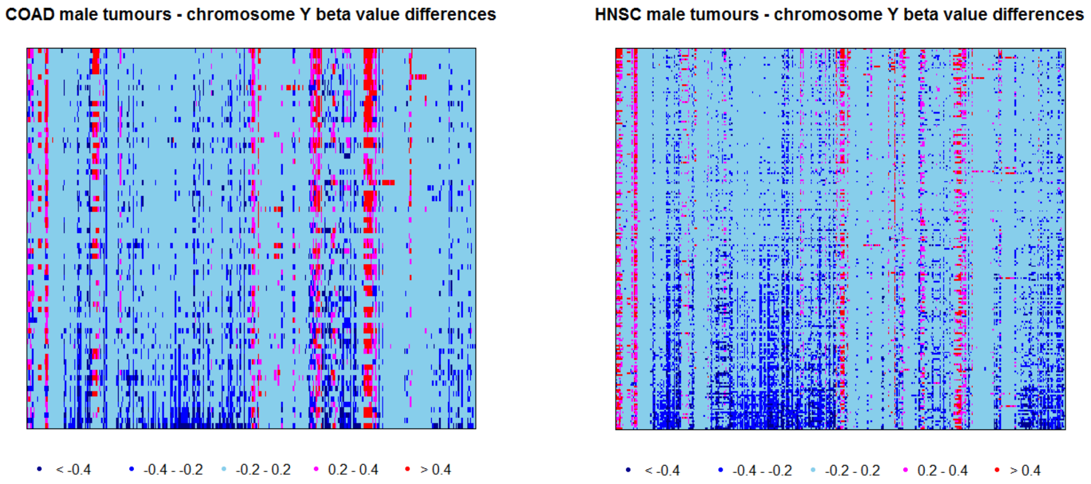


Figure 50 - heatmaps of hypo- and hyper-methylated tumour samples (COAD & HNSC)
Each column represents one of the 384 probes, which are ordered from left to right according to position on the Y chromosome. Each row represents a single COAD (left) or HNSC (right) tumour sample, and samples have been ordered from bottom to top in increasing order of the overall change in methylation levels across all 384 probes relative to normal samples (i.e. most hyper-methylated overall is at the top). Dark blue indicates that a sample was hypo-methylated at that probe and red indicates hyper-methylation.

The heatmaps show discrete bands of hyper-methylation common to large numbers of samples in all datasets, interspersed between more general areas of hypo-methylation. The hyper-methylated areas include sites targeted at the following genes: SRY, RPS4Y1 (not KIRC), ZFY, TBL1Y (in particular for COAD), TTTY15, NLGN4Y (in particular for COAD), TTTY14 (not KIRC), BCORL2, CYorf15A, and EIF1AY (in particular COAD).

A list of the CpG sites with the highest proportions of hyper-methylated samples is included in Appendix 2. 15 of these 36 sites are located in CpG islands, with a further 15 in north shores. 13 are also located within 200bp of a gene's transcriptional start site, with a further 10 within 1,500bp. 11 are located in gene bodies.

The regions of hypo-methylation are less uniform amongst samples compared with the hyper-methylated areas. Genes included in the most heavily hypo-methylated areas are PCDH11Y, AMELY, TBL1Y, TTTY16, RBMY3AP, TTTY1, NLGN4Y (not KIRC) and BCORL2.

A list of the CpG sites with the highest proportions of hypo-methylated samples is included in Appendix 3. This time less than half of the sites are near a transcriptional start site, and the most common island region is "ocean".

3.6.3 Summary

Across all three datasets, hypo-methylation is more common than hyper-methylation. The latter tends to be concentrated in discrete regions of the Y chromosome, interspersed among more general regions of hypo-methylation. CpG sites targeted at several genes are located within these regions of hypo- and hyper-methylation. COAD samples have the highest level of aberrant methylation on average, and KIRC samples have the lowest amount.

3.7 Conclusions

The key conclusions from my analyses of differential methylation between normal and tumour samples are as follows:

- 32 of the 416 Y chromosome probes were found to be subject to cross-hybridisation issues with other parts of the genome and have been removed, leaving 384 for further analyses;
- The 384 remaining probes are targeted at CpG sites within 55 genes, including 12 genes of particular interest;
- 18 samples appeared to have incorrectly labelled genders, and have therefore been removed;
- Between 40% and 50% of all tumour samples across the three datasets appear to have suffered some loss of the Y chromosome;
- Loss of the Y chromosome appears to have had a distorting effect on methylation measurements, in particular for CpG sites which are either very highly or very lowly methylated in normal samples;
- Batch effects were detected in the total intensity measurements, and have been (partially) corrected;
- COAD normal samples tended to be more highly methylated than their HNSC and KIRC counterparts;
- There are a small number of CpG sites which had significant methylation differences between different normal tissue samples;
- The methylation levels of tumour samples are generally more variable than those of normal samples;

- In all three datasets there are discrete regions of the Y chromosome which are hyper-methylated in sizeable subsets of tumours, and these regions are interspersed between larger regions of more general hypo-methylation;
- Tumour hyper-methylation is particularly prevalent in sites targeted at several genes, namely SRY, RPS4Y1, ZFY, TBL1Y, TTTY15, NLGN4Y, TTTY14, BCORL2, CYorf15A, and EIF1AY;
- Tumour hypo-methylation is also prevalent in sites targeted at a small number of genes, namely PCDH11Y, AMELY, TBL1Y, TTTY16, RBMY3AP, TTTY1, NLGN4Y and BCORL2.

I follow up on these hypo- and hyper-methylated genes in chapter 5, when I investigate the effects of both aberrant methylation and loss of Y chromosome on gene expression. In the next chapter, I analyse in detail the level of Y chromosome loss which has been incurred across the three datasets.

4. Copy number changes in the Y chromosome

In this chapter I investigate copy number changes to the Y chromosome (in particular loss).

In the first instance, I used the methylation data to estimate copy number, as set out in Feber's paper[14]. My analysis in chapter 3 hinted that there were sizeable proportions of samples in each dataset which appeared to have suffered some loss of the Y chromosome. I then used TCGA's copy number data to corroborate my findings.

An overview of the process I have followed is provided in figure 51:

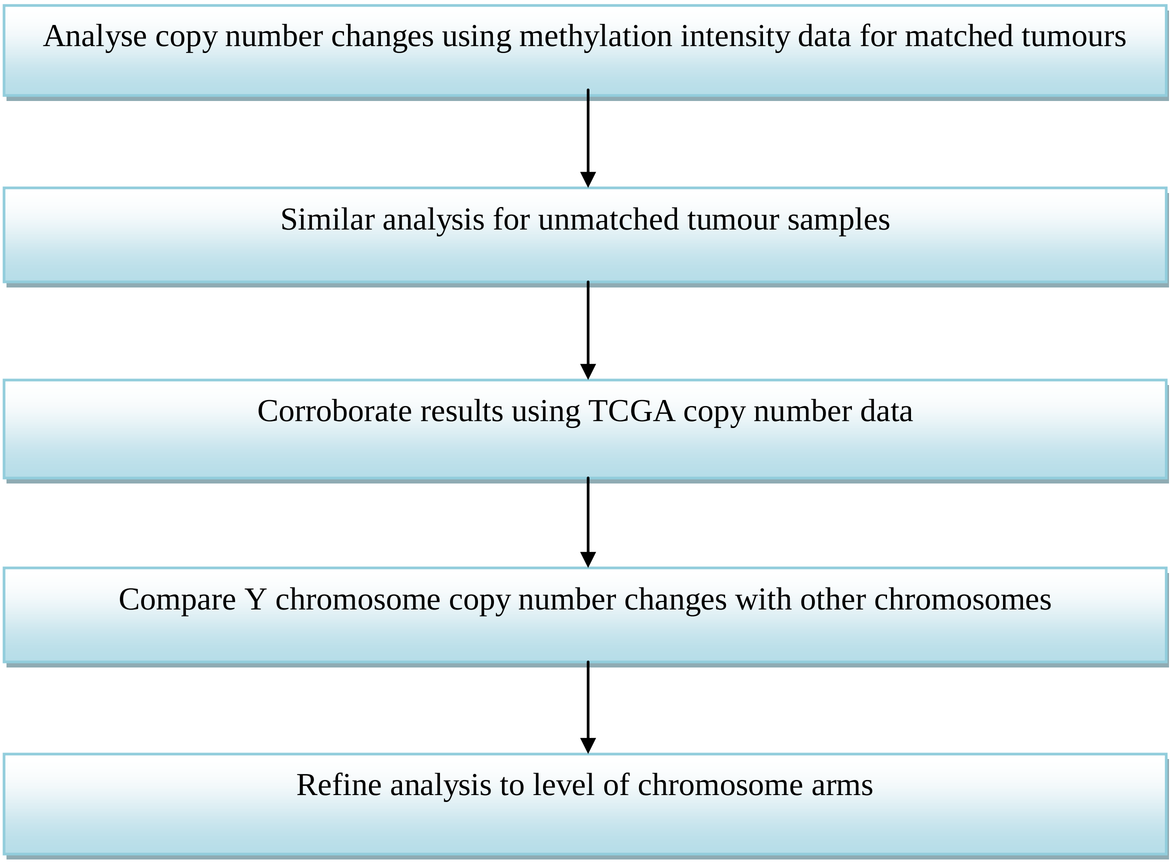

Analyse copy number changes using methylation intensity data for matched tumours
Similar analysis for unmatched tumour samples
Corroborate results using TCGA copy number data
Compare Y chromosome copy number changes with other chromosomes
Refine analysis to level of chromosome arms

Figure 51 - overview of experimental process for copy number analysis

I set out details of each of the above stages in the following sections.

4.1 Analysis of matched tumours using methylation data

I initially investigated copy number changes in matched tumour samples, for which I could make direct comparisons with the corresponding normal samples.

For each matched tumour / normal pair, I calculated, for each of the 384 probes, the ratio of the total intensity (methylated plus unmethylated) for the tumour sample over the normal sample. My logic was that the total intensity measurement provides a quantification of the amount of DNA present in a sample for the particular probe in question. Hence, the ratio of tumour / normal intensity gives an estimate of the copy number for the tumour sample relative to the normal sample.

Figure 52 contains a heatmap of the resultant total intensity ratios for all of the KIRC matched tumour samples:

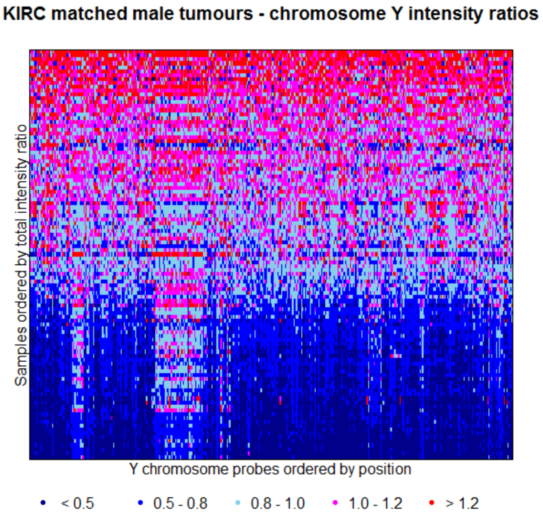


Figure 52 - heatmap of Y chromosome intensity ratios for KIRC matched tumours
For each KIRC matched tumour / normal pair, for each of the 384 Y chromosome probes, the ratio of the total intensity for the tumour sample over the normal sample was calculated. Each column of the heatmap represents one of the 384 probes, which are ordered from left to right according to their position on the Y chromosome (not to scale). Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of the total intensity ratio across all 384 probes. Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2) as shown in the legend.

Each column of the heatmap represents one of the 384 probes, which are ordered from left to right according to their position on the Y chromosome (not to scale). Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of the total intensity ratio across all 384 probes (i.e. those samples with the lowest overall ratios are located at the bottom). Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2).

The large band of dark blue at the bottom of the heatmap suggests that there is a sizeable proportion (39%) of KIRC matched tumour samples which appear to have lost at least 20% of their Y chromosome DNA across most of the entire length spanned by the 384 probes. There are a small number of probes for which there is little dark blue for any samples – these probes are discussed below.

Figure 53 shows similar heatmaps for the COAD and HNSC matched tumours:

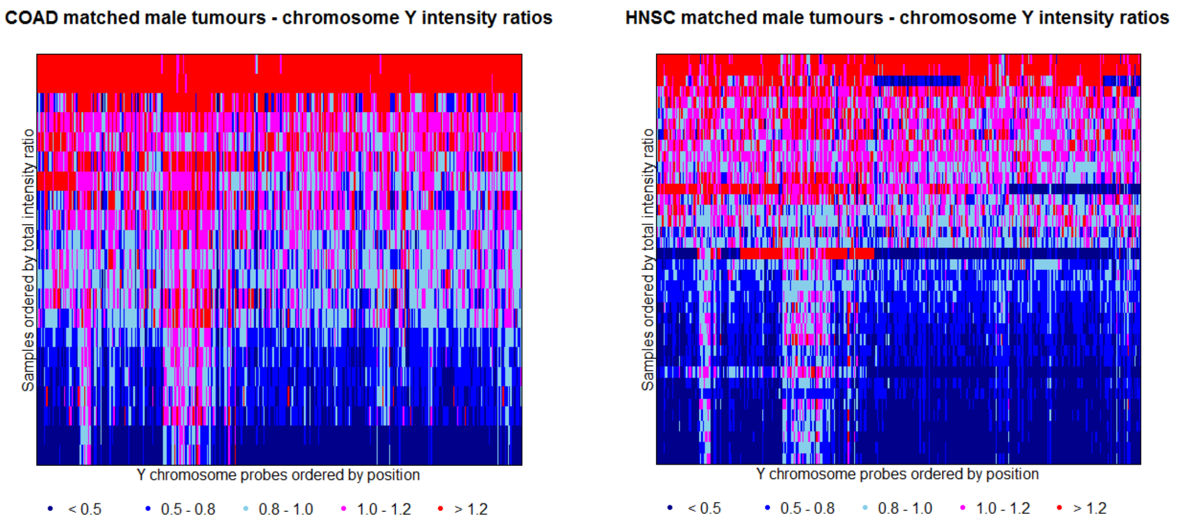


Figure 53 - heatmaps of Y chromosome intensity ratios for COAD & HNSC matched tumours
For each COAD (left) and HNSC (right) matched tumour / normal pair, for each of the 384 Y chromosome probes, the ratio of the total intensity for the tumour sample over the normal sample was calculated. Each column of the heatmap represents one of the 384 probes, which are ordered from left to right according to their position on the Y chromosome (not to scale). Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of the total intensity ratio across all 384 probes. Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2) as shown in the legend.

A similar pattern is observed for the COAD and HNSC datasets, with a slightly lower proportion (33%) of COAD samples and a slightly higher proportion (50%) of HNSC samples appearing to have suffered overall copy number loss of chromosome Y of at least 20%.

It occurred to me that the above results could have been caused simply by lower amounts of tumour DNA having been used (compared with normal samples). To check this was not the case, I produced similar heatmaps for other chromosomes. Figure 54 shows, as an example, the heatmap for chromosome 1 for the KIRC matched tumour samples, which are included in the same order form top to bottom based on their Y chromosome total ratios:

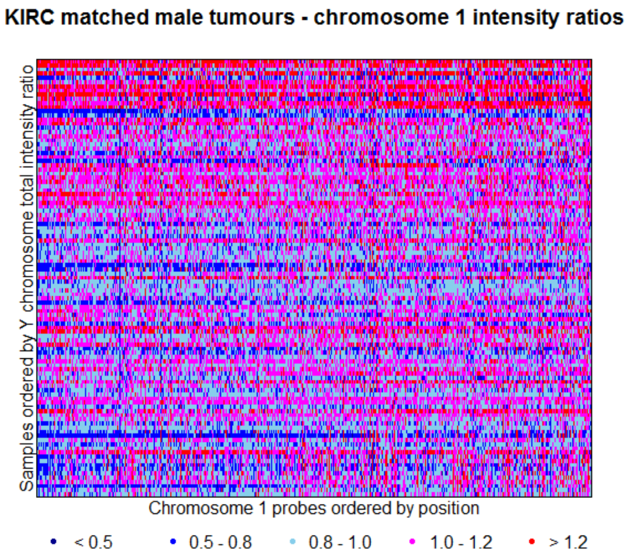


Figure 54 - heatmap of chromosome 1 intensity ratios for KIRC matched tumours
For each KIRC matched tumour / normal pair, for each of the chromosome 1 probes, the ratio of the total intensity for the tumour sample over the normal sample was calculated. Each column of the heatmap represents one of the probes, which are ordered from left to right according to their position on chromosome 1 (not to scale). Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of the total intensity ratio across all Y chromosome probes. Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2) as shown in the legend.

Unlike the heatmap shown in figure 52, there is no clear pattern of increasing intensity ratios from bottom to top. For example, there was no significant overlap between samples with the lowest 39% chromosome 1 ratios and the 39% of samples which appeared to have suffered at least 20% loss of the Y chromosome (chi-squared p-value = 0.1524).

Similar results were obtained for the COAD and HNSC datasets and for other chromosomes. Hence the results shown in figures 52 and 53 are not caused by technical artefacts.

4.1.1 Probes with apparently reduced copy number loss

A common feature within the heatmaps for all three matched tumour datasets (and also the unmatched samples – see next section) is the small subset of probes for which there appears to be much reduced copy number loss.

I inspected the intensity ratios for all samples / probes to find out more about the affected probes. There are two main groups of probes affected, one in the region between 6.11 and 6.17 mega-bases, and the other between 9.17 and 9.38 mega-bases. Both of these subsets of probes lie within the so-called "ampliconic" region of the Y chromosome[15].

The ampliconic region contains highly repetitive DNA sequences. For a sample of the affected probes, I checked where they aligned to the human genome using NCBI's nucleotide BLAST facility, and discovered that each probe aligned to multiple parts of the Y chromosome with high (>90%) homology. Hence, this would explain why the intensity measurements do not reduce for samples with copy number loss to the same extent as the measurements for probes in other parts of the chromosome.

4.1.2 Summary

A sizeable proportion (between one third and one half) of matched tumour samples in each dataset appeared to have suffered at least 20% loss of chromosome Y DNA. Around 10% of the probes are located in the ampliconic region of chromosome Y, and therefore gave misleading estimates of copy number loss.

4.2 Analysis of unmatched tumours using methylation data

I then considered the unmatched tumour samples. As I could not make a direct comparison for each tumour sample with its corresponding normal sample, I instead calculated a benchmark “normal” total intensity for each of the 384 probes against which I would compare all of the tumour samples. For this purpose I used the median total intensity across all normal samples for each probe. I then performed similar analyses as I had for the matched tumour samples.

Figure 55 contains a heatmap of the resultant total intensity ratios for all of the KIRC unmatched tumour samples:

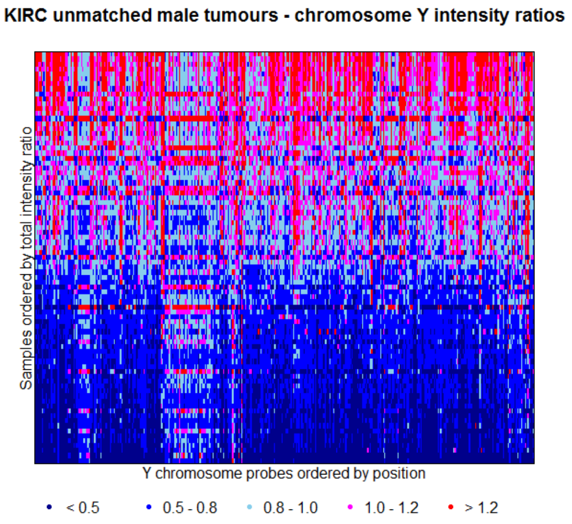


Figure 55 - heatmap of Y chromosome intensity ratios for KIRC unmatched tumours
For each KIRC unmatched tumour sample, for each of the 384 Y chromosome probes, the ratio of the total intensity for the tumour sample over the median total intensity across all normal samples was calculated. Each column of the heatmap represents one of the 384 probes, which are ordered from left to right according to their position on the Y chromosome (not to scale). Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of the total intensity ratio across all 384 probes. Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2) as shown in the legend.

Figure 56 shows similar heat-maps for the COAD and HNSC unmatched tumours:

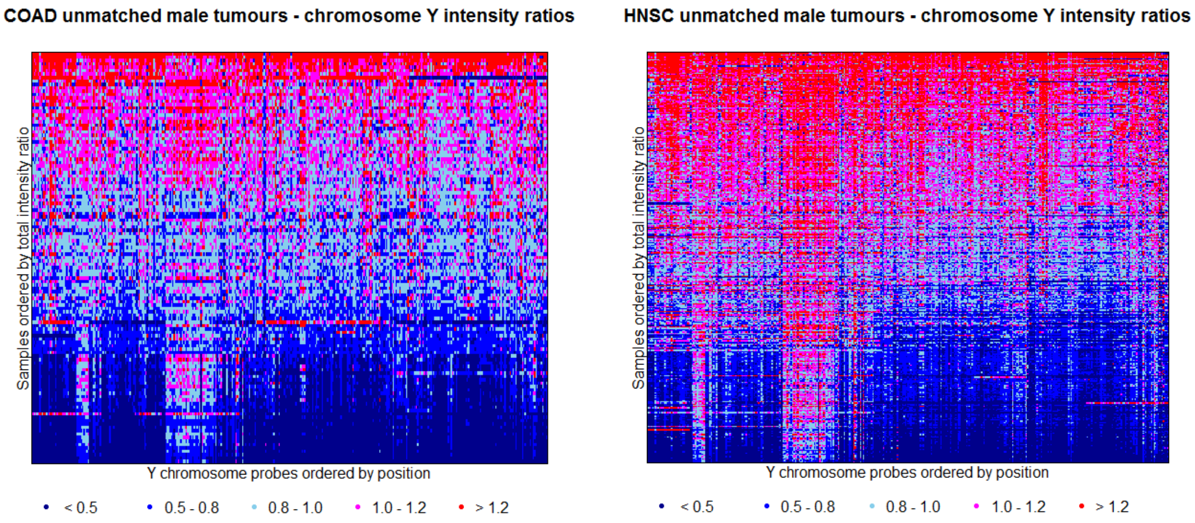


Figure 56 - heatmaps of Y chromosome intensity ratios for COAD & HNSC unmatched tumours
For each COAD (left) and HNSC (right) unmatched tumour sample, for each of the 384 Y chromosome probes, the ratio of the total intensity for the tumour sample over the median total intensity across all normal samples was calculated. Each column of the heatmap represents one of the 384 probes, which are ordered from left to right according to their position on the Y chromosome (not to scale). Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of the total intensity ratio across all 384 probes. Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2) as shown in the legend.

In all cases a similar pattern was observed as had been seen previously for the matched tumour samples. 46% of unmatched KIRC tumour samples, 38% of COAD samples and 34% of HNSC samples appear to have suffered overall copy number loss of chromosome Y of at least 20%.

Summary

Similar patterns of Y chromosome loss were observed for the unmatched tumour samples.

4.3 Corroboration of results using TCGA copy number data

The above analyses were all carried out using TCGA's level 2 methylation intensity data. To check the accuracy of the results, I also downloaded TCGA's level 3 copy number data, as described in Chapter 2. I then compared the two sets of results, separately for matched and unmatched tumour samples.

4.3.1 Comparison for matched tumour samples

TCGA's level 3 copy number data contains segmented results for each sample and chromosome – i.e. for each sample, each chromosome has been broken up into segments within which the copy number has been assessed to be invariant. So that I could compare these data against my methylation-based results, I produced an alternative segmentation based on the total intensity ratios derived from the methylation data using the DNAcopy package in R (default parameters).

For each matched tumour sample, I then plotted the segmented results from both the copy number data and my methylation analyses. Generally, for all datasets, the two sets of segmentation results were very similar - some examples for the KIRC dataset are shown in figure 57:

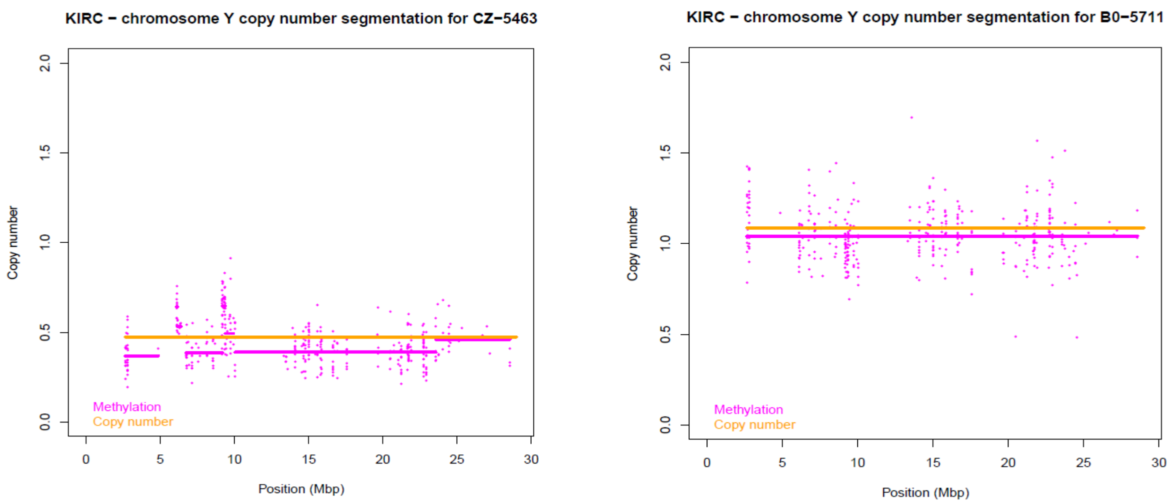


Figure 57 - example comparisons of segmented copy number results for KIRC matched tumours
Examples of Y chromosome segmented copy number data for two KIRC tumour samples. Both plots show position on the Y chromosome on the x-axis and copy number on the y-axis. In each case, the orange line represents TCGA's level 3 segmented copy number data. The pink line represents the segmentation calculated using the DNAcopy package in R, based on the methylation total intensity ratios.

I also calculated, for each set of segmented results, a copy number "index", being the average copy number across the whole length of the Y chromosome (for the methylation-based results I considered only the region spanned by the 384 probes). Figure 58 shows a plot of the resultant index figures for the KIRC matched tumour samples:

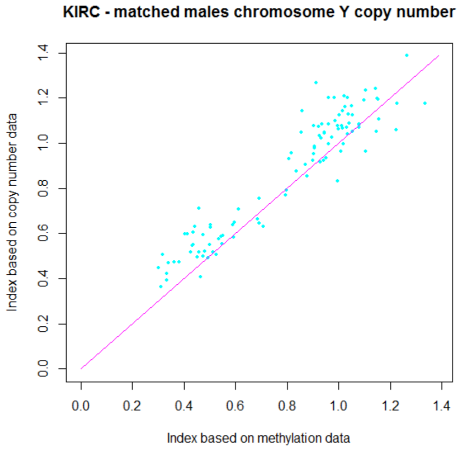


Figure 58 - comparison of Y chromosome copy number indices for KIRC matched tumours
Using the segmented copy number data, an average copy number "index" was calculated for the whole Y chromosome for all KIRC matched tumour samples. The plot shows the index value based on the methylation data (x-axis) against the index value based on TCGA's copy number data (y-axis) for each sample.

Figure 59 shows the comparable plots for the COAD and HNSC matched tumours:

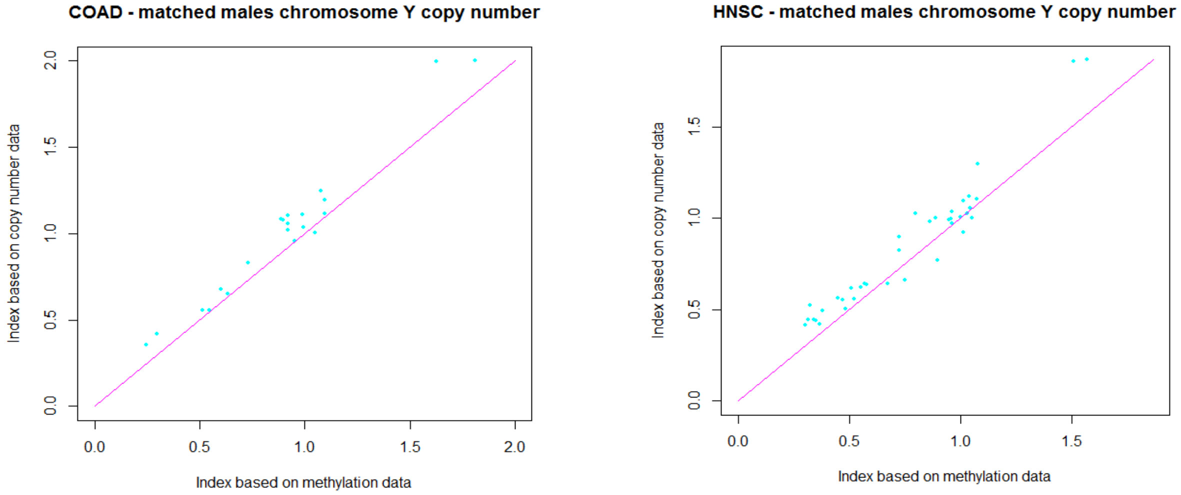


Figure 59 - comparison of Y chromosome copy number indices for COAD & HNSC matched tumours
Using the segmented copy number data, an average copy number "index" was calculated for the whole Y chromosome for all COAD (left) and HNSC (right) matched tumour samples. Each plot shows the index value based on the methylation data (x-axis) against the index value based on TCGA's copy number data (y-axis) for each sample.

In all cases there is a very good correlation between the index measurements (Pearson correlation = 0.94, 0.98 and 0.96 respectively, all p-values < 0.0001), although the methylation-based index values are generally slightly lower than those based on TCGA's copy number data.

4.3.2 Comparison for unmatched tumour samples

I then performed similar comparisons for the unmatched tumour samples. Once again, there was generally very good correspondence between TCGA's segmented copy number data and my methylation-based segmented results. Figures 60 and 61 show comparisons of the copy number index figures for the KIRC, COAD and HNSC unmatched tumour samples:

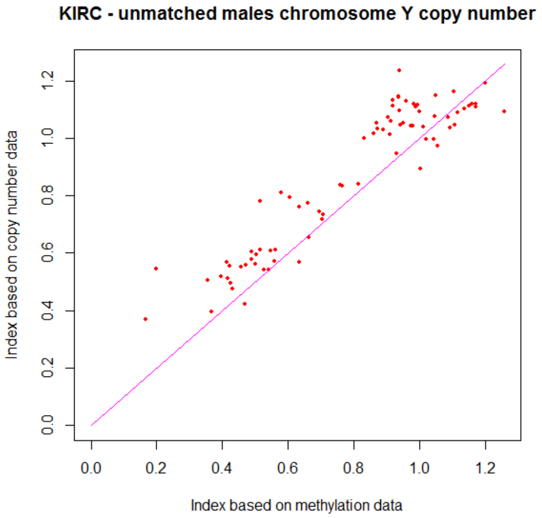


Figure 60 - comparison of Y chromosome copy number indices for KIRC unmatched tumours
Using the segmented copy number data, an average copy number "index" was calculated for the whole Y chromosome for all KIRC unmatched tumour samples. The plot shows the index value based on the methylation data (x-axis) against the index value based on TCGA's copy number data (y-axis) for each sample.

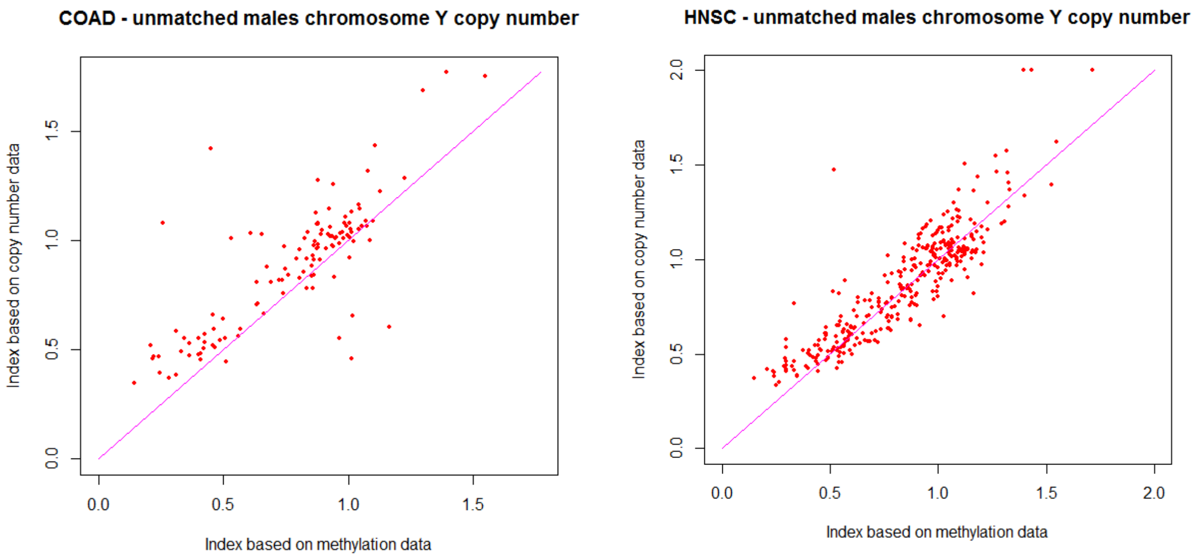


Figure 61 - comparison of Y chromosome copy number indices for COAD & HNSC unmatched tumours
Using the segmented copy number data, an average copy number "index" was calculated for the whole Y chromosome for all COAD (left) and HNSC (right) unmatched tumour samples. Each plot shows the index value based on the methylation data (x-axis) against the index value based on TCGA's copy number data (y-axis) for each sample.

In all cases there was again a very good correlation between the index measurements (Pearson correlation = 0.93, 0.79 and 0.89 respectively, all p-values < 0.0001), although there is a bit more variability for small numbers of COAD and HNSC samples.

4.3.3 Summary

The copy number results based on the methylation data were generally corroborated by TCGA's segmented copy number data.

4.4 Comparison with other chromosomes

I was then interested to know how the levels of Y chromosome loss compared with copy number changes for other chromosomes. For this purpose I used TCGA's segmented copy number data, and grouped all tumour samples together, rather than differentiating between matched and unmatched samples.

For each sample, I calculated the copy number index (as above) for each chromosome. Figure 62 shows a heatmap of the results for the KIRC tumour samples. The samples have been ordered according to their Y chromosome copy number indices, increasing from bottom to top. Each chromosome is represented by a separate column. The extent of Y chromosome loss is clearly much more substantial than for any other chromosome – in particular 8% of samples have a Y chromosome index value of less than 0.5.

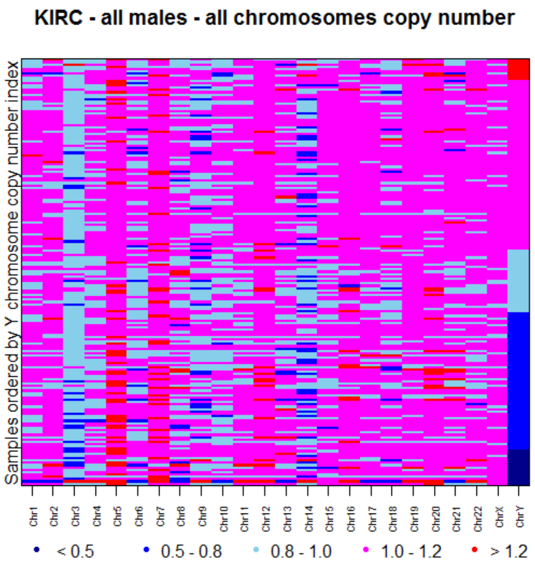


Figure 62 - heatmap of copy number indices by chromosome for KIRC male tumours
Using TCGA's segmented copy number data, an average copy number "index" was calculated across the whole length of each chromosome for all KIRC tumour samples. Each column of the heatmap represents one of the chromosomes in numerical order. Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of their Y chromosome index values. Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2) as shown in the legend.

Figure 63 shows similar heatmaps for the COAD and HNSC male tumours:

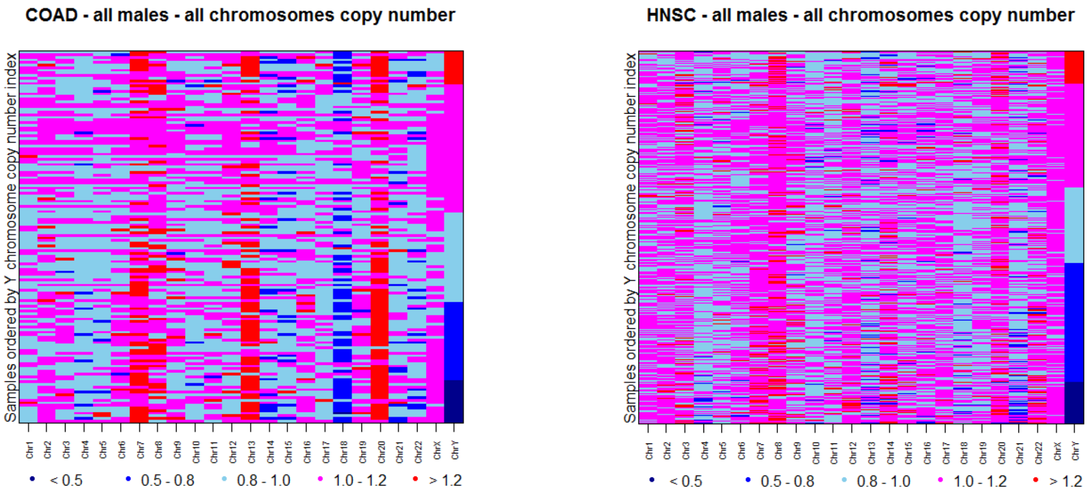


Figure 63 - heatmaps of copy number indices by chromosome for COAD & HNSC male tumours
Using TCGA's segmented copy number data, an average copy number "index" was calculated across the whole length of each chromosome for all COAD (left) and HNSC (right) tumour samples. Each column of both heatmaps represents one of the chromosomes in numerical order from left to right. Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of their Y chromosome index values. Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2) as shown in the legend.

Again, the Y chromosome has suffered the greatest copy number losses.

Summary

Across all three datasets, the extent of loss of the Y chromosome was greater than the loss incurred by any other chromosome.

4.5 Analysis of changes at chromosomal arm level

Next I refined the analysis in the previous section by looking at copy number changes at the level of chromosome arms, rather than whole chromosomes. Research has shown that copy number changes often occur at arm-level[98]. I repeated the analysis of the previous section, but this time calculating copy number indices for each chromosomal arm. Figure 64 shows a heatmap of the results for the KIRC tumour samples:

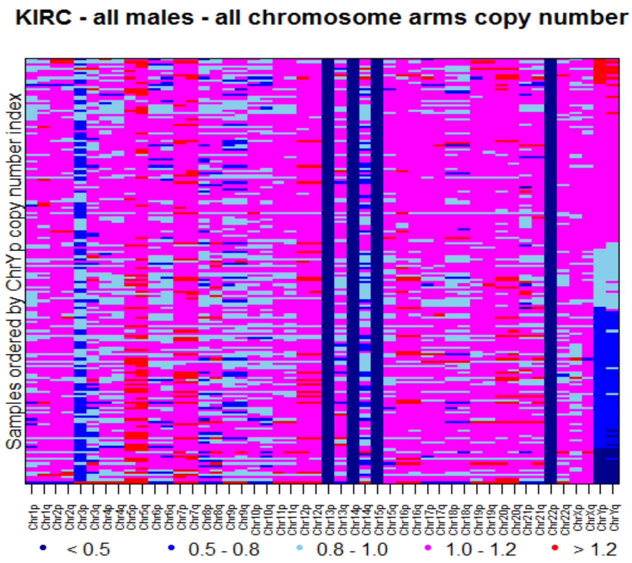


Figure 64 - heatmap of copy number indices by chromosome arm for KIRC male tumours
Using TCGA's segmented copy number data, an average copy number "index" was calculated across the whole length of each chromosome arm for all KIRC tumour samples. Each column of the heatmap represents one of the chromosome arms in numerical order. Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of their Y chromosome short arm index values. Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2) as shown in the legend.

The samples have been ordered according to their Y chromosome short-arm copy number indices, increasing from bottom to top. The extent of loss is similar in both Y chromosome arms, and is again more substantial than for any other chromosomal arm. However, there are some other chromosomal arms, notably 3p, which have also suffered losses of at least 20% in a large proportion of tumour samples.

Note that the solid blue columns for chromosome arms 13p, 14p, 15p and 22p are caused by a complete lack of data for these arms, rather than being indicative of significant copy number losses. These chromosomes are all acrocentric.

Figure 65 shows similar heatmaps for the COAD and HNSC male tumours:

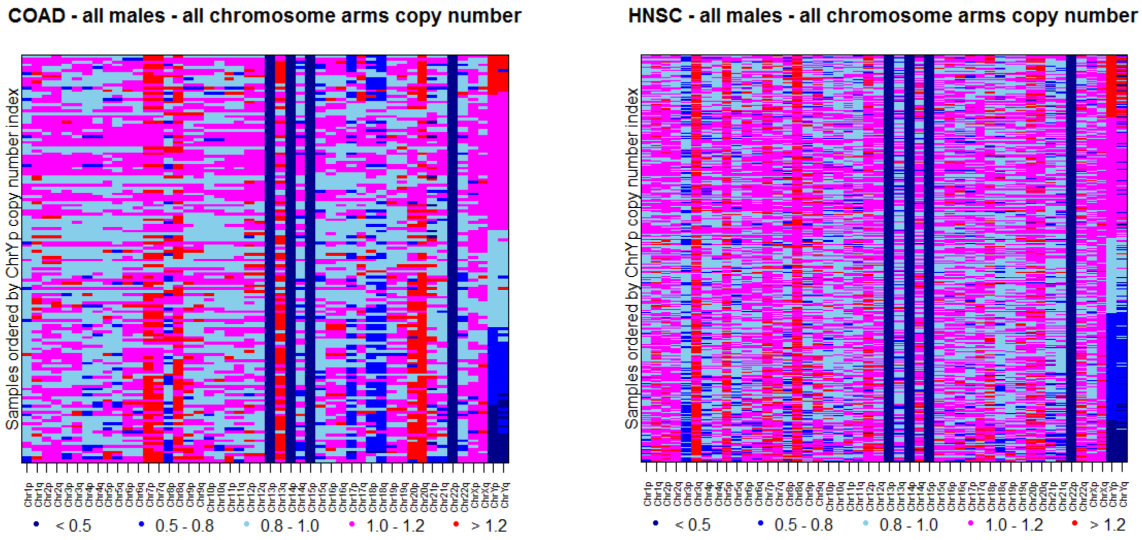


Figure 65 - heatmaps of copy number indices by chromosome arm for COAD & HNSC male tumours
Using TCGA's segmented copy number data, an average copy number "index" was calculated across the whole length of each chromosome arm for all COAD (left) and HNSC (right) tumour samples. Each column of the heatmap represents one of the chromosome arms in numerical order. Each row represents one of the tumour samples, which are ordered from bottom to top in increasing order of their Y chromosome short arm index values. Each ratio has been colour-coded according to four different thresholds (0.5, 0.8, 1.0 and 1.2) as shown in the legend.

Both Y chromosome arms have again suffered greater losses than any other chromosomal arms, although there are other arms which have suffered significant losses in individual cancers (both arms of chromosome 18 for COAD, and 3p for HNSC).

Summary

In all three datasets, both arms of the Y chromosome have suffered similar amounts of loss, and these losses are much greater than those for any other chromosomal arm. However, there are some other chromosome arms which have also been subject to significant levels of loss within one of more of the datasets.

4.6 Conclusions

The key conclusions from my analyses of Y chromosome copy number changes in male tumours are as follows:

- the methylation intensity data indicated that sizeable proportions of male tumour samples (between one third and one half) had suffered at least 20% loss of chromosome Y DNA in all three datasets;
- a small proportion of the 384 methylation probes appeared to cross-hybridise with multiple regions of the Y chromosome, and therefore gave misleading results for the copy number analyses;
- the methylation-based results were corroborated by TCGA’s level 3 copy number data;
- the Y chromosome was unique in the level of loss suffered across all chromosomes;
- both arms of the Y chromosome had been subject to similar levels of copy number loss;
- arm-level losses are also greater in the Y chromosome than for any other chromosome, although there are some other arms, particularly 3p in kidney and head and neck cancer, which have also suffered significant losses.

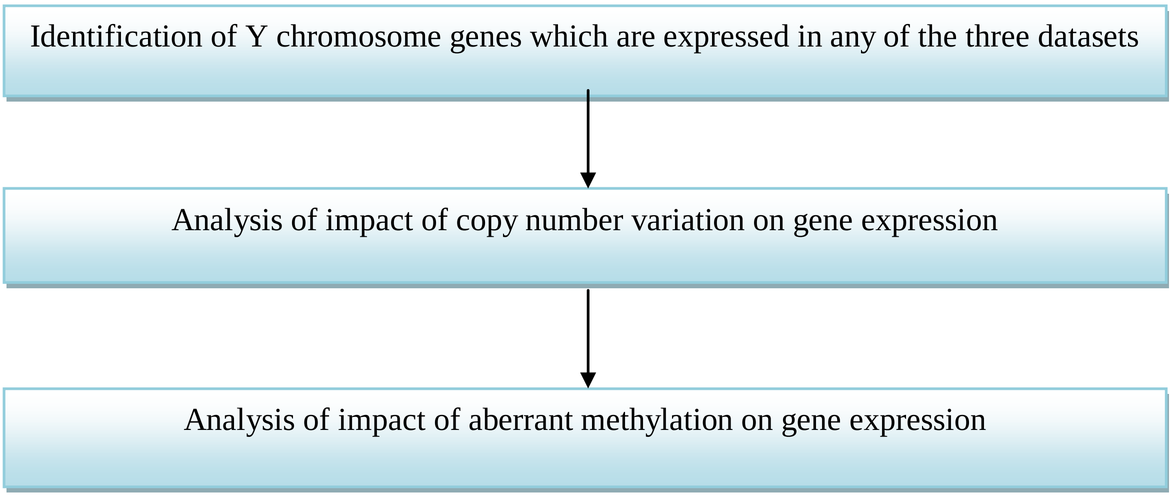
In the next chapter, I will consider the impact of both aberrant methylation and copy number loss on expression of Y chromosome genes.

5. Expression of Y chromosome genes

In the previous chapter I showed that significant subsets of tumours in all three datasets appear to have lost the Y chromosome in some cells. Furthermore, in chapter 3 I showed that there are several genes on the Y chromosome which are affected by aberrant methylation, in particular hyper-methylation, in some tumour samples. In this chapter I investigate the potential biological implications of these phenomena by analysing expression levels of Y chromosome genes.

My analyses are based on TCGA’s level 3 RNA-sequencing data. I have concentrated on samples and genes which are represented in both the RNA-seq and 450k methylation data.

An overview of the process I have followed is provided in figure 66:



Identification of Y chromosome genes which are expressed in any of the three datasets

Analysis of impact of copy number variation on gene expression

Analysis of impact of aberrant methylation on gene expression

Figure 66 - overview of experimental process for gene expression analyses

I set out details of each of the above stages in the following sections.

5.1 Y chromosome gene expression data

5.1.1 Data used

As described in chapter 2, I used TCGA's level 3 RNA-seq data produced using the Illumina HiSeq 2000 RNA Sequencing platform (Version 2) for the purposes of my gene expression analyses. These data provide genome-wide measures of gene expression (in the form of read counts) which have been normalised between samples.

I chose this version of the RNA-seq data as it has the best overlap with the 450k methylation data (out of the options available) in terms of samples covered. However, there are still samples for which methylation data were available but RNA-seq data were unavailable (and vice-versa). To ensure consistency with my methylation results, I only considered samples for which both methylation and RNA-seq data were available.

The numbers of samples which I have included in my gene expression analyses are set out in table 7 (numbers in brackets are those used for my previous methylation analyses):

	COAD	**HNSC**	**KIRC**
Tumour samples	134 (142)	372 (374)	188 (188)
Normal samples	11 (21)	15 (38)	18 (105)
Total	145 (163)	387 (412)	206 (293)

Table 7 - summary of data numbers for gene expression analyses

A small number of COAD and HNSC tumour samples were present in the methylation data but not in the RNA-seq data. However, much larger numbers of normal samples were not present in the RNA-seq data for all three datasets (especially KIRC).

5.1.2 Y chromosome genes which are expressed

I previously identified 55 Y chromosome genes which are represented on the 450k methylation platform. Expression data were also available for most of these genes. However, on inspection of the expression values, I discovered that only 16 genes were expressed in any samples within the three datasets.

15 of the 16 genes were expressed in all three datasets, albeit not always at the same level. The genes concerned are: RPS4Y1, ZFY, TBL1Y, PRKY, TTTY15, USP9Y, DDX3Y, UTY, TMSB4Y, NLGN4Y, NCRNA00185, TTTY14, CYorf15A, KDM5D, and EIF1AY. This list includes the two putative tumour suppressor genes and further 10 genes suggested as being important for male viability, which I highlighted in section 3.

One gene, SRY, was only expressed in the HNSC dataset.

My subsequent analyses concentrate on these 16 genes.

5.1.3 Summary

15 genes on the Y chromosome were expressed in all three datasets. The SRY gene was also expressed in the HNSC dataset only.

5.2 Analysis of impact of copy number variation on expression

I showed in chapter 4 that loss of the Y chromosome appeared to have occurred in sizeable proportions of male patients in all three datasets. I, therefore, concentrated initially on the extent to which chromosomal loss appears to impact on gene expression.

5.2.1 General association between copy number and gene expression

I first considered how expression levels of the 16 genes varied with general levels of Y chromosome copy number. To measure the latter, I used the copy number index calculated in chapter 4 based on TCGA's segmented copy number data.

Figure 67 contains a heatmap of the expression values for all the KIRC male tumour samples. For each of the 16 genes, samples have been categorised into expression value quartiles. Each row contains a different sample, and samples have been ordered (from bottom to top) in increasing order of their Y chromosome copy number index. Each column represents an individual gene, and genes have been ordered (from left to right) according to their position on the Y chromosome.

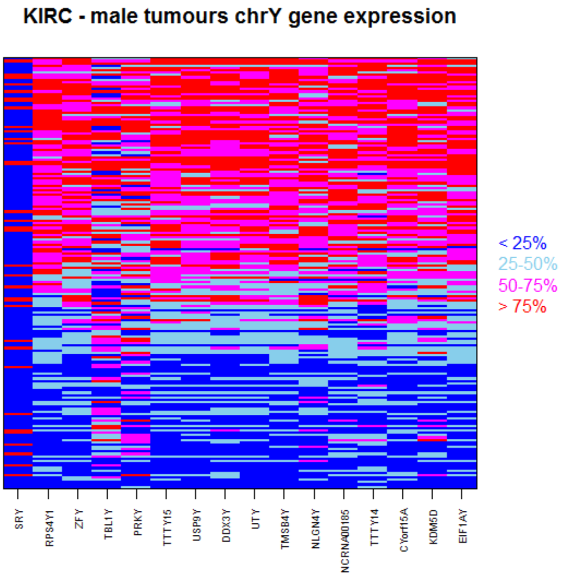


Figure 67 - heatmap of Y chromosome gene expression values for KIRC tumours
For each of the 16 genes being considered, all male KIRC tumour samples have been categorised into expression value quartiles (legend indicates colour-coding). Each row contains a different sample, and samples have been ordered (from bottom to top) in increasing order of their Y chromosome copy number index. Each column represents an individual gene, and genes have been ordered (from left to right) according to their position on the Y chromosome.

The heatmap shows that for most of the 16 genes, expression values within the KIRC dataset generally increase with increasing copy number. The three main exceptions to this pattern are genes SRY (not expressed), TBL1Y and PRKY.

Figure 68 shows corresponding heatmaps for the COAD and HNSC datasets:

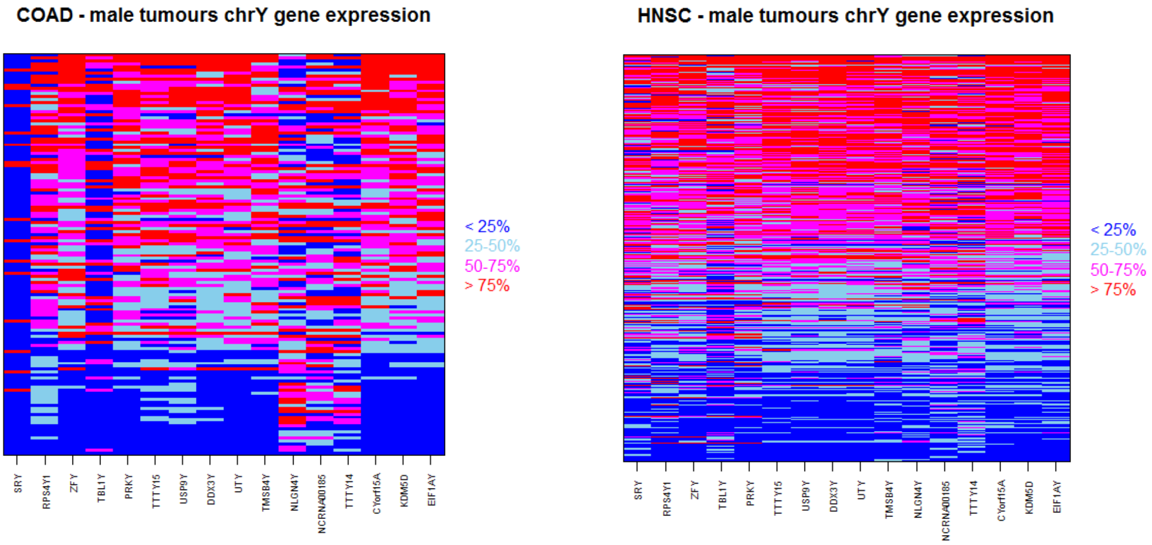


Figure 68 - heatmaps of Y chromosome gene expression values for COAD & HNSC tumours
For each of the 16 genes being considered, all male COAD (left) and HNSC (right) tumour samples have been categorised into expression value quartiles (legend indicates colour-coding). Each row contains a different sample, and samples have been ordered (from bottom to top) in increasing order of their Y chromosome copy number index. Each column represents an individual gene, and genes have been ordered (from left to right) according to their position on the Y chromosome.

For the HNSC dataset, expression values for all 16 genes generally increase with increasing copy number, although the pattern is less clear for TBL1Y in particular. Within the COAD dataset, there are five genes (SRY, TBL1Y, NLGN4Y, NCRNA00185 and TTTY14) for which there is no clear pattern of expression increasing with copy number.

5.2.2 Probe-level association between copy number and expression

To refine the above results, I then used the methylation intensity data to analyse the associations between expression and copy number at the level of individual probes.

Out of the 384 probes which I have previously used for my methylation analyses, 152 are targeted at CpG sites linked to the 16 genes under consideration. For each relevant probe, I plotted total intensity against gene expression, and calculated the (Pearson) correlation between the two values. Example plots for the ZFY gene are shown in figures 69 (KIRC) and 70 (COAD and HNSC):

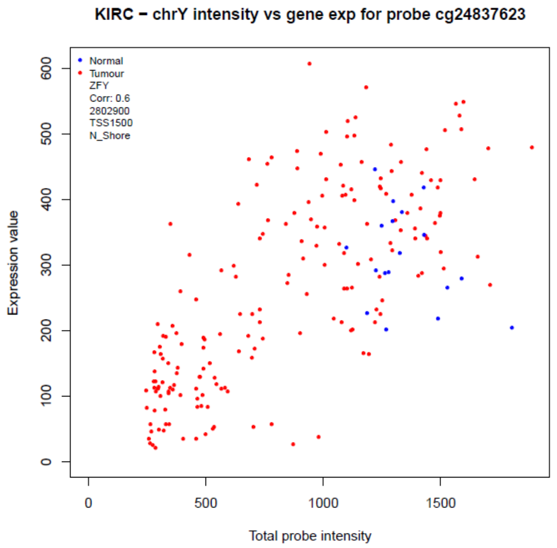


Figure 69 - example plot of total intensity vs ZFY gene expression (KIRC)
Example plot for methylation probe cg24837623 of total methylation intensity (x-axis) versus gene expression (y-axis) for all KIRC male samples. Tumour samples are plotted in red, and normal samples are plotted in blue.

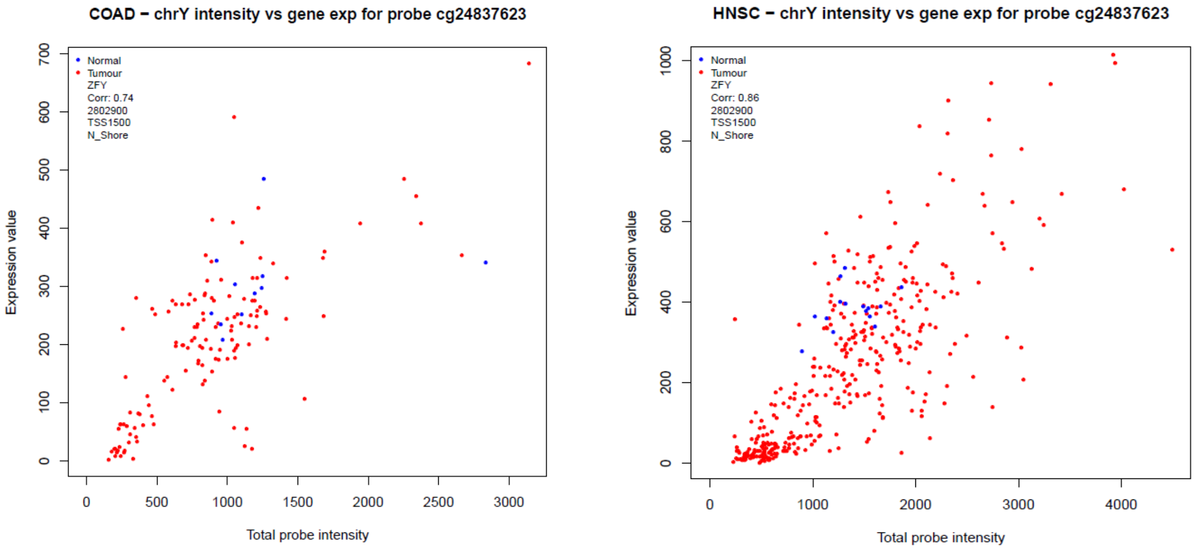


Figure 70 - example plots of total intensity vs ZFY gene expression (COAD & HNSC)
Example plots for methylation probe cg24837623 of total methylation intensity (x-axis) versus gene expression (y-axis) for all COAD (left) and HNSC (right) male samples. Tumour samples are plotted in red, and normal samples are plotted in blue.

For all three datasets there is a high correlation between the total intensity for probe cg24837623 and ZFY expression (r = 0.60, 0.74 and 0.66 respectively, all p-values < 0.0001). The same is true for other probes targeting the ZFY gene, although there is some variation in correlation values between probes. It is also true for many of the other 16 genes, but not for all. Table 8 shows details of the highest correlations observed for each gene across each of the three datasets:

			Highest correlation		
Gene	**Position (Mbp)**	**Number of 450k probes**	**COAD**	**HNSC**	**KIRC**
SRY	2.65	6	N/A	0.72	N/A
RPS4Y1	2.71	5	0.45	0.75	0.68
ZFY	2.80	18	0.78	0.86	0.66
TBL1Y	6.78	11	0.16	0.33	0.38
PRKY	7.14	9	0.62	0.75	0.42
TTTY15	14.77	7	0.58	0.72	0.74
USP9Y	14.81	1	0.54	0.72	0.65
DDX3Y	15.02	13	0.72	0.83	0.69
UTY	15.45	7	0.76	0.81	0.80
TMSB4Y	15.81	7	0.56	0.61	0.65
NLGN4Y	16.63	18	0.12	0.67	0.59
NCRNA00185	21.04	2	0.45	0.56	0.49
TTTY14	21.10	10	0.32	0.53	0.59
CYorf15A	21.73	11	0.66	0.76	0.78
KDM5D	21.87	8	0.64	0.79	0.56
EIF1AY	22.74	17	0.75	0.72	0.82

Table 8 - summary of probe intensity / gene expression correlations

The correlations in table 8 generally confirm the results from the earlier heatmap analyses, which show that copy number variation is an important factor in gene expression levels. All correlations were statistically significant ($p < 0.05$, and mostly < 0.0001) except for TBL1Y (COAD) and NLGN4Y (COAD).

5.2.3 Summary

Copy number variation of the Y chromosome is highly correlated with gene expression for most of the 16 genes across all three datasets. Anomalous genes include, in particular, TBL1Y (all datasets) and NLGN4Y (COAD).

5.3 Analysis of impact of aberrant methylation on expression

Of course, although copy number variation appears to have a significant influence on gene expression, it is not the only factor that can contribute to such variation. Therefore, building on my work in chapter 3, I next considered whether aberrant methylation of the Y chromosome may also impact on gene expression levels.

5.3.1 Genes whose expression may be affected by aberrant methylation

In chapter 3 I identified a number of CpG sites which were either hypo- or hyper-methylated in some tumour samples (relative to normal samples) for at least one of the datasets. Some of these sites are within the 16 genes which are expressed. In particular:

- Sites within SRY, RPS4Y1, ZFY, TBL1Y, PRKY, TTTY15, DDX3Y, NLGN4Y, TTTY14, CYorf15A and EIF1AY were hyper-methylated in some tumour samples; and
- Sites within TBL1Y and NLGN4Y were hypo-methylated in some tumour samples.

To determine whether aberrant methylation of any of these sites may be having an effect on gene expression, I produced, for each dataset, 2- dimensional and 3-dimensional plots of total probe intensity, beta value and gene expression for all of the sites within the genes mentioned above. I then scrutinised each of the graphs to identify any sites where hypo- or hyper-methylation appeared to be linked to unusually low / high gene expression. Figure 71 shows examples of these plots for one of the CpG sites within the RPS4Y1 gene (for the HNSC dataset):

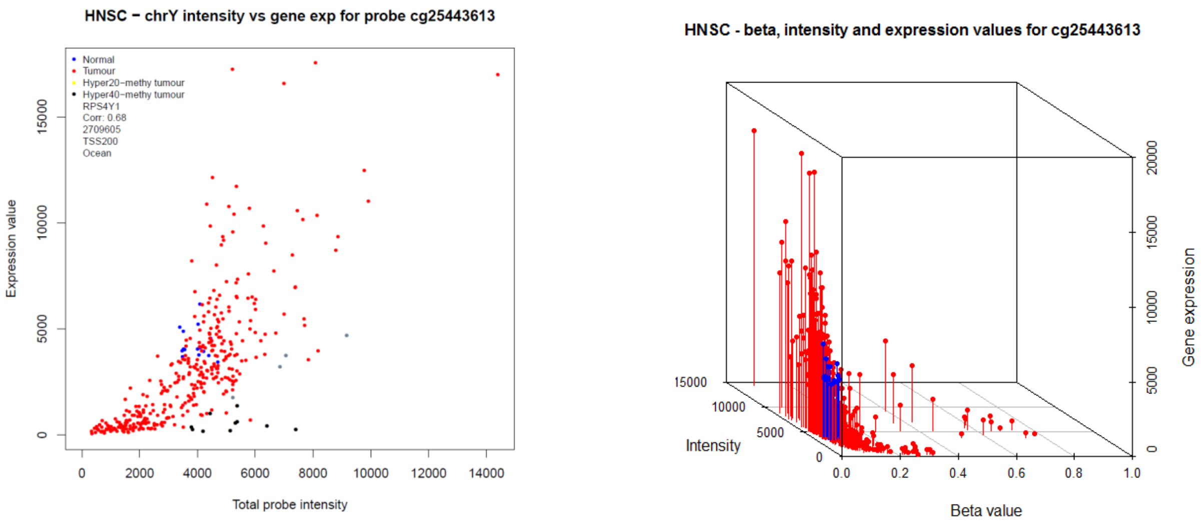


Figure 71 - example 2D & 3D plots of probe intensity, beta value and gene expression
Example 2D (left) and 3D (right) plots for methylation probe cg25443613. The 2D plot shows total methylation intensity (x-axis) versus gene expression (y-axis) for all HNSC male samples - non-hyper-methylated tumour samples are plotted in red, hyper-methylated tumour samples are plotted in black (40% threshold) or grey (20% threshold) and normal samples are plotted in blue. The 3D plot shows total methylation intensity plotted against gene expression and beta value – tumour samples are shown in red and normal samples are shown in blue.

The 2D plot on the left shows probe intensity plotted against gene expression, and has been colour-coded to show any hyper-methylated samples. The black dots represent hyper-methylated samples (using the 40% threshold), and the plot clearly shows a small number of samples which are hyper-methylated, and have very low gene expression values (< 2,000, below all the normal samples), although their total intensity levels are over 3,000, consistent with the normal samples.

The 3D plot on the right provides a different visualisation of the same pattern – the height of each “stick” indicates the level of gene expression, and the beta value measure is shown on the horizontal axis. Once again there is a small group of samples with higher than normal beta values and low expression.

By looking at all of the charts, I determined whether aberrant methylation of any of the sites I had identified appeared to be associated with unusual gene expression in at least one of the datasets. I concluded that methylation of many of the sites did not appear to be associated with aberrant gene expression. However, for five of the genes there were some sites which merited further analysis – the genes concerned are SRY, RPS4Y1, TBL1Y, NLGN4Y and TTTY14. I consider each of these genes in turn in the following sections.

5.3.2 Analysis of SRY gene methylation

SRY is located on the short arm of chromosome Y at around 2.65Mbp. It is a potentially interesting gene, since it only appears to be expressed in the HNSC dataset. Even then, the expression levels (as measured by normalised read counts) are very low compared to other genes, but the magnitude of the differences between the HNSC dataset and the other two datasets suggest that there is some SRY activity in the former.

The box-plots in figure 72 compare expression levels across the three datasets:

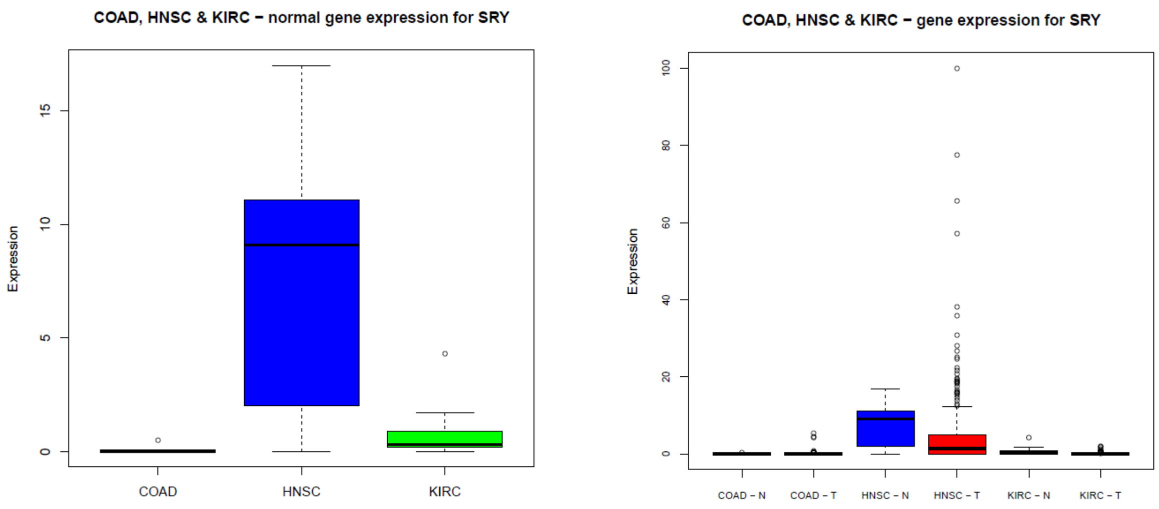


Figure 72 - comparison of SRY expression levels across datasets
Boxplots of expression values of the SRY gene for all male samples, broken down by the three datasets. The left hand plot shows the distribution of expression values for normal samples – COAD, HNSC and KIRC datasets are colour-coded red, blue and green respectively. The right hand plot shows the distribution of expression values for all samples, split by sample type and dataset. The distributions for normal samples are shown in blue, and those for tumour samples are shown in red.

HNSC tumour samples have generally lower SRY expression than normal samples (Kruskal-Wallis p-value = 0.0115), although there are some tumours with higher expression.

As I showed previously, SRY expression levels in the HNSC dataset appear to be influenced by copy number variation. To illustrate this further, figure 73 shows a plot of total probe intensity versus gene expression for one of the methylation probes targeted at SRY (the plot is colour-coded as shown in the legend – in particular, black dots represent tumour samples which are hyper-methylated at the 40% threshold level, and grey dots represent those hyper-methylated according to the lower 20% threshold):

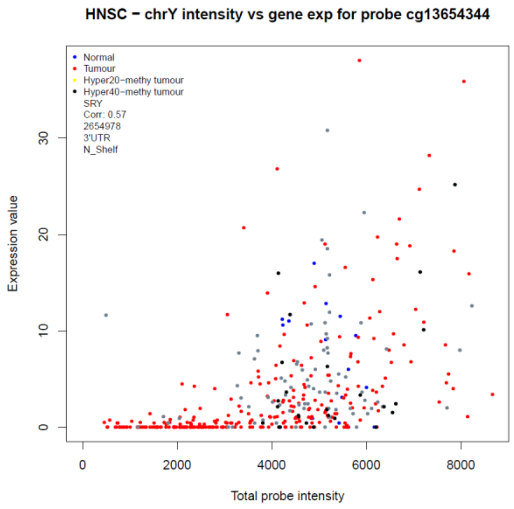


Figure 73 - example plot of total intensity and SRY expression (HNSC)
Scatterplot for methylation probe cg13654344 located within the SRY gene. The plot shows total methylation intensity (x-axis) versus gene expression (y-axis) for all HNSC male samples - non-hyper-methylated tumour samples are plotted in red, hyper-methylated tumour samples are plotted in black (40% threshold) or grey (20% threshold) and normal samples are plotted in blue.

In total, there are six methylation probes targeted at CpG sites within SRY. Methylation levels are generally slightly lower in the HNSC normal samples compared with the other two datasets. An example for probe cg13654344 (Kruskal-Wallis p-value < 0.0001) is shown in figure 74:

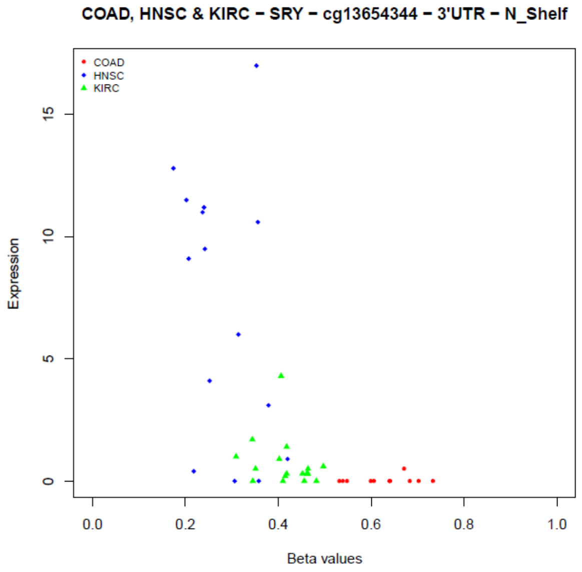


Figure 74 - plot of beta values for cg13654344 vs SRY gene expression
Scatterplot for all normal male samples of methylation beta value for probe cg13654344 (x-axis) versus expression value for the SRY gene (y-axis). COAD, HNSC and KIRC samples are shown in red, blue and green respectively.

This lower level of methylation could be a factor in the higher gene expression observed in the HNSC normal samples compared with the other datasets. There is a negative correlation between beta values and expression values (r = -0.61, p-value < 0.0001).

All six CpG sites are hyper-methylated in a subset of HNSC tumour samples. In particular, the four middle probes (cg04169747, cg11898347, cg27636129 and cg09595415), which are located near the transcriptional start site, are hyper-methylated together (with the odd exception) in around one quarter of the tumour samples (hyper-methylation is also observed in the COAD and KIRC datasets).

There is some evidence that this hyper-methylation may be associated with reduced gene expression. Taking probe cg27636129 as an example, figure 75 shows 2D and 3D plots of probe intensity, beta value and gene expression:

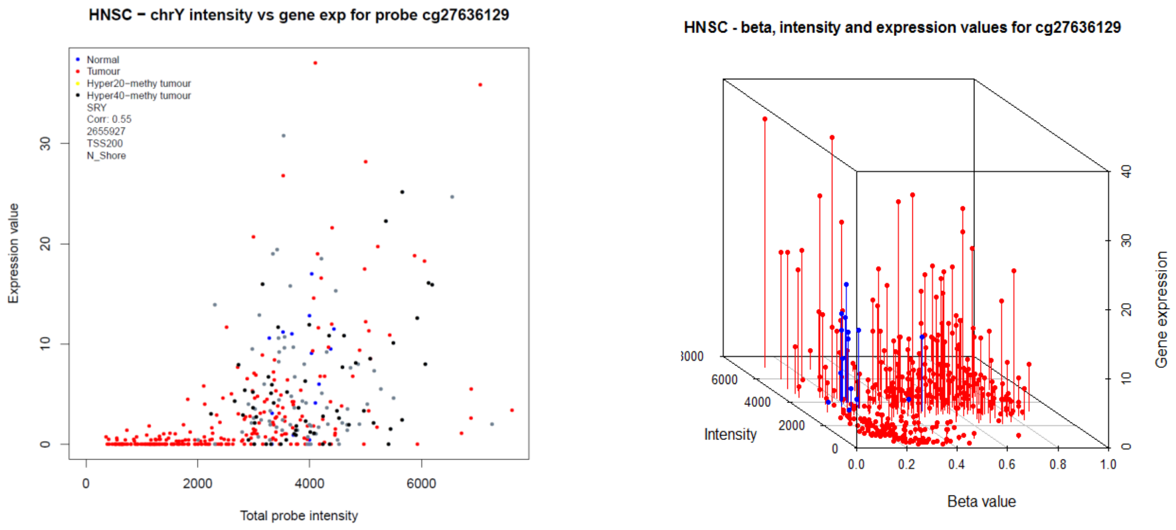


Figure 75 - plots of probe intensity, beta value and gene expression for cg27636129 (HNSC)
2D (left) and 3D (right) plots for methylation probe cg27636129 located within the SRY gene. The 2D plot shows total methylation intensity (x-axis) versus gene expression (y-axis) for all HNSC male samples - non-hyper-methylated tumour samples are plotted in red, hyper-methylated tumour samples are plotted in black (40% threshold) or grey (20% threshold) and normal samples are plotted in blue. The 3D plot shows total methylation intensity plotted against gene expression and beta value – tumour samples are shown in red and normal samples are shown in blue.

There is a subset of hyper-methylated cases which have low expression values (< 5) but do not have low total intensity values (> 3,000), which suggests that hyper-methylation may be a factor in reduced SRY gene expression for some HNSC tumour samples, in addition to reduced expression caused by chromosomal loss. However, there are other hyper-methylated samples which do not have low gene expression, which suggests that there may also be other mechanisms by which gene expression is reduced in tumours.

5.3.3 Analysis of RPS4Y1 gene methylation

RPS4Y1 is located close to SRY at around 2.71Mbp. It is expressed at slightly higher levels in HNSC normal samples compared with COAD and KIRC (Kruskal-Wallis p-value = 0.0002), and its expression is clearly impacted by copy number variation as shown earlier in this chapter. Figure 76 compares expression levels across the datasets:

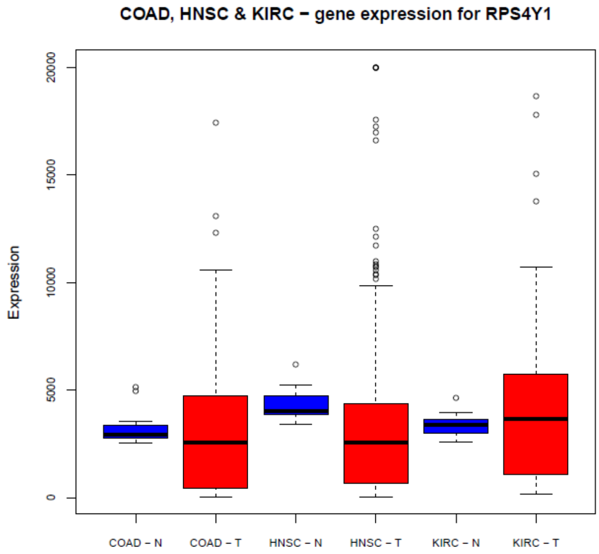


Figure 76 - comparison of RPS4Y1 expression levels across datasets
Boxplots of expression values of the RPS4Y1 gene for all male samples, broken down by sample type and dataset. The distributions for normal samples are shown in blue, and those for tumour samples are shown in red.

Expression levels in tumour samples are more variable than in normal samples for each dataset (Levene p-value = 0.0021, 0.0088 and < 0.0001 for COAD, HNSC and KIRC respectively). Also, expression levels in HNSC tumour samples are lower than in HNSC normal samples (Kruskal-Wallis p-value = 0.0022).

There are five methylation probes targeted at CpG sites within RPS4Y1, and three of these (cg01375382, cg25443613 and cg01311227), all of which are located near the transcriptional start site, are hyper-methylated in a small number of both COAD & HNSC tumour samples.

In both datasets, the same samples are hyper-methylated at each site, and these samples all show low gene expression levels. In figure 71 I showed example plots for cg25443613 in the HNSC dataset, and in figure 77 I show corresponding plots for the COAD dataset:

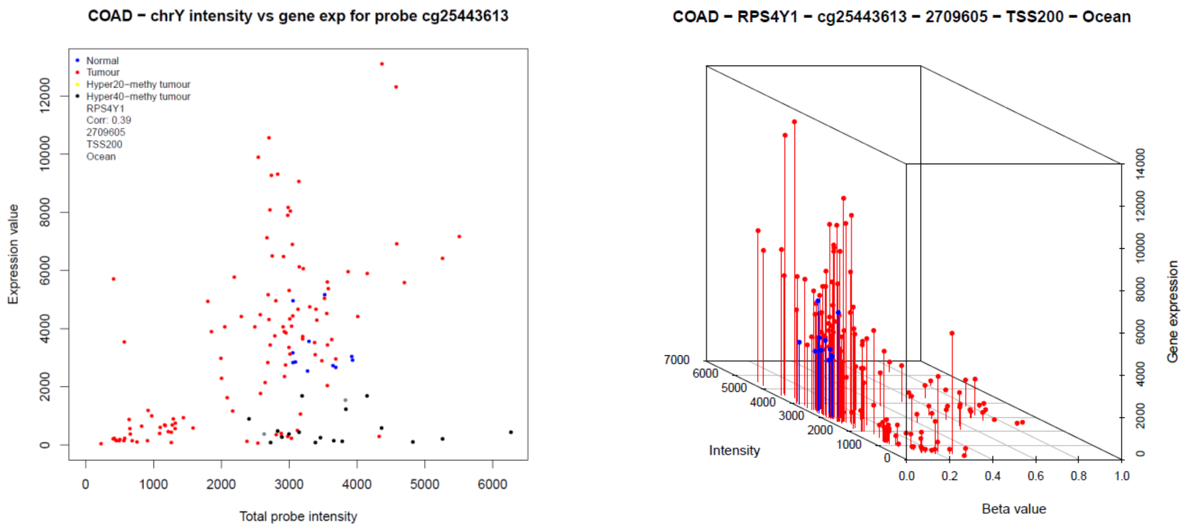


Figure 77 - plots of probe intensity, beta value and gene expression for cg25443613 (COAD)
2D (left) and 3D (right) plots for methylation probe cg25443613 located within the RPS4Y1 gene. The 2D plot shows total methylation intensity (x-axis) versus gene expression (y-axis) for all COAD male samples - non-hyper-methylated tumour samples are plotted in red, hyper-methylated tumour samples are plotted in black (40% threshold) or grey (20% threshold) and normal samples are plotted in blue. The 3D plot shows total methylation intensity plotted against gene expression and beta value – tumour samples are shown in red and normal samples are shown in blue.

For both COAD and HNSC datasets, there is a small subset of hyper-methylated tumour samples (shown in black) which have low expression values (< 2,000) but do not have low total intensity values (> 3,000). The plots for the other two probes (not included here) show a similar pattern, which indicates that for a small number of COAD and HNSC tumour samples, hyper-methylation may be another cause of reduced RPS4Y1 gene expression.

5.3.4 Analysis of TBL1Y gene methylation

TBL1Y is located further along the short arm of the Y chromosome at around 6.78Mbp. Like SRY, gene expression levels vary considerably between normal samples in the three datasets (Kruskal-Wallis p-value < 0.0001). In particular, there is no expression in COAD normal samples. The box-plots in figure 78 compare expression levels across the three datasets:

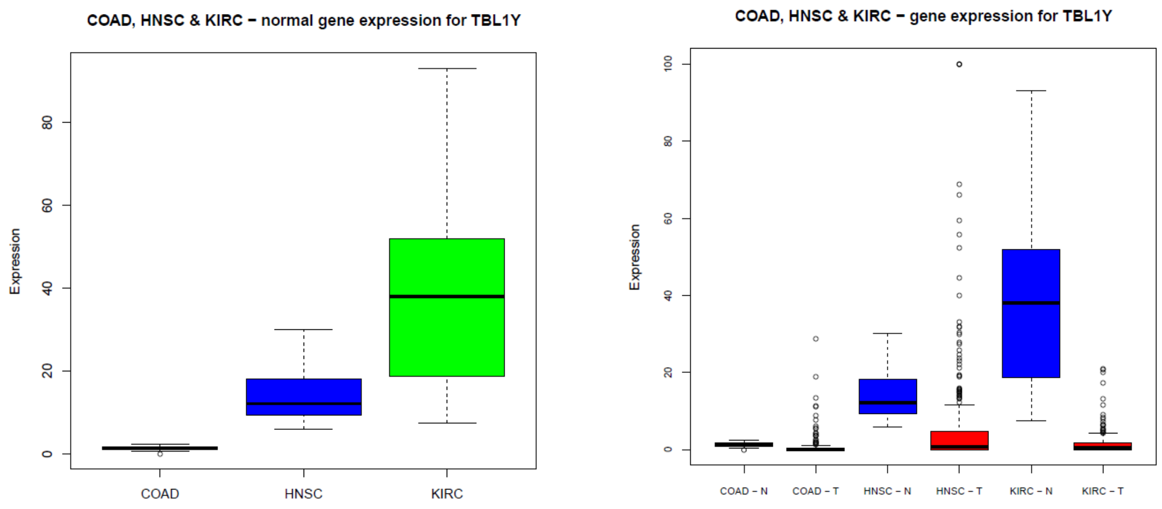


Figure 78 - comparison of TBL1Y expression levels across datasets
Boxplots of expression values of the TBL1Y gene for all male samples, broken down by the three datasets. The left hand plot shows the distribution of expression values for normal samples – COAD, HNSC and KIRC datasets are colour-coded red, blue and green respectively. The right hand plot shows the distribution of expression values for all samples, split by sample type and dataset. The distributions for normal samples are shown in blue, and those for tumour samples are shown in red.

As I showed previously, across all three datasets the correlation between gene expression and copy number variation is much lower for TBL1Y than for most of the other 16 genes. For the HNSC and KIRC datasets, expression is generally lower in tumour samples relative to normal samples (Kruskal-Wallis p-values < 0.0001 in both cases), although there are some tumour samples with higher expression, which appears to be related to higher copy number.

Methylation levels also vary between the datasets. In particular, out of the 11 CpG sites within TBL1Y, two (cg09728865 and cg15700967) have methylation levels which appear to correlate negatively with gene expression in normal samples (r = -0.74 and -0.52 respectively, p-values < 0.0001 and 0.0034 respectively), as shown in figure 79:

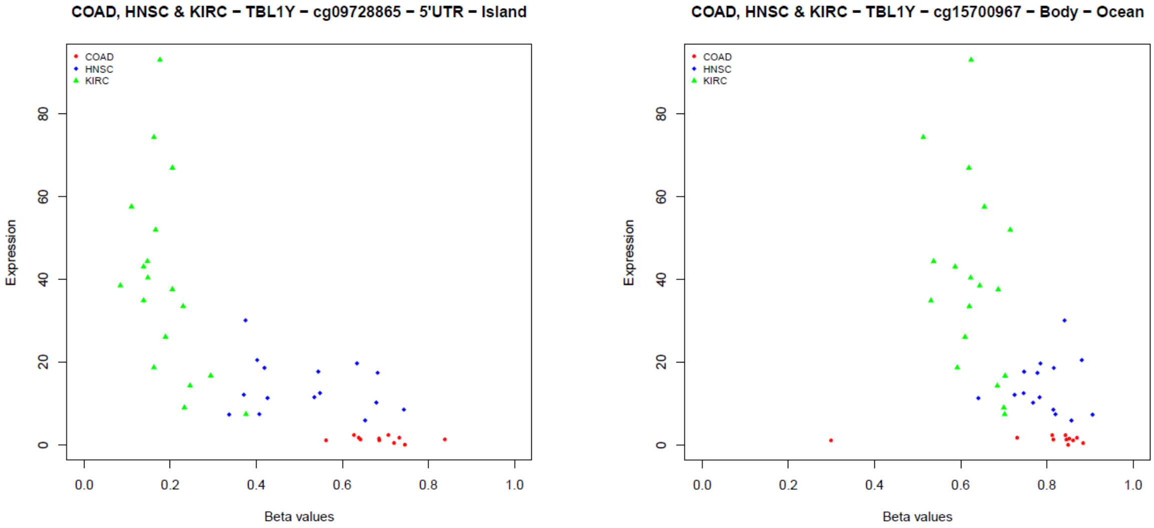


Figure 79 - plots of beta values for cg09728865 & cg15700957 vs TBL1Y expression
Scatterplots for all normal male samples of methylation beta value (x-axis) for probe cg09728865 (left) and cg15700967 (right) versus expression value for the TBL1Y gene (y-axis). COAD, HNSC and KIRC samples are shown in red, blue and green respectively.

The differences in methylation between the datasets are particularly clear for cg09728865, which is located within the 5’ untranslated region of TBL1Y.

There is another CpG site (cg08921682) located within the 5' untranslated region for which there appears to be a positive correlation between methylation and gene expression in normal samples (r = 0.63, p-value < 0.0001), as shown in figure 80:

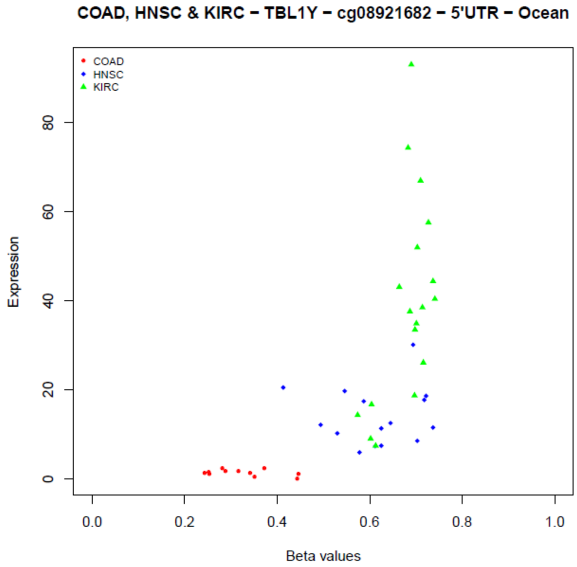


Figure 80 - plot of beta values for cg08921682 vs TBL1Y expression
Scatterplot for all normal male samples of methylation beta value for probe cg08921682 (x-axis) versus expression value for the TBL1Y gene (y-axis). COAD, HNSC and KIRC samples are shown in red, blue and green respectively.

The above figures suggest that methylation may be an important factor in regulating TBL1Y expression.

As I mentioned in chapter 3, there is evidence of both hypo- and hyper-methylation in tumour samples across all three datasets. CpG site cg15700967 in particular is prone to hypo-methylation, especially amongst COAD tumour samples, but this does not appear to impact on gene expression amongst COAD tumours. However, within the HNSC dataset, and to a lesser extent KIRC too, there is a subset of tumour samples for which this CpG site is hypo-methylated and for which TBL1Y gene expression is low. Scatterplots for the two datasets are shown in figure 81 (tumours hypo-methylated based on the 40% threshold are shown in green, and those hypo-methylated based on the 20% threshold are shown in light blue):

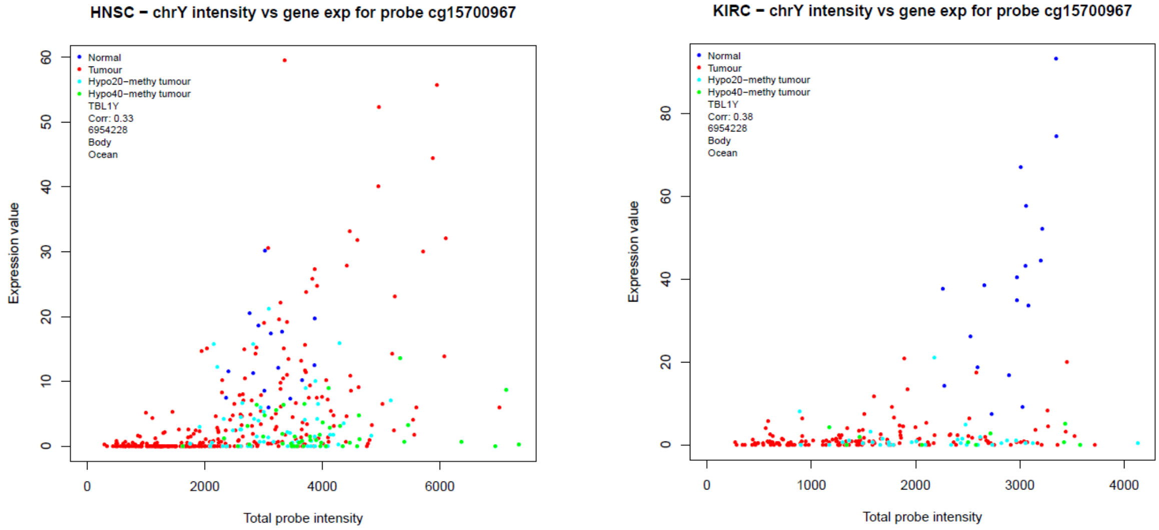


Figure 81 - plots of probe intensity versus gene expression for cg15700967 (HNSC & KIRC)
Scatterplots for methylation probe cg15700967 located within the TBL1Y gene. The plots show total methylation intensity (x-axis) versus gene expression (y-axis) for all HNSC (left) and KIRC (right) male samples - non-hypo-methylated tumour samples are plotted in red, hypo-methylated tumour samples are plotted in green (40% threshold) or light blue (20% threshold) and normal samples are plotted in blue.

For both the HNSC and KIRC datasets, there is a small subset of hypo-methylated tumour samples (shown in green) which have low expression values (< 5, lower than the normal samples) but do not have low total intensity values (> 2,000), which indicates that hypo-methylation may be an important factor in variation in TBL1Y expression in some HNSC and KIRC tumour samples. The effect of hypo-methylation on expression might explain why the correlation between copy number variation and expression is lower than for other genes.

I also observed hyper-methylation of three of the 11 CpG sites (cg02839557, cg01707559 and cg09728865) in subsets of samples in both the HNSC and KIRC datasets. However, this hyper-methylation does not appear to be associated with low gene expression.

5.3.5 Analysis of NLGNY gene methylation

NLGN4Y is located on the long arm of chromosome Y at around 16.63Mbp. Expression levels are similar for HNSC and KIRC, but lower in COAD samples (Kruskal-Wallis p-value = 0.0006 for normal samples), as shown in figure 82:

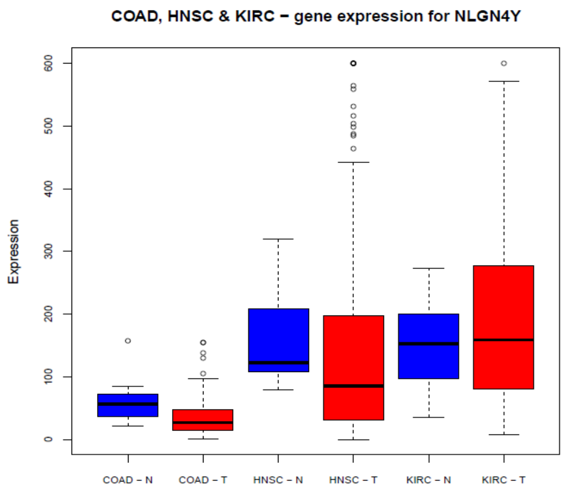


Figure 82 - comparison of NLGN4Y expression levels across datasets
Boxplots of expression values of the NLGN4Y gene for all male samples, broken down by sample type and dataset. The distributions for normal samples are shown in blue, and those for tumour samples are shown in red.

In both the COAD and HNSC datasets, expression levels are lower in tumour samples relative to normal samples (Kruskal-Wallis p-values = 0.0053 and 0.0405 respectively).

I previously showed that there is reasonable correlation between copy number and gene expression for the HNSC and KIRC datasets, but that the correlation for COAD is low.

There are 19 methylation probes targeted at CpG sites within NLGN4Y. Most of these CpG sites show consistent levels of methylation in normal samples across all three datasets. However, there is one site (cg09748856), located in the body of the gene, for which COAD normal samples are more highly methylated than normal samples in the other datasets. All of the HNSC and KIRC normal samples have very low methylation at this site, whereas the COAD normal samples are typically 50% methylated on average, as shown in figure 83:

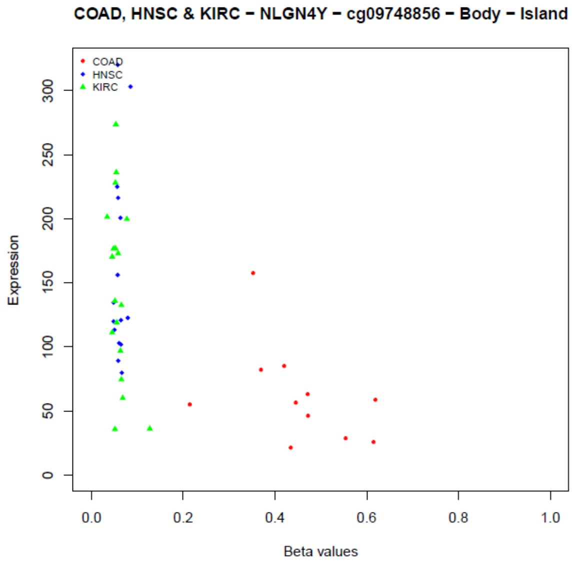


Figure 83 - plot of beta values for cg09748856 vs NLGN4Y expression

Scatterplot for all normal male samples of methylation beta value for probe cg09748856 (x-axis) versus expression value for the NLGN4Y gene (y-axis). COAD, HNSC and KIRC samples are shown in red, blue and green respectively.

Once again this may indicate that methylation is an important regulator of gene expression for NLGN4Y.

Hypo-methylation of tumour samples is a common feature for several of the CpG sites, especially in the COAD dataset. However, there is no obvious association between hypo-methylation and variation in gene expression.

Hyper-methylation of tumour samples is also a common feature across all three datasets. Again it is particularly prevalent in the COAD dataset, and this time there are three CpG sites (cg25518695, cg18113731 and cg19244032) for which hyper-methylation in COAD tumour samples does appear to be associated with lower gene expression. All three sites are located in the gene body, and there is high correlation between their beta values (i.e. all three sites tend to be hyper-methylated together within the same samples). An example of the association between beta value and gene expression is shown for cg25518695 in figure 84:

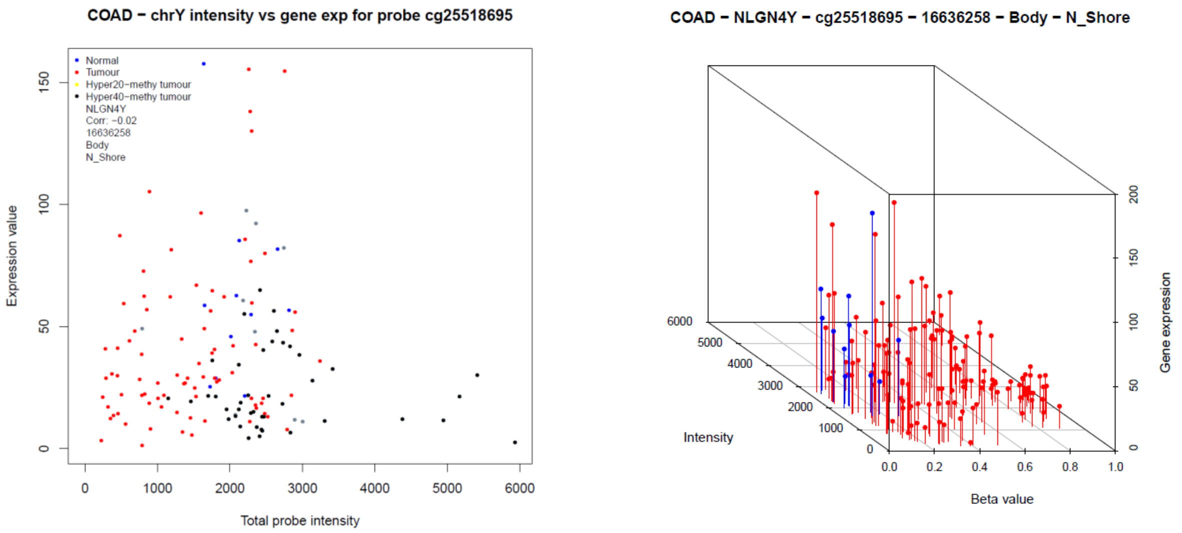


Figure 84 - plots of probe intensity, beta value and gene expression for cg25518695 (COAD)
2D (left) and 3D (right) plots for methylation probe cg25518695 located within the NLGN4Y gene. The 2D plot shows total methylation intensity (x-axis) versus gene expression (y-axis) for all COAD male samples - non-hyper-methylated tumour samples are plotted in red, hyper-methylated tumour samples are plotted in black (40% threshold) or grey (20% threshold) and normal samples are plotted in blue. The 3D plot shows total methylation intensity plotted against gene expression and beta value – tumour samples are shown in red and normal samples are shown in blue.

There is a small subset of hyper-methylated tumour samples (shown in black) which have low expression values (< 25, lower than the normal samples) but do not have low total intensity values (> 2,000), which indicates that hyper-methylation may be an important factor in reduced expression of NLGN4Y in a subset of COAD tumour samples.

5.3.6 Analysis of TTTY14 gene methylation

TTTY14 is located further along the long arm of the Y chromosome at around 21.21Mbp. It is expressed at slightly different levels in normal samples across the three datasets (Kruskal-Wallis p-value = 0.0002). In particular, HNSC normal samples have higher expression levels than their COAD and KIRC counterparts, as shown in figure 85:

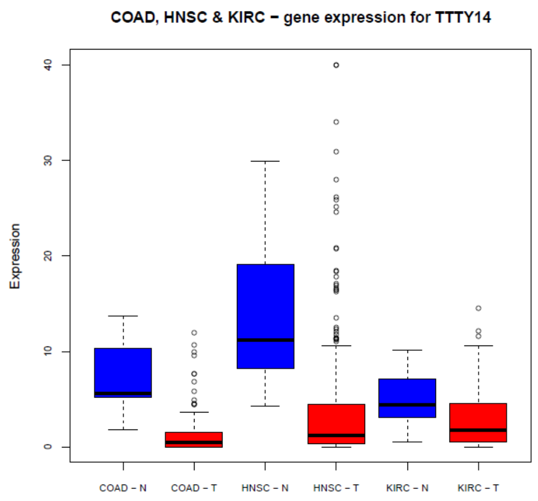


Figure 85 - comparison of TTTY14 expression levels across datasets
Boxplots of expression values of the TTTY14 gene for all male samples, broken down by sample type and dataset. The distributions for normal samples are shown in blue, and those for tumour samples are shown in red.

Figure 85 also shows that TTTY14 expression is generally lower in tumour samples (Kruskal-Wallis p-values < 0.0001 for COAD and HNSC and = 0.0013 for KIRC).

I showed previously that the correlations between copy number and TTTY14 expression are lower for all datasets compared with other genes (particularly for COAD), which suggests that there could be other mechanisms affecting expression levels.

There are 11 methylation probes targeted at CpG sites within TTTY14, and methylation levels are similar across all three datasets for each site.

There is no evidence of hypo-methylation in tumour samples, but hyper-methylation is a common feature for eight of the 11 sites in all three datasets, especially COAD and HNSC. Once again there is high consistency in methylation levels between the eight sites, and hyper-methylation is associated with low gene expression. A good example of this is site cg00212031, which is located near the transcriptional start site. Figure 86 shows the link between hyper-methylation and low gene expression for the COAD and HNSC datasets:

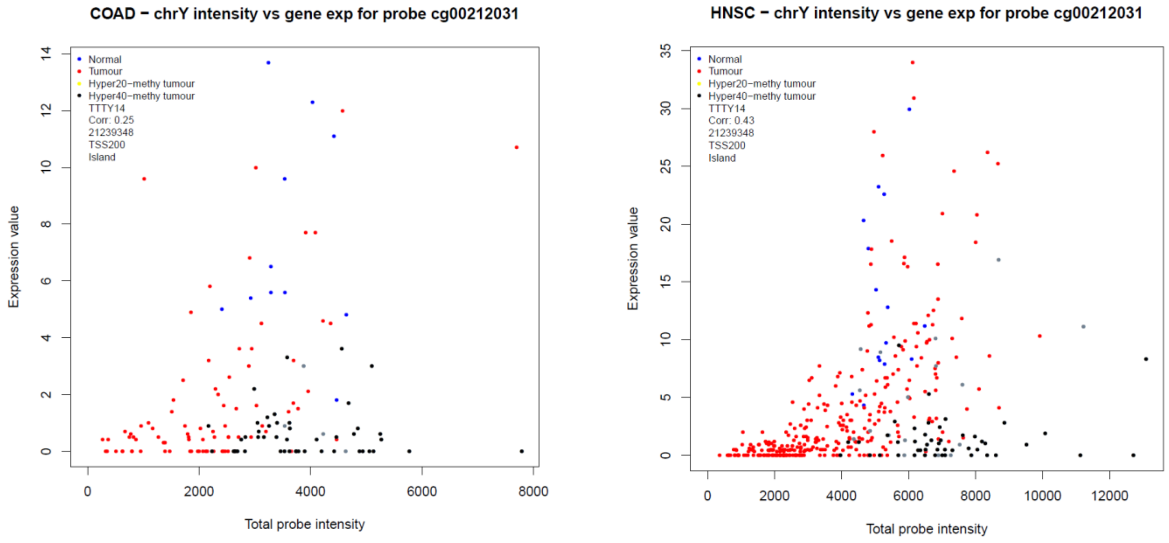


Figure 86 - plots of probe intensity, beta value and gene expression for cg00212031 (COAD & HNSC)
Scatterplots for methylation probe cg00212031 located within the TTTY14 gene. The plots show total methylation intensity (x-axis) versus gene expression (y-axis) for all COAD (left) and HNSC (right) male samples - non-hyper-methylated tumour samples are plotted in red, hyper-methylated tumour samples are plotted in black (40% threshold) or grey (20% threshold) and normal samples are plotted in blue.

There is a subset of hyper-methylated tumour samples which have low expression values (< 2, lower than the normal samples) but do not have low total intensity values (> 4,000 for COAD and > 6,000 for HNSC), which indicates that hyper-methylation may be an important factor in reduced expression of TTTY14 in subsets of both COAD and HNSC tumour samples.

5.3.7 Summary

For four of the 16 genes (SRY, RPS4Y1, NLGN4Y and TTTY14) there is evidence of hyper-methylation being associated with low gene expression in some tumour samples for at least one of the datasets. Furthermore, hypo-methylation of TBL1Y is also associated with low gene expression.

5.4 Conclusions

The key conclusions from my analyses of Y chromosome gene expression in male tumours are as follows:

- 15 Y chromosome genes (RPS4Y1, ZFY, TBL1Y, PRKY, TTTY15, USP9Y, DDX3Y, UTY, TMSB4Y, NLGN4Y, NCRNA00185, TTTY14, CYorf15A, KDM5D, and EIF1AY) are expressed in all three datasets;
- SRY gene is also expressed in the HNSC dataset, but not in COAD or KIRC;
- across all three datasets there is generally high correlation between copy number and gene expression;
- however the correlation is lower for some genes, in particular TBL1Y and NLGN4Y;
- in addition to the effects of copy number variations, for four of the 16 genes (SRY, RPS4Y1, NLGN4Y and TTTY14) there is evidence that hyper-methylation of tumours is associated with low gene expression for some samples in at least one of the datasets;
- hypo-methylation of TBL1Y is also associated with low gene expression for some samples in the HNSC and KIRC datasets

In the next chapter, I will consider the impact of differences in Y chromosome copy number, methylation and gene expression on patient survival for the three datasets.

6. Analysis of Y chromosome aberrations and survival

In this chapter I investigate whether there are any associations between the changes in the Y chromosome observed in the previous three chapters and patient survival. I concentrate, in the first instance, on copy number variations, as these are the most commonly observed alterations in all three datasets. Based on these results, I then look in more detail at the HNSC dataset only.

An overview of the process I have followed is provided in figure 87:

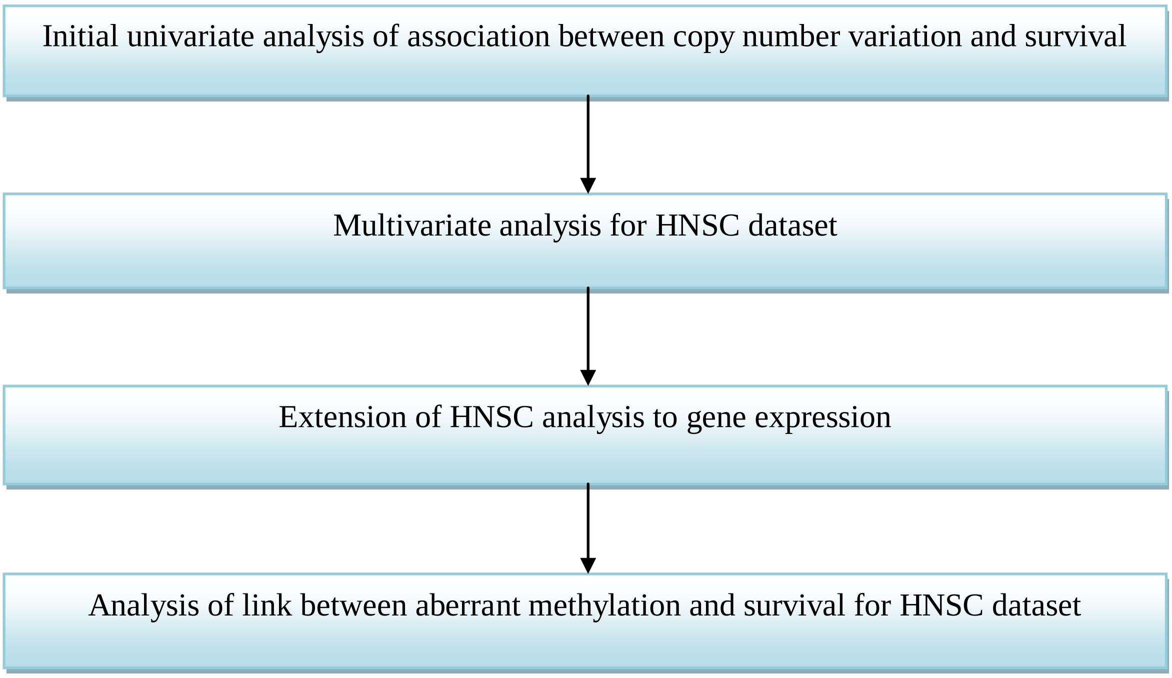

Initial univariate analysis of association between copy number variation and survival
Multivariate analysis for HNSC dataset
Extension of HNSC analysis to gene expression
Analysis of link between aberrant methylation and survival for HNSC dataset

Figure 87 - overview of experimental process for survival analyses

I set out details of each of the above stages in the following sections.

6.1 Initial univariate analysis of survival

In order to identify which, if any, of the datasets show an association between Y chromosome copy number variation and survival, I initially performed my analyses on a univariate basis, without allowing for the effect of any other factors which may influence survival. The TCGA clinical data include items "days to last follow up" and "days to death", both measured from date of initial diagnosis, which allowed me to investigate overall survival (i.e. death from any cause).

For the purposes of my analyses, I used the TCGA copy number index for the Y chromosome which I described in chapter 4. Figure 88 below shows a plot of the distribution of these index values for all KIRC male cases:

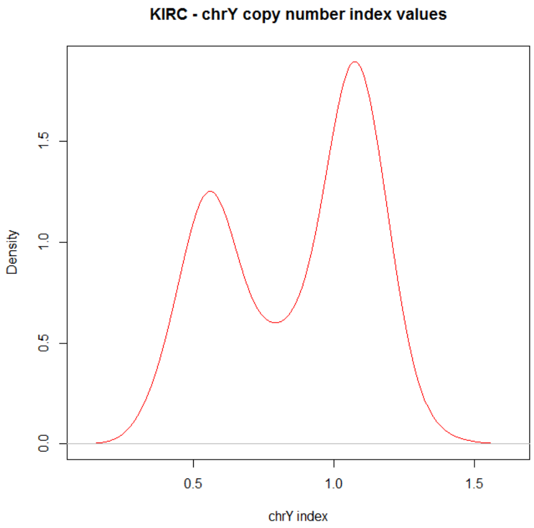


Figure 88 - density plot of KIRC Y chromosome copy number index values
Density plot of Y chromosome copy number index values for KIRC male tumour samples calculated using TCGA's copy number data. The plot shows the frequency distribution of the index values.

Figure 88 shows, in particular, that there is a clear bimodal distribution of Y chromosome copy number index values for the KIRC samples, with an initial peak at 0.56 and a second higher peak at 1.07, with an intervening trough at 0.79.

The bimodal distribution suggests that there are two separate subsets of samples, one which has lost the Y chromosome (the first peak) and one which has not (the second peak), and there is clearly some overlap between the two distributions. The reason why the first peak (representing samples which have lost the Y chromosome) is not at a lower value could be because not all tumour cells in any one sample have lost the Y chromosome. Also, the tumour samples may not be "pure" tumour and may contain some adjacent normal tissue which has not lost the Y chromosome.

TCGA provide estimates of how "pure" each tumour sample is (based on cell nuclei), and so I adjusted the copy number index values I had calculated to allow for this impurity. For this purpose I assumed that if the tumour sample, for example, was estimated to be 70% tumour, then the other 30% would be normal tissue which had a copy number of one. Figure 89 shows the index values for the KIRC samples after I made this adjustment:

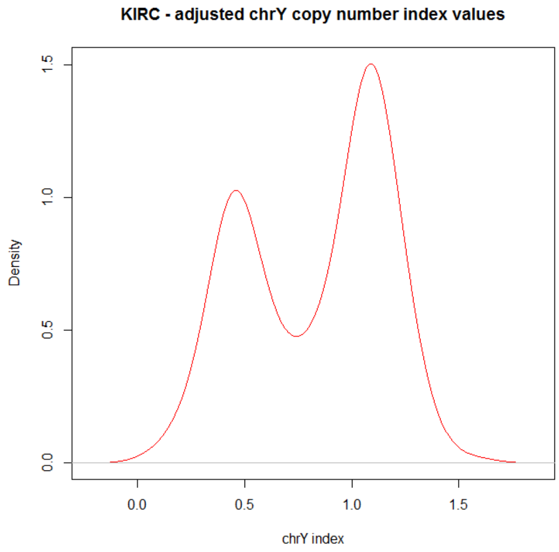


Figure 89 - density plot of KIRC adjusted Y chromosome copy number index values
Density plot of Y chromosome copy number index values for KIRC male tumour samples calculated using TCGA's copy number data after adjustment for sample tumour percentages. The plot shows the frequency distribution of the adjusted index values.

The peaks have now moved slightly to 0.46 and 1.09 respectively (i.e. further away from one, as expected), and the trough has moved to 0.74. Figure 90 contains similar plots for the COAD and HNSC cases:

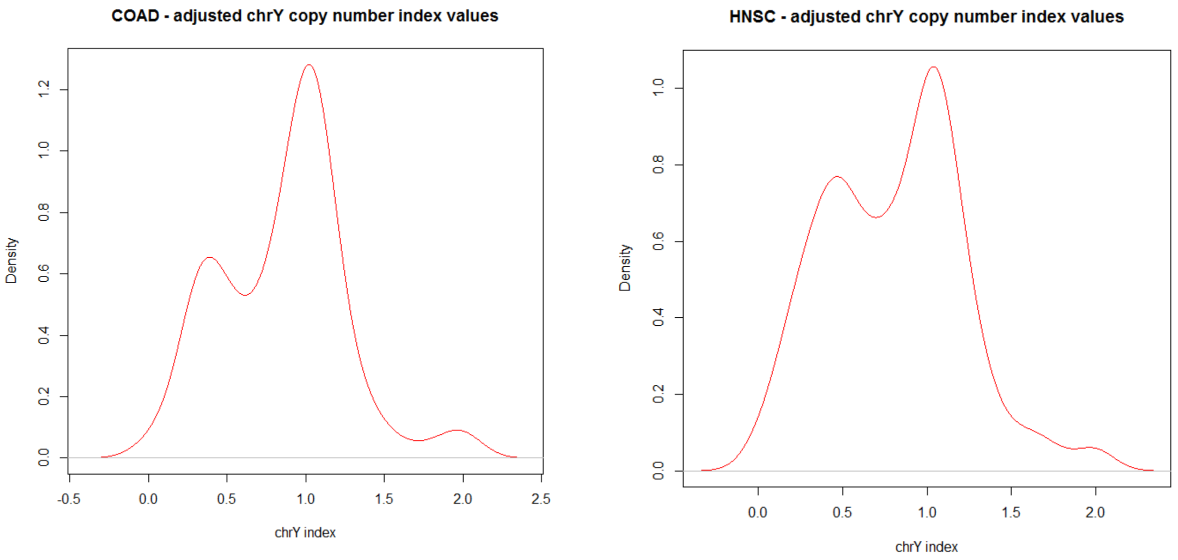


Figure 90 - density plots of COAD & HNSC adjusted Y chromosome copy number index values
Density plots of Y chromosome copy number index values for COAD (left) and HNSC (right) male tumour samples calculated using TCGA's copy number data after adjustment for sample tumour percentages. The plots show the frequency distributions of the adjusted index values.

Both plots again show bimodal distributions, though the first peak is lower for COAD cases. The COAD peaks occur at adjusted index values of 0.39 and 1.02, and the trough is at 0.62. The HNSC peaks are at 0.47 and 1.04, and the trough is at 0.71.

I then used the adjusted index values to perform Kaplan-Meier analyses of overall survival. In each dataset, I initially subdivided cases according to whether their adjusted index values were less than or greater than the value at which the trough in the bimodal distribution occurred, as this seemed a natural point to split the two distributions. The results for the COAD and KIRC datasets are shown in figure 91 (probability values calculated using log-rank test):

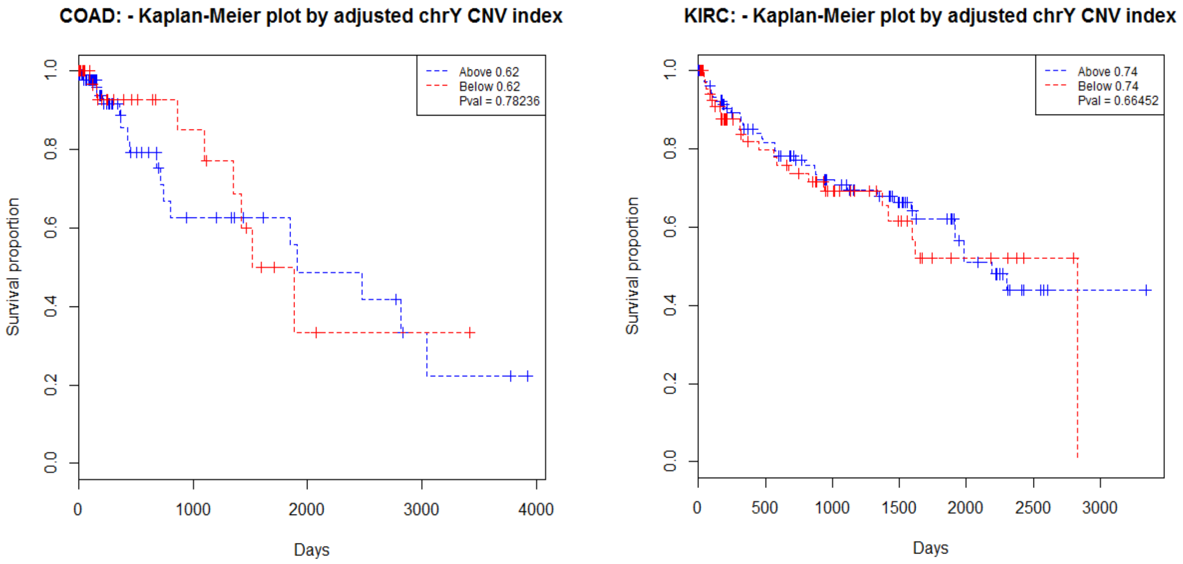


Figure 91 – COAD & KIRC Kaplan-Meier plots using single Y chromosome copy number index thresholds
Kaplan-Meier plots of overall survival for COAD (left) and KIRC (right) male tumour samples. In each case, samples have been subdivided into two groups based on threshold values of 0.62 and 0.74 respectively for the adjusted Y chromosome copy number index values.

For both datasets there is no association between loss of the Y chromosome and overall survival. However, it is possible that the accuracy of the analyses was affected by the group of samples which fell within the overlap of the two distributions of index values (as I cannot be certain whether cases which fall within this range have lost the Y chromosome or not).

To attempt to reduce the impact of this potential inaccuracy, I repeated the analysis by only considering those cases whose adjusted index values were lower than the value at which the first peak occurred or higher that the value at which the second peak occurred (i.e. ignoring cases which fell between the two peaks of the bimodal distribution). Figure 92 shows the results of these analyses for the COAD and KIRC datasets:

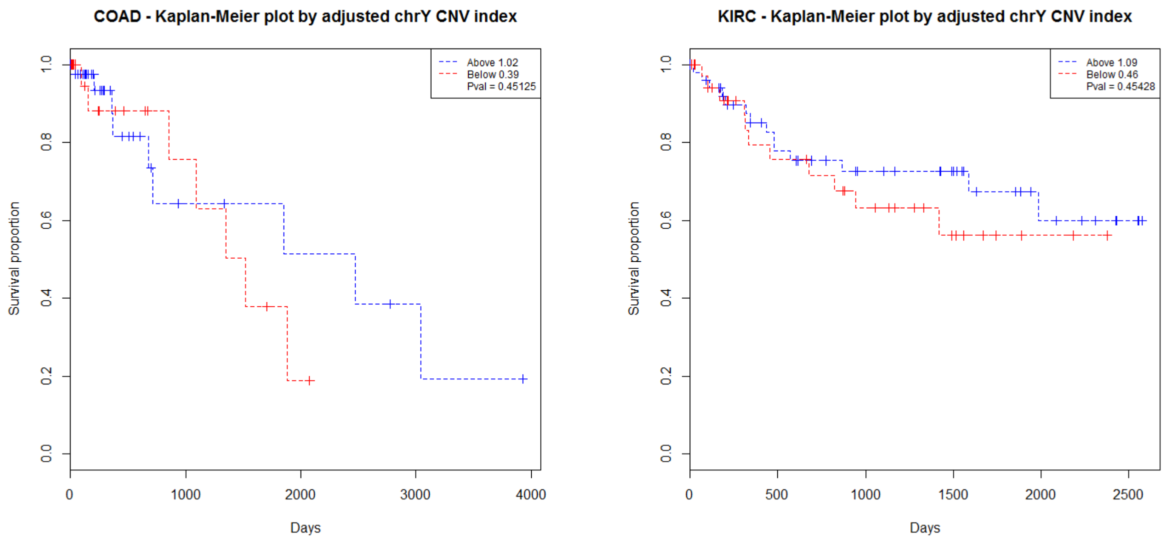


Figure 92 – COAD & KIRC Kaplan-Meier plots using dual Y chromosome copy number index thresholds
Kaplan-Meier plots of overall survival for COAD (left) and KIRC (right) male tumour samples. In each case, only samples with adjusted Y chromosome copy number index values above 1.02 and 1.09 respectively or below 0.39 and 0.46 respectively have been used.

Again the plots indicate that there is no obvious association between Y chromosome copy number and overall survival in the COAD and KIRC datasets (log-rank p-values > 0.05). However, the Kaplan-Meier plots in figure 93 show a different picture for the HNSC dataset:

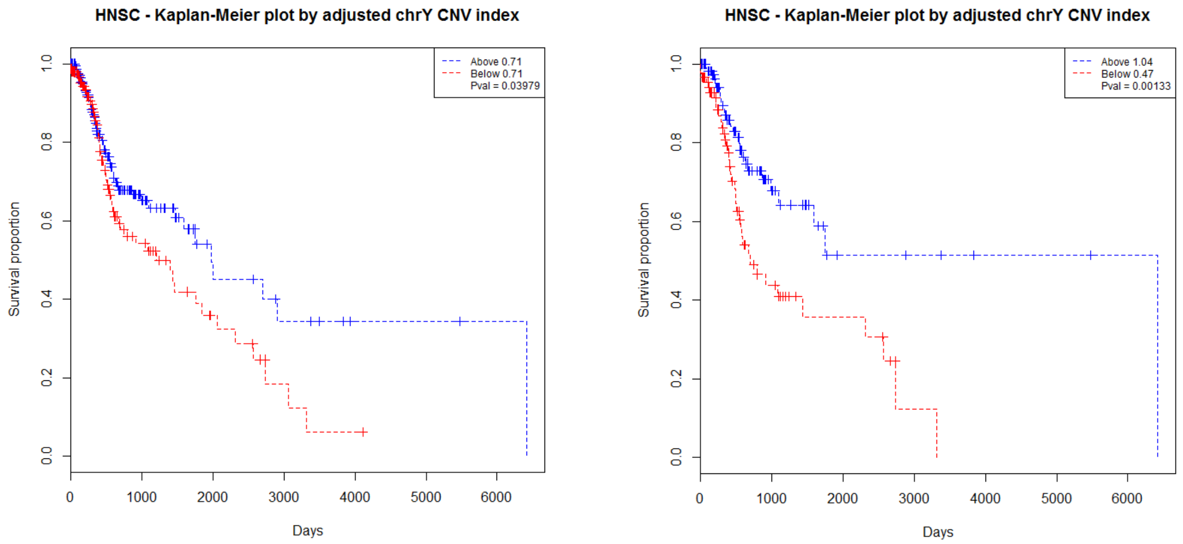


Figure 93 – HNSC Kaplan-Meier plots using single and dual Y chromosome copy number index thresholds
Kaplan-Meier plots of overall survival for HNSC male tumour samples. In the left hand plot, all samples have been subdivided into two groups based on a threshold value of 0.71 for the adjusted Y chromosome copy number index value. In the right hand plot, only samples with adjusted Y chromosome copy number index values above 1.04 or below 0.47 have been used.

This time there is a statistically significant link between reduced Y chromosome copy number and worse survival (log-rank p-value = 0.0398 based on splitting all samples using the value at which the trough occurred), and the association is stronger when the two extremes of the bimodal distribution are used (log-rank p-value = 0.0013). I, therefore, decided to concentrate my subsequent analyses on the HNSC dataset only, as set out in the following sections.

Summary

Univariate survival analysis indicated a possible link between loss of the Y chromosome and impaired overall survival for the HNSC dataset. No associations were observed for either the COAD or the KIRC dataset.

6.2 Multivariate survival analyses for HNSC dataset

Having discovered that there may be an association between loss of the Y chromosome and impaired overall survival in the HNSC dataset, I then refined my analyses of this dataset to allow for other factors which may influence the results.

6.2.1 Effect of HPV infection

Infection with human-papillomavirus (HPV) is an increasingly important factor in the development of a subset of head and neck cancers (in particular of the oropharynx)[13]. So HPV status is a natural item by which to sub-divide patients. The first question I asked, therefore, was whether there is any difference in overall survival between HPV positive and HPV negative patients within the HNSC male dataset.

The TCGA clinical data include an item indicating whether each sample has been categorised as HPV positive or negative. Of the 370 male patients in total, 85 had been categorised as HPV positive and 284 as HPV negative (data item was unavailable for one patient).

Based on this information, figure 94 shows that there was no obvious difference in overall survival between the two groups:

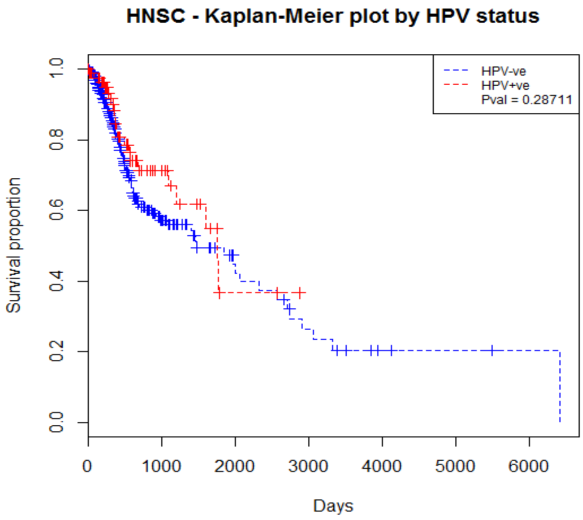


Figure 94 – Kaplan-Meier plot of overall survival for HNSC cases split by HPV status
Kaplan-Meier plot of overall survival for HNSC male tumour samples, with samples split according to HPV status

I then considered the adjusted chromosome Y copy number index values for each group. The distributions of these values are shown in figure 95:

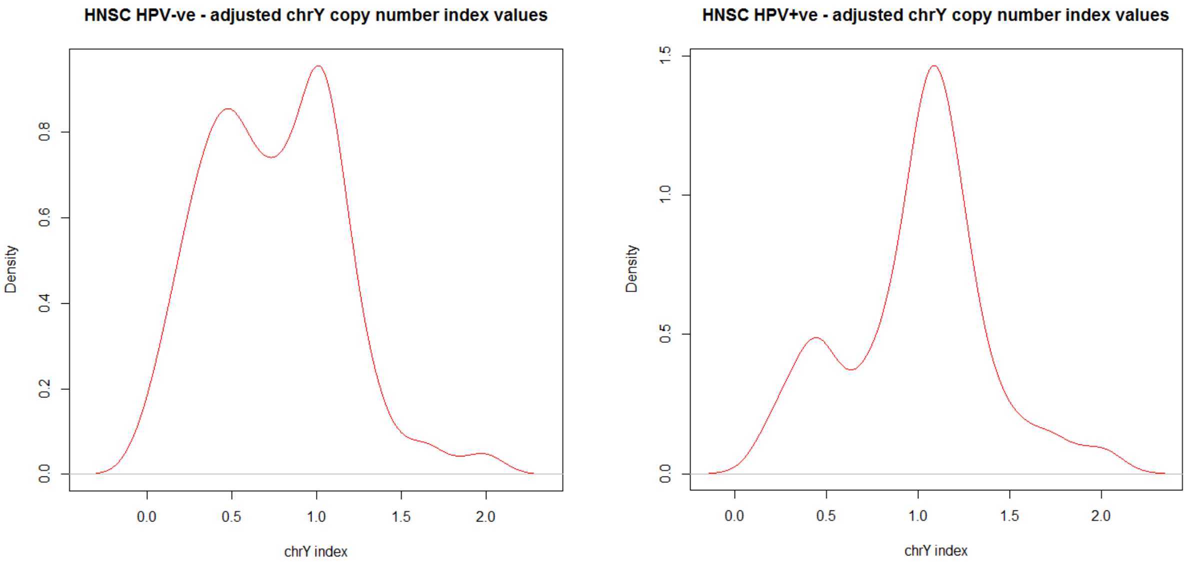


Figure 95 – density plots of adjusted Y chromosome copy number index values for HNSC HPV-ve and HPV+ve cases
Density plots of Y chromosome copy number index values for HNSC HPV-ve (left) and HPV+ve (right) male tumour samples calculated using TCGA's copy number data after adjustment for sample tumour percentages. The plots show the frequency distributions of the adjusted index values.

Again these is a clear bimodal distribution for both groups, although the first peak is a lot higher for HPV negative cases, which suggests that loss of the Y chromosome is more prevalent in these cases than in HPV positive cases.

I next considered whether there was an association between Y chromosome copy number variation for either HPV category. Figure 96 shows the resultant Kaplan-Meier plots using samples split by the values at which the troughs in the bimodal distributions occurred:

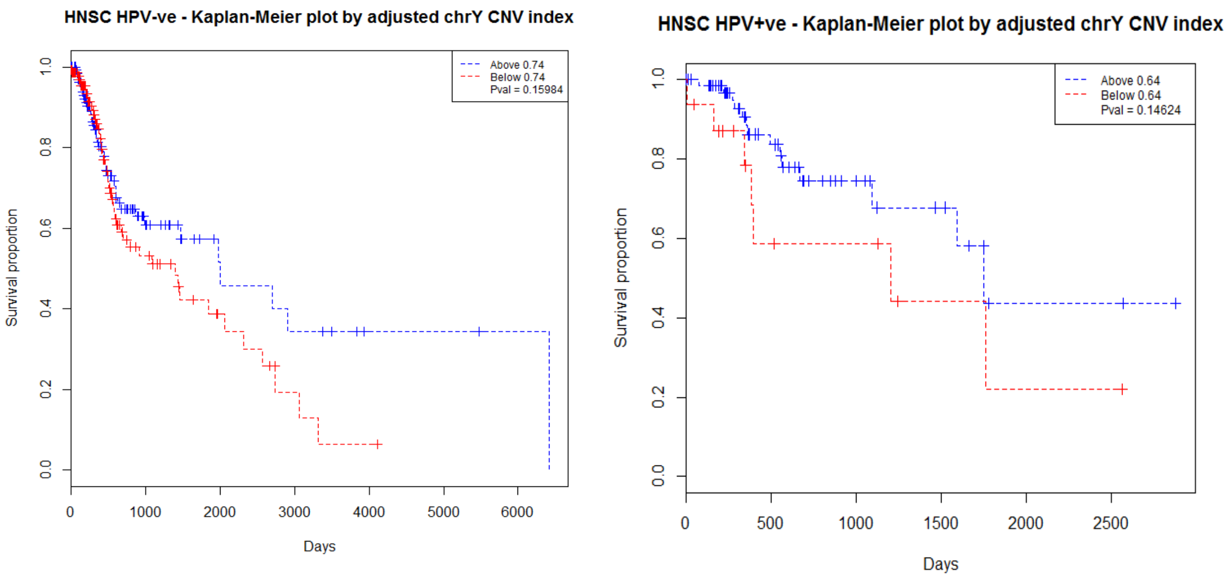


Figure 96 - Kaplan-Meier plots for HNSC HPV-ve & HPV+ve males using single Y chromosome copy number index threshold
Kaplan-Meier plots of overall survival for HNSC HPV-ve (left) and HPV+ve (right) male tumour samples. In each case, samples have been subdivided into two groups based on threshold values of 0.74 and 0.64 respectively for the adjusted Y chromosome copy number index values.

The above plots suggest that loss of the Y chromosome may be associated with impaired overall survival, especially for HPV negative cases, although neither result was statistically significant (log-rank p-values = 0.1598 and 0.1462 for HPV negative and HVP positive cases respectively).

I also performed Cox Proportional Hazards analyses for both groups, and there was a statistically significant result for HPV negative cases (hazard ratio = 0.52 - i.e. a higher index value implies better survival - p-value = 0.0192), but not for HPV positive cases.

I then repeated the Kaplan-Meier analyses using only samples whose adjusted index values were lower than the values at which the first peaks occurred or higher that the values at which the second peaks occurred. Figure 97 shows the resultant Kaplan-Meier plots:

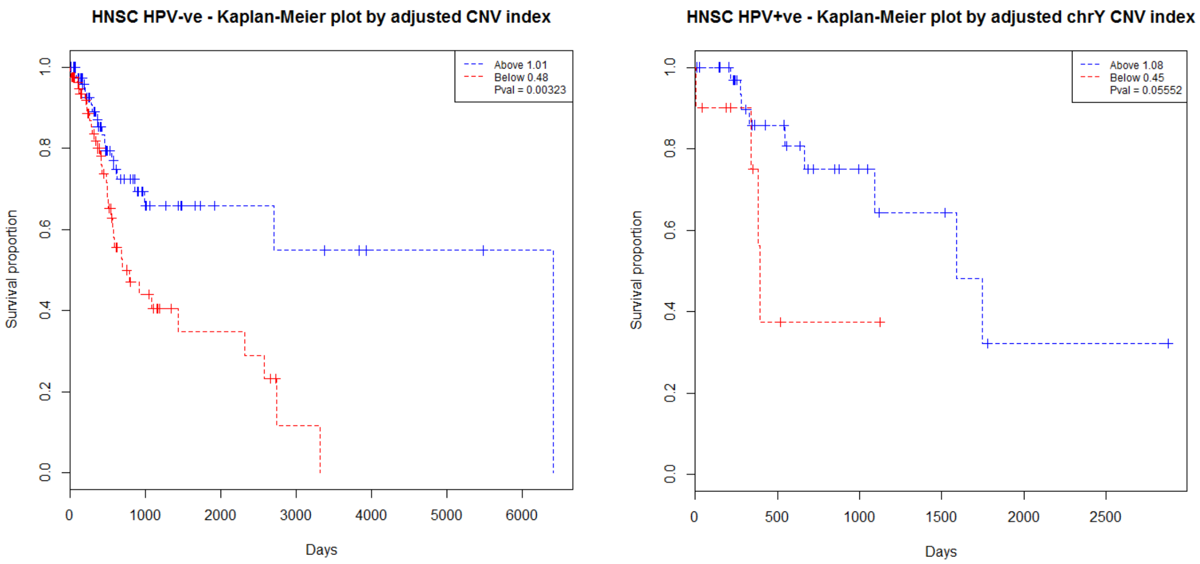


Figure 97 - Kaplan-Meier plots for HNSC HPV-ve and HPV+ve males using dual Y chromosome copy number index thresholds

Kaplan-Meier plots of overall survival for HNSC HPV-ve (left) and HPV+ve (right) male tumour samples. In each case, only samples with adjusted Y chromosome copy number index values above 1.01 and 1.08 respectively or below 0.48 and 0.45 respectively have been used.

This time there is a clear association between Y chromosome copy number and survival for the HPV negative patients (log-rank p-value = 0.0032), but not for the HPV positive patients. I, therefore, focussed my subsequent analyses on the HPV negative cases only.

I also discovered statistically significant associations for HPV negative cases using the adjusted copy number index for each individual arm of the Y chromosome (data not shown). Furthermore, these were the only chromosomal arms for which there were significant associations.

6.2.2 Other factors potentially affecting overall survival

There are numerous other factors which may affect overall survival. For head and neck cancers, the most commonly used information in a clinical setting is nodal status – i.e. the extent to which the cancer has spread to the lymph nodes. TCGA's clinical data includes an item "pathologic N" which indicates nodal status.

Other factors which may potentially affect survival and which are available for most samples in the TCGA HNSC clinical data are:

- Pathologic T – a standard indicator of tumour size and extent of invasion into nearby tissue;
- Perineural invasion – an indicator of whether the tumour has invaded areas surrounding nerves;
- Pathologic stage
- Histologic grade
- Age
- Smoking history – smoking is a known risk factor for head and neck cancer[89]. TCGA's data include an item indicating whether each patient has ever smoked, and if they have, whether they are still smoking, stopped within the last 15 years or stopped more than 15 years ago.
- Anatomic site – head and neck cancer covers several distinct regions, as shown in figure 98:

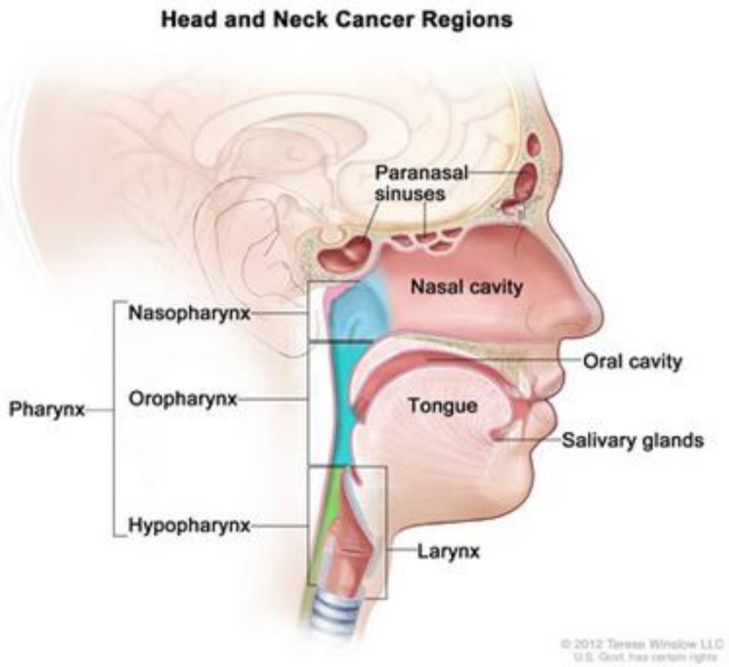


Figure 98 - diagram of head & neck cancer regions

I also included an indicator of whether each sample harboured a mutation of the TP53 gene. TP53 is very commonly mutated in head and neck cancer, and such mutations may be regarded as a sign of more general chromosomal instability[13]. For this purpose I downloaded TCGA’s level 2 somatic mutation data, as described in chapter 2.

6.2.3 Univariate Kaplan-Meier analyses using other factors

All of the above items were available for most, but not all, of the 284 HPV negative male patients. For each item I performed a separate, univariate Kaplan-Meier analysis of overall survival. In table 9 I set out the resultant (log-rank) probability values for each analysis using all HPV negative samples – for the Y chromosome analysis I split samples at the trough value of 0.74, which results in a 50/50 split:

Item	No. patients for which data available	Categories used for Kaplan-Meier analysis	P-value
Adjusted chrY index	284	< 0.74 vs > 0.74	0.1598
Pathologic N	247	N0/N1/N2/N3	< 0.0001
Pathologic T	261	T1/T2/T3/T4/T4a/T4b	0.0035
Perineural invasion	210	Yes / No	0.0115
Pathologic stage	259	Stage I/II/III/IV	0.1329
Histologic grade	276	G1/G2/G3/G4	0.2883
TP53 mutation	270	Yes / No	0.3366
Age (median split)	284	< median / > median	0.3515
Smoking history	276	As described above	0.5247
Anatomic site	284	Oral cavity / Larynx / Oropharynx *	0.6077
* one patient categorised as hypopharynx excluded			

Table 9 - summary of univariate survival analyses for HPV-ve males

The table shows, for each of the data items in the first column, the number of HNSC HPV-ve samples for which that item was available, the categories used to split samples for the Kaplan-Meier analysis of overall survival, and the log-rank p-value from the analysis.

The table shows that the association of pathologic N with survival is by far the most significant, followed by the associations with pathologic T and occurrence of perineural invasion.

Table 10 shows the corresponding results when the analysis was restricted to those cases whose adjusted Y chromosome copy number index values were lower than the value at which the first peak occurred or higher that the value at which the second peak occurred. This split samples into the lowest 29% and the highest 28% by index value.

Pathologic N still has the most significant association with overall survival, but it is now closely followed by the adjusted Y chromosome copy number index.

Item	No. patients for which data available	Categories used for Kaplan-Meier analysis	P-value
Adjusted chrY index	159	< 0.48 vs > 1.01	0.0032
Pathologic N	139	N0/N1/N2/N3	0.0028
Pathologic T	146	T1/T2/T3/T4/T4a/T4b	0.0061
Perineural invasion	119	Yes / No	0.0267
Pathologic stage	145	Stage I/II/III/IV	0.1881
Histologic grade	153	G1/G2/G3/G4	0.4975
TP53 mutation	155	Yes / No	0.3172
Age (median split)	159	< median / > median	0.2120
Smoking history	153	As described above	0.9479
Anatomic site	159	Oral cavity / Larynx / Oropharynx *	0.9248
* one patient categorised as hypopharynx excluded			

Table 10 - summary of univariate survival analyses for HPV-ve males split by dual Y chromosome copy number index thresholds

The table shows, for each of the data items in the first column, the number of HNSC HPV-ve samples for whom that item was available, the categories used to split samples for the Kaplan-Meier analysis of overall survival, and the log-rank p-value from the analysis. For these analyses, only samples with an adjusted Y chromosome copy number index of greater than 1.01 or less than 0.48 were used.

6.2.4 Multivariate analyses using Cox-proportional hazards

I then performed multivariate survival analyses using the Cox proportional hazards model, with adjusted Y chromosome copy number index and either pathologic N or pathologic T as my explanatory variables. In the former case, pathologic N retained a significant association with survival, but Y chromosome copy number did not. In the latter case, neither association was significant (at the 5% level), although the Y chromosome copy number association was border-line significant (Wald test p-value = 0.0548).

I next considered how Y chromosome copy number varied over the different pathologic N and pathologic T categories. Figure 99 shows that there is little variation in Y chromosome copy number between the pathologic N groups (Kruskal-Wallis p-value = 0.5073):

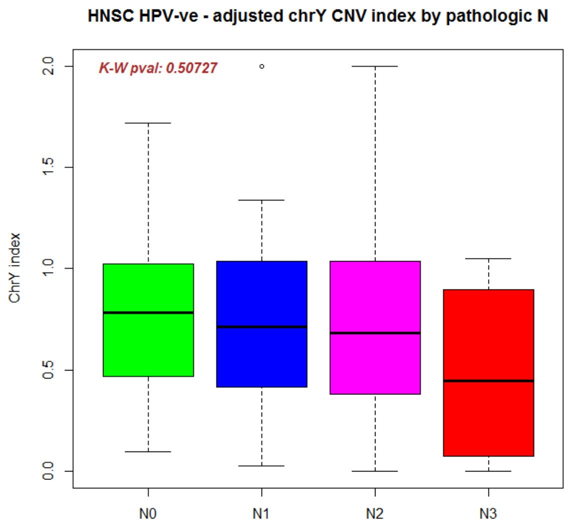


Figure 99 - boxplots of Y chromosome copy number by pathologic N
Boxplots of adjusted Y chromosome copy number index values for HNSC HPV-ve male tumours, subdivided by pathologic N data item.

However, figure 100 shows a statistically significant (Kruskal-Wallis p-value = 0.0002), pattern of Y chromosome copy number reducing with increasing pathologic T levels:

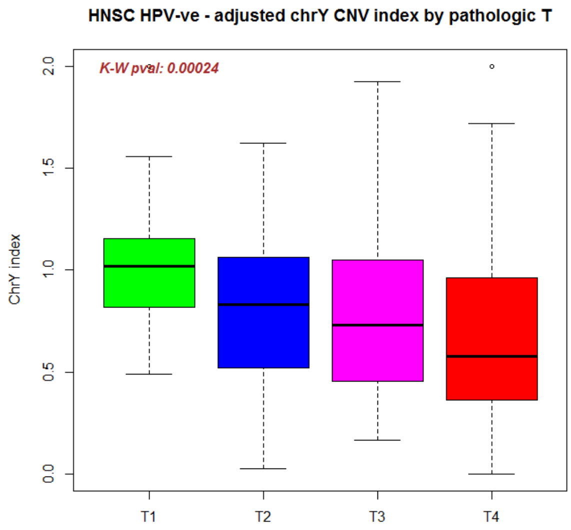


Figure 100 - boxplots of Y chromosome copy number by pathologic T
Boxplots of adjusted Y chromosome copy number index values for HNSC HPV-ve male tumours, subdivided by pathologic T data item.

Finally I performed a Cox proportional hazards analysis including only Y chromosome copy number index, TP53 mutation, age, smoking history and anatomic site as the explanatory variables, and using only samples whose adjusted Y chromosome copy number index values were lower than the value at which the first peak occurred or higher that the value at which the second peak occurred (samples were grouped on this basis). My reasoning for doing this is that the other variables (pathologic N etc) are essentially observed tumour outcomes whereas these five variables are potential, biological causes of those outcomes. Using this multivariate model, the grouping by Y chromosome index values was the only variable to show a border-line statistically significant association with overall survival (Wald test p-value = 0.0499).

6.2.5 Limitation of analyses

At this stage I should point out that there are a number of limitations to the survival analyses I have undertaken. First, TCGA did not create their datasets with the primary intention of performing detailed survival analyses on the data, and hence there was no proper control of patients as there would be in a full clinical trial. In particular, there is inconsistency in the treatment regimes which have been applied to the patients in question (details available within TCGA's clinical data). I have ignored this complication in my analyses as the treatment data were too varied (and sometimes missing) for them to be sensibly included. Furthermore, contact was lost with many patients at quite early time points, and so these people were censored and their contributions to the analyses are small.

Second, there are several factors which may influence survival rates, and I have been limited to those which are available within TCGA's clinical data (for example, useful data on alcohol consumption, another potential risk factor, was not available). Furthermore, even when a data item was available for most patients, it may have been missing for some, thereby reducing its reliability.

The number of potential factors is in itself a problem, as there are a limited number of samples, and so there is insufficient statistical power when several factors are tested simultaneously in a multivariate analysis.

Finally, for ease of presentation, I have in some cases created categorical variables from otherwise continuous, numerical variables, by using suitable cut-off values (e.g. for the copy number and median age). This is not ideal practice, but does enable graphical representation of trends in survival rates.

Despite the above issues, my analyses should provide, at the very least, reasonable indications of trends in survival.

6.2.6 Summary

For the HNSC dataset, loss of the Y chromosome is significantly associated with impaired overall survival in HPV negative cases (but not in HPV positive cases). Similar associations were observed for each arm of the Y chromosome, but not for any other chromosomal arm.

6.3 Analysis of association between gene expression and survival

I next considered whether there was also an association between reduced gene expression and impaired survival for the HPV negative cases. I expected that there probably would be, as I had previously shown in chapter 5 that for most of the 16 genes which were expressed, there was a strong correlation between Y chromosome copy number and expression level.

I performed univariate survival analyses using expression values for each of the 16 genes in turn as the explanatory variable. I initially produced Kaplan-Meier plots, splitting samples by whether their expression levels were below or above the median value (consistent with the fact that the trough of the bimodal distribution of adjusted Y chromosome copy number index values split cases into two equal groups.).

In a similar vein to my analyses using the copy number index values, I also performed alternative analyses whereby I excluded those samples for which expression values were close to the median. For this purpose I compared, for each gene separately, those samples with expression either below the 29th percentile or above the 72% percentile (consistent with the numbers of cases whose adjusted Y chromosome copy number index values were either lower than the value at which the first peak occurred or higher that the value at which the second peak occurred). In addition, I carried out Cox proportional hazards analyses based on expression levels for all samples.

I set out the results of all these analyses (log-rank probability values for the Kaplan-Meier analyses and hazard ratios and Wald probability values for the Cox proportional hazards analyses) in table 11.

In all cases, low expression was associated with worse overall survival, and for five genes (ZFY, PRKY, USP9Y, UTY and NLGN4Y) the results were statistically significant at the 5% level for both the Kaplan-Meier and Cox proportional hazards analyses. However, there is clearly a lot of variability in the statistical significance of the results depending on the method used to split the samples, which I attribute, in part, to noise in the expression data.

Gene	Log-rank p-value		Cox proportional hazards	
	Median split	<P29 / >P72	Hazard ratio	P-value
SRY	0.0993	0.1087	0.9683	0.1095
RPS4Y1	0.0502	0.4216	1.0000	0.2033
ZFY	0.1198	0.0108	0.9989	0.0339
TBL1Y	0.8633	0.5461	0.9800	0.1986
PRKY	0.0470	0.0320	0.9990	0.0492
TTTY15	0.0620	0.0180	0.9981	0.0825
USP9Y	0.0049	0.0014	0.9991	0.0216
DDX3Y	0.0141	0.0213	0.9997	0.0610
UTY	0.0033	0.0113	0.9987	0.0155
TMSB4Y	0.0360	0.0377	0.9910	0.1131
NLGN4Y	0.0004	0.0012	0.9979	0.0156
NCRNA00185	0.4065	0.1502	0.9891	0.2772
TTTY14	0.4751	0.8682	0.9805	0.3892
CYorf15A	0.0723	0.2325	0.9990	0.2159
KDM5D	0.0024	0.0297	0.9996	0.1875
EIF1AY	0.5004	0.4061	1.0000	0.8728

Table 11 - results of survival analyses based on gene expression

The second and third columns show, for each of the 16 Y chromosome genes, the results of Kaplan-Meier analyses of overall survival for HNSC HPV-ve male samples split by median expression and also with samples split into those whose expression was lower than the 29th percentile or greater than the 72nd percentile. The final two columns show the results of Cox proportional hazards analyses for all HNSC HPV-ve male samples using gene expression as the explanatory variable.

The Kaplan-Meier analyses based on comparing samples whose expression levels were below the 29th percentile against those whose expression levels were above the 72nd percentile were intended to reduce the effect of noise in the data by ignoring samples whose categorisation could be influenced by a small variation in expression level. Based on these analyses, expression levels for nine of the 16 genes had a statistically significant association with survival. However, there are four genes (RPS4Y1, TBL1Y, TTTY14 and EIF1AY) for which there was little association.

It may be the case that the analyses for these four genes were affected by samples whose expression had been reduced by another mechanism. In particular, I observed in chapter 5 that expression levels of RPS4Y1 and TTTY14 (hyper-methylated) and TBL1Y (hypo-methylated) may have been affected by abnormal methylation. There was also some evidence of hyper-methylation affecting EIF1AY expression levels for a very small subset of samples.

Summary

For each of the 16 genes, low levels of expression were generally associated with worse overall survival for HPV negative cases, although the results were not always statistically significant. There were four genes in particular, for which the association with survival was much weaker than for the other genes.

6.4 Aberrant methylation and survival

Finally, for RPS4Y1, TBL1Y and TTTY14, I investigated whether there was any difference in survival between patients who may have had reduced gene expression as a result of aberrant methylation, and those whose expression appeared to be reduced as a result of chromosomal loss. For each gene I performed univariate Kaplan-Meier analyses, with HPV negative patients whose expression levels were below the median split into two groups:

1. Hypo-methylated (TBL1Y) or hyper-methylated (RPS4Y1 and TTTY14) based on the 40% threshold (probes cg15700967, cg25443613 and cg00212031 respectively); and
2. Not hypo- / hyper- methylated

The resultant plots for TBL1Y and TTTY14 are shown in figure 101:

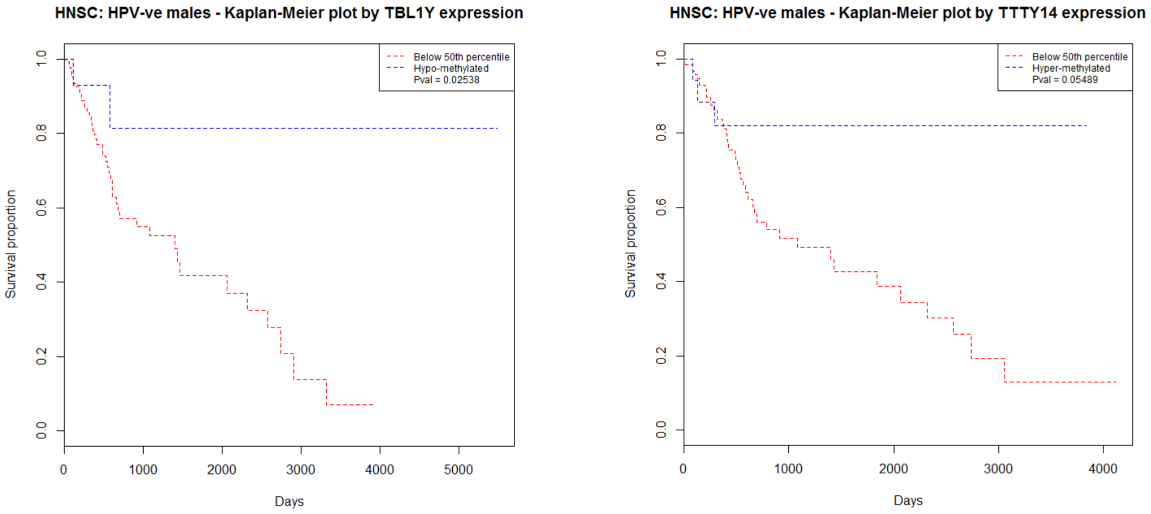


Figure 101 - Kaplan-Meier plots based on methylation for TBL1Y and TTTY14

Kaplan-Meier plots of overall survival for HNSC HPV-ve male tumour samples with below median expression values for TBL1Y gene (left) and TTTY14 gene (right). In the left hand plot, samples have been split into those which were hypo-methylated (40% threshold) and those which were not. In the right hand plot, samples have been split into those which were hyper-methylated (40% threshold) and those which were not.

For TBL1Y, those patients which had experienced hypo-methylation enjoyed statistically significant better survival (log-rank p-value = 0.0254) than those which were not subject to aberrant methylation. The result for hyper-methylation of TTTY14 was similar, but not quite statistically significant (log-rank p-value = 0.0549). A similar association was also observed for RPS4Y1, but there were too few hyper-methylated samples for the results to be statistically significant.

Summary

Patients whose expression of RPS4Y1, TBL1Y and TTTY14 may have been reduced by aberrant methylation appear to enjoy better survival rates relative to those patients who have suffered loss of the Y chromosome.

6.5 Conclusions

The key conclusions from my survival analyses are as follows:

- there appears to be no association between loss of the Y chromosome and overall survival in the COAD and KIRC datasets;
- however, there does appear to be an association in the HNSC dataset;
- in particular, for HPV negative head and neck cancer patients there is a link between loss of the Y chromosome and impaired overall survival;
- the association for HPV negative patients also applied separately to each arm of the Y chromosome, and these results were unique amongst all chromosomal arms;
- for HNSC HPV negative patients nodal status is by far the most significant factor influencing survival;
- but amongst potential causes of HNSC tumours, loss of the Y chromosome is the most significant factor;
- associations between gene expression and survival are generally consistent with those for loss of the Y chromosome;
- however, for RPS4Y1, TTTY14 and TBL1Y in particular, reduced expression through aberrant methylation may be associated with improved survival compared to patients whose expression is reduced by chromosomal loss.

In my final results chapter I look in more detail at the analyses for the HNSC dataset.

7. Further analyses of head and neck tumours

In this chapter I investigate in greater detail the associations between loss of the Y chromosome and key risk factors for male head and neck tumours.

I decided to concentrate on chromosomal loss as this is by far the most common Y chromosome abnormality I have observed in tumours. I chose the HNSC dataset as it is the only one of the three datasets I have analysed for which TCGA provide useable data on patients' smoking histories, and for which there appears to be a link between Y chromosome loss and impaired survival.

An overview of the process I have followed is provided in figure 102:

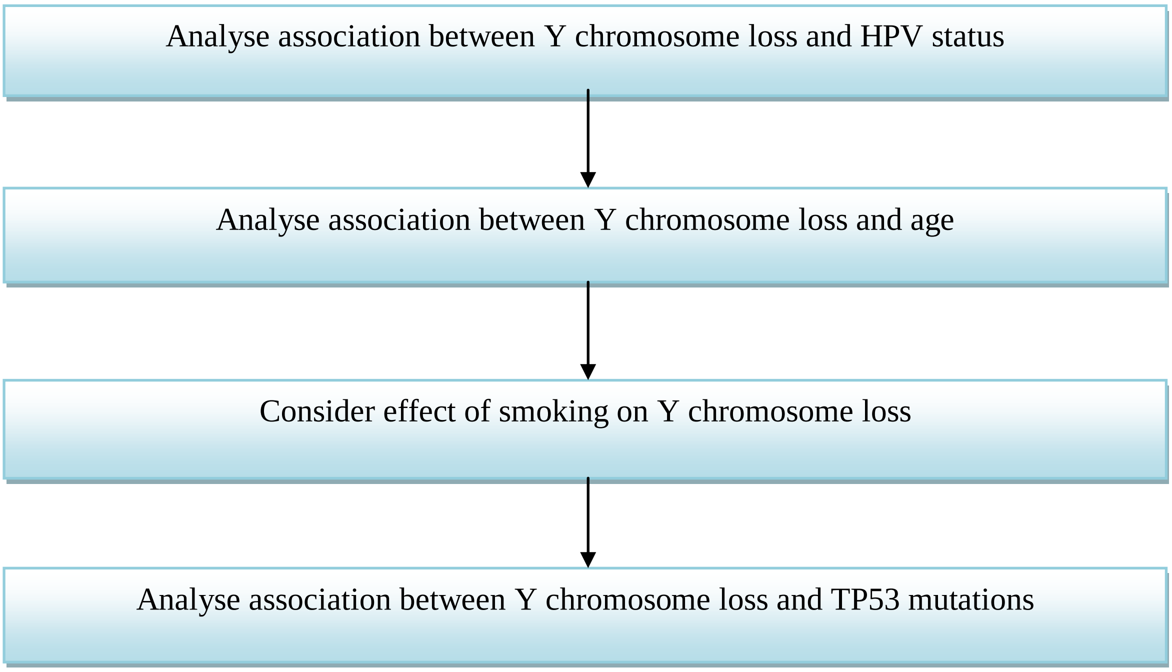

Analyse association between Y chromosome loss and HPV status
Analyse association between Y chromosome loss and age
Consider effect of smoking on Y chromosome loss
Analyse association between Y chromosome loss and TP53 mutations

Figure 102 - overview of experimental process for further analyses of HNSC tumours

I set out details of each of the above stages in the following sections.

7.1 Analysis of Y chromosome loss by HPV status

As I mentioned in the previous chapter, out of the 370 male tumour samples for which both methylation and copy number data were available, 369 also had information on HPV infection status . 85 (23.0%) of these had been classified as HPV positive, and the other 284 as HPV negative. This compares with 12 (8.9%) out of 135 female samples which had been categorised as HPV positive. Hence the prevalence of HPV infection was more than twice as high in the male patients.

7.1.1 Loss of Y chromosome and HPV status

I also mentioned previously that loss of the Y chromosome appeared to be more prevalent in HPV negative cases than in HPV positive cases, and this is shown in figure 103:

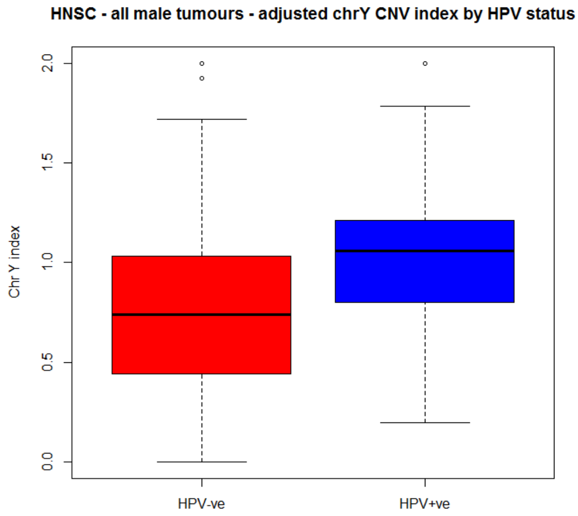


Figure 103 - comparison of Y chromosome loss in HPV-ve & HPV+ve tumours
Boxplots of adjusted Y-chromosome copy number index values for all HNSC males split by HPV status. The boxplot for HPV-ve cases is shown in red and the boxplot for HPV+ve cases is shown in blue.

Whilst there is evidence of some loss of Y chromosome in HPV positive tumours, it is much more common in HPV negative tumours (Kruskal-Wallis p-value < 0.0001).

7.1.2 Breakdown of HPV status by anatomical site

There are five main sub-categories of head and neck cancer by anatomical site:- hypophayrnx, nasopharynx, oropharynx, larynx and oral cavity[88]. TCGA's data contain samples from all of these sites except nasopharynx. Table 12 shows the breakdown of the 369 male tumour samples by HPV status and anatomical site:

Site	HPV-ve	HPV+ve	Total
Hypopharynx	1	5	6
Oropharynx	20	49	69
Larynx	88	6	94
Oral cavity	175	25	200
Total	284	85	369

Table 12 - breakdown of HNSC tumour samples by HPV status and anatomical site

HPV negative tumours are primarily found in the oral cavity and the larynx, whereas oropharyngeal and hypopharyngeal tumours are more commonly HPV positive. In figure 104 I compare, for each HPV subgroup, the extent of Y chromosome loss across the different sites (hypopharynx excluded due to small numbers):

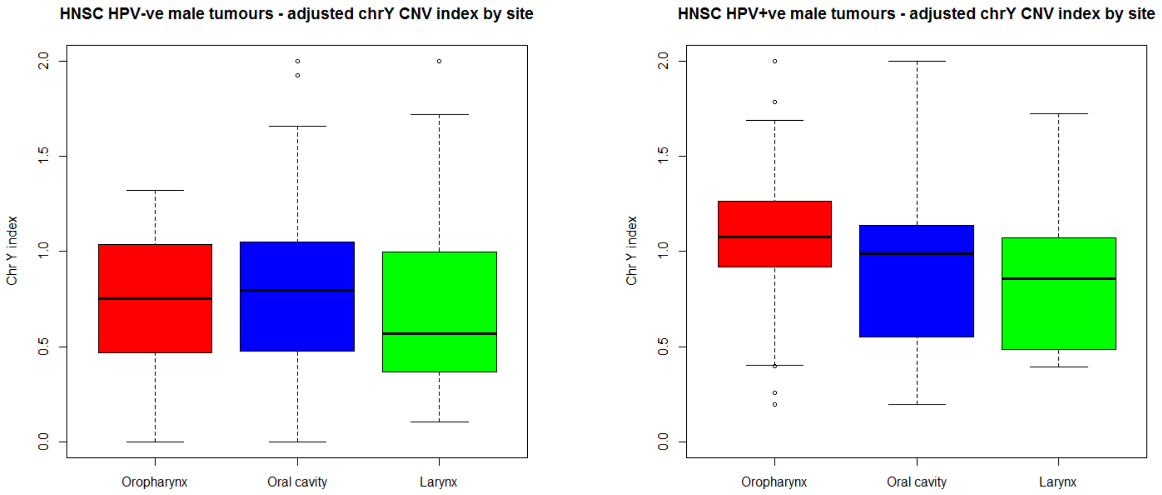


Figure 104 - comparison of Y chromosome loss in different HNSC anatomical sites
Boxplots of adjusted Y chromosome copy number index values for HNSC males split by anatomical site. The left hand plot shows HPV-ve cases and the right hand plot shows HPV+ve cases. The anatomical site categories are colour-coded red (oropharynx), blue (oral cavity), and green (larynx).

Amongst HPV negative tumours, the extent of Y chromosome loss is similar across the three sites shown, albeit slightly greater in laryngeal tumours (Kruskal-Wallis p-value = 0.1979). In the HPV positive tumours, oropharyngeal tumours generally show the lowest levels of Y chromosome loss, whilst laryngeal tumours again show the greatest level of loss (although this is only a small subset of samples, and is not statistically significant – Kruskal-Wallis p-value = 0.1446).

The higher level of Y chromosome loss in HPV negative tumours is evident in all three sites, although this is only statistically significant for oropharygeal cases (Kruskal-Wallis p-value = 0.0019).

7.1.3 Summary

Loss of Y chromosome is more prevalent in HPV negative male patients, and this pattern is observed across different anatomical sites.

7.2 Analysis of Y chromosome loss by age

I next considered whether loss of the Y chromosome was associated with age.

7.2.1 Analysis of age profile by HPV status

Figure 105 compares the age profiles of the HPV negative and HPV positive patients:

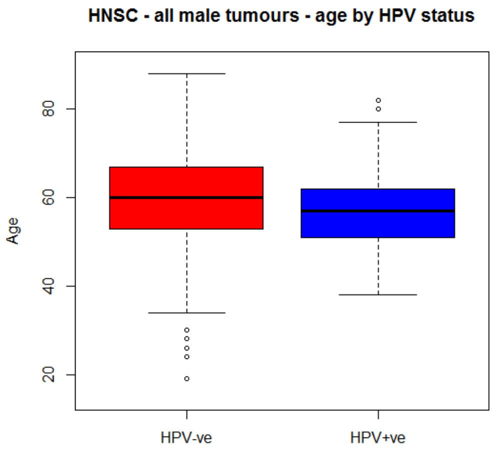


Figure 105 - comparison of age profiles of HPV-ve & HPV+ve tumours
Boxplots of ages for all HNSC males split by HPV status. The boxplot for HPV-ve cases is shown in red and the boxplot for HPV+ve cases is shown in blue.

Figure 105 shows that the HPV positive male patients are slightly younger on average (57.3 years) than the HPV negative patients (61.0 years), and this difference was statistically significant (Kruskal-Wallis p-value = 0.0084). Age was measured at date of initial cancer diagnosis.

In figure 106 I compare, for each HPV group, the age profiles across the different sites (hypopharynx again excluded):

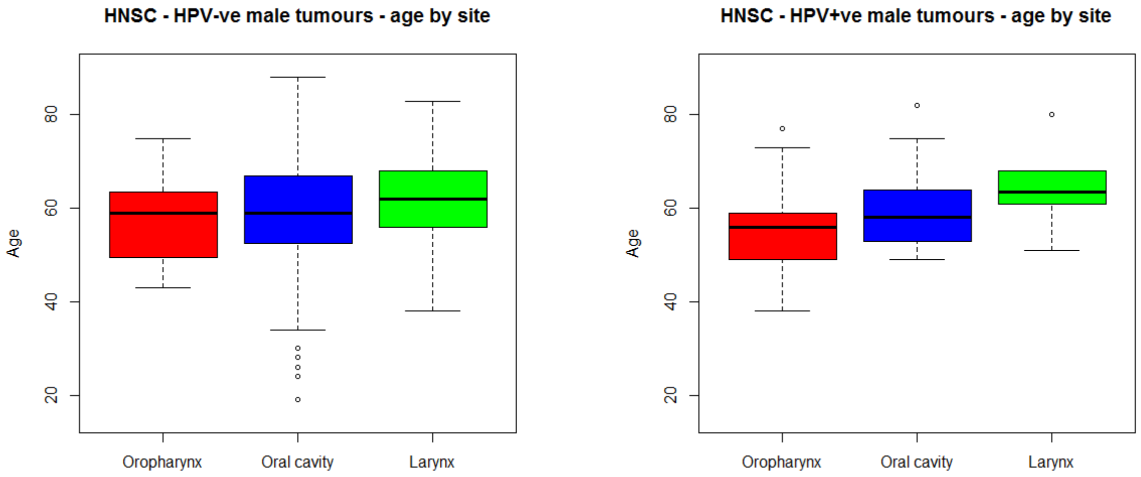


Figure 106 - comparison of age profiles in different HNSC anatomical sites
Boxplots of ages for HNSC males split by anatomical site. The left hand plot shows HPV-ve cases and the right hand plot shows HPV+ve cases. The anatomical site categories are colour-coded red (oropharynx), blue (oral cavity), and green (larynx).

For both HPV negative and HPV positive patients, those with laryngeal tumours tended to be slightly older on average and those with oropharyngeal tumours tended to be the youngest, although these differences were only statistically significant for the HPV positive cases (Kruskal-Wallis p-value = 0.0177).

7.2.2 Variation of Y chromosome loss with age

I then plotted age against adjusted Y chromosome copy number index for all the male patients. This is shown in figure 107, with colour-coding according to HPV status:

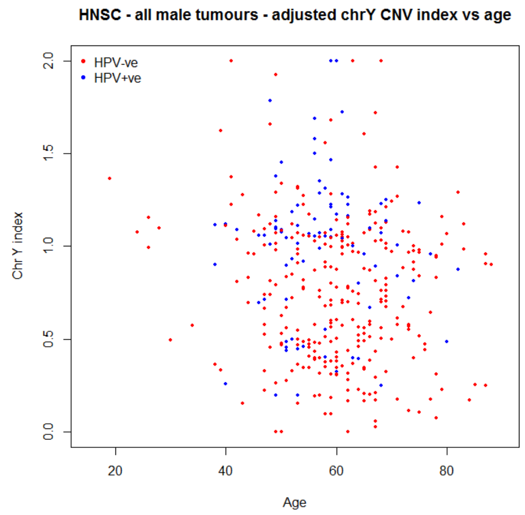


Figure 107 - plot of age vs Y chromosome copy number for HNSC tumours
Scatterplot of age (x-axis) versus adjusted Y chromosome copy number index value (y-axis) for all HNSC male tumours. HPV-ve cases are shown in red and HPV+ve cases are shown in blue.

There was a weak correlation (-0.118, p-value = 0.0230) between age and adjusted Y chromosome copy number index (i.e. loss of Y chromosome tended to increase slightly with age). Separate graphs for HPV negative and HPV positive tumours are set out in figure 108 (colour-coded by anatomical site):

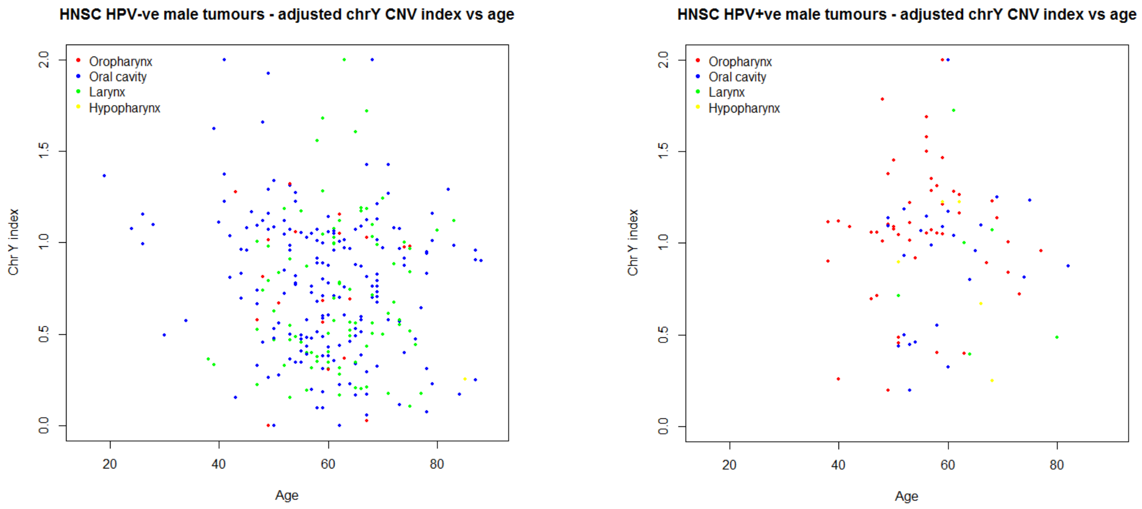


Figure 108 - plots of age vs Y chromosome copy number for HPV-ve and HPV+ve tumours
Scatterplots of age (x-axis) versus adjusted Y chromosome copy number index value (y-axis) for HNSC male tumours. The left hand plot shows HPV-ve cases and the right hand plot shows HPV+ve cases. Cases are colour-coded by anatomical site: red (oropharynx), blue (oral cavity), green (larynx), and yellow (hypopharynx).

A similar, small level of negative correlation (-0.115, p-value = 0.0515) was observed in the HPV negative tumours, whereas there was no correlation in the HPV positive tumours.

7.2.3 Summary

There was a weak correlation between loss of the Y chromosome and age, in particular for HPV negative patients.

7.3 Analysis of Y chromosome loss by smoking history

Smoking is a key risk factor for head and neck cancer[89], and TCGA provide useful data on smoking history for most HNSC patients. I, therefore, decided to extend my analysis to consider the association of smoking with loss of the Y chromosome.

7.3.1 Summary of TCGA smoking data

There is evidence that the influence of smoking on cancer occurrence diminishes over time once someone has stopped smoking[4,99]. TCGA categorise patients into four subgroups on this basis, as I mentioned in the previous chapter. Data on smoking history was available for 276 of the 284 HPV negative patients and for 84 of the 85 HPV positive patients, as summarised in table 13:

Smoking history	HPV-ve	HPV+ve	Total
Never smoked	41 (14.9%)	24 (28.6%)	65 (18.1%)
Stopped >15 years ago	39 (14.1%)	12 (14.3%)	51 (14.2%)
Stopped <15 years ago	82 (29.7%)	25 (29.8%)	107 (29.7%)
Current smoker	114 (41.3%)	23 (27.4%)	137 (38.1%)
Total	276	84	360

Table 13 - breakdown of HNSC male tumour samples by HPV status and smoking history

Table 13 shows that smoking was more prevalent amongst HPV negative patients (41.3% vs 27.4% are current smokers, chi-squared p-value = 0.0186).

Frequency of smoking is another, potential risk factor for head and neck cancer[4]. TCGA also provide some information on amount of cigarettes smoked, but this information was not available for 75 patients. Based on the data that were available, I observed no association between frequency of smoking and Y chromosome loss (data not shown).

7.3.2 Association of smoking history and loss of Y chromosome

I then considered the extent to which loss of the Y chromosome was associated with smoking history. In figure 109 I compare, for each HPV subgroup, the extent of Y chromosome loss across the four smoking categories:

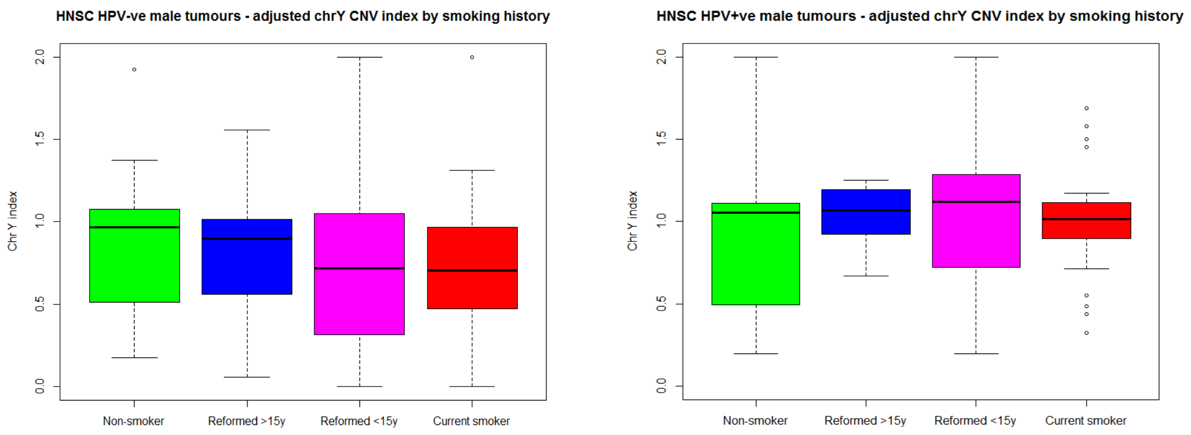


Figure 109 - comparison of Y chromosome loss by smoking history
Boxplots of adjusted Y chromosome copy number index values for HNSC males split by smoking history category. The left hand plot shows HPV-ve cases and the right hand plot shows HPV+ve cases. The smoking history categories are colour-coded green (non-smoker), blue (former smoker who stopped more than 15 years ago), pink (former smoker who stopped within the last 15 years) and red (current smoker).

For the HPV negative patients, there is a trend (Kruskal-Wallis p-value of 0.0682) for loss of the Y chromosome to increase across the subgroups (from non-smokers to current smokers). There is, however, no obvious trend amongst the HPV positive cases.

7.3.3 Summary

Smoking was more prevalent amongst HPV negative patients, and there was a trend for Y chromosome loss to be associated with smoking for these patients.

7.4 Analysis of Y chromosome loss by TP53 mutations

The final risk factor that I considered was whether each patient harboured a mutation of the TP53 gene. For this purpose I used TCGA’s somatic mutation data, which was available for 270 of the 276 HPV negative and 78 of the 84 HPV positive male patients for whom smoking data were also available.

7.4.1 Summary of TP53 mutation data

The breakdown of the male patients by HPV status, smoking history and TP53 mutation status is set out in table 14:

Smoking history	HPV-ve		HPV+ve		Total	
	TP53-ve	TP53+ve	TP53-ve	TP53+ve	TP53-ve	TP53+ve
Never smoked	5	35	20	3	25	38
Stopped >15 years ago	14	25	7	3	21	28
Stopped <15 years ago	12	67	17	6	29	73
Current smoker	21	91	13	9	34	100
Total	52	218	57	21	109	239

Table 14 - breakdown of HNSC tumours by HPV status, smoking history & TP53 status

Table 14 shows that 239 (68.7%) out of the 348 male patients for which data were available had a mutation of TP53. However the frequency of mutation was much greater (chi-squared p-value < 0.0001) in HPV negative cases (80.7%) than in HPV positive cases (26.9%). For the former group, there was no clear evidence of mutational frequency being linked to smoking, whereas for HPV positive cases there was a trend (not statistically significant) for mutational frequency to be greater in smokers.

7.4.2 Association of Y chromosome loss with TP53 mutations

I then considered whether loss of the Y chromosome was associated with TP53 mutations. Initially I compared, for each HPV subgroup, the extent of Y chromosome loss between patients with and without a TP53 mutation, as shown in figure 110:

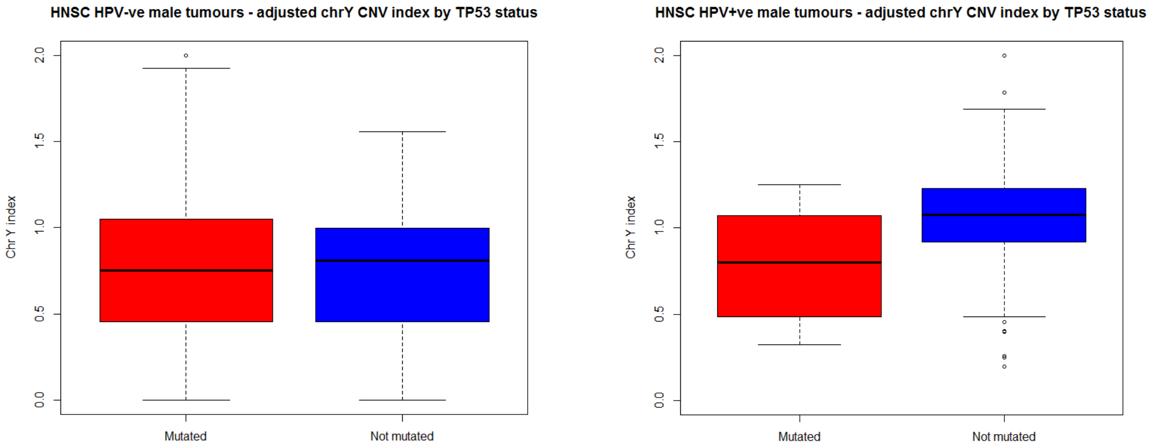


Figure 110 - comparison of Y chromosome loss by TP53 mutations
Boxplots of adjusted Y chromosome copy number index values for HNSC HPV-ve (left) and HPV+ve (right) males split by TP53 mutation status. The boxplots for cases with a TP53 mutation are shown in red, and the boxplots for non-mutated cases are shown in blue.

For HPV negative patients, the extent of Y chromosome loss did not appear to be associated with TP53 mutational status (Kruskal-Wallis p-value = 0.7688). However, HPV positive patients who harboured a TP53 mutation appeared to suffer greater loss of the Y chromosome than those who did not have a mutation (Kruskal-Wallis p-value = 0.0116).

For each HPV subgroup, I then further sub-divided the data by smoking history. The comparisons for HPV negative patients are shown in figure 111:

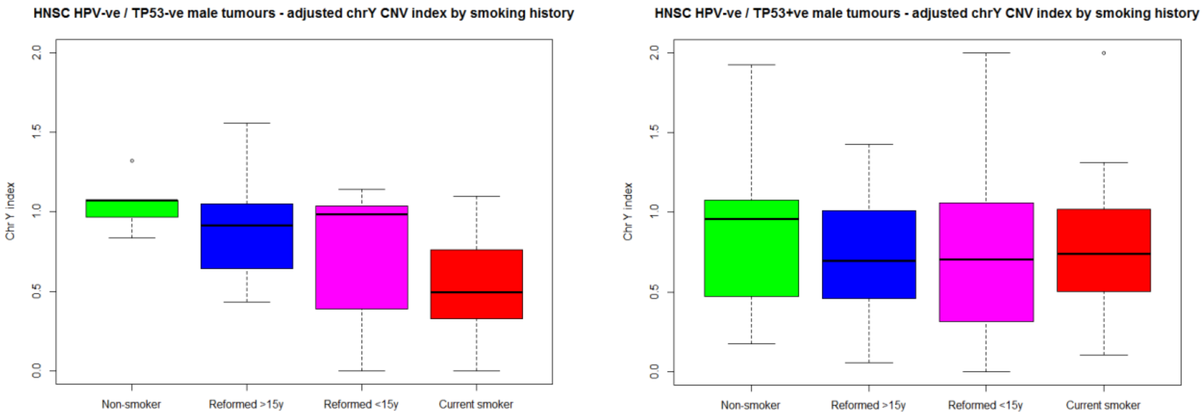


Figure 111 - comparison of Y chromosome loss by TP53 status & smoking history for HPV-ve patients
Boxplots of adjusted Y chromosome copy number index values for HNSC HPV-ve males split by smoking history category. The left hand plot shows cases with no TP53 mutation and the right hand plot shows cases with a TP53 mutation. The smoking history categories are colour-coded green (non-smoker), blue (former smoker who stopped more than 15 years ago), pink (former smoker who stopped within the last 15 years) and red (current smoker).

For the 52 HPV negative patients who did not harbour a TP53 mutation, there was a clear association between loss of the Y chromosome and smoking (Kruskal-Wallis p-value = 0.0047). However, there was no association for the 218 patients who did have a TP53 mutation (Kruskal-Wallis p-value = 0.4561).

Figure 112 shows the corresponding comparisons for the HPV positive patients:

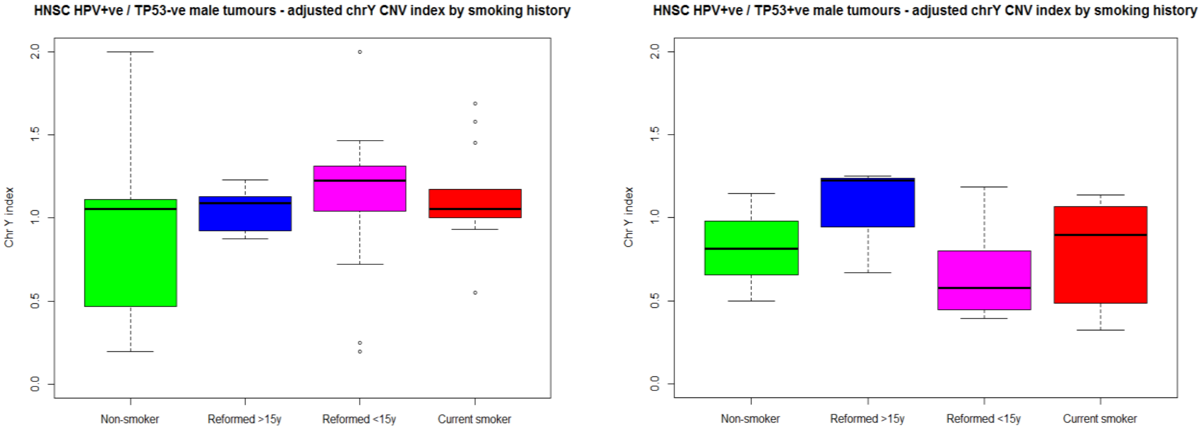


Figure 112 - comparison of Y chromosome loss by TP53 status & smoking history for HPV+ve patients
Boxplots of adjusted Y chromosome copy number index values for HNSC HPV+ve males split by smoking history category. The left hand plot shows cases with no TP53 mutation and the right hand plot shows cases with a TP53 mutation. The smoking history categories are colour-coded green (non-smoker), blue (former smoker who stopped more than 15 years ago), pink (former smoker who stopped within the last 15 years) and red (current smoker).

In both cases there was no obvious association between loss of the Y chromosome and smoking (Kruskal-Wallis p-values = 0.3568 and 0.3196 respectively).

For the HPV negative cases with no TP53 mutation, I also carried out the same comparison for each arm of the Y chromosome, as shown in figure 113:

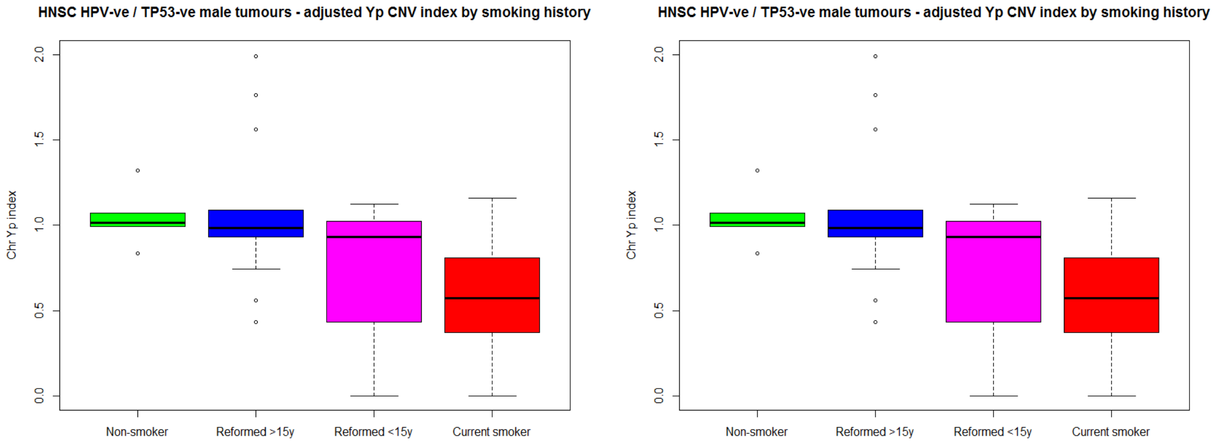


Figure 113 - comparison of Y chromosome arm-level loss by smoking history for HPV-ve/TP53-ve patients
Boxplots of adjusted Y chromosome arm-level copy number index values for HNSC HPV-ve males with no TP53 mutation split by smoking history category. The left hand plot shows results for the short arm (Yp) and the right hand plot shows results for the long arm (Yq). The smoking history categories are colour-coded green (non-smoker), blue (former smoker who stopped more than 15 years ago), pink (former smoker who stopped within the last 15 years) and red (current smoker).

Similar, statistically significant associations with smoking were observed for both the short and long arms of the Y chromosome (Kruskal-Wallis p-values = 0.0055 and 0.0086 respectively). Furthermore, I performed the same analyses for every chromosome arm, and discovered that statistically significant associations were unique to the Y chromosome.

7.4.3 Summary

HPV positive patients with a TP53 mutation had generally higher levels of Y chromosome loss than their counterparts with no mutation. However, the levels of Y chromosome loss for HPV negative patients were similar regardless of TP53 mutation status. For HPV negative patients with no TP53 mutation there was a clear association between smoking and loss of chromosome Y. This association was observed in both arms of the Y chromosome, and was not observed in any other chromosome.

7.5 Conclusions

The main conclusions from my analyses of the associations of Y chromosome copy number changes and key risk factors in male head and neck tumours are as follows:

- HPV infection was more than twice as prevalent in male than female HNSC patients;
- loss of the Y chromosome was more prevalent in HPV negative male tumours than in HPV positive tumours;
- and this difference was observed across different anatomical sites;
- there was a minor correlation between loss of the Y chromosome and age, in particular for HPV negative tumours;
- a history of smoking was more prevalent amongst HPV negative patients than amongst HPV positive patients;
- for HPV negative patients there was a trend for loss of the Y chromosome to be associated with smoking;
- mutations of TP53 were much more frequently observed in HPV negative patients than in HPV positive patients;
- for HPV positive patients, there was some evidence that TP53 mutations are more frequently observed in smokers than non-smokers;
- for HPV positive patients, loss of the Y chromosome was greater on average when there was a TP53 mutation;
- however, the same association was not observed in HPV negative patients;
- for HPV negative patients with no TP53 mutation, there was a clear association between loss of the Y chromosome and smoking;

- furthermore, the same association with smoking was observed independently in both arms of the Y chromosome;
- and this association was unique to the Y chromosome.

I will discuss all of my results in the next chapter.

8. Discussion

To my knowledge this is the first time that a detailed, multi-‘omic analysis of the male-specific region of the Y chromosome has been performed using TCGA, or indeed any, dataset. By limiting my research solely to the Y chromosome, thereby excluding all female samples, I have been able to discover features which are potentially overlooked in studies which follow unisex, genome-wide approaches. These are discussed in the following sections.

8.1 Methylation analyses

There has been very little research into aberrant methylation of the Y chromosome in human cancer. My analyses have yielded results which are largely consistent between the three datasets, and which are in line with previous, more general findings of methylation in cancer.

First, methylation levels are generally similar in normal samples across the three datasets. However, there were some differences for a small proportion of individual CpG sites, which suggests that methylation could be an important regulator of gene transcription for some Y chromosome genes.

Second, methylation levels are more variable amongst tumour samples than amongst normal samples, indicating that, consistent with the rest of the genome, the Y chromosome methylome is disrupted during the tumorigenic process. Furthermore, this disruption appears not to occur in an entirely random manner, as hyper-methylated CpG sites tend to be concentrated in relatively small, discrete regions of the Y chromosome, and are interspersed between much larger regions of general hypo-methylation[34].

Hyper-methylation of the Y chromosome in tumour samples occurs predominantly in or near CpG islands and close to the transcriptional start sites of genes, which implies that it could be an important causative agent in aberrant gene expression (see below). Several genes are affected by hyper-methylation, including the putative tumour suppressor gene ZFY[6] and other genes which have been suggested as important for male viability, namely RPS4Y1, TBL1Y, NLGN4Y, CYorf15A and EIF1AY[5].

The SRY gene is also hyper-methylated in a large number of tumour samples, in particular within the HNSC dataset. This is of particular interest for the following reasons:

- SRY is expressed in normal HNSC samples but not in COAD or KIRC samples (see later);
- Methylation levels are lower in normal HNSC samples compared with COAD and KIRC; and
- Previous research has suggested that SRY expression is regulated by methylation in prostate cancer, and can be restored using de-methylating agents[22].

Hypo-methylation is more common than hyper-methylation in all three datasets, but generally occurs at a lower level. Once again, CpG sites within a small number of genes are affected, including the afore-mentioned TBL1Y and NLGN4Y.

My analyses also revealed some interesting, technical issues with TCGA’s methylation data. First, around 10% of the Y chromosome probes appear to cross-hybridise with other parts of the genome and are, therefore, unreliable. Problems with cross-hybridisation are a known feature of the Illumina platform[100].

Second, methylation levels appear to be distorted by copy number changes, consistent with previous research[14]. This was clearly evident in cases which had lost the Y chromosome, but there was also some evidence (data not shown) that high copy number may also have a distorting effect for some probes.

8.2 Copy number analyses

Loss of the Y chromosome has previously been reported for all three cancer types studied[25,26,27]. My analyses, using both TCGA's methylation and copy number data, confirm that the Y chromosome is frequently lost in each dataset, and that loss generally occurs at the whole chromosome level, with both arms being lost in similar proportions. 35%, 42% and 39% of COAD, HNSC and KIRC male patients respectively appear to have suffered some loss of the Y chromosome, making this the most common copy number aberration in male tumours.

For each dataset there was a clear bimodal distribution of copy number index values, which suggests that in each case there are two distinct sub-populations of patients – one who have lost the Y chromosome and another who have not. The first peaks of the distributions occurred at 0.39, 0.47, and 0.46 for COAD, HNSC and KIRC samples respectively, and the second peaks occurred at 1.02, 1.04 and 1.09 respectively. In each case there was an intervening trough at 0.62, 0.74 and 0.74 respectively. The two sub-populations merge together between the two peaks, and so for patients whose index values fall in this range, it is not clear which sub-population they actually belong to. The fact that the peaks are not sharper and that there is considerable overlap between the two distributions will in part be due to some inaccuracy within the data

Although the Y chromosome was frequently lost in each dataset, I rarely found that it was completely lost in any individual sample – for example, the peak of the distribution for samples which had lost the Y chromosome occurred at an index value of between 0.39 and 0.47 rather than zero, even after making allowance for the proportion of cells in each sample which TCGA actually believed to be non-cancerous. Aside from any inaccuracy in the data measurements, there are a couple of possible explanations for this observation. First, the assessment of tumour purity on which I based my calculations may not be precise. Second, and probably more importantly, Y chromosome loss may not occur in every tumour cell, but may instead be restricted to a subset of cells.

Whether Y chromosome loss is an important factor in tumorigenesis, or simply a by-stander effect of little consequence, has been a matter of debate. For example, research in acute myelogenous leukaemia has suggested that it is primarily a consequence of ageing, with no causal effect[101]. However, other research using a prostate cancer cell-line in a mouse model has shown that reintroducing a lost Y chromosome can suppress tumour formation[102]. These conflicting results raise the possibility that the importance of Y chromosome loss to cancer development may depend on the anatomical location in which it occurs.

On a technical level, my work also confirms that total intensity measurements generated by the Illumina 450k methylation platform can be used to give very good estimates of copy number levels[14]. Furthermore, I discovered that a small number of Y chromosome probes on the 450k platform appear to cross-hybridise with other regions within the Y chromosome, and therefore give potentially unreliable measurements.

8.3 Gene expression analyses

My analyses revealed that 15 Y chromosome genes are expressed in all three datasets. These include all 12 of the genes identified in the literature as potentially being important for male viability[5], along with three others, namely TTTY14, TTTY15 and NCRNA00185. One further gene, SRY, was only expressed in the HNSC dataset.

As might be expected, expression levels of each gene were associated with copy number variations. For most of the genes, these associations were strong. However, for a small number of genes, the associations were weaker, and in these cases, there was evidence that aberrant methylation, both hyper- and hypo-, may also be a contributory factor in reduced expression.

The fact that all 12 of the genes put forward by Bellott[5] as being essential for male viability were expressed in all of the three datasets suggests that their functionality is important across different tissue types. Furthermore, each has an X-chromosome homolog which escapes X-inactivation in females, suggesting that maintenance of total expression levels across the X/Y homologs is important.

Using gene ontologies based on their X-chromosome homologs, Bellott[5] showed that at least eight of these genes are likely to play roles in several fundamental biological processes, as set out in the Venn diagram in figure 114 taken from his paper:

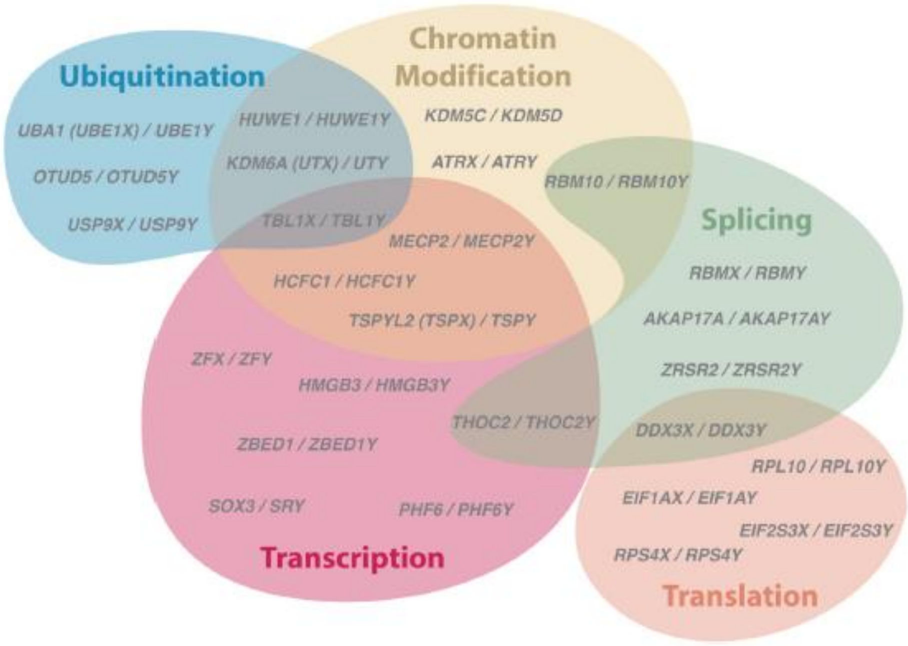


Figure 114 - Y chromosome gene ontologies

This diagram is taken from Bellott[5] and shows categorisations of genes on the Y chromosome using the ontologies of their X-chromosome counterparts

Several of these genes have also been implicated in other pathologies[16]. For example:

1. DDX3Y, KDM5D, RPS4Y1, TMSB4Y, USP9Y and UTY have been shown to code for minor-histocompatability antigens which are believed to play a role in graft-versus-host disease or graft rejection after sex-mismatched bone marrow transplants.
2. TBL1Y and NLGN4Y have been implicated in the four-fold higher incidence of autism in males relative to females.
3. Over-expression of DDX3Y, EIF1AY, RPS4Y1 and USP9Y has been implicated in gender differences in the incidence of cardiovascular disease.

Turning to the other three genes which were expressed in all three datasets, the fact that two genes of the TTTY family (TTTY14 and TTTY15) were expressed is potentially interesting in that these are, by definition, genes whose functionality is supposedly limited to the testes. Little is known about them, though they are believed to code for non-coding RNAs – indeed TTTY14 is also known as NCRNA000185 (http://www.ncbi.nlm.nih.gov/gene/83869).

Finally, SRY is potentially the most interesting of all the 16 genes as it was only expressed in the HNSC dataset. SRY is also known as the testis-determining-factor, and is the key gene responsible for specification of the male gender[16]. As such, it is mainly expressed in male-specific tissues such as the testes. However, it has also been shown to be expressed in other tissues, such as the brain[103]. Hence, it is possible that SRY has other male-specific roles beyond its main sex-determination function.

8.4 Head and neck cancer analyses

My analyses of overall survival provided some potentially interesting results for the HNSC dataset.

First, I discovered that there was an association between loss of the Y chromosome and impaired overall survival in the HNSC dataset. No such association was evident in the COAD and KIRC datasets, which raises the possibility that Y chromosome loss may contribute to the pathogenesis of head and neck cancer.

Previous research has reported that there may be an association between loss of the Y chromosome and impaired survival in head and neck cancer[27]. However, that research was based on a much smaller dataset than the one I have used.

Also, consistent with previous research[28], my results suggest that Y chromosome loss in head and neck cancer is not simply a by-product of the ageing process, as there was only a very weak correlation between the extent of loss and age.

To investigate the survival association in more detail, I decided to refine my analyses by first splitting samples between those which had been assessed by TCGA as HPV positive and those which had been assessed as HPV negative. I believe this is a reasonable subdivision of the data to make for the following reasons. HPV infection was first discovered in some tumours of the oropharynx in 1983[104], and since then research has indicated that it is a primary cause of a subset of head and neck cancers[105]. In particular, HPV positive tumours differ from HPV negative tumours in several respects such as epidemiology, molecular biology and outcome[106] – for example:

- They tend to occur at younger ages, especially in males, and patients with HPV positive tumours tend to have lower consumption of tobacco and alcohol;
- HPV positive tumours are much more prevalent in certain anatomical sites – in particular the oropharynx;
- The ways in which key signalling pathways are disrupted are different – for example, the TP53 gene is the most commonly mutated gene in HPV negative tumours, potentially leading to disruption of downstream pathways involved with important biological processes such as cell-cycle control, apoptosis, DNA repair and metabolism. However, the frequency of TP53 mutation in HPV positive tumours is much lower. Instead, the p53 protein product of TP53 is often degraded by the viral oncoprotein E6;

- Patients with HPV positive tumours, especially of the oropharynx, have been found to have better survival compared to patients with HPV negative tumours.

The first three of the above characteristics were all observed in my analyses, although the difference in outcome was not seen.

Detection of HPV infection is a potential issue, as the techniques used for this purpose are not necessarily 100% reliable[107]. These techniques include p16 immunohistochemistry (the p16 protein is used as a surrogate marker for HPV infection, since the viral oncoprotein E7 degrades the RB1 gene which in turn leads to overexpression of p16), in-situ hybridisation for high risk HPV, quantitative polymerase chain reaction for either viral mRNA or DNA, and detection of viral mRNA using RNA-seq analysis[107]. In their determination of HPV statuses, TCGA used the last of these techniques[12], which is regarded as being an accurate method for identifying HPV positive samples[107]. They aligned mRNA from each sample to the HPV transcriptome, and then, based on the distribution of results, they categorised any sample that had at least 1,000 aligned reads as being HPV positive. Whilst this threshold was an arbitrary choice, it appears to be reasonable based on the data. They also performed other techniques for determining HPV status, and observed good concordance of results[12].

For both HPV positive and HPV negative tumours I again observed bimodal distributions of Y chromosome copy number index values, indicating that both groups had sub-populations which had lost the Y chromosome. However the first peak (of tumours which had lost the Y chromosome) was much higher in HPV negative cases, suggesting that loss of the Y chromosome is much more prevalent in HPV negative cases.

Whilst there was some evidence that loss of the Y chromosome may be associated with worse overall survival in HPV positive patients, the result was not statistically significant. However, I did discover a statistically significant association in HPV negative patients. This applied to the whole chromosome and also to each arm separately. Furthermore, I observed that the association was unique to the Y chromosome, not being found in any other chromosomal arm. I also discovered a significant association between extent of Y chromosome loss and pathologic T categorisation.

I then further subdivided samples depending on whether or not they harboured a mutation of the TP53 gene. TP53 is the gene most frequently mutated in head and neck cancer (in particular HPV negative tumours)[106]. TP53 mutation is believed to be an early event in carcinogenesis and potentially has multiple downstream impacts on important signalling pathways[108]. Consistent with previous research I observed that TP53 mutation was much more frequent in HPV negative tumours than in HPV positive tumours[106].

My analyses using TP53 mutation data provided two other pieces of evidence supporting the possibility that Y chromosome loss could be important in tumour formation for HPV negative head and neck cancer patients. First, I showed that the extent of Y chromosome loss is independent of TP53 mutation. Conversely, for HPV positive cases, Y chromosome loss was significantly greater in patients who harboured a TP53 mutation. This implies that for HPV negative cases, the Y chromosome can be lost in the absence of TP53 mutation.

Second, for those HPV negative patients who did not harbour a TP53 mutation, I observed a statistically significant link between Y chromosome loss and a history of smoking. This result applied separately to both arms of the Y chromosome, and was again unique amongst all chromosomal arms. The corollary from this is that smoking could be a causative factor in Y chromosome loss, consistent with research published in 2015[4].

I chose to subdivide samples by TP53 mutational status as TP53 is the most commonly mutated gene in head and neck cancer[106], it is involved in numerous important signalling pathways, and it is also mutated in other types of cancer (e.g. lung cancer)[109]. Analysis of TCGA mutation data for 3,281 tumours from 12 different cancer types (including head and neck) found that TP53 was the most frequently mutated gene, occurring in 42% of all samples[109]. Other research has found it to be mutated in different cancer types, such as epithelial, mesenchymal and haematological malignancies[110].

Given the high prevalence of mutation across multiple cancer types, and the large number of critical signalling pathways which it affects, TP53 has earned the reputation as the "guardian of the genome", as it is important for the maintenance of genomic stability[108]. Analysis using the COSMIC database of somatic mutations in cancer has shown that 22.38% of the 67 autosomal genes in the TP53 pathway contain mutations which have been causally implicated in cancer, including ATM. MDM2 and CDKN2A[110]. This represented a 11.15-fold enrichment compared with the corresponding figure when all genes were considered. Figure 115 overleaf shows a diagram of the TP53 pathways (taken from Stracquadanio[110]), including the genes included in the COSMIC list.

Although TP53 is clearly a very important gene in head and neck cancer, there are, however, a variety of different mutations which occur, and they can have different downstream effects. TP53 is generally regarded as a tumour suppressor gene, and most loss-of-function mutations occur in its DNA-binding domain, which inhibits its function as a transcription factor[108].

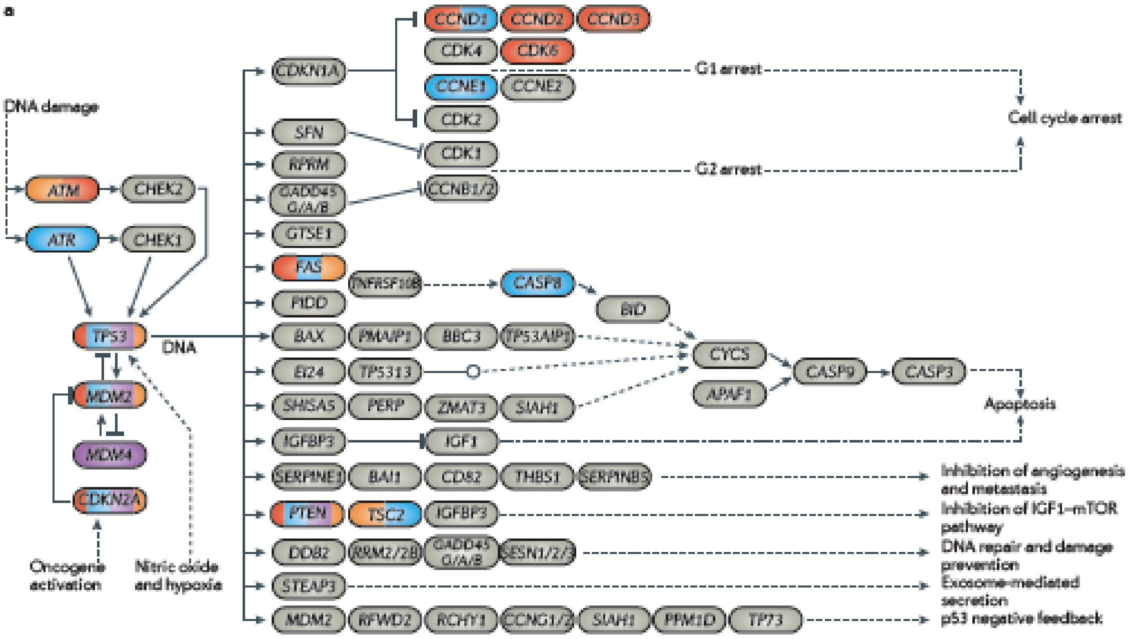


Figure 115 - schematic diagram of TP53 pathways and genes causally implicated in cancer
This diagram is taken from Stracquadanio[110] and shows the p53 pathway as annotated by Kyoto Encyclopedia of Genes and Genomes. Genes for which mutations have been causally implicated in cancer (COSMIC) are coloured according to the relevant cancer type(s) - blue (epithelial cancers), red (leukaemia or lymphoma), purple (mesenchymal) or orange (other).

However, the precise mechanisms by which TP53 performs its tumour suppression role are unclear – for example, research in mice has suggested that suppression of TP53’s cell-cycle arrest, apoptosis and senescence functionality does not promote tumour formation[108]. To add to the complexity, other research has shown that certain TP53 mutations may actually cause gain-of-function which in turn promotes tumour progression[108]. Hence, whilst my observations of the associations between the existence of TP53 mutation and Y chromosome loss are potentially interesting, further detailed analysis of the different types of mutation, and its downstream consequences, would be necessary to refine the findings.

Some research, including that of TCGA[12], has indicated that HPV negative patients with a TP53 mutation suffer worse survival than those patients with no mutation[111]. However, I did not observe such an association for the HPV negative male patients. This could be because

TCGA's (and other) analyses included both males and females, and hence the impact of Y chromosome loss on survival was reduced.

Gross et al[112] used TCGA's data to investigate pairs of genomic events and their associations with survival. They discovered that for HPV negative cases TP53 mutation was often accompanied by loss of the short arm of chromosome 3 (3p), and that patients who had suffered both of these events had worse prognosis. However, once again males and females were combined. Furthermore, the Y chromosome was excluded from the analysis. For HPV negative males I observed frequent loss of 3p, but although there was some evidence that patients who had lost 3p had worse survival, the result was not statistically significant (data not shown).

TCGA[12] also identified a subset of HPV negative tumours which appeared to be driven by mutations and which had low level copy number changes (the so-called "M" class[113]). These patients were observed to have favourable outcomes. However, once again, male and female patients were not segregated, and the Y chromosome was not included in the analysis.

It is interesting to speculate why Y chromosome loss should impact on survival for head and neck cancer patients but not for colon or kidney cancer patients.

It could simply be the case that the result I have observed is purely a quirk of the data and will never be validated in another dataset (see below). Alternatively, it might be that there are other factors at play in the colon and kidney datasets, and that there were insufficient numbers of samples to observe a similar result.

More intriguingly, it is also possible that a gene (or genes) on the Y chromosome are particularly important in the context of head and neck cancer. A prime candidate would be

SRY, which was only expressed in HNSC samples. However, my survival analyses based on gene expression levels suggest that low SRY expression on its own is not linked to worse overall survival. Also, the fact that low expression caused by aberrant methylation in RPS4Y1, SRY, TBL1Y and TTTY14 was associated with better survival compared with patients who had lost the Y chromosome, suggests that loss of expression in these genes on their own is not important. An alternative theory could, therefore, be that loss of expression in a combination of genes, possibly including SRY, may be the important factor.

Loss of some Y chromosome genes has been implicated in tumour formation / progression for other cancer types. For example, loss of the KDM5D gene on the long arm has been associated with resistance to chemotherapy in prostate cancer cell-lines[114].

Loss of UTY has been linked to increased cell proliferation in urothelial bladder cell-lines[115]. UTY is one of the two putative tumour suppressor genes mentioned in Davoli[6]. It is a homolog of UTX (also known as KDM6A) on the X chromosome, a gene which escapes X-inactivation in females. UTX is known to act as a demethylase of lysine 27 on histone 3 (H3K27), and research has shown that UTY may also share this enzymatic activity, albeit at a reduced level[116]. Furthermore, research using head and neck cancer cell-lines has shown that aberrant methylation of H3K27 is linked to dysregulated squamous cell differentiation[117].

TCGA discovered that 19% of head and neck tumours harboured an inactivating mutation in the NOTCH1 gene[12], which is also involved in squamous cell differentiation. So it is possible that loss of UTY may play a role in dysregulation of squamous cell differentiation, which is known to be important in the pathogenesis of squamous cell carcinomas[117].

Another characteristic of head and neck tumours is their ability to evade immune surveillance[106]. The mechanisms by which this evasion is achieved include the downregulation of antigen presenting molecules (such as TAP1/2), the expression of receptors which inhibit immune response (such as PD-L1), and the over-production of immune-suppressive cytokines in the tumour micorenvironment[106].

Therapeutic targeting of tumour immune suppression is an active area of cancer research. Recently Ferris et al[118] reported that treatment with the drug nivolumab resulted in increased overall survival of patients with chemotherapy-resistant, recurrent head and neck squamous cell carcinoma. Nivolumab is an inhibitor of a molecule called programmed death 1 (PD-1), which when bound by its ligand programmed death ligand 1 (PD-L1), suppresses the action of immune system T-cells. PD-L1 is often expressed by head and neck tumours[106].

It is possible that loss of the Y chromosome may contribute to the ability of tumours to evade immune detection. Research using mouse models[119], showed that the Y chromosome is able to regulate, on a genome-wide basis, the expression of genes in immune cells via epigenetic mechanisms. Whilst that research was focussed on autoimmune diseases, there is the possibility that the same principle could be extended to tumour cells – for example, loss of the Y chromosome may result in dysregulated expression of proteins such as PD-L1. Obviously this is highly speculative at the present time, but it could be another mechanism through which loss of the Y chromosome impacts on the progression of male head and neck cancer.

Further research will be required to attempt to identify which Y chromosome gene / genes are the most important, or whether it is the loss of them all collectively that is important in male, HPV negative head and neck cancer formation. However, my results open up the possibility that the Y chromosome may, in some way, be a useful biomarker of survival for such patients.

8.5 Further work

Whilst my results have raised the possibility that loss of the Y chromosome may be a contributory factor in the pathogenesis of head and neck cancer, further work will be required to try and validate my findings.

I have searched online for other datasets which I could use to check my findings in relation to head and neck cancer. Unfortunately, so far I have not been able to find a suitable dataset. Therefore, our group, in conjunction with our collaborators, are proposing to create our own data using tissue samples which have already been collected. We will use these, in particular, to investigate using FISH analysis the extent to which the Y chromosome is lost, and to perform survival analyses. In our analyses we will also include some oral dysplasia samples, which will enable us to investigate whether Y chromosome loss is an early event in tumorigenesis.

We are also intending to carry out FISH analyses on selected, male head and neck cancer cell-lines to investigate Y chromosome loss. For cell-lines which have lost the Y chromosome, we then intend to check that expression of a panel of selected genes is missing. If the Y chromosome is still intact, we will check whether expression of selected genes is depressed as a result of promoter hyper-methylation. Furthermore, we will investigate the phenotypic consequences of re-introducing these genes into the cell-lines. Expression of those genes with the most striking results will then be analysed in the tissue samples, with a view to assessing their potential utility as biomarkers of survival.

8.6 Conclusions

In conclusion, aberrant methylation and copy number loss of the Y chromosome were frequently observed in each dataset, and were often associated with reduced gene expression.

For HPV negative head and neck cancer cases, Y chromosome loss was associated with impaired survival, which raises the possibility that it could be used as a biomarker for such patients. However, further work is required to validate my findings.

Appendices

Appendix 1 – 32 CpG sites removed from methylation analyses

CpG site ID	Infinium type	Position (Mbp)	Gene name	Gene location	Island region	CpG count
cg01053349	II	3.45	TGIF2LY	TSS1500	Ocean	8
cg03515901	II	3.45	TGIF2LY	5'UTR	Ocean	17
cg04477336	II	3.45	TGIF2LY	Body	Ocean	19
cg27539833	I	3.45	TGIF2LY	3'UTR	Ocean	12
cg09703571	II	4.87	PCDH11Y	TSS1500	N_Shore	21
cg10465579	II	4.87	PCDH11Y	TSS1500	Island	22
cg08455548	II	4.87	PCDH11Y	TSS1500	Island	23
cg15295597	I	4.87	PCDH11Y	5'UTR	Island	30
cg02494853	I	4.87	PCDH11Y	5'UTR	Island	29
cg00223952	II	4.87	PCDH11Y	5'UTR	N_Shore	25
cg02432075	II	4.87	PCDH11Y	5'UTR	Island	38
cg13884608	II	5.61	PCDH11Y	3'UTR	Ocean	5
cg08045599	II	7.14	PRKY	TSS1500	N_Shore	24
cg04964672	I	7.43	nil	nil	Island	27
cg01086462	II	7.43	nil	nil	S_Shore	3
cg25059696	II	7.43	nil	nil	S_Shelf	4
cg02624968	II	10.04	nil	nil	N_Shore	18
cg04462340	I	13.91	nil	nil	S_Shore	13
cg09093035	II	14.07	nil	nil	Island	43
cg03258315	II	14.08	nil	nil	Island	23
cg26520468	I	14.1	nil	nil	N_Shore	15
cg05213048	I	14.1	nil	nil	Island	53
cg04023335	I	14.1	nil	nil	Island	47
cg09197443	II	14.11	nil	nil	Island	27
cg15183843	II	14.53	nil	nil	N_Shelf	7
cg02288797	II	14.53	nil	nil	N_Shore	22
cg09730640	II	14.53	nil	nil	S_Shore	31
cg02233183	II	16.63	NLGN4Y	TSS1500	N_Shore	9
cg27265812	II	16.63	NLGN4Y	Body	N_Shore	15
cg03706273	I	16.64	NLGN4Y	1stExon	Island	19
cg09300505	II	16.94	NLGN4Y	Body	Island	34
cg05939513	II	16.94	NLGN4Y	Body	S_Shore	29

Appendix 2 – hyper-methylated CpG sites

CpG site ID	Infinium type	Position (Mbp)	Gene name	Gene location	Island region	CpG count
cg04169747	II	2.66	SRY	5'UTR	N_Shelf	14
cg11898347	II	2.66	SRY	TSS200	N_Shore	9
cg27636129	II	2.66	SRY	TSS200	N_Shore	9
cg09595415	II	2.66	SRY	TSS200	N_Shore	8
cg01375382	II	2.71	RPS4Y1	TSS200	Ocean	21
cg25443613	II	2.71	RPS4Y1	TSS200	Ocean	21
cg01311227	II	2.71	RPS4Y1	1stExon	Ocean	23
cg14170959	II	2.80	ZFY	TSS1500	N_Shore	13
cg11131351	I	2.80	ZFY	TSS1500	N_Shore	11
cg14972466	II	2.80	ZFY	TSS1500	N_Shore	11
cg02002345	II	6.78	TBL1Y	TSS1500	N_Shore	21
cg27355713	II	6.78	TBL1Y	TSS200	N_Shore	40
cg04042030	I	6.78	TBL1Y	TSS200	Island	44
cg02839557	I	6.78	TBL1Y	TSS200	Island	44
cg01707559	I	6.78	TBL1Y	TSS200	Island	52
cg25032547	I	14.77	TTTY15	TSS1500	Ocean	13
cg27443332	II	16.63	NLGN4Y	Body	N_Shore	14
cg27214488	II	16.64	NLGN4Y	TSS200	N_Shore	9
cg25518695	I	16.64	NLGN4Y	Body	N_Shore	28
cg18113731	II	16.64	NLGN4Y	Body	Island	34
cg19244032	II	16.64	NLGN4Y	Body	Island	35
cg10990737	II	16.64	NLGN4Y	Body	S_Shore	30
cg03244189	I	21.24	TTTY14	Body	Island	36
cg13845521	II	21.24	TTTY14	Body	Island	52
cg10811597	II	21.24	TTTY14	Body	Island	43
cg11816202	II	21.24	TTTY14	TSS200	Island	34
cg00212031	I	21.24	TTTY14	TSS200	Island	32
cg15345074	I	21.24	TTTY14	TSS200	Island	38
cg11684211	II	21.24	TTTY14	TSS1500	Island	40
cg06628792	II	21.24	TTTY14	TSS1500	Island	27
cg00121626	II	21.66	BCORL2	Body	N_Shore	16
cg00876332	II	21.66	BCORL2	Body	Island	29
cg15794778	II	21.67	BCORL2	Body	Island	24
cg25756647	I	21.73	CYorf15A	TSS1500	N_Shore	14
cg01988452	II	22.74	EIF1AY	TSS1500	N_Shore	5
cg13308744	II	22.74	EIF1AY	TSS1500	N_Shore	5

Appendix 3 – hypo-methylated CpG sites

CpG site ID	Infinium type	Position	Gene name	Gene location	Island region	CpG count
cg15431336	II	4.87	PCDH11Y	5'UTR	S_Shore	2
cg13805219	II	6.14	nil	nil	S_Shelf	7
cg15849038	II	6.17	nil	nil	N_Shore	12
cg17972491	II	6.74	AMELY	TSS1500	Ocean	5
cg15700967	II	6.95	TBL1Y	Body	Ocean	16
cg14778208	II	7.57	TTTY16	Body	Ocean	1
cg05098815	II	9.45	RBMY3AP	Body	Ocean	1
cg08053115	II	9.61	TTTY1	Body	Ocean	3
cg18163559	II	9.74	nil	nil	N_Shelf	8
cg02107461	II	14.10	nil	nil	N_Shelf	2
cg07795413	II	16.73	NLGN4Y	5'UTR	Ocean	5
cg14151065	II	19.68	nil	nil	N_Shore	24
cg06065495	II	21.67	BCORL2	TSS1500	S_Shore	10

References

1. Ferlay J, Shin H-R, Bray F, et al: Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int. J. Cancer 127: 2893–2917, 2010

2. Cook MB, Dawsey SM, Freedman ND, et al: Sex disparities in cancer incidence by time period and age. Cancer Epidemiol Biomarkers Prev. 18(4): 1174-1182, 2009

3. Forsberg LA, Rasi C, Malmqvist N, et al: Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nature Genetics 46(6): 624-628, 2014

4. Dumanski JP, Rasi C, Lonn M, et al: Smoking is associated with mosaic loss of chromosome Y: Science 347: 81-83, 2015

5. Bellott DW, Hughes JF, Skaletsky H, et al: Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature 508(7497): 494-499, 2014

6. Davoli T, Xu AW, Mengwasser KE, et al: Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns to Shape the Cancer Genome”. Cell 155(4): 948-962, 2013

7. Jones PA, Baylin SN: The epigenomics of cancer. Cell 128(4): 683-692, 2007

8. Collins FS, Barker AD: Mapping the cancer genome. Sci Am, 00368733, 296(3), 2007

9. The Cancer Genome Atlas research Network: Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499: 43-49, 2013

10. The Cancer Genome Atlas research Network: Comprehensive molecular characterization of human colon and rectal cancer. Nature 487: 330-337, 2012

11. Toyota M, Ahuja N, Ohe-Toyota M, et al: CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci USA 96: 8681-8686, 1999

12. The Cancer Genome Atlas research Network: Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517: 576-582, 2015

13. Leemans CR, Braakhuis BJ, Brakenhoff RH: The molecular biology of head and neck cancer. Nat Rev Cancer 11, 9-22, 2011

14. Feber A, Guilhamon P, Lechner M, et al: Using high-density DNA methylation arrays to profile copy number alterations. Genome Biol 15:R30, 2014

15. Bachtrog D: Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet 14: 113-124, 2013

16. Jangravi Z, Alikhani M, Arefnezhad B, et al: A fresh look at the male-specific region of the human Y chromosome. J Proteome Res 12: 6-22, 2013

17. Lahn BT, Page DC: Four evolutionary strata on the human X chromosome. Science 286: 964-976, 1999

18. Hinch AG, Altemose N, Noor N, et al: Recombination in the human pseudoautosomal region PAR1. PLoS Genet 10(7): e1004503, 2014

19. Flaquer A, Fischer C, Wienker TF: A new sex-specific genetic map of the human pseudoautosomal regions (PAR1 and PAR2). Hum Hered 68: 192-200, 2009

20. Ross MT, Grafham DV, Coffey AJ, et al: The DNA sequence of the human X chromosome. Nature 434: 325-337, 2005

21. Veerappa AM, Padakannaya P, Ramachandra NB: Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. Funct Integr Genomics 13: 285-293, 2013

22. Dasari VK, Deng D, Perinchery G, et al: DNA methylation regulates the expression of Y chromosome specific genes in prostate cancer. J Urol 167: 335-338, 2002

23. Yao L, Ren S, hang M, et al: Identification of specific DNA methylation sites on the Y-chromosome as biomarker in prostate cancer. Oncotarget: 1-11, 2015

24. Duijf PHG, Schultz N, Benezra R: Cancer cells preferentially lose small chromosomes. Int J Cancer 132(10): 2316-2326, 2013

25. Klatte T, Rao PN, de Martino M, et al: Cytogenetic profile predicts prognosis of patients with clear cell renal cell carcinoma. J Clin Oncol 27(5): 746-753, 2009

26. Ali RH, Marafie MJ, Bitar MS, et al: Gender-associated genomic differences in colorectal cancer: clinical insight from feminization of male cancer cells. Int J Mol Sci 15: 17344-17365, 2014

27. Bergamo NA, Silva Veiga LC, Reis PP, et al: Classic and molecular cytogenetic analyses reveal chromosomal gains and losses correlated with survival in head and neck cancer patients. Clin Cancer Res 11: 621-631, 2005

28. Silva Veiga LC, Bergamo NA, Reis PP, et al: Loss of Y-chromosome does not correlate with age at onset of head and neck carcinoma: a case-control study. Braz J Med Biol Res 45(2): 172-178, 2012

29. Kido T, Lao Y-FC: Roles of the Y chromosome genes in human cancers. Asian J Androl 17: 373-380, 2015

30. Ludwig JA, Weinstein JN: Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer 5: 845-856, 2005

31. Poste G: Bring on the biomarkers. Nature 469:156-157, 2011

32. Bird A: DNA methylation patterns and epigenetic memory. Genes Dev 16: 6-21, 2002

33. Jones PA: Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13: 484-492, 2012

34. Bergman Y, Cedar H: DNA methylation dynamics in health and disease. Nat Struct Mol Biol 20(3): 274-281, 2013

35. Satterlee JS, Schubeler D, Ng H-H: Tackling the epigenome: challenges and opportunities for collaboration. Nat Biotechnol 28(10): 1039-1044, 2010

36. Moarefi AH, Chedin F: ICF syndrome mutations cause a broad spectrum of biochemical defects in DNMT3B-mediated *de novo* DNA methylation. J Mol Biol 409: 758-772, 2011

37. Hashimshony T, Zhang J, Keshet I, et al: The role of DNA methylation in setting up chromatin structure during development. Nat Genet 34: 187-192, 2003

38. Gal-Yam EN, Egger G, Iniguez L, et al: Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proc Natl Acad Sci USA 105: 12979-12984, 2008

39. Jones PA: The DNA methylation paradox. Trends Genet 15(1): 34-37, 1999

40. Laurent L, Wong E, Li G, et al: Dynamic changes in the human methylome during differentiation. Genome Res 20: 320-331, 2010

41. Baylin SB, Jones PA: A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Cancer 11: 726-734, 2011

42. Weisenberger DJ, Siegmund KD, Campan M, et al: CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet 38: 787-793, 2006

43. Noushmehr H, Weisenberger DJ, Diefes K, et al: Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 17: 510-522, 2010

44. Keshet I, Schlesinger Y, Farkash S, et al: Evidence for an instructive mechanism of *de novo* methylation in cancer cells. Nat Genet 38: 149-153, 2006

45. Cedar H, Bergman Y; Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet 10: 295-304, 2009

46. Widschwendter M, Fiegl H, Egle D, et al: Epigenetic stem cell signature in cancer. Nat Genet 39: 157-158, 2007

47. Vire E, Brenner C, Deplus R, et al: The Polycomb group protein EZH2 directly controls DNA methylation. Nature 439: 871-874, 2006

48. Williams K, Christensen J, Pederesen MT, et al: TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature 473: 343-348, 2011

49. Saito Y, Liang G, Egger G, et al: Specific activation of microRNA-127 with downregulation of the proto-oncogene BCL6 by chromatin-modifying drugs in human cancer cells. Cancer Cell 9(6): 435-443, 2006

50. Hansen KD, Timp W, Bravo HC, et al: Increased methylation variation in epigenetic domains across cancer types. Nat Genet 43: 768-775, 2011

51. Berman BP, Weisenberger DJ, Aman JF, et al: Regions of focal DNA hypermethylation and long-range hypo-methylation in colorectal cancer coincide with nuclear-lamina associated domains. Nat Genet 44: 40-48, 2012

52. Gaudet F, Hodgson JG, Eden A, et al: Induction of tumors in mice by genomic hypomethylation. Science 300: 489-492, 2003

53. Laird PW: Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet 11: 191-203, 2010

54. Gronbaek K, Muller-Tidlow C, Perini G, et al: A critical appraisal of tools available for monitoring epigenetic changes in clinical samples from patients with myeloid malignancies. Haematologica 97(9): 1380-1388, 2012

55. Frommer M, McDonald LE, Millar DS, et al: A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci USA 89(5): 1827-1832, 1992

56. Bibikova M, Le J, Barnes B, et al: Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics 1: 177-200, 2009

57. Bibikova M, Barnes B, Tsan C, et al: High Density DNA methylation array with single CpG site resolution. Genomics 98: 288-295, 2011

58. Takai D, Jones PA: Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci USA 99: 3740-3745, 2002

59. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, et al: The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 41: 178-185, 2009

60. Hansen KD, Aryee M: minfi: Analyze Illumina's 450k methylation arrays. R package version 1.0.0.

61. Davis S, Du P, Bilke S, et al: methylumi: Handle Illumina methylation data. R package version 2.0.13.

62. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN-900051-07-0, URL http://www.R-project.org/, 2011

63. Bock C: Analysing and interpreting DNA methylation data. Nat Rev Genet 13: 705-719, 2012

64. Irizarry RA, Hobbs B, Collin F, et al: Exploration, normalization and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2): 249-264, 2003

65. Ritchie ME, Silver J, Oshlack A, et al: A comparison of background correction methods for two-colour microarrays. Bioinformatics 23(20): 2700-2707, 2007

66. Triche Jr TJ, Weisenberger DJ, van den Berg D, et al: Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res 41(7): e90, 2013

67. Bolstad BM, Irizarry RA, Astrand m, et al: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2): 185-193, 2003

68. Yang YH, Dudoit S, Luu P, et al: Normalization for cDNA microarray data. In Microarrays: optical technologies and informatics 4266: 141-152, 2001

69. Dedeurwaerder S, Defrance M, Calonne E, et al: Evaluation of the Infinium Methylation 450K technology. Epigenomics 3(6): 771-784, 2011

70. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ et al (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium Methylation data. Bioinformatics 28, 729-730

71. Touleimat N, Tost J: Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 4(3): 325-341, 2012

72. Maksimovic J, Gordon L, Oshlack A: SWAN: Subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biol 13:R44, 2012

73. Teschendorff AE, Marabita F, Lechner M, et al: A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. Bioinformatics 29(2): 189-196, 2013

74. Valavanis I, Sifakis EG, Gergiadis P, et al: A composite framework for the statistical analysis of epidemiological DNA methylation data with the Infinium Human Methylation 450K BeadChip. IEEE J Biomed Health Inform 18(3): 817-823, 2014

75. Sun Z, Chai HS, Wu Y, et al: Batch effect correction for genome-wide methylation data with Illumina platform. BMC Med Genomics 4:84, 2011

76. http://media.affymetrix.com/support/technical/datasheets/genomewide_snp6_datasheet.pdf (2009)

77. The Cancer Genome Atlas research Network: Integrated genomic analyses of ovarian carcinoma. Nature 474: 609-615, 2011

78. Venkatraman ES, Olshen AB: A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23(6): 657-663, 2007

79. http://support.illumina.com/sequencing/sequencing_instruments/hiseq_2000.html

80. Wang K, Singh D, Zeng Z, et al: MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 38(18): e178, 2010

81. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323, 2011

82. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14): 1754-1760, 2009

83. Cibulskis K, Lawrence MS, Carter SL, et al: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31: 213-219, 2013

84. Brenner H, Kloor M, Pox CP: Colorectal cancer. Lancet 383(9927): 1490-1502, 2014

85. Coppede F: Epigenetic biomarkers of colorectal cancer: Focus on DNA methylation. Cancer Lett 342: 238-247, 2014

86. Hinoue T, Weisenberger DJ, Lange CPE, et al: Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res 22: 271-282, 2012

87. Argiris A, Karamouzis MV, Raben D, et al: Head and neck cancer. Lancet 371(9625): 1695-1709, 2008

88. Riaz N, Morris LG, Lee W, at al: Unravelling the molecular genetics of head and neck cancer through genome-wide approaches. Genes Dis 1: 75-86, 2014

89. Maasland DHE, van den Brandt PA, Kremer B, et al: Alcohol consumption, cigarette smoking and the risk of subtypes of head and neck cancer: results from the Netherlands Cohort Study. BMC Cancer 14: 187, 2014

90. Smeets SJ, Braakhuis BJM, Abbas S, et al: Genome-wide DNA copy number alterations in head and neck squamous cell carcinomas with or without oncogene-expressing human papillomavirus. Oncogene 25: 2558-2564, 2006

91. Ang KK, Harris J, Wheeler R, et al: Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med 363: 24-35, 2010

92. Ricketts CJ, Linehan WM: Intratumoral heterogeneity in kidney cancer. Nat Genet 46(3): 214-215, 2014

93. Bhatt JR, Finelli A: Landmarks in the diagnosis and treatment of renal cell carcinoma. Nat Rev Urol 11: 517-525, 2014

94. Brugarolas J: Molecular genetics of clear-cell renal cell carcinoma. J Clin Oncol 32(18): 1968-1976, 2014

95. Gerlinger M, Horswell S, Larkin J, et al: Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat Genet 46(3): 225-233, 2014

96. Maher ER: Genomics and epigenomics of renal cell carcinoma. Semin Cancer Biol 23: 10-17, 2013

97. Robinson MD, Oshlack A: A scaling normalization method for differential expression of RNA-seq data. Genome Biol 11:R25, 2010

98. Mermel CH, Schumacher SE, Hill B, et al: GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 12: R41, 2011

99. Thun MJ, Carter BD, Feskanich D, et al: 50-year trends in smoking-related mortality in the United States. N Engl J Med 368(4): 351-364, 2012

100. Price EM, Cotton AM, Lam LL, et al: Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin 6:4, 2013

101. Wong AK, Fang B, Zhang L, et al: Loss of the Y chromosome: an age-related or clonal phenomenon in acute myelogenous leukemia / myelodysplastic syndrome? Arch Pathol Lab Med 132: 1329-1332, 2008

102. Vijayakumar S, Garica D, Hensel CH, et al: The human Y chromosome suppresses the tumorigenicity of PC-3, a human prostate cancer cell-line, in athymic nude mice. Genes Chromosomes Cancer 44: 365-372, 2005

103. Mayer A, Lahr G, Swaab DF, et al: The Y-chromosomal genes SRY and ZFY are transcribed in adult human brain. Neurogenetics 1: 281-288, 1998

104. Syrjanen KJ, Pyrhonen S, Syrjanen SM, et al: Immunohistochemical demonstration of human papillomavirus (HPV) antigens in oral squamous cell lesions. Br. J. Oral Surg. 21: 147-153, 1983

105. Bhatia A, Burtness B: Human papillomavirus-associated oropharyngeal cancer: defining risk groups and clinical trials. J. Clin. Oncol. 33(29): 3243-3251, 2015

106. Kang H, Kiess A, Chung CH: Emerging biomarkers in head and neck cancer in the era of genomics. Nat. Rev. Clin. Oncol. 12: 11-26, 2015

107. Spence T, Bruce J, Yip KW, et al: HPV associated head and neck cancer. Cancers 8,75, 2016

108. Zhou G, Liu Z, Myers JN: TP53 mutations in head and neck squamous cell carcinoma and their impact on disease progression and treatment response. J. Cell. Biochem. 9999: 1-11, 2016

109. Kandoth C, McLellan MD, Vandin F, et al: Mutational landscape and significance across 12 major cancer types. Nature 502: 333-339, 2013

110. Stracquadanio G, Wang X, Wallace MD, et al: The importance of p53 pathway genetics in inherited and somatic cancer genomes. Nat. Rev. Cancer 16: 251-265, 2016

111. Guo G, Califano JA: Molecular biology and immunology of head and neck cancer. Surg. Oncol. Clin. N. Am. 24(3): 397-407, 2015

112. Gross AM, Orosco RK, Shen JP, et al: Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss. Nat. Rev. Genet. 46(9): 939-943, 2014

113. Ciriello G, Miller ML, Aksoy BA, et al: Emerging landscape of oncogenic signatures across human cancers. Nat. Genet.45(10): 1127-1133, 2013

114. Komura K, Jeong SH, Hinohara K, et al: Resistance to docetaxel in prostate cancer is associated with androgen receptor activation and loss of KDM5D expression. Proc. Natl. Acad. Sci. USA 113(22): 6259-6264, 2016

115. Ahn J, Kim KH, Park S, et al: Target sequencing and CRISPR/Cas editing reveal simultaneous loss of UTX and UTY in urothelial bladder cancer. Oncotarget 7(39): 63252-63260, 2016

116. Walport LJ, Hopkinson RJ, Vollmar M, et al: Human UTY (KDM6C) is a male-specific N$^{\varepsilon}$-methyl lysyl demethylase. J. Biol. Chem. 289(26): 18302-18313, 2014

117. Gannon OM, de Long LM, Endo-Munoz L, et al: Dysregulation of the repressive H3K27 trimethylation mark in head and neck squamous cell carcinoma contributes to dysregulated squamous differentiation. Clin. Cancer Res. 19(2): 428-441, 2012

118. Ferris RL, Blumenschein Jr G, Fayette J, et al: Nivolumab for recurrent squamous-cell carcinoma of the head and neck. N. Engl. J. Med. 375: 1856-67, 2016

119. Case LK, Wall EH, Dragon JA, et al: The Y chromosome as a regulatory element shaping immune cell transcriptomes and susceptibility to autoimmune disease. Cancer Res. 23: 1474-1485, 2013