EPIGENOMICS OF SARCOMAS

Paul Guilhamon

UCL Cancer Institute

University College London

This thesis is submitted for the degree of Doctor of Philosophy

DECLARATION

I, Paul Guilhamon, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Paul Guilhamon

20 November 2014

PUBLICATIONS

The following publications have resulted from the work presented in this thesis and that conducted through collaboration on other projects. The corresponding abstracts and title pages can be found in the *Appendices*.

1. **P Guilhamon,** M Eskandarpour, D Halai, GA Wilson, A Feber, AE Teschendorff, V Gomez, A Hergovich, R Tirabosco, MF Amary, D Baumhoer, G Jundt, MT Ross, AM Flanagan, S Beck, Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2, *Nature Communications,* 4, 2166, (2013)

2. **P Guilhamon**, LM Butcher, N Presneau, GA Wilson, A Feber, DS Paul, MT Ross, AM Flanagan, S Beck, Assessment of tumour xenografts as a discovery tool for cancer epigenomics (Under Review)

3. A Feber, **P Guilhamon**, M Lechner, T Fenton, GA Wilson, C Thirlwell, TJ Morris, AM Flanagan, AE Teschendorff, JD Kelly, S Beck, Using high-density DNA methylation arrays to profile copy number variations, *Genome Biology* (2014)

4. DS Paul, **P Guilhamon**, A Karpathakis, C Thirlwell, A Feber and S Beck, Assessment of RainDrop BS-seq as a method for large-scale, targeted bisulfite sequencing, *Epigenetics* (2014)

5. PJ Smith, AP Levine, J Dunne, **P Guilhamon**, M Turmaine, G Sewell, N O'Shea, R Vega, JC Paterson, D Oukrif, S Beck, SL Bloom, M Novelli, M Rodriguez-Justo, AM Smith, AW Segal, Mucosal transcriptomics implicates under expression of *BRINP3* in the pathogenesis of ulcerative colitis, *Inflammatory Bowel Diseases* (Accepted)

6. SK Stewart, TJ Morris, **P Guilhamon**, H Bulstrode, M Bachman, S Balasubramanian, S Beck, oxBS-450K: A method for analysing hydroxymethylation using 450K BeadChips, *Methods* (Accepted)

7. J Charlton, RD Williams, NJ Sebire, S Popov, G Vujanic, T Chagtai, M Maschietto, M Alcaide-German, T Morris, LM Butcher, **P Guilhamon**, S Beck, K Pritchard-Jones, Comparative methylome analysis identifies new tumour subtypes and biomarkers for transformation of nephrogenic rests into Wilms tumour (In Revision)

ABSTRACT

Isocitrate dehydrogenase (IDH) genes 1 and 2 are frequently mutated in acute myeloid leukemia (AML), lower-grade glioma (LGG), and cholangiocarcinoma (CC). In these three malignancies, mutant *IDH* status is associated with increased 2-hydroxyglutarate (2-HG) production and a DNA hypermethylation phenotype, implicating altered epigenome dynamics in the aetiology of these cancers. Here I show that the *IDH* variants in chondrosarcoma (CS) are also associated with a hypermethylation phenotype, supporting the role of mutant IDH-produced 2-HG as an inhibitor of TET-mediated DNA demethylation. The associated gene expression profile is also investigated, highlighting the need for a better understanding of DNA methylation-mediated transcriptional regulation. The generated methylation data is additionally harnessed to reveal novel copy number variants in CS.

Meta-analysis of the AML, LGG, CC and CS methylation data identifies cancer-specific effectors within the retinoic acid receptor activation pathway among the hypermethylated targets. By analysing sequence motifs surrounding hypermethylated sites across the four cancer types, and using chromatin immunoprecipitation and western blotting, I identify the transcription factor EBF1 as an interaction partner for TET2, in the first description of a targeted demethylation pathway.

In an effort to assess whether patient-derived tumour xenografts (PDXs) are suitable models for epigenetic research in rare and common cancers, such as osteosarcoma (OS) and colon cancer, respectively, I compare PDXs to their matched patient tumour and reveal that an average of only 2.7% of the assayed methylome undergoes major methylation changes with xenografting. In addition, no further changes are identified in subsequent PDX generations, making these models highly suitable for expansion of rare tumours and preclinical drug screening. Finally I propose a model to inform future study design and statistically dilute those methylation shifts identified in PDXs.

ACKNOWLEDGEMENTS

For his steadfast support and guidance, I am immensely grateful to my supervisor
Prof. Stephan Beck, who, throughout these four years, was a fantastic advisor:
Always calm, always patient and seemingly never tiring,
His constant excitement for science is inspiring.
I would also like to thank Prof. Adrienne Flanagan, my second faculty
Advisor, whose counsel and samples were both of great quality.
Thanks also to Dr. Mark Ross, for welcoming me at Illumina,
And guiding me through sequencing with impressive stamina,
And I am grateful to Dr. Peter Campbell on my thesis
Committee for his advice on results and analysis.
Many thanks to the Medical Genomics team, in random order:
Lee 'Eli' Butcher, Christopher Bell, and Andrew 'Fever' Feber
Harpreet 'HD' Dibra and Dirk Paul, the 'Practical Gangster',
Tiffany 'Tiff' Morris, Gareth Wilson, and Manu Libertini,
Chrissie Thirlwell, Helena Caren, and Pawan Dhami,
Sabrina 'DJ Rina' Stewart and Tosin Taiwo, the 'weekender',
Charles Breeze, the Anna Ks, and Matthias Lechner.
Thanks to the Sarcoma group for their great support and bits of tumour:
Nadege Presneau, Dina Halai, and Malihe Eskandarpour
And to my collaborators Valenti Gomez and Alexander Hergovich,
For the gels and membranes and teaching me which is which.
For funding, housing and feeding me,
My thanks to Illumina and the MRC.
To all those friends I could always count on,
From Shropshire, Wales, and Puddington,
From Dublin and Manchester,
And also from Dorchester
From Bristol, Oxford, London,
Hull, Basingstoke and Trenton,
Merci encore à tous ces amis de Paris et du L.I.,
De Vendée et NYC, et du Brésil aussi.

A Papa, Maman, Charles, Laure, Alice et Louis,
Pour votre présence et votre soutien, MERCI !

CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations and Acronyms

2-HG	2-hydroxyglutarate
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
AML	Acute Myeloid Leukaemia
BS-seq	Bisulfite sequencing
CC	Cholangiocarcinoma
CpG	Cytosine-guanine dinucleotide
CRISPR	Clustered regularly interspaced short palindromic repeats
CS	Chondrosarcoma
DBD	DNA binding domain
DMR	Differentially methylated region
DNMT	DNA methyltransferase
DREME	Discriminative regular expression motif elicitation
DSBH	Double stranded beta helix
ES	Ewing's Sarcoma
FBS	Foetal bovine serum
FH	Fumarate hydratase
FM	Formaldehyde
HELP	HpaII tiny fragment enrichment by ligation-mediated PCR
HLH	Helix loop helix
IN	Input
IP	Immunoprecipitation
IPA	Ingenuity Pathway Analysis
LGG	Lower grade glioma
MAD	Median absolute deviation
MDS	Multidimensional scaling
MeDIP	Methylated DNA immunoprecipitation
MEME	Multiple expectation maximisation for motif elicitation
MUT	IDH mutant
MVP	Methylation variable position
OS	Osteosarcoma
PBS	Phosphate buffered saline

PCR	Polymerase chain reaction
PDB	Protein Data Bank
PI	Protease inhibitor
RRBS	Reduced representation bilsufite sequencing
RT	Room temperature
SDH	Succinate dehydrogenase
SVD	Singular value decomposition
TBE	Tris/Borate/EDTA
TDG	Thymine-DNA glycosylase
TSS	Transcription start site
UTR	Untranslated region
WB	Western blot
WGBS	Whole-genome bisulfite sequencing
WT	IDH wild-type
α-KG	α-Ketoglutarate

LIST OF APPENDICES

1 INTRODUCTION

1.1 Epigenetics

Epigenetics can be defined as "the study of heritable changes in gene expression that are not due to changes in DNA sequence"[2]. These changes in expression are governed by various processes including histone modifications (e.g. acetylation), non-coding RNAs (e.g. miRNAs), and DNA methylation. The work presented here focuses on the latter and its role in bone sarcomas.

1.1.1 Structure of the methylome

DNA methylation is characterised by the addition of a methyl group to the 5-position on the cytosine ring to form 5-methylcytosine (5mC) and is most commonly studied in the context of cytosine-guanine dinucleotides (CpG), although non-CpG methylation is receiving increasing attention, especially with regards to stem cell differentiation[3]. There are ~28 million CpGs, mostly methylated, in the haploid human genome.

Interspersed throughout the genome are regions termed CpG islands (CGIs). These are bioinformatically defined as sequences over 200 bp in length with an elevated GC content (>50%) and high CpG density (>0.6), and tend to be unmethylated[4,5]. The haploid human genome contains over 25,000 CGIs, half of which are associated with gene promoters while the rest are evenly distributed between intragenic and intergenic regions[6]. The regions 2 kb upstream and downstream of a CGI are called CpG shores and are the regions with the highest methylation variability in cancer[7], while those extending from the shores are termed CpG shelves[8]. These are the definitions used for the analysis described in this thesis, but it should be noted that regions with GC content and CpG density that are high but do not reach the thresholds described above could well have similar functional roles to those described in the next section. Finally, recent studies have shown for the first time with artificial CpG islands that both GC content and CpG density are necessary functional components of these regions[9]: an island with high CpG density and GC content led to nearby histones carrying both activating and repressing marks, characteristic of a poised chromatin state. When the GC content was maintained but the CpG sites removed, no histone marks were found. Interestingly, when the CpG frequency was left unchanged but the GC content was reduced, no histone marks were found either; however, this GC-poor island was largely methylated, indicating that the GC content was necessary to protect the island from methylation and thus allow the recruitment of factors involved in histone modification. Thus, although these investigations are still at an early stage, it appears that the particular composition of CpG islands is precisely linked with their function, and further research might soon lead to the enumeration of more specific parameters to define these epigenomic domains.

1.1.2 Function of the methylome

In addition to those CGIs in promoter regions that are evidently in close proximity to transcription start sites, many in inter- or intragenic locations are also associated with transcription initiation. This is exemplified by the *Air* transcript, initiated within a CGI in the second intron of *Igf2r*, and necessary to the silencing of the paternal allele[6].

This role of CGIs as transcription initiation regions has been explored from various angles. Firstly, little sequence conservation has been observed among CGIs aside from the constraints in nucleotide content mentioned above and a recurring lack in promoter elements such as the TATA box, but this is possibly compensated for by the GC-richness of mammalian transcription factor binding sites[10]. Secondly, CGIs are nucleosome-deficient and genes with CGIs in their promoters seem not to require chromatin remodelling complexes such as SWI/SNF[11], potentially making them more amenable to transcriptional regulation. Thirdly, CGI chromatin is enriched for histone marks associated with active transcription (e.g. H3K4me3)[12].

While most of the human genome is methylated, CGIs tend to be unmethylated, and how a change in that methylation state affects the ability of the associated promoter to initiate transcription has been the focus of numerous studies, reaching as far back as the early 1980s. The widely accepted theory is that methylation of a promoter island is associated with the stable downregulation of transcription[6]. The mechanism for this process is thought to be either through methylated loci preventing the binding of transcription factors or the recruitment of methyl-binding proteins that lead to a change in chromatin state. In either case, methylation of the promoter CGI is often regarded as necessary but insufficient to ensure silencing: necessary in X inactivation, for example, as when DNA methylation is inhibited, genes are reactivated in a fraction of the cells[13]; but also insufficient as it sometimes only occurs as a locking mechanism after silencing histone marks, such as H3K27me3, are in place.

Methylation of inter- and intragenic CGIs is much more common with up to 35% of intragenic islands being prone to methylation[4]. Although their function in this situation is less clear, it is thought that gene body methylation could serve to silence the transcription of non-coding RNAs from these intragenic transcription start sites (TSS) that would have silenced the expression of the associated gene; this is supported by gene body methylation often being associated with active transcription.

1.1.3 DNA demethylation

DNA methylation is established and maintained by DNA methyltransferases (DNMTs): DNMT3A and DNMT3B, modulated by DNMT3L, catalyse *de novo* methylation; DNMT1 maintains methylation levels when it is directed by a specialised protein, ubiquitin-like plant homeodomain and RING finger domain 1 (UHRF1), to hemimethylated sites in newly synthesised DNA strands during replication[14].

This mode of methylation maintenance forms the basis for the process of passive demethylation: successive replication cycles without the action of DNMT1 prevent the symmetrical conservation of methylation and eventually lead to loss of methylation. However, the observation of global and rapid demethylation events, such as in developing primordial germ cells (PGCs), could not be explained by a passive loss of 5mC[15], triggering a search for active demethylation mechanisms. One such mechanism, supported by multiple studies, is based on the oxidation of 5mC by a family of dioxygenases, the ten-eleven translocation (TET) proteins.

The TET enzymes are responsible for the iterative conversion of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) in an α-ketoglutarate-dependent manner[16]. These have been shown to be intermediates in the cytosine demethylation pathway[17](**Figure 1.1**). There are 3 proteins in the human TET family (TET1, 2 and 3) that are expressed in a tissue- and developmental stage-dependent manner. For example, TET1 and TET2 are found to be crucial in the maintenance of pluripotency in embryonic stem cells[18]. The crystal structure of TET2 has recently been solved[19], and is described in further detail in *Chapter 4*.

Multiple pathways have been suggested for the second stage in the active demethylation pathway, to restore an unmethylated cytosine in the position occupied by one of the oxidised derivatives of 5mC, including direct conversion through dehydroxymethylation of 5hmC or decarboxylation of 5caC[20]. However, the mechanism with most experimental supporting evidence to date involves excision of 5fC or 5caC by thymine-DNA glycosylase (TDG) followed by DNA polymerase-mediated replacement of the nucleotide with dCTP and ligation of the nicked DNA[21].

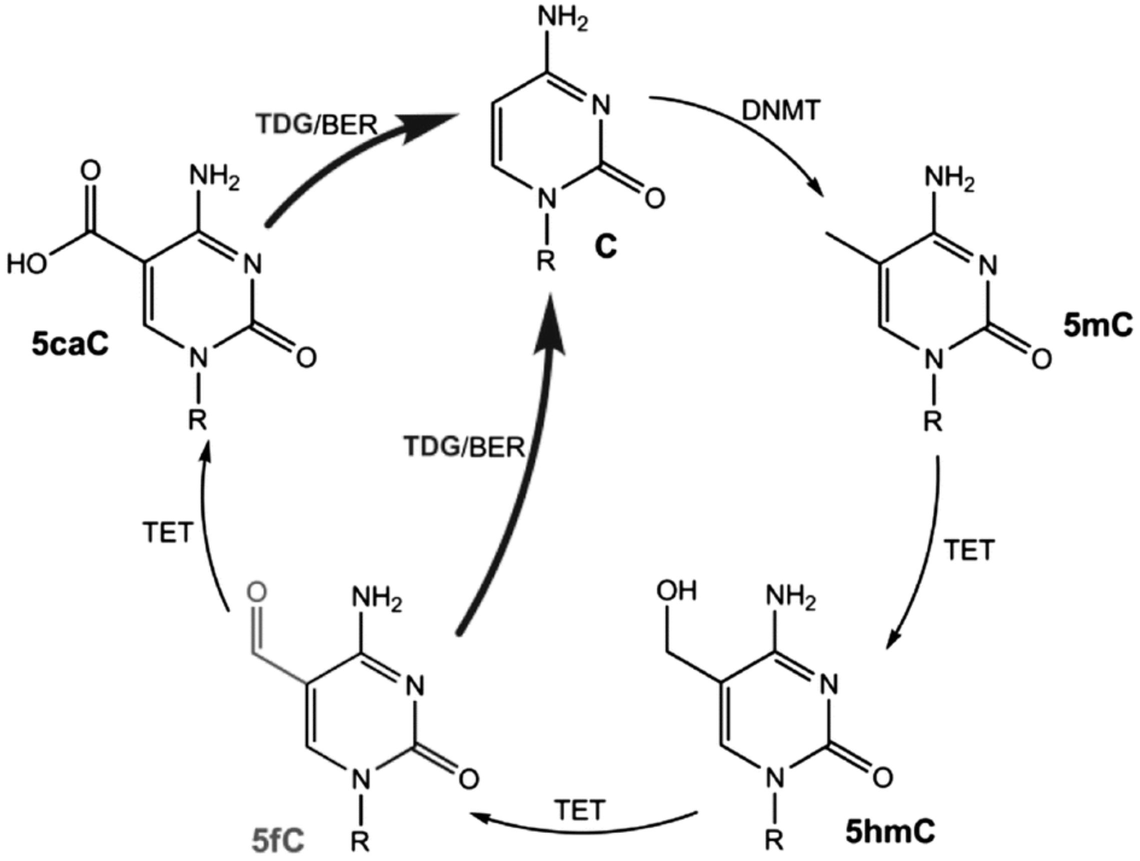


Figure 1.1: Active demethylation pathway

Cytosine is methylated by DNMT and demethylated through oxidation by TET followed by excision and repair by TDG and the BER pathway, respectively. This figure was adapted from Xu et al.[22]

1.1.4 IDH mutations and DNA hypermethylation in cancer

The cancer-state epigenetic phenotype is often characterised by the silencing of tumour suppressor genes through hypermethylation of their promoters, and/or overexpression of oncogenes due to hypomethylation[23]. Over the past few years, a particular hypermethylation phenotype has been detected in certain cancer types, associated with mutations in the isocitrate dehydrogenase family of enzymes.

Cytosolic isocitrate dehydrogenase 1 (IDH1) and its mitochondrial counterparts IDH2 and IDH3, are involved in a major pathway of cellular metabolism: the Krebs cycle (or TCA cycle). Their normal enzymatic activity consists in the irreversible conversion of their substrate isocitrate into α-ketoglutarate (α-KG), producing CO_2 and NADH or NADPH as side products. These enzymes have been found mutated at a high frequency in certain cancers, such as acute myeloid leukaemia (AML)[24] and glioma[25], with mutations in IDH1 Arg132 and IDH2 Arg140 and Arg172 accounting for >90% of reported cases[26]. Biochemical studies showed that these mutations reduce the enzyme's ability to bind isocitrate and they were thus initially considered to act through a dominant-negative loss-of-function mechanism. This theory was supported by the majority of affected tumours being heterozygous for these aberrations.

However, subsequent experiments[27] showed that mutant IDH enzymes had gained a different function and were in fact converting α-KG into R-2-hydroxyglutarate (2-HG)(**Figure 1.2**). Further studies demonstrated that 2-HG competitively inhibits α-KG-dependent dioxygenases, including histone demethylases and the TET dioxygenases[28]: 2-HG binds to the proteins where α-KG would normally reside. In glioma, for example, *IDH* mutations are associated with an increase in histone methylation and a decrease in genome-wide 5hmC levels[28]. Moreover, introduction of a mutant *IDH1* in primary human astrocytes was sufficient to reproduce this epigenetic phenotype[29].

To date, gain-of-function mutations in the IDH proteins and an associated DNA hypermethylation phenotype have been observed in low-grade glioma[30] (LGG), AML[30], cholangiocarcinoma[31] (CC), spindle cell hemangiomas[32], at low frequency in other malignancies and in chondrosarcoma (CS), as described in further detail in *Chapter 3*.

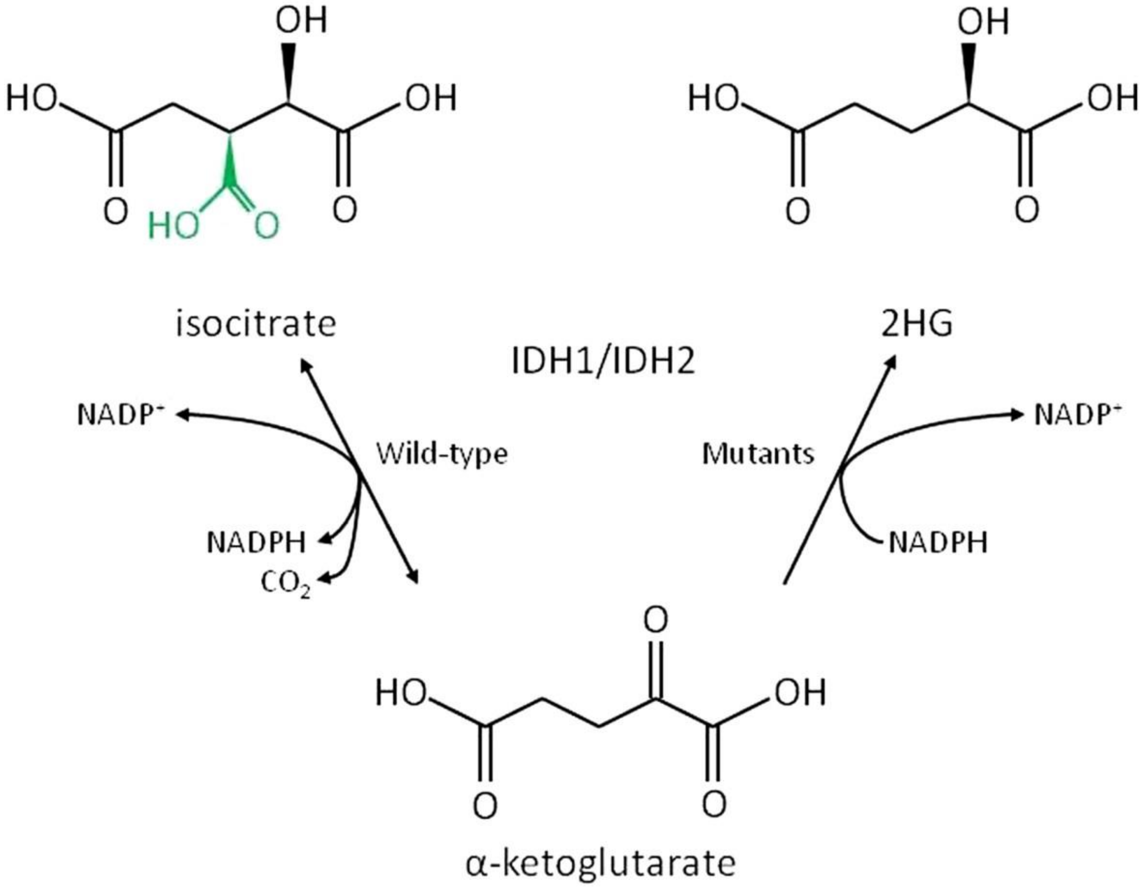


Figure 1.2: Function of wild-type and mutant IDH enzymes

Wild-type IDH proteins function as part of the TCA cycle to catalyse the interconversion of isocitrate and α-KG. When mutated at specific residues, the enzymes convert α-KG to 2-HG in a NADPH-dependent manner. Original figure from Reitman et al.[33]

1.2 Epigenetics toolkit

The technological advancements of recent years have noticeably improved our ability to study the epigenome while taking into account the aims as well as the limitations of each individual project, such as cost, sample number and sample quality. While pyrosequencing has been widely used for the validation of methylation at base-level resolution for over a decade, new technologies involving both microarrays and next-generation sequencing have made high-throughput methylome profiling a near-routine operation. Four of these methods were at the core of the results presented here, including the Illumina Infinium HumanMethylation450 BeadChips (450K), pyrosequencing, methylated DNA Immunoprecipitation followed by next-generation sequencing (MeDIP-seq), and the novel RainDrop-BSseq.

1.2.1 450K array

1.2.1.1 Probe design

The 450K array[8] is an expansion of the earlier Illumina 27K, which uses so-called Type I chemistry: each locus is targeted by two probes, one for the methylated and one for the unmethylated version of the CpG site. These differ at their 3' end, where the methylated version of the probe hybridises to the cytosine (C) (protected from bisulfite conversion by its methyl group), while the unmethylated probe matches the thymine (T) produced as a result of the conversion of the unmethylated C. For the 450K array, an additional probe type was introduced, with Type II chemistry. This assay uses a single probe per locus, and its 3' ends complements the base directly upstream of the target C; a single-base extension then adds the complementary guanine (G) or adenine (A), depending on the methylation state of the original C (**Figure 1.3**).

1.2.1.2 450K content

The array targets 485,577 loci, including 65 SNPs and over 3,000 non-CpG sites. These cover a range of genomic and epigenomic regions. For instance, 99% of RefSeq genes, each covered by an average of 17.2 probes, are targeted across the entire gene region: transcription start sites (TSS), gene body, 5' and 3' untranslated regions (UTR), and 1st Exon (**Figure 1.4**). In addition, 86-96% of CpG islands, shelves and shores, over 80,000 predicted enhancers and a number of other intergenic regions are also probed.

1.2.1.3 Analysis

A number of analysis pipelines have been developed to extract biological information from the raw data produced by the 450K array; they vary in their handling of the different probe types, batch effects, normalisation procedures and the various statistical tests and thresholds used[34].

The 450K array was used in this study to profile methylation variation associated with *IDH* mutations in chondrosarcoma and xenografting of osteosarcoma tumours. The analyses of 450K array data used here are covered in detail in the relevant chapters (*Chapter 2*, *Chapter 3* and *Chapter 5*).

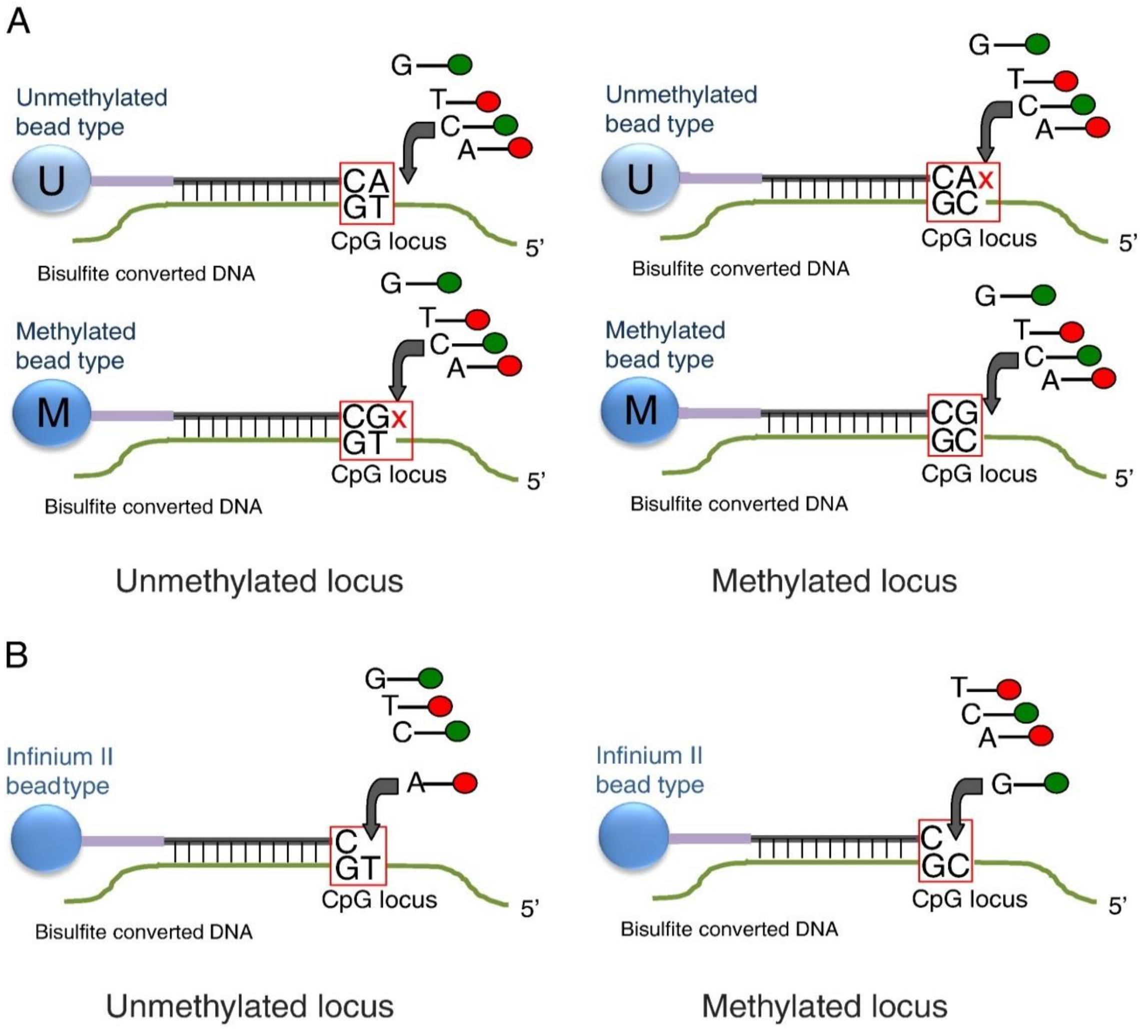


Figure 1.3: Probe types on the 450K array

The 450K array contains both Type I (A) and Type II (B) probes. Type I chemistry uses two probes per locus, one for unmethylated and one for methylated, while the Type II requires only one probe per target CpG, with a single-base extension to determine the methylation state. Original figure from Bibikova et al.[8]

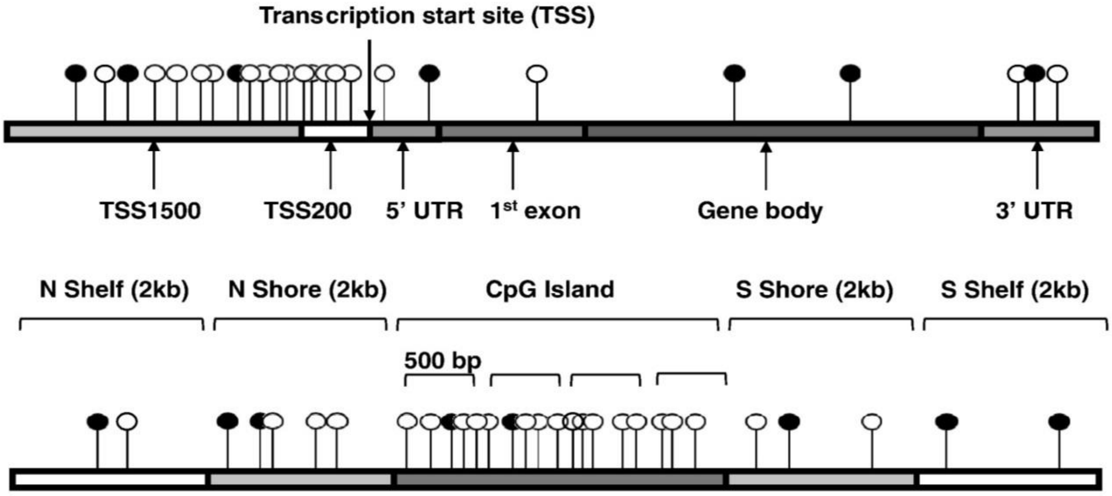


Figure 1.4: Genomic and epigenomic regions on the 450K array

A range of (epi)genomic regions are probed on the array, including TSSs, UTRs, gene bodies, and CpG islands, shelves and shores. Original figure from Bibikova et al.[8]

1.2.2 Pyrosequencing

Pyrosequencing is a sequencing-by-synthesis method that allows real-time determination of a template sequence. It has been used in a wide range of genetic applications from routine genotyping[35], to massively parallel sequencing on the 454 Life Sciences platform[36]. Pyrosequencing has also become an integral part of the epigenetic toolkit in order to validate either new methods[1,37] or novel differentially methylated loci associated with a particular phenotype[38], and it is even used for diagnostic purposes to detect abnormal methylation characteristic of specific diseases[39].

When used for methylation analysis, pyrosequencing is performed with prior bisulfite conversion of the DNA. The method relies on an enzymatic cascade that results in an emission of light correlated with the proportion of template DNA with a particular base at each position (**Figure 1.5**). Nucleotides are added sequentially to the reaction vessel, and, if they are incorporated into the extending strand, a pyrophosphate (PP_i) is released and used by an ATP sulfurylase to generate ATP, which is in turn used as substrate by the luciferase enzyme to emit light[40].

In the present study, pyrosequencing was used to validate regions of differential methylation between *IDH* mutant and wild-type samples, and the specific loci that were targeted, as well as the primer design and reagents used, are detailed in *Chapter 3*.

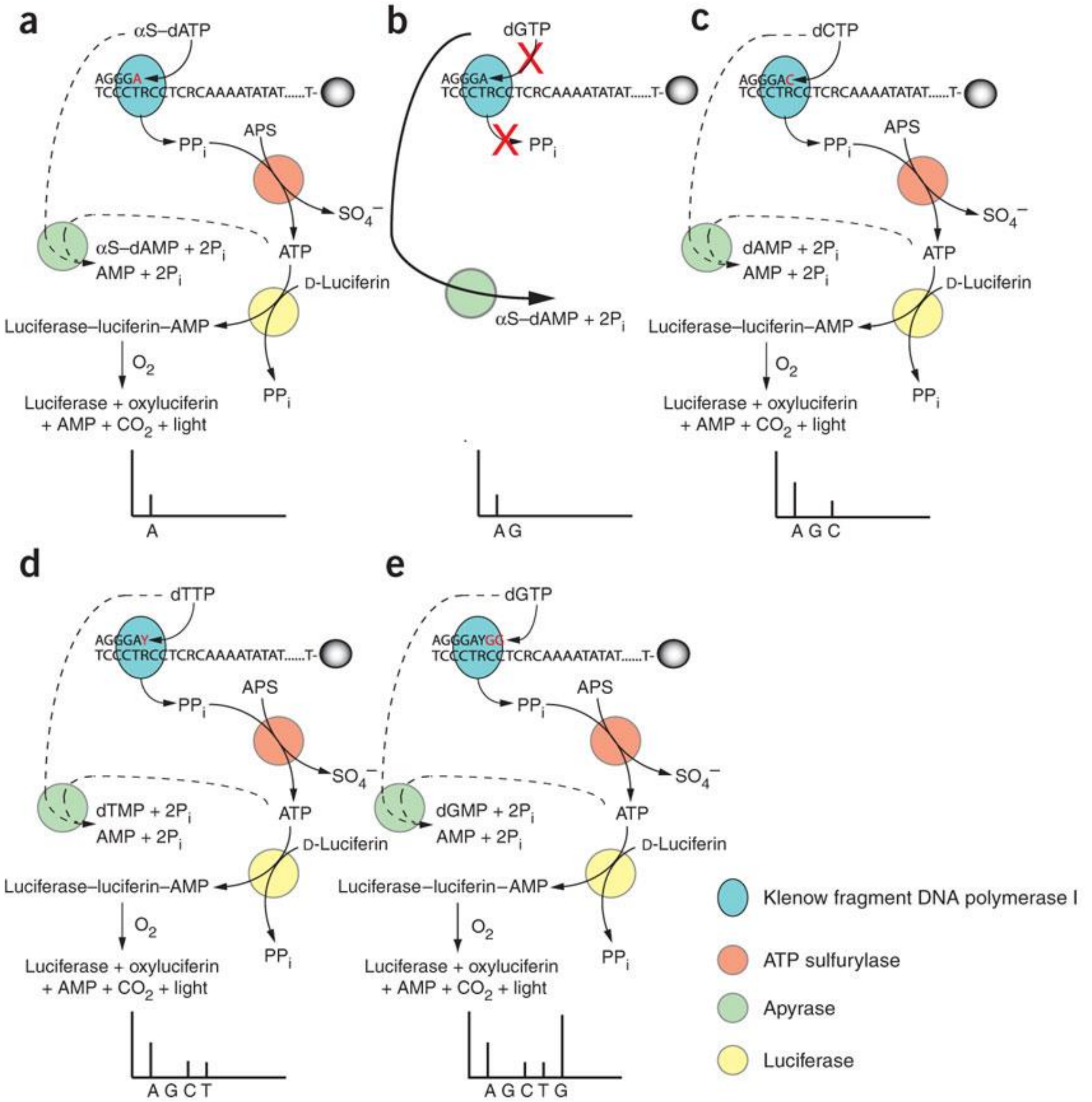


Figure 1.5: Pyrosequencing reaction

a) 2'- Deoxyadenosine- 5'- O- (1- thiotriphosphate) (αS-dATP) is used instead of dATP as the latter would be directly used a substrate for the luciferase reaction. Incorporation of the nucleotide releases PP$_i$, which is in turn used to make ATP; a luciferase-luciferin-AMP complex is formed by using the ATP, which, in the presence of oxygen, leads to the release of light proportionally to the amount of available PP$_i$ in the reaction vessel.

b) A nucleotide that cannot be incorporated into the extending strand is degraded by an apyrase

c,d) The ‘R’ in the template strand represents a methylation variable position: its complementary base on the extending strand is part of a CpG. The next nucleotide to be incorporated could then be either a C (if the locus was originally methylated) or a T (if the locus was originally unmethylated). Both C and T are incorporated in approximately equal amounts, each generating a light signal that is half of that measured for the first A, present in every extending strand in the reaction vessel.

e) Two consecutive G bases are incorporated, generating a light signal twice that of the first A.

Original figure from Tost et al.[40]

1.2.3 MeDIP-seq

MeDIP-seq combines immunoprecipitation of methylated DNA with next-generation sequencing. MeDIP uses an antibody against 5mC to capture the methylated fraction of the genome (**Figure 1.6**), with a resolution ranging from 100-300 bp, depending on the chosen insert size. It can be achieved reliably with as little as 50 ng of starting material[41], and can be automated with minimal hands-on time[42]. These advantages set MeDIP-seq apart from the other enrichment-based technologies, MBD-seq and methylCap-seq, for which no protocol exists with such low amounts of input DNA. MeDIP-seq is also achievable at considerably lower cost than whole-genome bisulfite sequencing (WGBS)[43], and permits a much more extensive coverage of the genome than the more affordable reduced representation bisulfite sequencing (RRBS)[43].

With its genome-wide coverage and, thus, its ability to capture regions of the genome not historically targeted by arrays, MeDIP-seq has been widely used in methylation profiling: from the first whole-genome methylation profile of a mammalian genome[44], to the analysis of a variety of tissues, including peripheral blood cells[45] and nerve sheath tumours[46].

However, this assay also displays noteworthy limitations. First among those is the resolution afforded by a MeDIP-seq experiment: it is limited to 100-300 bp and thus is far from the base–level resolution achievable with other sequencing and array technologies; although this is suitable to explore methylation of regions such as islands and shores, it is less applicable to deciphering the epigenetic regulation of small enhancer loci, or subtle differences in the methylation levels of adjacent regions. Another drawback of MeDIP-seq is its bias towards sequences of high methylation density; regions with densities lower than 1.5% can be missed entirely and erroneously considered as unmethylated[41]. In addition, as the procedure only enriches for the methylated portion of the genome, unmethylated regions can only be deduced through an elimination process, making the reliability of the method intrinsically dependent on sequencing depth. Although this last issue can be circumvented by combining MeDIP-seq with methylation-sensitive restriction enzyme sequencing (MRE-seq), albeit at a higher cost.

MeDIP-seq was used in this study to validate results obtained with the 450K array in the comparison of patient tumours with their corresponding xenografts. The specific protocol and analysis methods used are detailed in the relevant chapters (*Chapter 2* and *Chapter 5*). The next-generation sequencing aspect of this experiment was conducted at the Illumina facilities in Cambridge, UK, as the industrial placement involved in the MRC-CASE studentship that supported this work. This not only ensured that the experiment was performed quite literally according to the manufacturer's instructions, but also provided unparalleled access and guidance with regards to operating the company's systems, including the HiSeq 2000 and MiSeq platforms.

1.2.4 RainDrop-BSseq

High-throughput targeted resequencing through the combination of microdroplet polymerase chain reaction (PCR) and next-generation sequencing was developed by RainDance Technologies and first adapted to methylation analysis by Komori et al.[47] and Herrmann et al.[48] This method, termed Raindrop-BSseq, was further refined for the purposes of this project[1] and in a subsequent study to reduce the required amount of input DNA by Paul et al[49].

Other technologies available for methylation profiling can be categorised as follows: 1) WGBS that provides single base resolution and whole-genome coverage, but is still prohibitively expensive for large sample cohorts and small research groups; 2) pyrosequencing, which also achieves single-base resolution but does not scale up well to cover extended portions of the genome and is time-consuming to optimise; 3) enrichment-based methods such as MeDIP-seq that can provide genome-wide information for small sample cohorts but at relatively low resolution; and 4) array-based methods like the 450K array, which is applicable to large groups of samples, resolves individual CpG sites, and provides genome-wide, but not whole-genome, coverage and is intrinsically biased towards particular regions. Thus, RainDrop-BSseq bridges the gap between existing methods to allow targeted, base-resolution, and high-throughput methylation profiling of large sample cohorts.

The detailed protocol and analysis performed here are presented in the relevant chapters (*Chapter 2* and *Chapter 3*). Briefly, DNA samples are bisulfite-converted and a primer library is designed to amplify selected regions; using RainDance Technologies instruments, each primer pair is coupled to a template fragment within a single droplet on a microfluidic chip, effectively transforming each droplet into a micro-PCR tube and allowing up to 4,000 amplification reactions to take place in parallel. The amplified products are then pooled into a single library for each sample and a second round of amplification allows the incorporation of indices for next-generation sequencing on the Illumina MiSeq (**Figure 1.7**).

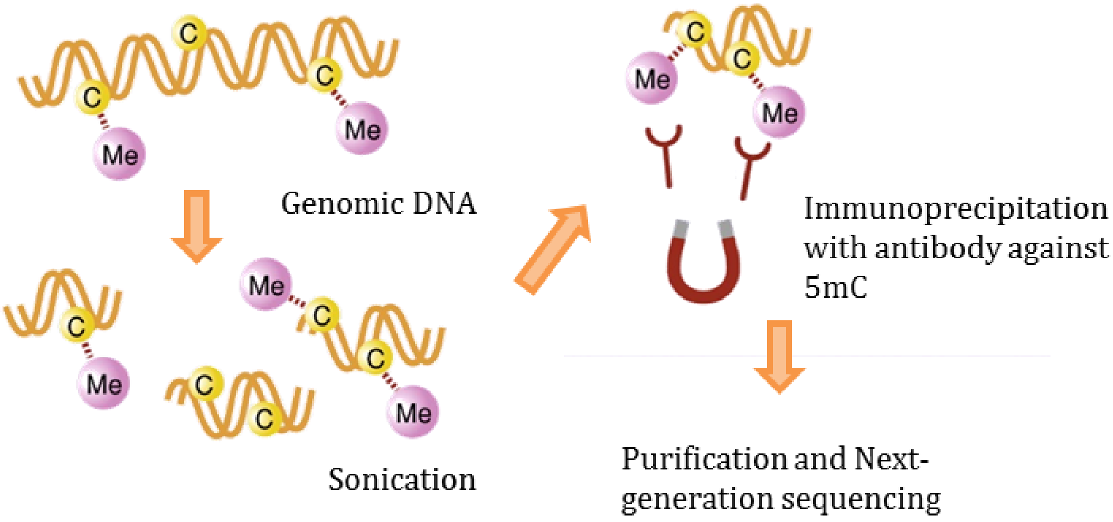

C
Me
Genomic DNA
Sonication
Immunoprecipitation with antibody against 5mC
Purification and Next-generation sequencing

Figure 1.6: Experimental workflow for MeDIP-seq

Genomic DNA is first sonicated into fragments of ~100 bp; methylated regions are then immunoprecipitated using an antibody raised against methylated cytosines, and the enriched portion is subsequently purified and analysed on a next-generation sequencing platform. Figure adapted from Denk et al.[50]

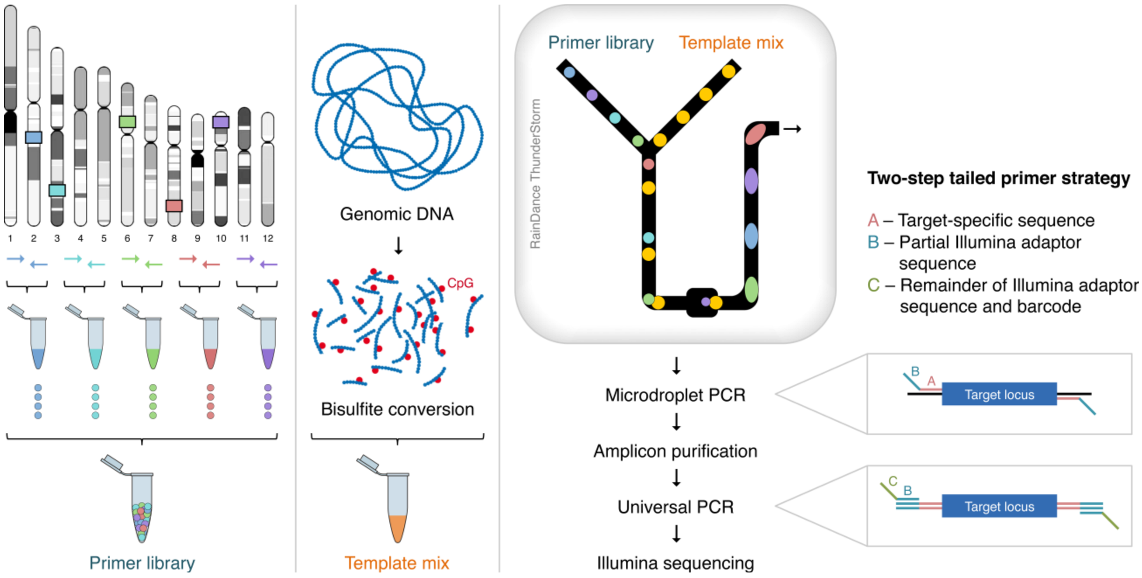


Figure 1.7: Workflow for methylation analysis by RainDrop-BSseq

The sample DNA is bisulfite converted and this template is then merged with individual primer pairs in microdroplets. After the subsequent PCR amplification of the amplicons, the droplets are destabilised to release the products, which are then purified with magnetic beads or columns. An additional round of PCR incorporates the sequencing barcodes for each sample. Figure courtesy of Dirk S. Paul.

1.3 Bone sarcomas

1.3.1 Overview

Malignant bone cancers are sarcomas of bone, cartilage and associated tissues, of which the three major types are osteosarcoma (OS), chondrosarcoma (CS), and Ewing's sarcoma (ES): together they represent over 90% of all bone cancers diagnosed in patients 15 to 29 years old [51]. In the UK, incidence rates of bone sarcoma have remained relatively stable since the 1970s (Cancer Research UK), as shown in **Figure 1.8**.

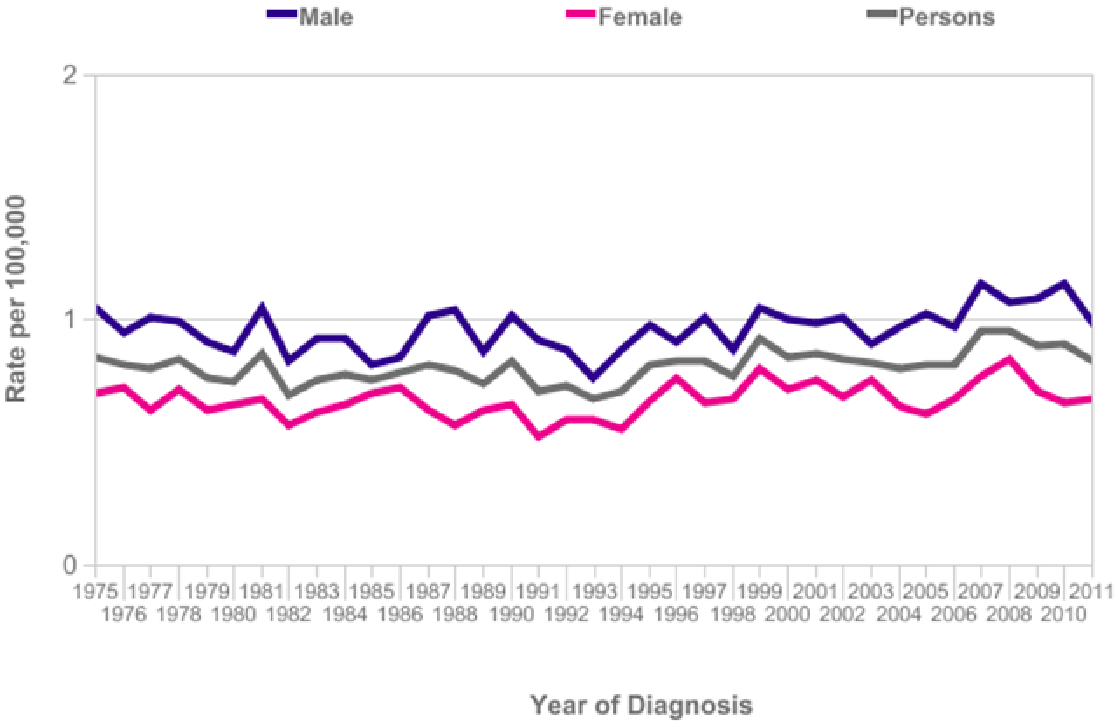


Figure 1.8: UK age-standardised incidence rates of bone sarcoma per 100,000

The incidence of bone sarcoma has remained stable overall since 1975 in the UK, but shows a slight dip followed by an increase. This increase is most likely due to improved diagnostic techniques rather than an actual rise in incidence. Original figure from Cancer Research UK.

1.3.2 Osteosarcoma

Osteosarcoma is the most common type of primary bone cancer and the 3rd most common childhood cancer with an age-standardised incidence in the UK of 8 and 6 per million in males and females, respectively[52]. It follows a bimodal age distribution, with the early and major peak in adolescents and young adults (5-20 years old) where it correlates with an important growth spurt. The second peak of incidence is in adults over 65 years of age, at which point it most often occurs as a second malignancy due to adverse effects of treatment, such as radiotherapy. The 5-year survival rate (**Figure 1.9**) improved from ~20% in 1970 to ~65% in 1975 with the introduction of preoperative chemotherapy; little progress has been made since then, however, and the current 5-year survival rate stands at 68%, highlighting the need for new therapeutic strategies.

Three cancer predisposition syndromes are known to increase the incidence of OS: 3% of OS patients are affected by Li-Fraumeni syndrome[53], characterised by a mutation in *TP53*, while hereditary retinoblastoma (*RB1*) and Rothmund-Thomson syndrome (*RECQL*) are also found in a significant number of OS cases. In addition to these hereditary syndromes, various environmental factors can contribute to an enhanced risk of osteosarcoma development: ionizing radiation, for example, has been implicated as a causative factor in ~3% of osteosarcomas.

Although the genetic traits of OS have been well documented[54-57], with chromosomal gains (1p, 1q, 6p, 8q, and 17p) and losses (3q, 6q, 9, 10, 13, 17p) as well as mutations in crucial tumour suppressor genes (*RB1, MDM2, CDKN2A…*) and oncogenes, the epigenetic aspects of OS have received less attention.

In 2009, Sadikovic et al.[58] described the integration of genomic and epigenomic profiles of two osteosarcoma cell lines with gene expression, using normal osteoblasts to provide baseline levels. The authors described a large number of under- or overexpressed genes in both cell lines that were also hyper- or hypomethylated, respectively. From gene network analysis, the authors concluded that the aberrantly expressed genes affected four major biological pathways and that the epigenetics and genomic imbalance seemed to have a cumulative role in the deregulation of gene networks in the cell lines. Although the data presented provided an interesting initial look at the epigenetic state of osteosarcoma, it remains an *in vitro* study and as such requires validation from work on primary tumours.

A crucial issue with bone sarcomas, however, is the scarcity of tissue samples available for analysis. In order to avoid the bias induced by using cell lines, a possible alternative is tumour xenografting. Validating the use of patient-derived tumour xenografts (PDXs) for epigenetic studies of rare cancers, such as OS, and in pre-clinical drug trials, in collaboration with the OncoTrack Consortium, is discussed in *Chapter 5*.

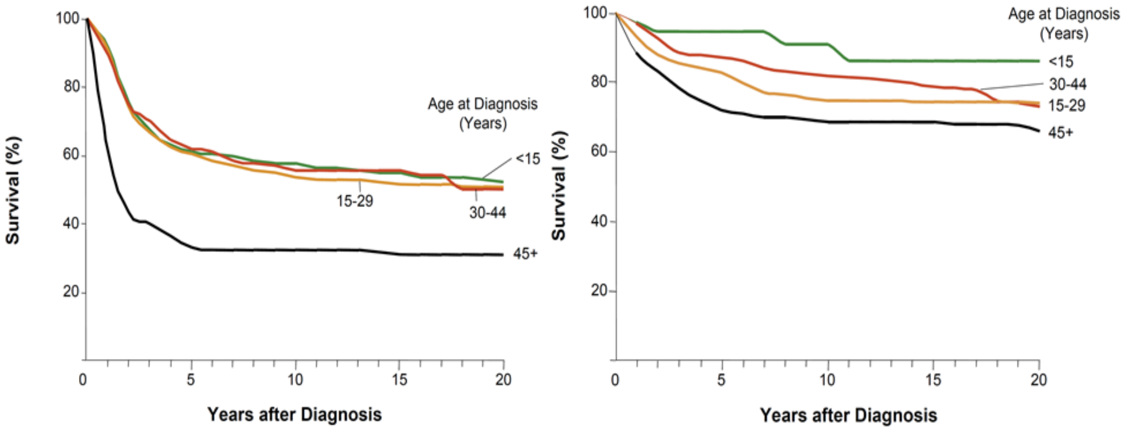


Figure 1.9: Osteosarcoma and chondrosarcoma survival rates

Survival rates of osteosarcoma (left) and chondrosarcoma (right) in the US from 1975 to 2000. Original figure from Bleyer et al.[51]

1.3.3 Chondrosarcoma

Chondrosarcomas are also infrequently found malignancies of the bone with an incidence of 80 cases per year in the UK. They are a heterogeneous group of tumours with highly diverse features and range from slow-growing lesions to highly aggressive metastasising sarcomas. CS is a cancer of the cartilage that generally arises around the pelvic bones, shoulder bones and the upper parts of the arms and legs. It can occur at any age, although its incidence is higher in later life (>50); various sub-classifications of CS exist such as central, peripheral, clear cell, mesenchymal, and dedifferentiated. Currently, the preferred treatment for chondrosarcoma is wide surgical excision; successful use of adjuvant chemotherapy has been reported in a few cases, but its use in mainstream treatment of CS remains ill defined[59]. The 5-year survival rate (**Figure 1.9**) varies depending on the stage and grade of the tumour: for conventional chondrosarcoma patients, it ranges from 48% to 60%; in the case of highly aggressive, dedifferentiated CS, however, it can be as low as 10% after one year[60].

Various factors can increase the risk of CS incidence, although a precise mechanism for initial development has yet to be determined. Patients with benign bone tumours such as chondromas or osteochondromas present a slightly increased risk of developing CS, for example[53]. Other syndromes, such as Ollier's disease or Maffucci's syndrome can also lead to CS.

The genetics of CS have also been extensively studied, and recent reports[61,62] have described mutations or rearrangements in cartilage collagen gene *COL2A1* (37%), the RB1 pathway (33%), *TP53* (20%), and Hedgehog signalling (18%). The presence of *IDH1* or *IDH2* gain-of-function mutations in over 50% of central chondrosarcomas[63] suggested that these tumours might be under a level of epigenetic regulation, as observed in other malignancies with frequent *IDH* mutations, such as AML. The identification and analysis of a potential hypermethylation phenotype, as well as its consequences on gene expression, associated with *IDH* mutations in CS forms the basis of the experiments discussed in *Chapter 3*. In addition to this epigenetic work, the potential of the 450K array to be used as a SNP array and provide copy number information, was utilised to uncover novel copy number variants (CNVs) in chondrosarcoma, and these are also discussed in *Chapter 3*.

1.4 Aims of the project

1.4.1 Genomic and epigenomic analysis of chondrosarcoma

IDH gain-of-function mutations are found in 50% of central CS, but the methylome of *IDH* mutant (MUT) relative to wild-type (WT) tumours had not yet been characterised. My first objective was, using a combination of microarrays and sequencing-based techniques, to determine whether *IDH* mutation in CS is associated with a similar DNA hypermethylation profile as previously found in other malignancies, and what its functional consequences are in terms of transcriptional regulation. In addition, I aimed to uncover novel and recurrent genomic alterations in CS by harnessing data generated for methylation profiling (*Chapter 3*).

1.4.2 Meta-analysis of *IDH*-mutant cancers

By combining the methylation data generated for CS with those from AML, glioma, and cholangiocarcinoma already in the public domain, I aimed to investigate whether the hypermethylation phenotype affects the same loci in all four cancers, and use the location of hypermethylated sites as a means to better understand the process of targeted demethylation (*Chapter 4*).

1.4.3 PDXs as a tool for epigenomic studies

Through the use of xenografts generated in OS and colon cancer, the final aim of this thesis was to quantify the suitability of PDXs to serve as proxies for patient tumours for both fundamental epigenetic research and pre-clinical drug screening (*Chapter 5*).

2 Materials and Methods

2.1 Chondrosarcoma

2.1.1 Samples

The material was obtained from the Stanmore Musculoskeletal Biobank, the approval for which was provided by the Cambridgeshire 1 Research Ethics Committee (Reference Number: 09/H0304/78).

IDH mutations were tested and validated in the Flanagan laboratory by at least two of the following techniques including Sequenom MassARRAY, capillary sequencing, exome sequencing, and a custom-made Taqman array[63,64].

A total of 71 patient samples and 4 technical blood controls were used for various analyses (**Table 2.1**).

2.1.2 450K array

2.1.2.1 DNA extraction

DNA was extracted from tumour samples using a QIAamp DNA Mini Kit (QIAGEN) according to the manufacturer’s instructions.

2.1.2.2 Bisulfite conversion

Bisulfite conversion of DNA for methylation profiling was performed using the EZ DNA Methylation kit (Zymo Research) according to the manufacturer’s instructions, on 500 ng from tumour samples. Conversion efficiency was assessed by quantitative PCR (qPCR), on the Applied Biosystems 7300 with default settings.

Samples	IDH Mutation Status	450K Dataset	Gene Expression Dataset	450K Sample Name	450K analysis plots sample name	Gene Expression Sample Name	Age	Sex	Grade	Site	2HG (ng/ml)	RainDrop-Bsseq	Pyrosequencing	SNP Array
MUT 6 FF/13361	Mutant	Test	1	MUT 6 FF/13361	MUT.19	MUT_S00022335	50	M	G4	Pelvis	79900	Validation	N/A	N/A
MUT 19 FF/13004	Mutant	Test	1	MUT 19 FF/13004	MUT.37	MUT_S00022355	70	F	G2	Humerus	30900	Validation	MUT A	N/A
MUT 3 FF/13340	Mutant	Test	N/A	MUT 3 FF/13340	MUT.17	N/A	72	M	G2	foot	34100	N/A	N/A	MUT 3 FF/13340
MUT 10 FF/13007	Mutant	Test	1	MUT 10 FF/13007	MUT.31	MUT_S00022313	62	F	G2	hand	30900	Validation (x2)	MUT B	MUT 10 FF/13007
MUT 1 FF/13952	Mutant	Test	N/A	MUT 1 FF/13952	MUT.18	N/A	71	M	G2	Pelvis	32800	N/A	N/A	N/A
MUT 2 FF/13924	Mutant	Test	N/A	MUT 2 FF/13924	MUT.20	N/A	36	F	G4	Pelvis	9830	Validation	N/A	N/A
MUT.exWT 5 FF/14082	Mutant	Test	N/A	MUT.exWT 5 FF/14082	MUT.26	N/A	77	M	G2	Tibia	32500	Validation	N/A	MUT.exWT 5 FF/14082
MUT 16 FF/13138	Mutant	Test	1	MUT 16 FF/13138	MUT.35	MUT_S00022342	73	M	G2	Pelvis	17400	Validation	N/A	MUT 16 FF/13138
MUT 11 FF/13432	Mutant	Test	N/A	MUT 11 FF/13432	MUT.28	N/A	69	F	G3	hand	42600	Validation	N/A	N/A
MUT 18 FF/13008	Mutant	Test	N/A	MUT 18 FF/13008	MUT.29	N/A	65	M	G2	calcaneus	33400	N/A	N/A	N/A
MUT 9 FF/13359	Mutant	Test	N/A	MUT 9 FF/13359	MUT.32	N/A	81	M	G2	Femur	28700	N/A	N/A	N/A
MUT.exWT 12 FF/13398	Mutant	Test	N/A	MUT.exWT 12 FF/13398	MUT.30	N/A	81	F	G2	Tibia	28600	Validation	N/A	N/A
MUT 17 FF/13402	Mutant	Test	N/A	MUT 17 FF/13402	MUT.36	N/A	66	M	G4	Femur	102000	N/A	N/A	MUT 17 FF/13402
MUT 14 FF/13364	Mutant	Test	1	MUT 14 FF/13364	MUT.34	MUT_S00023112	70	M	G4	Femur	22100	N/A	N/A	MUT 14 FF/13364
MUT.exWT 6 FF/13334	Mutant	Test	N/A	MUT.exWT 6 FF/13334	MUT.22	N/A	78	M	G2	Femur	1460	N/A	N/A	N/A
WT 1 FF/13676	Wild-Type	Test	N/A	WT 1 FF/13676	WT.1	N/A	69	F	G2	Rib	313	N/A	N/A	N/A
WT 7 FF/13662	Wild-Type	Test	N/A	WT 7 FF/13662	WT.7	N/A	73	F	G4	Pelvis	35.4	Validation	N/A	N/A
WT 4 FF/13664	Wild-Type	Test	N/A	WT 4 FF/13664	WT.4	N/A	31	M	G2	Scapula	136	N/A	N/A	N/A
WT 8 FF/12262	Wild-Type	Test	N/A	WT 8 FF/12262	WT.8	N/A	67	F	G1	Femur	2030	N/A	N/A	N/A
WT.exMUT 5 FF/13343	Wild-Type	Test	N/A	WT.exMUT 5 FF/13343	WT.6	N/A	50	M	G1	Humerus	8230	Validation	WT A	N/A
WT 2 FF/13139	Wild-Type	Test	1	WT 2 FF/13139	WT.5	WT_S00022333	10	F	G1	Femur	162	N/A	N/A	WT 2 FF/13139
WT 3 FF/13927	Wild-Type	Test	N/A	WT 3 FF/13927	WT.3	N/A	44	M	G2	sternum	66.2	N/A	N/A	N/A
WT.exMUT 13 FF/13663	Wild-Type	Test	N/A	WT.exMUT 13 FF/13663	WT.15	N/A	66	M	G2	sternum	138	N/A	N/A	N/A
WT 11 FF/13679	Wild-Type	Test	1	WT 11 FF/13679	WT.10	WT_S00022375	65	F	G1	Pelvis	32.5	N/A	N/A	N/A
WT 9 FF/13427	Wild-Type	Test	1	WT 9 FF/13427	WT.11	WT_S00019661	30	F	G2	Tibia	86.6	Validation	WT B	WT 9 FF/13427
WT 10 FF/13446	Wild-Type	Test	1	WT 10 FF/13446	WT.9	WT_S0022394	60	M	G4	Femur	42.4	Validation	N/A	WT 10 FF/13446
WT 14 FF/13006	Wild-Type	Test	N/A	WT 14 FF/13006	WT.14	N/A	71	M	G4	Tibia	73.4	Validation	N/A	WT 14 FF/13006
Blood Control 1	N/A	N/A	N/A	BC1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Blood Control 2	N/A	N/A	N/A	BC2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Blood Control 3	N/A	N/A	N/A	BC3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Blood Control 4	N/A	N/A	N/A	BC4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MUT_S00022895	Mutant	N/A	1	N/A	N/A	MUT_S00022895	N/A	N/A	G1	N/A	N/A	N/A	N/A	N/A
MUT_S00023115	Mutant	N/A	1	N/A	N/A	MUT_S00023115	27	M	G4	Pelvis	99200	N/A	N/A	N/A
MUT_S00021034	Mutant	N/A	1	N/A	N/A	MUT_S00021034	N/A	N/A	G2	N/A	N/A	N/A	N/A	N/A
MUT_S00021429	Mutant	N/A	1	N/A	N/A	MUT_S00021429	N/A	N/A	G4	N/A	N/A	N/A	N/A	N/A
WT_S00018175	Wild-Type	N/A	1	N/A	N/A	WT_S00018175	73	M	G4	Pelvis	873	N/A	N/A	N/A
MUT_S00022892	Mutant	N/A	1	N/A	N/A	MUT_S00022892	N/A	N/A	G1	N/A	N/A	N/A	N/A	N/A
MUT_S00021192	Mutant	N/A	1	N/A	N/A	MUT_S00021192	N/A	N/A	G2	N/A	N/A	N/A	N/A	N/A
MUT_S00021459/60	Mutant	N/A	1	N/A	N/A	MUT_S00021459/60	59	M	G2	Femur	N/A	N/A	N/A	N/A
MUT_S00022886	Mutant	N/A	1	N/A	N/A	MUT_S00022886	N/A	N/A	G2	N/A	N/A	N/A	N/A	N/A
MUT_S00022351	Mutant	N/A	1	N/A	N/A	MUT_S00022351	39	M	G2	Humerus	11600	N/A	N/A	N/A
MUT_S00022512	Mutant	N/A	1	N/A	N/A	MUT_S00022512	51	M	G4	Femur	N/A	N/A	N/A	N/A
MUT_S00022367	Mutant	N/A	1	N/A	N/A	MUT_S00022367	67	F	G2	Femur	N/A	N/A	N/A	N/A
WT_S00023122	Wild-Type	N/A	1	N/A	N/A	WT_S00023122	28	F	G1	Pelvis	N/A	N/A	N/A	N/A
MUT_120000563	Mutant	Replication	2	MUT_120000563	N/A	MUT_120000563	69	F	G3	Lung	N/A	N/A	N/A	N/A
WT_120001764	Wild-Type	Replication	2	WT_120001764	N/A	WT_120001764	64	M	G2	Pubic Ramus	N/A	N/A	N/A	N/A
WT_GER50	Wild-Type	Replication	2	WT_GER50	N/A	WT_GERMANY 50	24	M	G1	Rib	N/A	N/A	N/A	N/A
MUT_120001152	Mutant	Replication	2	MUT_120001152	N/A	MUT_120001152	64	M	G2	Ishchium	N/A	N/A	N/A	N/A
WT_GER93	Wild-Type	Replication	2	WT_GER93	N/A	WT_GERMANY 93	87	F	G2	Sternum	N/A	N/A	N/A	N/A
MUT_GER104	Mutant	Replication	2	MUT_GER104	N/A	MUT_GERMANY 104	60	F	G4	Femur	N/A	N/A	N/A	N/A
WT_GER69	Wild-Type	Replication	2	WT_GER69	N/A	WT_GERMANY 69	72	F	G2	Hip Joint	N/A	N/A	N/A	N/A
WT_GER84	Wild-Type	Replication	2	WT_GER84	N/A	WT_GERMANY 84	80	M	G3	Kidney	N/A	N/A	N/A	N/A
MUT_GER125	Mutant	Replication	2	MUT_GER125	N/A	MUT_GERMANY 125	70	F	G2	Femur	N/A	N/A	N/A	N/A
MUT_GER64	Mutant	Replication	2	MUT_GER64	N/A	MUT_GERMANY 64	66	M	G2	Hip Joint	N/A	N/A	N/A	N/A
WT_120001316	Wild-Type	Replication	2	WT_120001316	N/A	WT_120001316	72	F	G4	Tibia	N/A	N/A	N/A	N/A
MUT_120000960	Mutant	Replication	2	MUT_120000960	N/A	MUT_120000960	32	M	G2	Femur	N/A	N/A	N/A	N/A
WT_GER31	Wild-Type	Replication	2	WT_GER31	N/A	WT_GERAMNY 31	30	F	G3	Femur	N/A	N/A	N/A	N/A
MUT_GER131	Mutant	Replication	2	MUT_GER131	N/A	MUT_GERMANY 131	72	F	G3	Femur	N/A	N/A	N/A	N/A
MUT_GER49	Mutant	Replication	2	MUT_GER49	N/A	MUT_GERMANY 49	72	M	G2	Humerus	N/A	N/A	N/A	N/A
MUT_120001357	Mutant	Replication	2	MUT_120001357	N/A	MUT_120001357	67	M	G4	Femur	N/A	N/A	N/A	N/A
WT_GER83	Wild-Type	Replication	2	WT_GER83	N/A	WT_GERMANY 83	78	M	G4	Hip Joint	N/A	N/A	N/A	N/A
WT_09-1225	Wild-Type	Replication	N/A	WT_09-1225	N/A	N/A	50	M	G2	Neck	98.4	N/A	N/A	N/A
MUT_GER81	Mutant	Replication	2	MUT_GER81	N/A	MUT_GERMANY 81	32	F	G4	Chest Wall	N/A	N/A	N/A	N/A
WT_120000656	Wild-Type	Replication	2	WT_120000656	N/A	WT_120000656	29	M	G3	Clavicle	N/A	N/A	N/A	N/A
MUT_GER48	Mutant	Replication	2	MUT_GER48	N/A	MUT_GERMANY 48	50	M	G2	Hip Joint	N/A	N/A	N/A	N/A
MUT_120001050	Mutant	Replication	2	MUT_120001050	N/A	MUT_120001050	77	M	G4	Tibia	N/A	N/A	N/A	N/A
MUT_GER129	Mutant	Replication	2	MUT_GER129	N/A	MUT_GERMANY 129	31	F	G2	Hip Joint	N/A	N/A	N/A	N/A
MUT_120001779	Mutant	Replication	2	MUT_120001779	N/A	MUT_120001779	62	M	G2	Chest Wall	N/A	N/A	N/A	N/A
WT_GERMANY 86	Wild-Type	N/A	2	N/A	N/A	WT_GERMANY 86	N/A	N/A	N/A	N/A	N/A	Replication	N/A	N/A
18222	Mutant	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Replication	N/A	N/A
18228	Mutant	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Replication	N/A	N/A
17929	Mutant	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Replication	N/A	N/A
12633	Mutant	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Replication	N/A	N/A
18065	Mutant	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Replication	N/A	N/A
18224	Mutant	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Replication	N/A	N/A

Table 2.1: Chondrosarcoma samples

List of samples used for experiments on chondrosarcoma, including methylation, SNP, and gene expression arrays, RainDrop-BSseq, and pyrosequencing. Clinical information such as age, sex, tumour grade and *IDH* mutation status are also included. Dataset numbers refer to different sample batches.

The qPCR mix for each sample consisted of 1µl DNA, 4.375 µl water, and 6.25 µl MESA Blue qPCR Mastermix Plus for SYBR Assay (Eurogentec). Each sample was run in triplicate. Actin primers for both converted and unconverted DNA were used, with sequences shown below:

Actin_Converted:	F (5'>3')	TGGTGATGGAGGAGGTTTAGTAAGT
	R (5'>3')	AACCAATAAAACCTACTCCTCCCTTAA
Actin_Unconverted:	F (5'>3')	TGGTGATGGAGGAGGCTCAGCAAGT
	R (5'>3')	AGCCAATGGGACCTGCTCCTCCCTTGA

Conversion efficiency was calculated according to the following formulae:

$\%Unconverted = \frac{100\%}{1+2^{\Delta Ct}}$, with ΔCt = AvgCt(Unconverted) – AvgCt(Converted)

$\%Converted = 100\% - \%Unconverted$

2.1.2.3 Array Processing

For each sample, a total of 500 ng of bisulfite converted DNA was used. The Infinium HumanMethylation450 BeadChips (Illumina)[8] were processed as per manufacturer’s recommendations by the UCL Genomics facility. Raw data was then obtained in the form of Illumina IDAT files.

2.1.2.4 Analysis

The raw output from the 450K BeadChips was processed using GenomeStudio software (Illumina) and uploaded to GEO (accession number GSE40853). The non-normalised and non-background corrected data and array annotation were exported as text files from GenomeStudio and all subsequent analysis was performed using the R statistical software v2.15.0 (http://www.R-project.org) with R packages[65-68] and custom scripts (*Appendices*). Quality control of the data resulted in removal of any probes that did not pass a detection p-value threshold of 0.01 across all samples; a final dataset of 27 samples (12 WT and 15 MUT) and 472,655 probes were available for analysis in the test set.

A principal component analysis[69] of the data was performed to identify the principal components of variation. Unsupervised consensus clustering was conducted on the top probes selected using a median absolute deviation (MAD) estimator, which provides a more robust measure of variance than standard deviation. I selected the top 150 most variable positions corresponding to a lower-end threshold of MAD = 0.5. Thus, these selected probes show substantial variance with methylation differences across many samples on the order of 50% methylation changes. I also performed consensus clustering on more methylation variable positions (MVPs) by lowering the MAD threshold to include 300 and 500 probes, with identical results, demonstrating robustness to the choice of threshold.

A Wilcoxon rank sum test was used for supervised analysis; p-values obtained from the latter were adjusted for multiple testing (Benjamini-Hochberg[70]) and only probes with p-value ≤ 0.001 were used in the clustering. A further filter of absolute (Δ (medianβ)) ≥ 0.35 was used to compensate for the Wilcoxon rank sum test not taking into account absolute difference in methylation between the groups, and to narrow down our search to differences with higher potential for functional effect.

The MVPs used to separate the validation sample sets (n = 24, 10 WT and 14 MUT) were selected based on the same method used for the filtering of MVPs in the initial dataset, specifically ordering them by: 1) increasing adjusted p-value and then 2) decreasing absolute median difference between the MUT and WT groups.

The statistical significance of the observed percentage enrichments for genomic and epigenomic features among the 3,057 MVPs was calculated on the basis of 1,000 repetitions of a random selection of 3,057 probes from the overall probe set (472,655 probes) used in the analysis. The aforementioned features correspond to the official annotation of the 450K BeadChips, and were extracted using GenomeStudio.

Copy number variation was assessed using the ChAMP[71,72] package, filtering the data for detection p-value (0.01), and setting the bead cut-off at 0.05. GISTIC[73] analysis was performed using hg19 as the reference genome, the output of the ChAMP CNV module as the segmentation file, the full array probe coordinates as the markers file, and with the following settings: amp/del threshold = 0.3, removeX = yes, join segment file = 4, qv threshold = 0.05, confidence level = 0.95. The full analysis script can be found in the *Appendices*.

2.1.3 RainDrop-BSseq

2.1.3.1 Overview

For the validation and replication using targeted microdroplet PCR bisulfite sequencing (RainDrop-BSseq), sample preparation and bisulfite conversion were carried out as described above. The parallel amplification of target loci was performed by RainDance Technologies (Lexington, MA, USA) and the subsequent sequencing by Illumina. The RainDance technology allows massively parallel amplification of specific DNA fragments by conducting PCR reactions in pico litre droplets on integrated microfluidic chips. The produced library (one for each sample) was then separately subjected to a second round of PCR to incorporate the sequencing indices. The libraries for all samples were pooled and sequenced on the Illumina MiSeq.

2.1.3.2 Protocol for 2nd PCR

- The samples were quantified by running 1 µl on an Agilent Bioanalyzer
- The samples were diluted to 2.5 ng/µl
- The following PCR master mix was prepared in a PCR tube:

Platinum HiFi Buffer (10x)	3.25 µl
$MgSO_4$ (50 mM)	0.88 µl
dNTPs (10 mM)	0.88 µl
Betaine (4 M)	2.50 µl
DMSO	1.25 µl
Primer pairs (5 µM)	2.50 µl
1st PCR template (2.5 ng/ µl)	4.00 µl
Platinum HiFi Taq (5 units/ µl)	0.50 µl
Water	9.24 µl

- The samples are PCR thermal cycled with the following program:
    - 94°C – 2 minutes
    - 94°C – 30 seconds
    - 56°C – 30 seconds
    - 68°C – 60 seconds
    - Repeat steps 2-4 9 times
    - 68°C – 10 minutes
    - 4°C – Forever
- Purify over a Qiagen MinElute column
- Quantify the reaction by running 1 µl on an Agilent Bioanalyzer

2.1.3.3 Analysis

Raw sequencing reads from the microdroplet PCR were trimmed to 60bp as recommended by Krueger et al.[74] and fastq_quality_trimmer from the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) was used to trim lower quality bases from the ends of sequence reads (threshold set at 30); reads were trimmed down to a minimum length of 20 bp, and removed if shorter.The alignment was conducted using Bismark[75], specifically designed for mapping bisulfite converted sequence reads. Finally, methylation states were determined using the Bismark methylation_extractor and custom Perl scripts. CpG sites covered by 10 sequencing reads or more and with methylation scores between 0-20% (unmethylated) or 80-100% (methylated) were selected. The alignment and filtering of reads were conducted by Dr. Gareth Wilson. The sequencing data is available from GEO (accession number GSE40853).

2.1.4 Pyrosequencing

Pyrosequencing validation was conducted using PyroMark Gold Q96 (QIAGEN) reagents and the PyroMark Q96 MD pyrosequencer as per manufacturer's instructions.

2.1.4.1 Bisulfite conversion

DNA bisulfite conversion was performed with the EZ DNA Methylation kit (Zymo Research). Eluted DNA was normalised to 10 ng/µl, and the conversion efficiency was assessed by qPCR as described above.

2.1.4.2 Target loci and Primer Design

Pyrosequencing targets were selected to overlap with probed sites on the 450K array and with targets of RainDrop-BSseq, as shown in **Table 2.2**. The primers were designed using PyroMark Assay Design software (QIAGEN) and ordered from Sigma-Aldrich.

450K Probe ID	Genomic Feature	Epigenomic Feature	Hg18 Coordinate
cg08924430	TSS1500	N_Shore	4:106286501
cg10884288	5'UTR	N_Shore	7:4888722
450K Probe ID	**Forward Primer**	**Reverse Primer**	**Sequencing Primer**
cg08924430	AGGGGGTTATTAGTGAGAAATTTAT	ACATACCCTTAATACTTTAAAAACCTATAT	GGTTAAAGTAAATAGAAGGT
cg10884288	AAGGAAGGGTTTAGTTTTTGATG	TTTCCCTCCTACTAAAAATAATAATAAATT	GAGTATAGGGAGTGAG

Table 2.2: Pyrosequencing targets

The selected targets for pyrosequencing with their (epi)genomic feature descriptions and the corresponding primer sequences for amplification and sequencing.

2.1.4.3 PCR

A standard PCR programme with touchdown annealing temperatures was used to increase specificity and reduce the risk of primer dimers. The PCR mix consisted of the following components for each DNA sample, using the GoTaq HotStart (Promega) and 10 µl of the PCR product was then used for each pyrosequencing reaction.

• 5x GoTaq Buffer	10 µl
• MgCl2 (25 mM)	5 µl
• dNTPs (10 mM)	2 µl
• Taq (5 u/ul)	0.25 µl
• H2O	28.75 µl
• Primers (10 µM)	2 µl
• DNA	2 µl

2.1.4.4 Checking for methylation bias in the amplification

A standard curve for each primer set is created by mixing fully methylated and fully unmethylated DNA standards in varying proportions to achieve 0%, 25%, 50%, 75% and 100% methylated DNA. These are then subjected to pyrosequencing to ensure that the read-out of the methylation score as determined by the sequencer is proportional to the methylation level of the sample. The standard curves created for sites 1 and 2 are shown in **Figure 2.1**. The fully methylated and unmethylated standards were produced respectively using in vitro methylation with M.SssI and whole-genome amplification using the REPLI-g kit (Qiagen). The plots show that although the amplification seems unaffected by the methylation level of the sample (linear standard curve with $R^2 > 0.9$), the pyrosequencing slightly underestimates that methylation level, possibly due to incomplete methylation of the standard used.

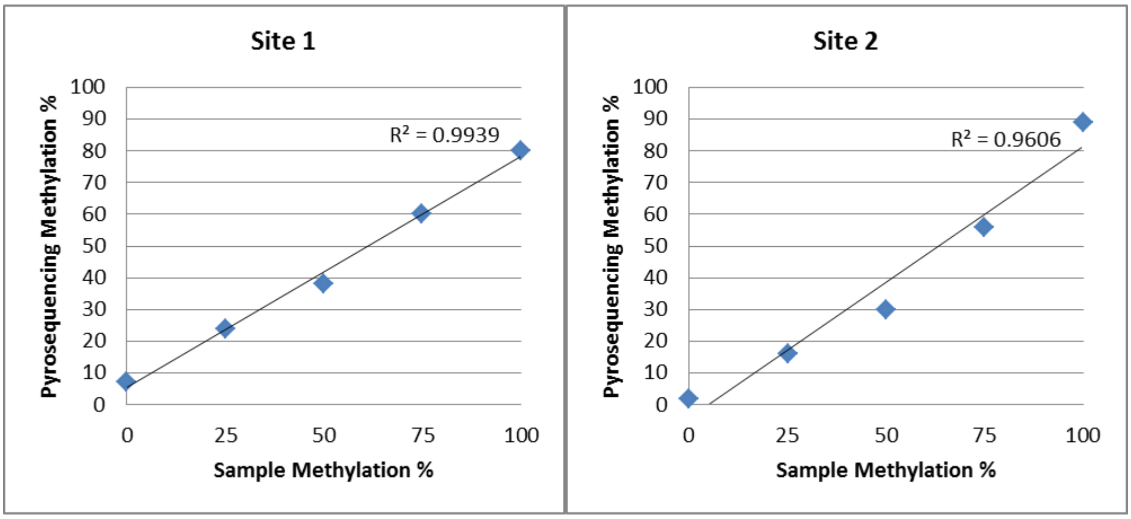


Figure 2.1: Pyrosequencing standard curve

Methylation bias in each primer pair for amplification was assessed by amplifying DNA standards of varying methylation levels through the pyrosequencing reaction. As shown by the high R^2 (> 0.9) in each plot, the primers amplify these regions in a manner unbiased by the methylation level of the sample DNA, but the methylation score observed through pyrosequencing slightly underestimates that of the sample.

2.1.5 SNP array

2.1.5.1 Processing

From each of 10 CS samples, 300 ng DNA was extracted as described for the 450K array above. The samples were run on the HumanCytoSNP-12 BeadChips (Illumina) by UCL Genomics, according to the manufacturer’s instructions. The raw data was extracted using GenomeStudio (Illumina).

2.1.5.2 Analysis

The R values extracted from GenomeStudio were analysed using the DNAcopy[76] R package and custom scripts (*Appendices*). The output from DNAcopy was then processed through GISTIC as described above for the 450K array data.

2.1.6 Gene expression array

2.1.6.1 RNA extraction

RNA was extracted from 46 frozen tumour samples with the following protocol:

- Frozen sections were thawed in Qiazol (Qiagen) on ice
- Sections were homogenised
- Samples were incubated for 30 min at RT
- Chloroform (200 µl/ml Qiazol) was added, tubes shook vigorously for 20s and incubated for 5 min.
- All samples were centrifuged at 14,000 g for 15 min at 4°C.
- Aqueous phase was taken (or repeated phenol/chloroform steps if supernatant was not clear)and an equal volume of 70% ethanol was added;
- Samples were mixed and loaded onto an RNeasy column (Qiagen).
- RNA was purified according to the manufacturer's instructions.

2.1.6.2 Array preparation and processing

Gene expression was measured using the HumanHT-12 v4 Expression BeadChips (Illumina). Sample total RNA conversion to biotin-labelled cRNA was performed using the TargetAmp Nano Labeling Kit for Illumina Expression BeadChips (Epicentre), as per manufacturer's instructions. Following amplification and labelling the cRNA was purified using the ZR-96 Clean & Concentrator kit (Zymo Research), to allow parallel purification of up to 96 samples, as well as concentration of the RNA to meet the requirements of downstream hybridisation to the HT-12 BeadChips. Integrity of the RNA was assessed both before and after labelling using the RNA Nano 6000 assay on the Bioanalyser (Agilent), and concentration was measured using the Nanodrop ND-1000 after labelling, as the biotin tag interferes with RNA migration thus affecting Bioanalyser readings. Hybridisation of the samples to the arrays, followed by staining and scanning, was performed at UCL Genomics.

2.1.6.3 Analysis

The data was exported from GenomeStudio without background correction or normalisation and analysed using the limma[77] R package, following the analysis protocol outlined by Ritchie et al.[78] The analysis results are described in detail in *Chapter 3* and the full analysis script is available in the *Appendices*. Briefly, the probe signals were background corrected using negative control probes and quantile normalised using negative and positive control probes with the limma *neqc* function. Probes were then filtered according to their annotation quality, with removal of those labelled as “No match” or “Bad”. Unsupervised clustering was performed on the 500 most variable probes as determined by the interquartile range. Supervised clustering was conducted on the top 500 probes as determined by either p-value or log fold change (logFC) calculated with the *topTable* limma function.

2.2 EBF1-TET2 interaction

2.2.1 Meta-analysis of publically available datasets

2.2.1.1 Data

For the meta-analysis, I used the published list of differentially methylated genes for AML (n = 398, 347 WT and 51 MUT), significantly differentially methylated genes (Wilcoxon p-value ≤ 0.001, |Δβ| ≥ 0.35) for LGG (n = 81, 32 WT and 49 MUT) and CC (n = 50, 31 WT and 19 MUT) and the data I generated for CS, with a further restriction to sites found in gene promoters and CpG islands/shores.

2.2.1.2 Gene, pathway, and motif analysis

The Ingenuity Pathway Analysis (IPA) Functional Analysis identified the biological functions that were most significant to the dataset. A right-tailed Fisher's exact test was used to calculate a p-value determining the probability that each biological function assigned to that dataset is due to chance alone. This p-value was further adjusted (Benjamini-Hochberg[70]) for multiple testing. Canonical pathways analysis identified the pathways from the IPA library of canonical pathways that were most significant to the dataset. Molecules from the dataset that were associated with a canonical pathway in the Ingenuity Knowledge Base were considered for the analysis. Fisher's exact test was used to calculate a p-value determining the probability that the association between the genes in the dataset and the canonical pathway is explained by chance alone. This p-value was further adjusted (Benjamini-Hochberg[70]) for multiple testing. The motif analysis was conducted using the online multiple expectation maximisation for motif elicitation (MEME) suite of tools[79]: FASTA sequences were downloaded from the UCSC Genome Browser, and used for input in the MEME-Chip tool of the MEME suite; parameters were set to default except for the number of repetitions (set to 'Any number of repetitions'), motif width (min=4, max=15), and maximum number of motifs to find (20).

2.2.2 Chromatin immunoprecipitation (ChIP)

2.2.2.1 Cell culture

The ChIP experiments were performed on the CS cell line SW1353 (ATCC HTB-94, IDH2 R172S), cultured until passage 11, in RPMI-1640 medium (Lonza), with Lglutamine and sodium bicarbonate, and supplemented with 10% foetal bovine serum (FBS). Cell culture flasks (25, 75, or 175 cm^3) were used for this adherent cell lines and incubated at 37°C, 5% CO_2, and 100% humidity.

2.2.2.2 Chromatin extraction

Each chromatin extraction was performed on a single SW1353 frozen pellet of ~10^7 cells. The process involved the following steps:

2.2.2.2.1 Preparations

- Prepared 1% (or 2%) Formaldehyde (FM) in media (from 37% stock – 135 µl FM+ 4.865 ml media).
- The volume of each buffer was aliquoted and protease inhibitors (PI; frozen at 10x) were added. Lysis buffer was left to dissolve/clear up at room temperature (RT) before aliquoting. All buffers placed on ice.
    - Lysis buffer (for 500 ml):

1% SDS	25 ml 20% SDS solution
10 mM EDTA, pH 8.0	10 ml 0.5M
50 mM Tris-HCl, pH 8.0	25 ml 1M
ddH2O	460 ml

    - Hypotonic buffer:
        - 10 mM Tris/HCl pH 7.2
        - 2 mM $MgCl_2$
        - 0.5% Triton X100
- The Bioruptor UCD-200 sonicator (Diagenode) was filled with ice and a large beaker with ice-water. Sonicator-adaptors were placed on ice.
- The centrifuge was pre-cooled.

2.2.2.2.2 Cell collection and cross-linking

- Cells were resuspended in 1ml 1%FM media in an eppendorf tube.

- Cell suspension was transferred to 15 ml falcon tubes and another 4 ml 1%FM was added. Fixation was conducted for 20 min at RT.
- Fixation was stopped by adding 550 µl 1.25 M glycine (Sigma-Aldrich; final concentration 0.125 M) and cell suspension was nutated on platform for 5 min at RT.
- Cells were spun at 447 g for 4 min and media removed.
- Cells were resuspended in 1 ml cold phosphate buffered saline (PBS) and transferred to an eppendorf tube. Tubes were spun at 447 g 4 min at 4°C, supernatant was removed and this wash step was repeated one more time. Cells were kept on ice.
- The cell pellet was resuspended in 300 µl hypotonic buffer with PI by pipetting up and down, and incubated on ice for 10 min.
- Tubes were spun at 699 g for 5 min at 4°C and the supernatant was discarded.
- Nuclei were resuspended in 300 µl of lysis buffer with PI by pipetting up and down and the number of cells was normalised ($5x10^6$-$1x10^7$ cells/tube) and aliquoted into Diagenode sonication tubes. Each tube was topped up to 300 µl with lysis buffer. Samples were incubated on ice for 30 min (10-30 min).

2.2.2.2.3 Sonication

- Before sonicating, the lysate was transferred to the Diagenode sonication tubes if not already in these tubes.
- Lysate was sonicated on ice to reduce DNA length to between 100 and 300 bp for sequencing or up to 500 bp for qPCR analysis on the Bioruptor:
    - Max output, 30 sec on-off cycles for 10 min; 4 x 10 min cycles
- The tube was spun 15 min at 13,400 g, 4°C to remove debris. The supernatant was transferred to a new 1.5 ml eppendorf tube.
- The sonication efficiency of the lysate was checked on reverse cross-linked chromatin:
    - Added to a PCR strip:

- lysate equivalent to about 100,000 cells (6 µl if 5x10^6 cells)
- 1 µl Proteinase K (20 mg/ml)
- 50 µl Tris-EDTA 25 mM
- Incubated:
- 55°C 15'
- 100°C 15'
- 4°C (to cool down)
- Ran on a 1.5% agarose gel.

2.2.2.3 ChIP

Automated ChIP was performed using the Auto-ChIP Kit and IPure kit (Diagenode) and the SX-8G IP-Star (Diagenode) according to the manufacturer's instructions. For each reaction, 3 µg and 7 µg of TET2 (sc-136926; Santa Cruz) and EBF1 (clone 1G8, Abnova) antibodies, respectively, were used. The following steps were followed for each experiment:

- To 50 µl of the sheared chromatin, were added:
  - 5 µl protease inhibitor mix (200x)
  - 450 µl ChIP Buffer H
  - (This 0.5 ml sheared chromatin mix could then be used for 4 IPs + input.)
- Preparation Buffer H + Ab:
  - x µl antibody + 100-x µl ChIP Buffer H
- Preparation IPure Elution Buffer (for each reaction):

Buffer A	96 µl
Buffer B	4 µl
Total Volume	100 µl

- Reagents were dispensed as below on an ice block. One strip per IP:

1	IPure EB (IN)	90ul
2	Empty	
3	Magnetic beads	10 µl
4	ChIP Buffer H	50 µl
5	ChIP Buffer H	50 µl
6	ChIP Buffer H + Ab	100 µl
7	Sheared Chrom Mix	100 µl
8	Wash Buffer H1	100 µl
9	Wash Buffer H2	100 µl
10	Wash Buffer H3	100 µl
11	Wash Buffer H4	100 µl
12	IPure EB (IP)	100 µl

- Tips were loaded onto the tip block of the IP Star and the strips were added. If only running 8 samples, the right block was used, with samples number 1-8 placed from left to right. Protocol called “ChIP IPure8 100vol” in the ChIP Ab coating folder was run.
- 10 µl input was added to well 1 and the caps were placed on the PCR strip: reverse crosslinking was then performed at 65°C for 4 hours.
- Ipure cleanup (in plate): Using a magnet, well 12 (IP) and well 1 (IN) were transferred to new tubes and placed on row 12 of the tube block.
    - Well 1: 100 µl Buffer C
    - Well 2: 100 µl Isopropanol, 15 µl beads, 2 µl carrier (+ 100µl MeDIP/input)
    - Well 3: 100 µl Wash Buffer 1
    - Well 4: 100 µl Wash Buffer 2
    - Well 5: Empty (pending elution)
    - Well 6: Empty (pending elution)

- Cleaned up DNA was collected in well 6. The eluted volume was adjusted to 50 µl with buffer C.

2.2.2.4 ChIP-qPCR

For each qPCR, 2 µl of the eluted DNA from the ChIP is used. The target regions and negative control region are referred to here by the nearest gene; qPCR primers for these sites were designed using NCBI Primer Blast and manufactured by Sigma-Aldrich:

CCND2	F:	5’-GTTTCTGCTCGAGGATCACA-3’,
	R:	5’GGGAGAGGTGGGTATTAGGA-3’
FABP3	F:	5’-CCTGGGGCTTCCTATTTCG-3’
	R:	5’-TGCCGCTTTAAATAGCCCTC-3’
FBRSL1	F:	5’-TACGCGCTGCATGAATCAAT-3’
	R:	5’-CTGGTGGGGTTTTCTGAGC-3’
OOEP	F:	5’-TATGGTCGATGATGCTGGTG-3’
	R:	5’-GGGTCTCTCAGTTCCTGCAC-3’

The *OOEP* primer set was used as negative control. qPCR was performed on the Applied Biosystems 7300. Enrichments were assessed by normalising qPCR results to the mock IgG IP.

2.2.2.5 ChIP-seq

For each sample, the sequencing library was prepared using the Microplex Library Preparation Kit (Diagenode) and sequenced on the MiSeq (Illumina) according to manufacturer’s instructions.

2.2.3 Co-immunoprecipitation

Immunoprecipitation of TET2 and EBF1 was performed with 3 µg of anti-TET2 antibody (sc-136926, Santa Cruz Biotechnology) and protein A sepharose in cell lysates of SW1353 cells with 5% of the lysate taken as input control before IP.

2.2.3.1 Preparations

- 1.1 ml per sample of lysis buffer was prepared (30mM HEPES, 20mM β-glycerophosphate, 20mM KCl, 1mM EGTA, 2mM NaF, 1mM Na_3VO_4, 1% TX100, 1mM benzamidine, 4µM leupeptin, 5mM PMSF, 1mM DTT at pH 7.4).
- Labelled 1.5ml Eppendorf tubes were prepared and pre-chilled on ice for at least 15 min.

2.2.3.2 Collection and lysis of cells

- Two flasks of suspension cells were pooled into 50ml falcons.
- Cells were gently washed with ice-cold PBS containing Ca^{2+} and Mg^{2+} (from Invitrogen DPBS) – about 4ml.
- Cell suspensions were pooled to end up with one falcon per condition.
- Tubes were spun down and the supernatant taken off; 1 ml fresh PBS was added to the pellet and the suspension was transferred to a 1.5 ml eppendorff tube.
- The tubes were spun down and PBS aspirated.
- 400 µl of lysis buffer were added to each tube.
- Tubes were incubated for 1h (cell lysis step), on roller in cold cabinet.
- In the meantime, the centrifuge was down to 4°C.
- The samples were spun for 10 min at maximum speed at 4°C.
- New tubes were pre-chilled on ice.
- Supernatant was transferred into the new pre-chilled tubes.
- A 50 µl aliquot of each sample was taken as input lysate control; 20 µl of Lammli buffer was added and sample was boiled for 5 min at 95°C; it was

then left to cool down and stored at -20°C until ready to run the samples by SDS-PAGE.

- IPs were topped up to 1,000 µl.
- 5 µl were taken for the Bradford assay and mixed with 1 ml of a 1 in 5 dilution of Bradford reagent (also made for the blank); 900 µl of the lowest concentration sample were used and a proportional volume of the others.

2.2.3.3 Pre-clearing and immunoprecipitation

- A second set of new tubes was pre-chilled on ice; 15 µl of Protein A/G-sepharose were added to each tube.
    - Protein A was used for rabbit antibody.
- A similar amount of each sample was transferred into the pre-chilled tubes containing Protein A/G-sepharose.
- Tubes were placed in 50 ml Falcon tubes and placed on RM5 rocker in cooling cabinet for 1-2 hours.
- In the meantime, a third round of new tubes was pre-chilled on ice.
- Samples were spun down for 5 min at 4°C at 500 g.
- The supernatant was transferred into pre-chilled tubes containing the antibody.
- The tubes were placed in 50 ml Falcon tubes and placed on RM5 rocker in cooling cabinet overnight.
- Beads were added (30 µl) and the samples incubated for 5 hours on rocker in cold cabinet.
- Tubes were spun for 3 min at 500 g, 4°C; the supernatant was collected and constituted the non-bound fraction.
- Each sample (bead pellet) was washed three times with 1 ml of wash buffer (20 mM Tris pH 8.0, 1M NaCl, 10% glycerol, 1% NP-40, 5mN EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 50mM NaF, 20 mM β-glycerophosphate, 1mM Na_3VO_4) ; Samples were spun 3 min, at 500 g at 4°C between each wash.

- After the final wash, the supernatant was taken off without touching the beads; 40 µl Lammli buffer were added; the samples were boiled for 5 min.

2.2.3.4 Western blotting

Detection of EBF1 and TET2 by western blotting (WB) was conducted with 1:500 and 1:1,000 dilutions of EBF1 (Abnova, H00001879-M02) and TET2 (Abcam, ab94580) antibodies respectively, with 5% milk as blocking agent in TBST (TBS with 0.1% Tween 20). Secondary antibodies from GE Healthcare were used at 1:5,000 dilutions.

2.3 Osteosarcoma xenografts

2.3.1 Samples and xenografting

OS patient-derived tumour xenografts (PDXs) were generated from tumour samples received directly from surgery (Stanmore RNOH). Samples were washed in PBS and cut to the appropriate size (~2-3mm^3). Under isoflurane anaesthesia delivered via a nasal attachment tube, tumour fragments were inserted subcutaneously in one or both flanks of the mice. In total, 14 female SCID mice (3-6 weeks old) were kept at the UCL Animal Housing facility in individually-ventilated cages, and monitored at least twice a week for the duration of the experiment. Procedures were followed as described in the project license (delivered by the UK Home Office PPL 70/6666) and, when necessary, animals were sacrificed according to an approved schedule 1 protocol. Tumour growth was measured using digital measuring callipers. Tumours were snap-frozen in liquid nitrogen after excision. The two OS patient tumour samples were obtained from the Stanmore Musculoskeletal Biobank, satellite to the UCL Biobank for Health and Disease, with ethics approval EC17/14 ("Using tissue surplus to diagnostic requirements to study the biology of benign and malignant tumours, and identify the cell of origin of these tumours").

2.3.2 450K array

2.3.2.1 Overview

The experimental aspects of the methylation array analysis of the OS PDXs were nearly identical to that described above for the CS samples. Briefly, genomic DNA was extracted from PDX samples using the QIAamp DNA Mini Kit (QIAGEN) according to manufacturer's instructions. The bisulfite conversion of the DNA was performed using the EZ DNA Methylation kit (Zymo Research) on 500 ng. Conversion efficiency was assessed by quantitative PCR (qPCR), as described above. The Infinium HumanMethylation450 BeadChips were processed by UCL Genomics as per manufacturer's instructions.

2.3.2.2 Analysis

The raw data obtained from the 450K arrays was processed from the IDAT files through to normalisation with BMIQ[80] using the ChAMP[71] pipeline, and all subsequent analysis was performed with the R statistical software v3.0.2 and custom scripts (*Appendices*). Quality control of the array data included removal of probes for which any sample did not pass a 0.01 detection p-value threshold, bead cut-off of 0.05, and removal of probes on the sex chromosomes. Probes passing the detection p-value threshold of 0.01 in the mouse-only sample were also removed from downstream analysis in all xenografts to avoid confounding signal from any mouse DNA. The genomic and epigenomic features used are those annotated on the array and enrichments were calculated on the basis of 1,000 repetitions of a random selection of probes from the overall probe set used in the analysis. The data was deposited into GEO under accession GSE59352.

2.3.3 MeDIP-seq

Methylated DNA Immunoprecipitation (MeDIP) followed by next-generation sequencing was also carried out on 200 ng DNA extracted from the seven PDXs and two patient tumours.

2.3.3.1 Sonication

- DNA was resuspended in 85 µl EB buffer (Qiagen).
- The sample was sonicated using the Bioruptor UCD-200 sonicator (Diagenode) for 1h (4 x 15 min cycles on high, 30s on/off cycles).
- The undiluted sample was run on a High Sensitivity DNA bioanalyzer chip (Agilent), aiming for a peak of 180-230 bp.
- The sample was re-sonicated if necessary.

2.3.3.2 Library preparation

Library preparation for MeDIP-seq was conducted using the NEBNext kit (New England BioLabs), according to the manufacturer's instructions. The product was eluted in 41 µl EB. Fragment length was checked again by running 1 µl of undiluted sample on a High Sensitivity DNA chip.

2.3.3.3 Immunoprecipitation

The MeDIP portion of the procedure was carried out following the steps below, with the Auto MeDIP kit (Diagenode), SX-8G IP-Star (Diagenode) and IPure kit (Diagenode):

- The 'Antibody (Ab) Dilution' was prepared as follows

Antibody	1 µl
Ultra pure Water	15 µl
Total Volume	16 µl

- The 'Antibody Mix' was prepared as follows

Diluted Ab (from Step 1)	2.40 µl
MagBuffer A (5X)	0.60 µl

MagBuffer C	2.00 µl

- The'Incubation Mix' for one IP (and Input) was prepared as follows:

MagBuffer A (5X)	24 µl
MagBuffer B	6 µl
Lambda Spike cocktail	1 µl + 1 µl
Ultra pure Water	18 µl
Adapter Ligated DNA	40 µl
Total Volume/reaction	90 µl

- The mix was incubated at 100°C for 10 min.
- MagBuffer A (1X) was prepared:

MagBuffer A (5X)	25 µl
Ultra pure Water	100 µl

- IPure Elution Buffer was prepared:

Buffer A	115.4 µl
Buffer B	4.6 µl

- The IP-Star was loaded:
    - The IP Star was switched on approximately 10 min before loading to equilibrate Peltier heat/cool blocks to 4°C.
    - The protocol MeDIP for IPure.HLD was selected.
    - The following reagents were loaded into a 12-strip tube (1 strip tube per IP):
    - Well 1: Empty
    - Well 2: Empty
    - Well 3: Empty
    - Well 4: 50 µl IPure Elution Buffer
    - Well 5: 50 µl 1xMagBuffer A + 10 µl beads
    - Well 6: 50 µl 1xMagBuffer A

  - Well 7: 75 µl Denatured Incubation mix + 20 µl 1xMagbuffer A + 5 µl antibody mix
  - Well 8: 100 µl MagWashBuffer1
  - Well 9: 100 µl MagWashBuffer1
  - Well 10: 100 µl MagWashBuffer1
  - Well 11: 100 µl MagWashBuffer2
  - Well 12: 50 µl IPure Elution Buffer
- After MeDIP, the following reagents were loaded into a NUNC U96 plate:
  - Well 1: 100 µl Buffer C
  - Well 2: 100 µl Isopropanol, 15 µl beads, 2 µl carrier (+ 100µl MeDIP/input)
  - Well 3: 100 µl Wash Buffer 1
  - Well 4: 100 µl Wash Buffer 2
- 50 µl of purified DNA was eluted.

2.3.3.4 Size selection

- A 100 ml 2% Tris/Borate/EDTA (TBE) agarose gel was prepared for each DNA sample.
- The PCR amplified DNA was mixed with 3 µl 6X loading dye and loaded on the gel, leaving space either side for 50 bp ladders.
- Gel electrophoresis was carried out in freshly prepared TBE buffer at 100 volts for 100 min.
- Following gel electrophoresis, the gel was transferred onto a UV transilluminator.
- A strip of aluminium foil was placed beneath the lane containing the DNA sample to prevent UV crosslinking of DNA.
- With a clean scalpel, the desired 50 bp library size range was excised: 250-300 bp, 300-350 bp, 350-400 bp; the agarose was dissolved in 2 ml lo-bind tubes.

- DNA libraries were extracted and purified using a MinElute Gel DNA extraction kit (QIAGEN), according to the manufacturer's instructions, eluting in 12 µl EB.

2.3.3.5 Sequencing

Libraries were sequenced on a HiSeq 2000 (Illumina), according to the manufacturer's instructions.

2.3.3.6 Analysis

The sequencing data was uploaded to GEO (GSE59352) and processed from fastq files using the MeDUSA[81] pipeline. The reads were aligned separately to both the hg19 and mm10 genomes, with all redundant and unpaired reads removed. After assessment of the levels of likely contamination from mouse DNA, those reads aligning only to human or to both human and mouse were kept for downstream analysis with custom scripts (*Appendices*).

3 Genomic and Epigenomic Analysis of Chondrosarcoma

Most of the results presented in this chapter, relating to the epigenetics of chondrosarcoma, an assessment of RainDrop-BSseq, and the derivation of copy number data from methylation arrays, have been published[1,49,72] and the corresponding abstracts and title pages can be found in the *Appendices*.

All clinical information and samples, including 2-HG measurements and *IDH1/2* mutation status were provided by Prof. Adrienne M. Flanagan et al. The RainDrop-BSseq library design was conducted in collaboration with RainDance Technologies and the microdroplet PCR was carried out by the company in Boston, MA; I performed the subsequent sequencing at and with the assistance of Illumina (Cambridge, UK), as part of my industrial placement, which was a requirement of the MRC CASE studentship supporting this work. Sequencing reads were filtered and aligned by Dr. Gareth Wilson.

3.1 Introduction

Over half of central CS harbour a somatic heterozygous gain-of-function mutation in *IDH1* and/or *IDH2*[63]. These mutant enzymes' 2-HG production leads to the competitive inhibition of, among others, the TET demethylases and subsequent DNA hypermethylation. Although this CpG island hypermethylation phenotype has been previously observed in other malignancies[30], this represents the first characterisation of its presence and effects in CS. I describe them here using methylation microarrays, with validation by two sequencing-based methods and replication in an independent set of samples. The transcriptional profiles of CS samples with or without *IDH* mutations were also investigated with microarrays and are discussed in detail in this chapter. In addition, the data obtained from methylation arrays was harnessed to investigate copy number alterations in the analysed samples[72] and uncover novel genomic features of CS.

3.2 DNA Methylation and *IDH* mutation in Chondrosarcoma

An initial sample cohort of 12 IDH-wild-type (WT) and 15 *IDH*-mutant (MUT) CS were profiled using the 450K array. After quality control of the raw data (*Chapter 2*), including removal from the whole dataset of any probe not passing a 0.01 detection p-value threshold, a final set of 472,655 β values was used in the downstream analysis.

3.2.1 Unsupervised analysis

The data was initially analysed in an unsupervised manner to determine the impact of *IDH* mutations on the observed methylome relative to other clinical factors as well as potential technical artefacts. I performed a principal component analysis using singular value decomposition (SVD)[69] in order to determine which factors had the highest impact on the variation in the data (**Figure 3.1**).

The first component, PC-1, is significantly correlated with both mutation status and 2-HG levels, with Kruskal-Wallis p-values of 2×10^{-6} and 6.8×10^{-4}, respectively. In addition, tumour grade also appears to affect the first few components. This can be explained by the uneven distribution of grades within the two groups, as grade I and grade III tumours were only present in the WT and MUT groups, respectively: *IDH* mutation is linked to increased DNA methylation and, as certain tumour grades are only found in one of the MUT or WT groups, the SVD analysis is only highlighting here an indirect link between tumour grade and DNA methylation. Moreover, previous studies[63,64] have found no correlation between tumour grade and *IDH* mutation and the present sample distribution seems to have occurred by chance.

Other available clinical factors, such as age at disease presentation and patient sex, were not significantly associated with principal components. Technical effects were also minimal with only a weak contribution of Sentrix ID, the 450K chip identifier, to the fifth component. These results support an effect of *IDH* mutation on the methylome via abnormal production of 2-HG.

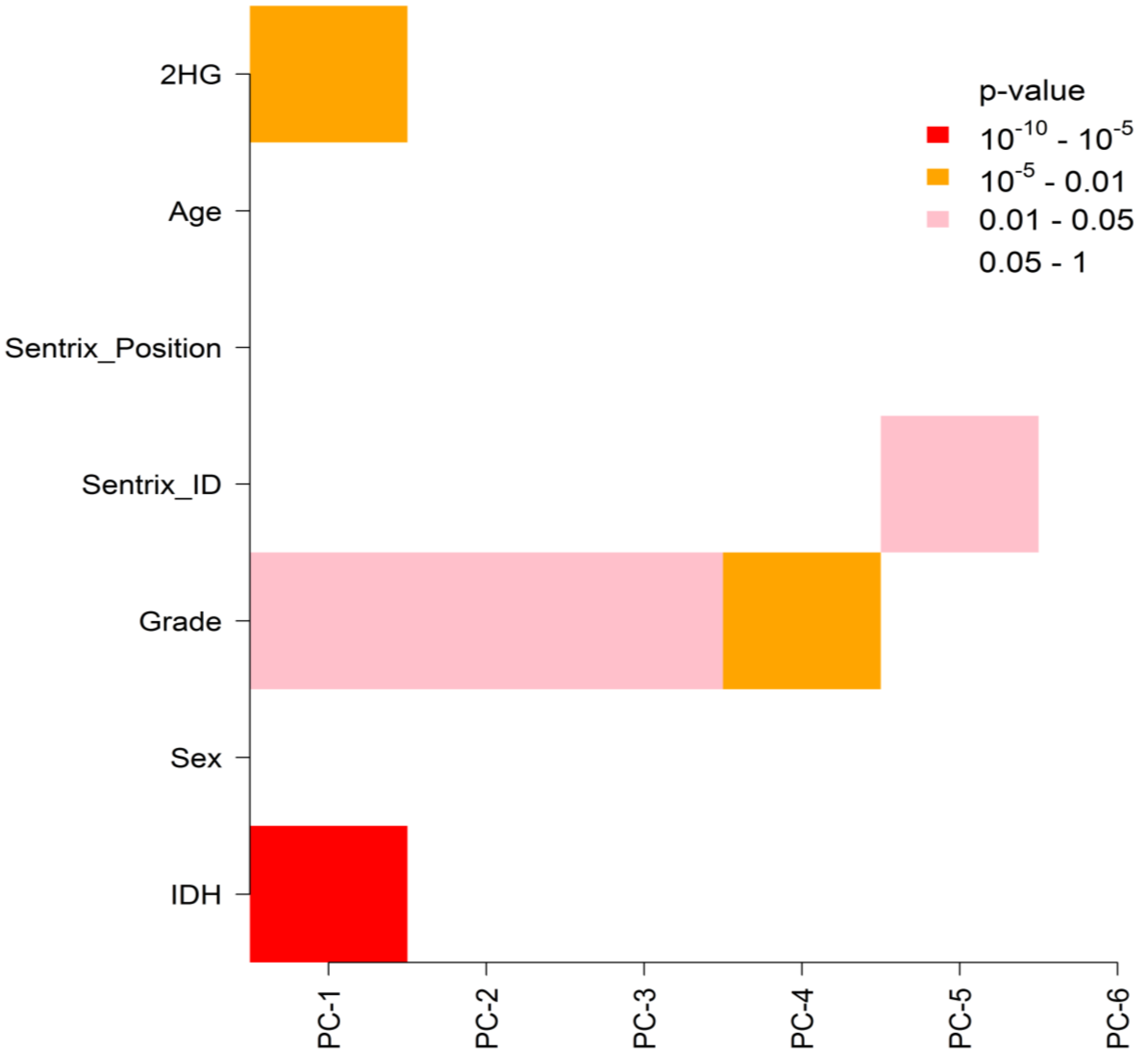


Figure 3.1: Principal component analysis

PC-k refers to the kth principal component. Here the first two components associate most strongly with *IDH* mutation status and 2-HG levels. Original figure from Guilhamon et al.[1]

Additional supporting evidence of *IDH* mutation as the main effector of the observed methylation variability was provided by unsupervised consensus clustering (**Figure 3.2**). I used the median absolute deviation (MAD) estimator, a more robust measure of variance than standard deviation, to select the 150 most variable probes on the array, irrespective of sample mutation status, and performed the clustering based solely on the methylation values of these probes. These 150 probes correspond to a threshold of MAD ≥ 0.5, indicating that they show methylation differences across samples in the order of 50% methylation changes.

The resulting plot shows four clusters, with the first two as 92% WT (1 MUT and 12 WT) and clusters three and four made up exclusively of MUT samples. I performed this analysis again by lowering the MAD threshold to include additional probes (300 and 500) and obtained identical results, demonstrating robustness of the clustering.

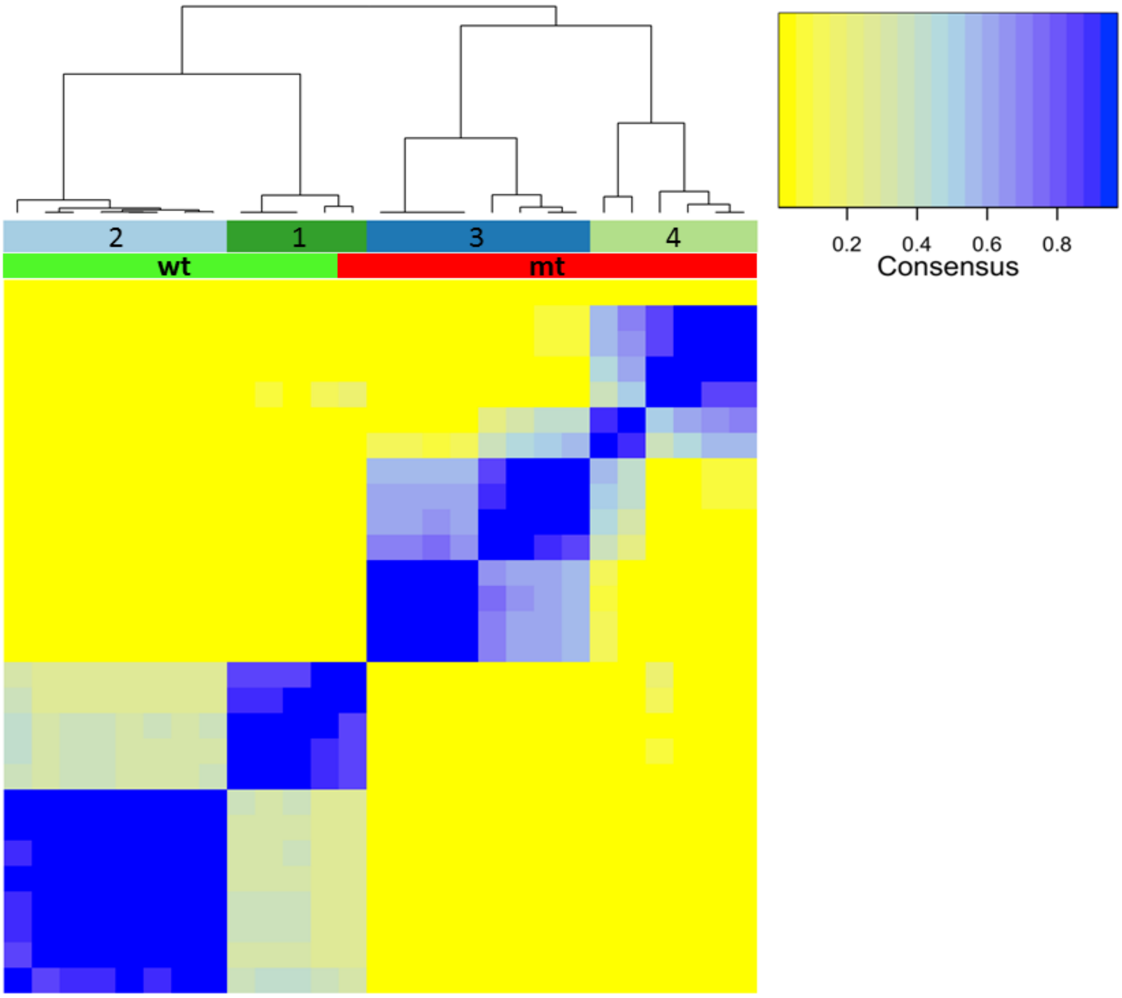


Figure 3.2: Unsupervised consensus clustering

Clustering performed on 150 most variable probes as determined by MAD (threshold MAD ≥ 0.5). Clusters one and two contain mainly WT samples, while clusters three and four are exclusively MUT. Original figure from Guilhamon et al.[1]

3.2.2 Supervised clustering

After establishing that the main contributor to the variation in methylation among CS samples was *IDH* mutation status, I conducted a supervised analysis using a Wilcoxon rank-sum test to ascertain both the directionality of methylation change between WT and MUT sample groups and the loci showing the largest and most significant changes. In order to narrow down the number of identified MVPs to those with the highest potential for functional consequences and to compensate for the Wilcoxon test not incorporating the absolute differences in methylation between groups in its results, I applied an absolute Δβ filter of 0.35 ($|\Delta\beta| \geq 0.35$) in addition to a statistical significance filter of p-value ≤ 0.001. Based on these filters, a total of 3,057 MVPs were available for downstream analysis.

Plotting the absolute median β difference between WT and MUT (**Figure 3.3**) shows that the vast majority of these 3,057 probes are hypermethylated in the *IDH* mutant group relative to the wild-type (99.5%, 3,042/3,057), providing further evidence that the hypermethylation phenotype observed in other cancers carrying these mutations can also be found in CS.

Hierarchical clustering of the samples (**Figure 3.4**) defines three distinct groups, indicated on the plot by the coloured bands labelled 'Cluster': from left to right, an exclusively WT cluster displaying low methylation levels, (median β = 0.16) a MUT group with intermediate to high β values (median β = 0.55) and a second MUT cluster with high methylation scores (median β = 0.75). The mutation status also correlates strongly with 2-HG levels, with a Spearman rank correlation coefficient ρ of 0.84 and p-value equal to 3.62×10^{-8}. This suggests, that not only is the hypermethylation phenotype observed here comparable to that previously found in other malignancies[24,29,31], but the mechanism linking it to *IDH* mutation is also similar, and involves abnormally elevated levels of 2-HG production. Other clinical factors, such as age and sex, are not significantly correlated with *IDH* status or methylation profile.

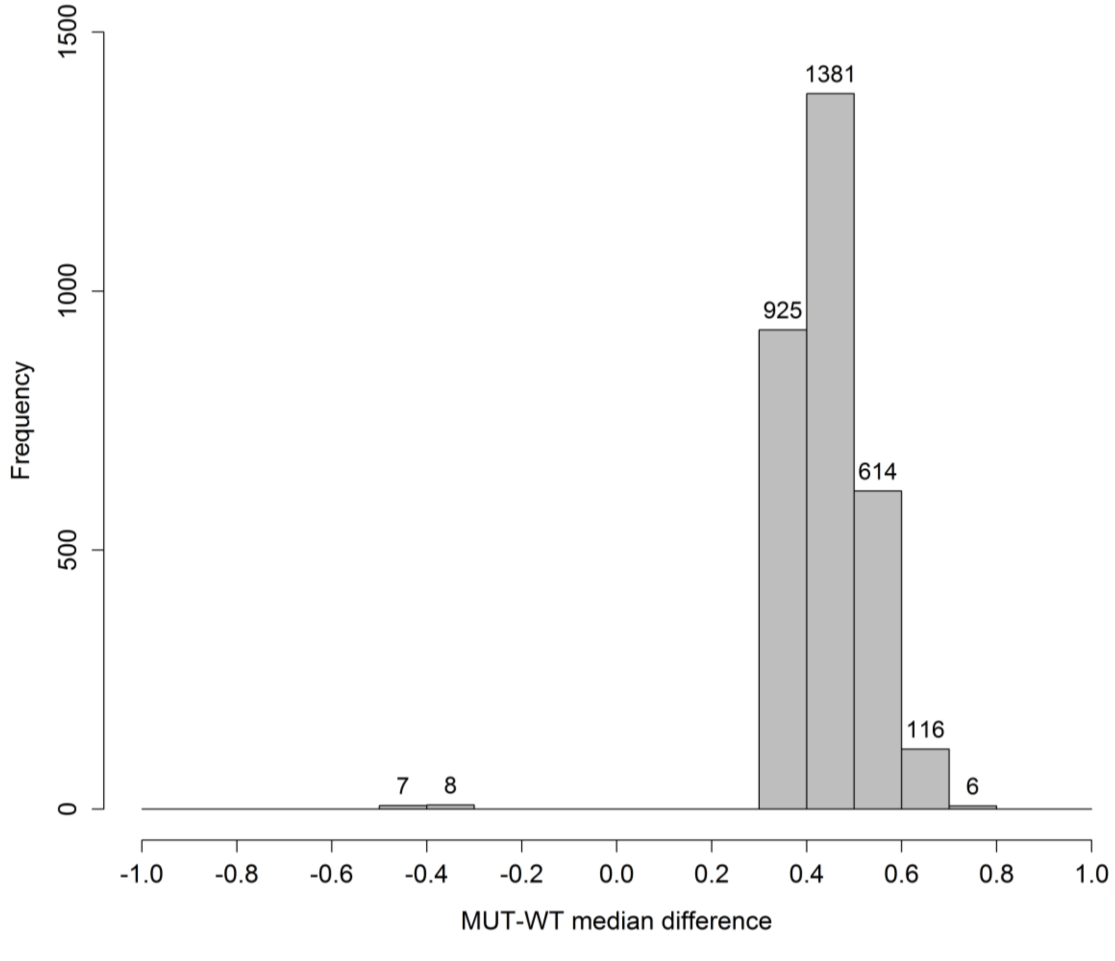


Figure 3.3: Frequency distribution of median β value differences between MUT and WT groups in selected probes

Of the top 3,057 probes, 3,042 (99.5%) are hypermethylated in MUT relative to WT samples. Original figure from Guilhamon et al.[1]

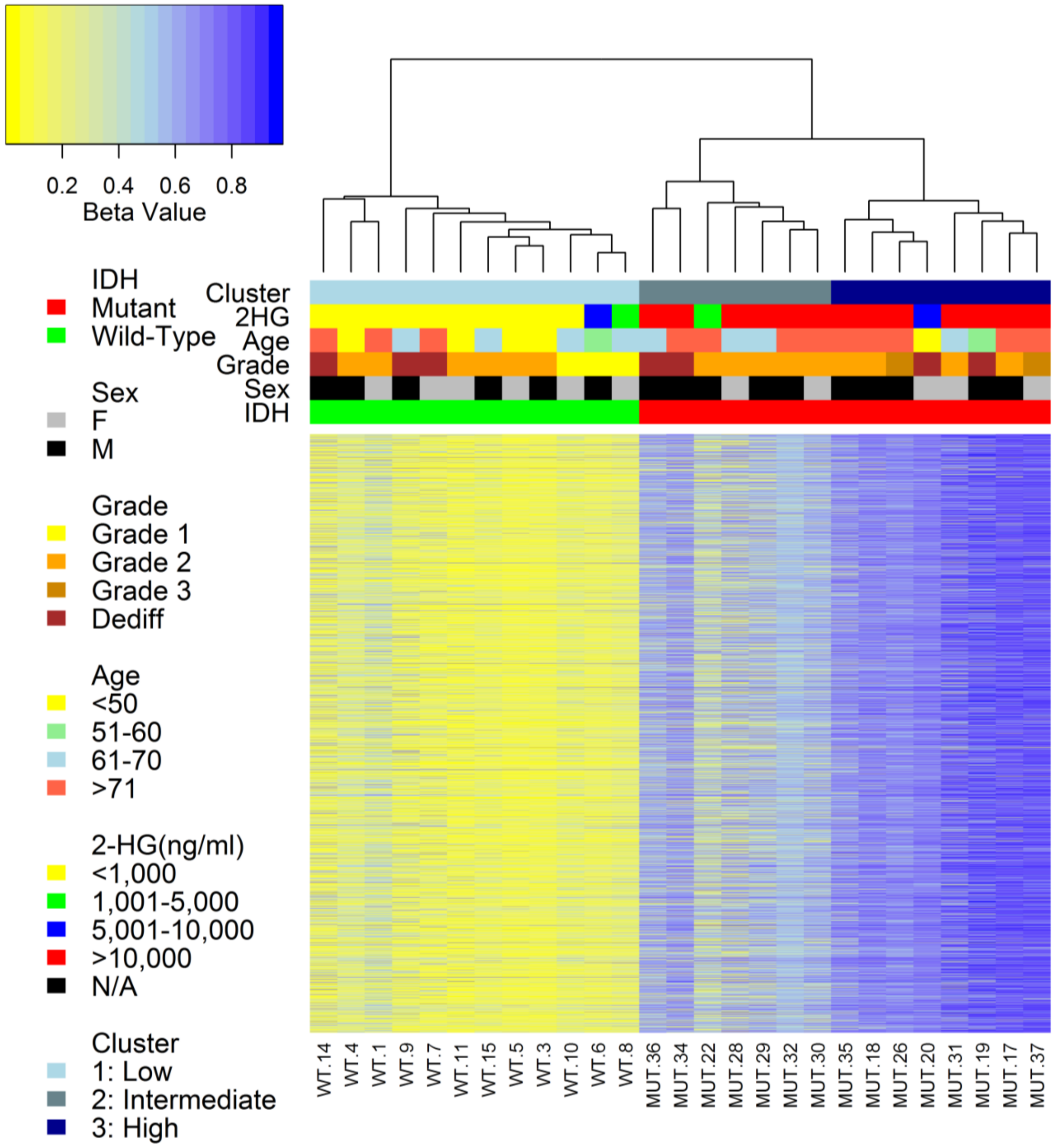


Figure 3.4: Hierarchical clustering of the top 3,057 MVPs between MUT and WT

The 3,057 MVPS were selected based on a Wilcoxon rank-sum test (p ≤ 0.001) and an additional filter for large differences in methylation levels between the two sample groups (|Δß| ≥ 0.35). The samples separate into three groups: low/unmethylated WT cluster (1), intermediate/high methylation MUT cluster (2), and high methylation MUT cluster (3). The measured 2-HG levels positively correlate with *IDH* mutation and hypermethylation, while other clinical factors such as age and sex show no correlation. Original figure from Guilhamon et al.[1]

3.2.3 Genomic and epigenomic distribution of *IDH* mutation-associated MVPs

The 450K methylation array is annotated in such a way that each probe can be classified by its position within a particular genomic and/or epigenomic feature. The annotation chosen by Illumina for this array defines CpG islands as DNA sequences with a GC content of 50% or higher and an observed/expected CpG ratio of 0.6 or more, while CpG shores are mapped to the 2 kb regions directly surrounding the islands, both upstream and downstream[8]. Aiming to find whether a pattern existed in the distribution of significant MVPs across the genome, I calculated the enrichment of the annotated features (**Figure 3.5**) through a random resampling strategy (p-value ≤ 0.001) (*Chapter 2*).

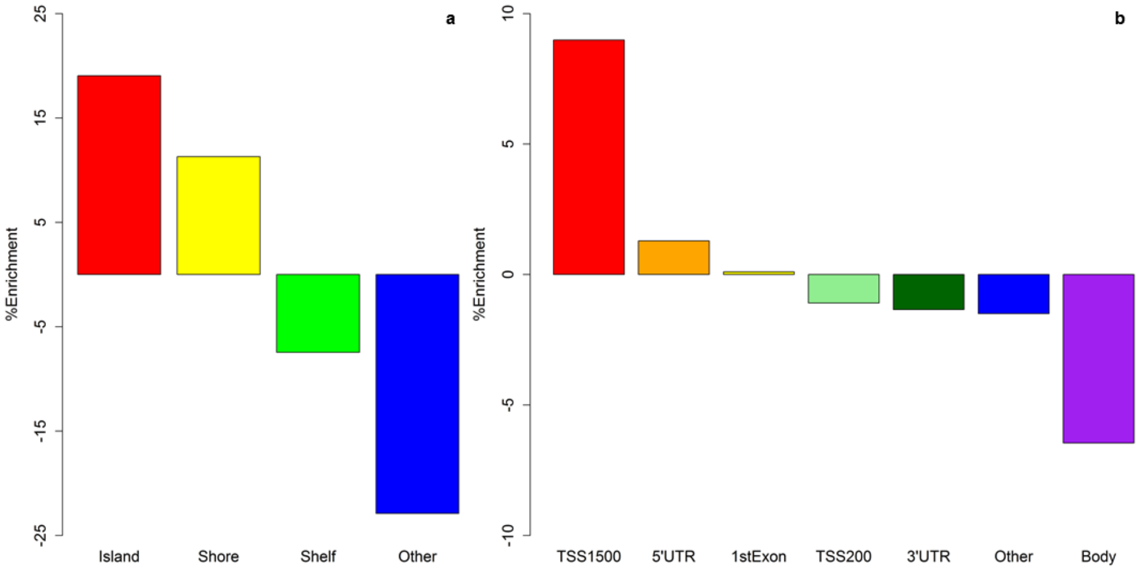


Figure 3.5: Feature enrichment in 3,057 *IDH* mutation-associated MVPs

a) Percentage enrichment of epigenomic features determined by random resampling (p-value ≤ 0.001) indicates that CpG islands and shores are enriched for by 19.1% and 11.3%, respectively.
b) Percentage enrichment of genomic features determined by random resampling (p-value ≤ 0.001). TSS1500 = 1,500 bp upstream of transcription start site; TSS200 = 200 bp upstream of transcription start site; the TSS1500 region is enriched for by 9%, while probes located within the gene body are depleted by 6.4%.
Original figure from Guilhamon et al.[1]

The identified MVPs are over-represented in CpG islands and shores by 19.1% and 11.3%, respectively as well as in promoter-associated genomic features such as transcription start sites (9%). High levels of methylation in CpG-dense regions associated with gene promoters are known to affect transcriptional regulation[82] and these *IDH* mutation-associated changes in methylation are thus likely to play a functional role in tumour development.

3.2.4 Validation

The results obtained from the 450K array were validated using two additional methods: RainDrop-BSseq and pyrosequencing.

3.2.4.1 RainDrop-BSseq

A total of 27 target regions were analysed in each sample, amounting to 500 individual amplicons, and covering 212 of the CpG sites present on the array. Each target region was centred on a particular CpG site and the targets were selected based on the following criteria: they were in genes central to the *IDH* mechanism, such as *IDH1/2* and *TET1/2/3*, in genes previously known to play a role in CS development, as *GLI2/3* in the Hedgehog signalling pathway, or among the most hypermethylated probes identified in the supervised analysis. The latter category was comprised of probes present in genes such as *TP73*, *CUL1* and *ACCN4*. The selected sites were hypermethylated in MUT relative to WT, each was one of multiple loci found hypermethylated in the affected gene and they covered a range of genomic and epigenomic features.

As RainDrop-BSseq used recently developed technology for a novel application, various issues arose in the process, only some of which could be remedied at the analysis stage. In the latter category was the occasional mismatch of amplicon length and sequencing read length: I used 150 bp paired-end reads for sequencing, adequate for the longer amplicons, but these ranged from 70 to 200 bp in length, which translated into the sequences for shorter amplicons carrying a number of unidentified bases at each end and being classified as poor quality by automated alignment and read-filtering pipelines. This particular problem was partially solved by using more stringent trimming of the sequencing reads for alignment purposes. The other two issues I encountered could not, however, be corrected *a posteriori*. The first was a simple PCR clean-up issue: the products from the first round PCR conducted by RainDance Technologies had not been cleaned up in a way that excluded small primer dimers and these then amplified in the second round PCR, resulting in numerous short (20-30 bases) and uninformative sequences. The second issue was due to a technical flaw in the instrument used for the microdroplet PCR (**Figure 3.6**). The library was designed in a tiled manner, with overlapping amplicons in order to adequately cover the entire target region. This means that in some instances the reverse primer for a particular amplicon hybridised well downstream of the forward primer for the next amplicon, and when these were inaccurately placed together in a droplet, they amplified only the short region between them. Taking the following two regions as an example: chr2:121386496-121386610 and chr2:121386600-121386739; here the primer at chr2:121386610 was accidentally combined with the one at chr2:121386600 so that only the ten bases between them were amplified instead of the two full amplicons these primers were designed for.

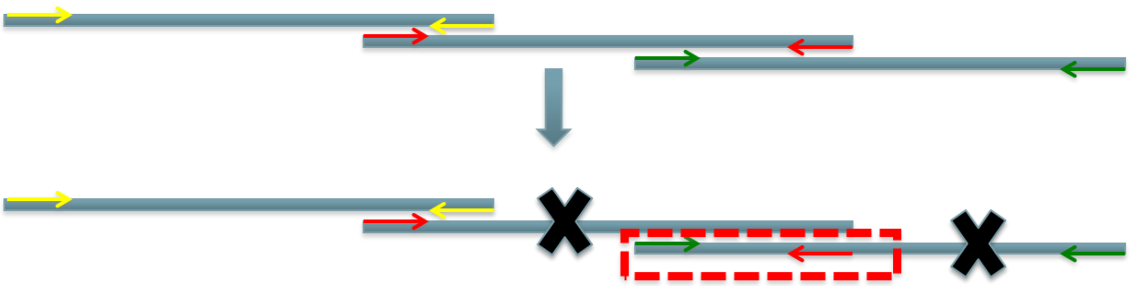


Figure 3.6: Inaccurate coupling of PCR primers in RainDrop-BSseq

Inadvertant mismatching of primers from overlapping tiled amplicons produced short, unexpected sequences.

Despite these issues, using RainDrop-BSseq enabled me to validate the methylation score of a vastly superior number of sites than would have been possible through pyrosequencing alone (**Figure 3.7a**). A total of 16 samples were used in the validation set, including 11 MUT and 5 WT and 890 CpG sites across that entire cohort were covered at a sufficient depth (≥10 reads) to call their methylation score with confidence. Of those 890 loci, 98.8% (426/431) and 95.5% (429/449), in the WT and MUT groups, respectively, matched the methylation state determined with the 450K array .

3.2.4.2 Pyrosequencing

One of the most widely-used methods for validation of DNA methylation since the early 2000s has been pyrosequencing[35,37,38]. Briefly, the method involves designing primer pairs to amplify selected regions of bisulfite-converted DNA, and sequencing these fragments in a light-emitting reaction: tri-phosphate nucleotides are added sequentially to the reaction; if the nucleotide matches the template and hybridises, its incorporation releases a pyrophosphate which is then used in ATP generation, consequently providing the required substrate for the luciferase enzyme to emit light and indicate inclusion of the nucelotide.

Although well-established and characterised, pyrosequencing suffers from being notoriously challenging to optimise (~50% of designed primer pairs successfully amplify bisulfite-converted DNA in a methylation-unbiased manner) as well as low-throughput, considering the comparatively large amount of data generated by current discovery tools (in this case, the 450K array). Nevertheless, it provided a useful benchmark to estimate the reliability of RainDrop-BSseq and the 450K array, both relatively recent technologies at the time.

Two loci overlapping with those tested by both 450K and RainDrop-BSseq were tested in two WT and two MUT samples (**Figure 3.7b**). The results obtained with the different platforms agreed in every case, with a mean cross-platform difference of β = 0.09 (min = 0.01, max = 0.19). The validated methylation values ranged from low to high methylation, demonstrating the robustness of the chosen methods.

3.2.5 Replication

3.2.5.1 RainDrop-BSseq

In addition to the validation set described above, I used RainDrop-BSseq to ascertain the methylation scores of selected sites in six previously untested MUT samples (**Figure 3.7a**), and compared those to the scores obtained with the same method and at the same loci, but in different samples: 94.3% (352/373) displayedmatching methylation levels.

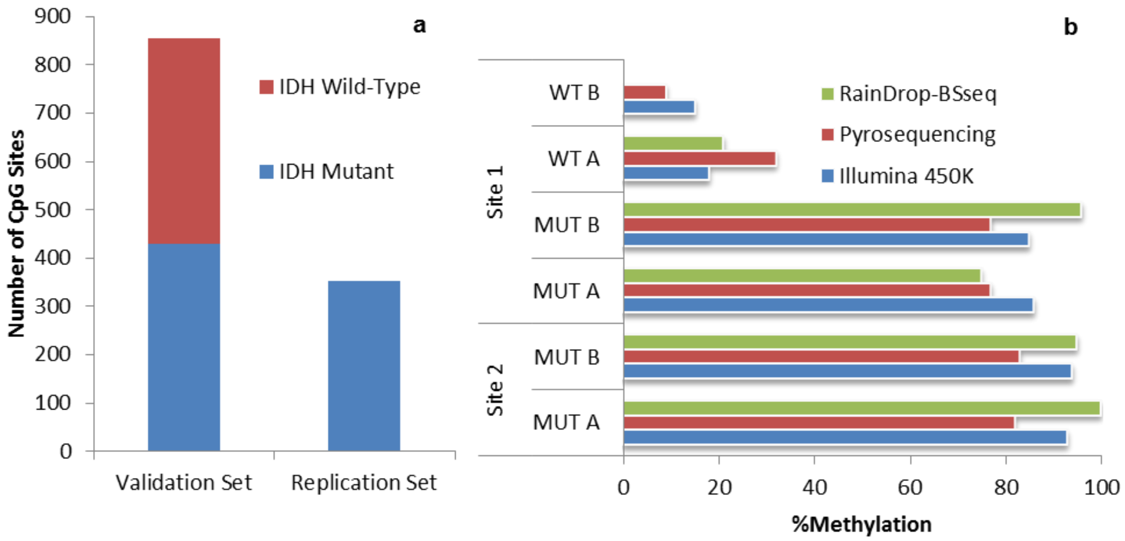


Figure 3.7: Validation and replication with RainDrop BS-seq and pyrosequencing

a) Cumulative bar chart of MVPs validated by RainDrop-BSseq. Of the selected MVPs, 855 across 16 samples (11 IDH MUT + 5 IDH WT) validated, representing 98.8% and 95.5% in the WT and MUT groups, respectively. Bar chart of MVPs validated by RainDrop-BSseq in the replication set (n=6). Of the selected MVPs, 352 (94.3%) matched the values measured in the validation set.
b) Cross-platform validation for MVPs at two different genomic sites. MVPs at the two sites represented a range of methylation levels (low, intermediate, high). All three methods produced similar results, with measurements at each site within 19% (max beta difference) of each other.
Original figure from Guilhamon et al.[1]

3.2.5.2 450K methylation array

I additionally ran a further cohort of 24 independent samples on 450K arrays. These 14 MUT and 10 WT samples were used to validate the hypermethylation profile originally identified as opposed to individual CpG methylation levels. Using the top 500 MVPs from the discovery set, as ordered first by increasing p-value and second by decreasing absolute median difference between sample groups, this replication set was separated into clusters by *IDH* mutations status, with 92% (22/24) of samples being accurately classified (**Figure 3.8**). The MUT cluster in this cohort is clearly hypermethylated relative to the WT samples, thus replicating the previously identified profile.

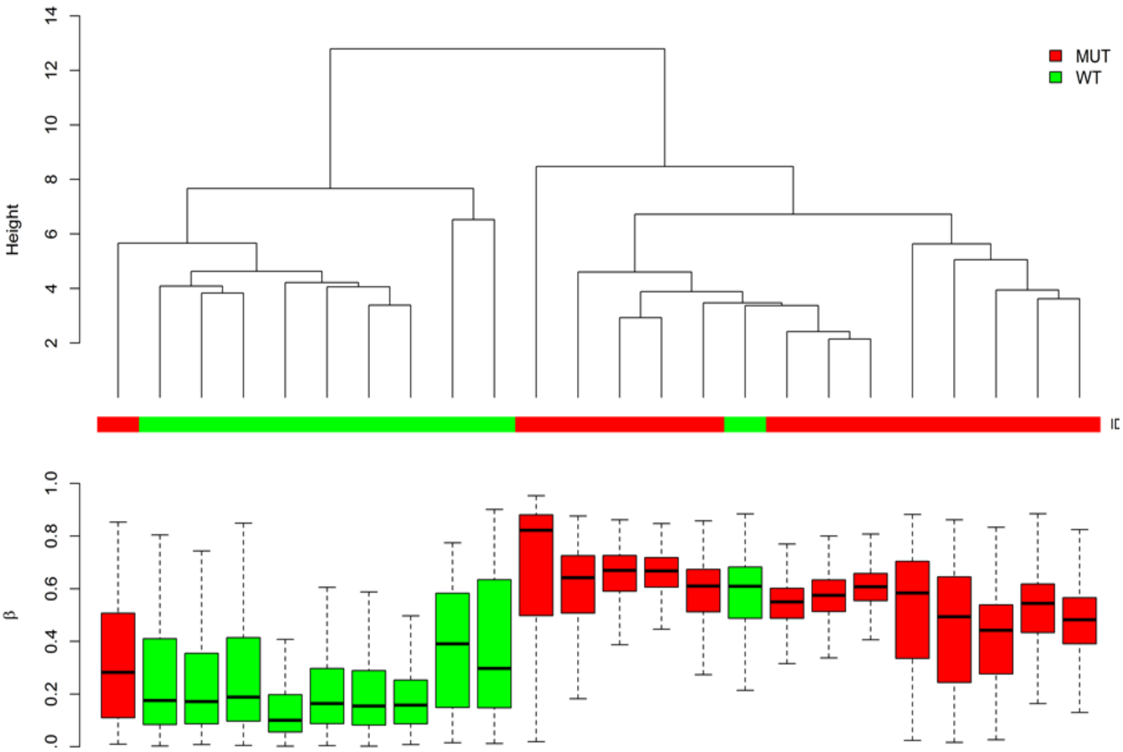


Figure 3.8: Replication of *IDH*-related hypermethylation phenotype in an independent cohort

Supervised hierarchical clustering of the 24 samples in the 450K replication set using the top 500 MVPs from the first cohort (top). The samples separate into two clusters, with 92% (22/24) of samples grouped in the correct cluster. Box plots of β-values from these 500 MVPs (bottom). The MUT sample group is clearly hypermethylated relative to the WT. Original figure from Guilhamon et al.[1]

3.2.6 Integration with gene expression

In addition to methylation profiling, 46 CS samples (30 *IDH* MUT and 16 *IDH* WT) were analysed for gene expression on the Illumina HumanHT-12 v4 Expression BeadChips. A total of 32 of these samples overlap with the 450K sample cohort. Sample preparation and analysis protocols are described in detail in *Chapter 2* and the full analysis script can be found in the *Appendices*.

3.2.6.1 Unsupervised analysis

Quality control of the raw data showed no significant difference in the proportion of expressed probes between the MUT and WT cohorts (Welch Two Sample t-test p-value = 0.12). After background correction and normalisation, I performed multidimensional scaling (MDS) of the entire dataset (47,210 probes, 46 samples) to check for similarity within groups (**Figure 3.9**), and the samples did not cluster by either *IDH* status or processing batch (samples were processed in two groups).

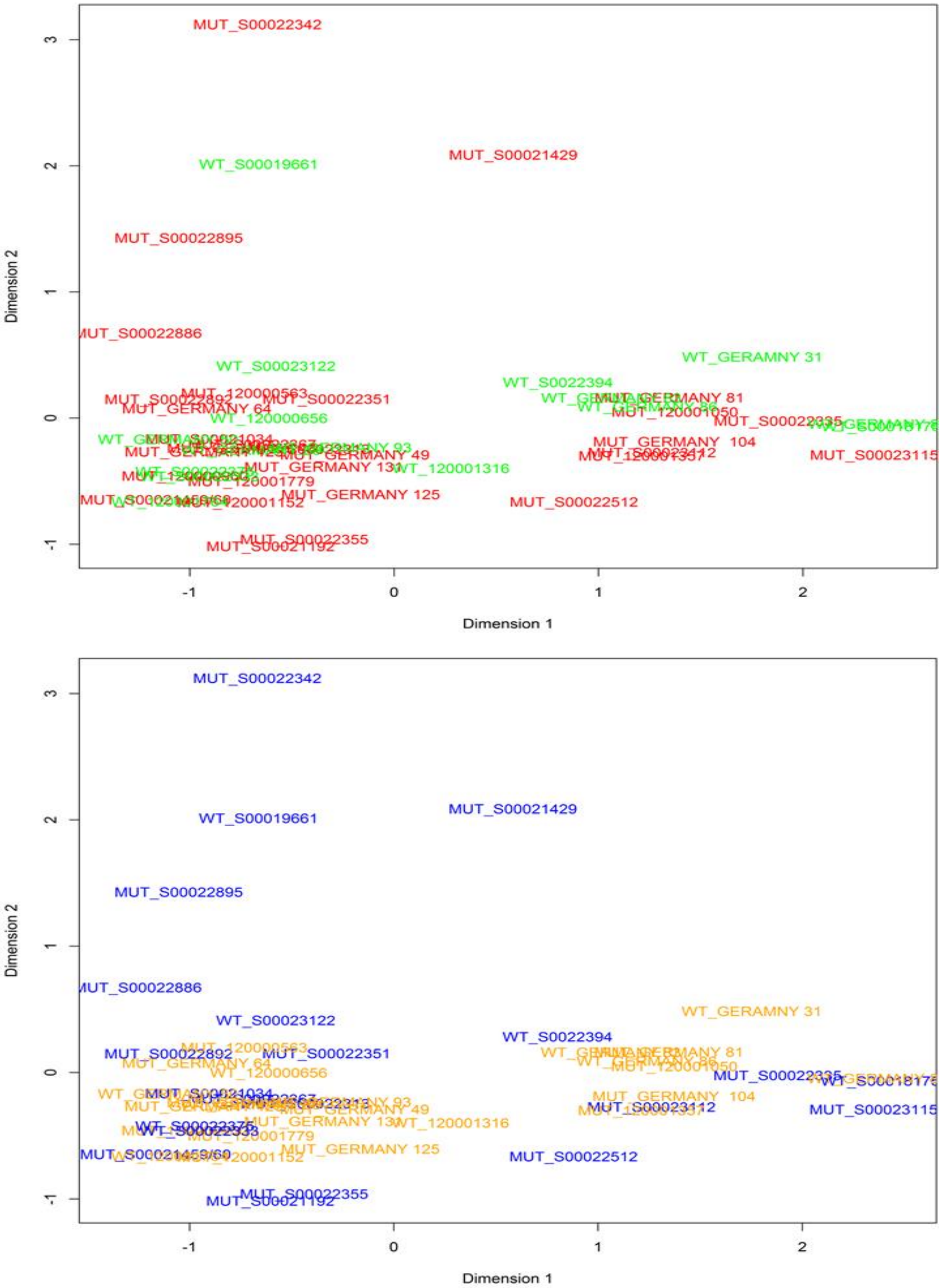


Figure 3.9: MDS of normalised data

Samples do not cluster by either *IDH* mutation status (top panel; MUT = red, WT = green), or processing batch (bottom panel; Batch 1 = blue, Batch 2 = orange).

Annotation quality information for the HumanHT-12 v4 array[83] provides a score for each probe based on its sequence to quantify its reliability ('Perfect', 'Good', 'Bad', 'No match'). Probes with multiple genomic matches, for example, are annotated as 'Bad'. Removing all probes labelled as 'Bad' or 'No match' yielded a dataset of 34,463 probes. Hierarchical clustering of the 500 most variable probes from this dataset (**Figure 3.10**), as determined by calculating the interquartile range, shows no clustering of CS samples by *IDH* mutation status. These results suggest a lack of differential expression between MUT and WT samples.

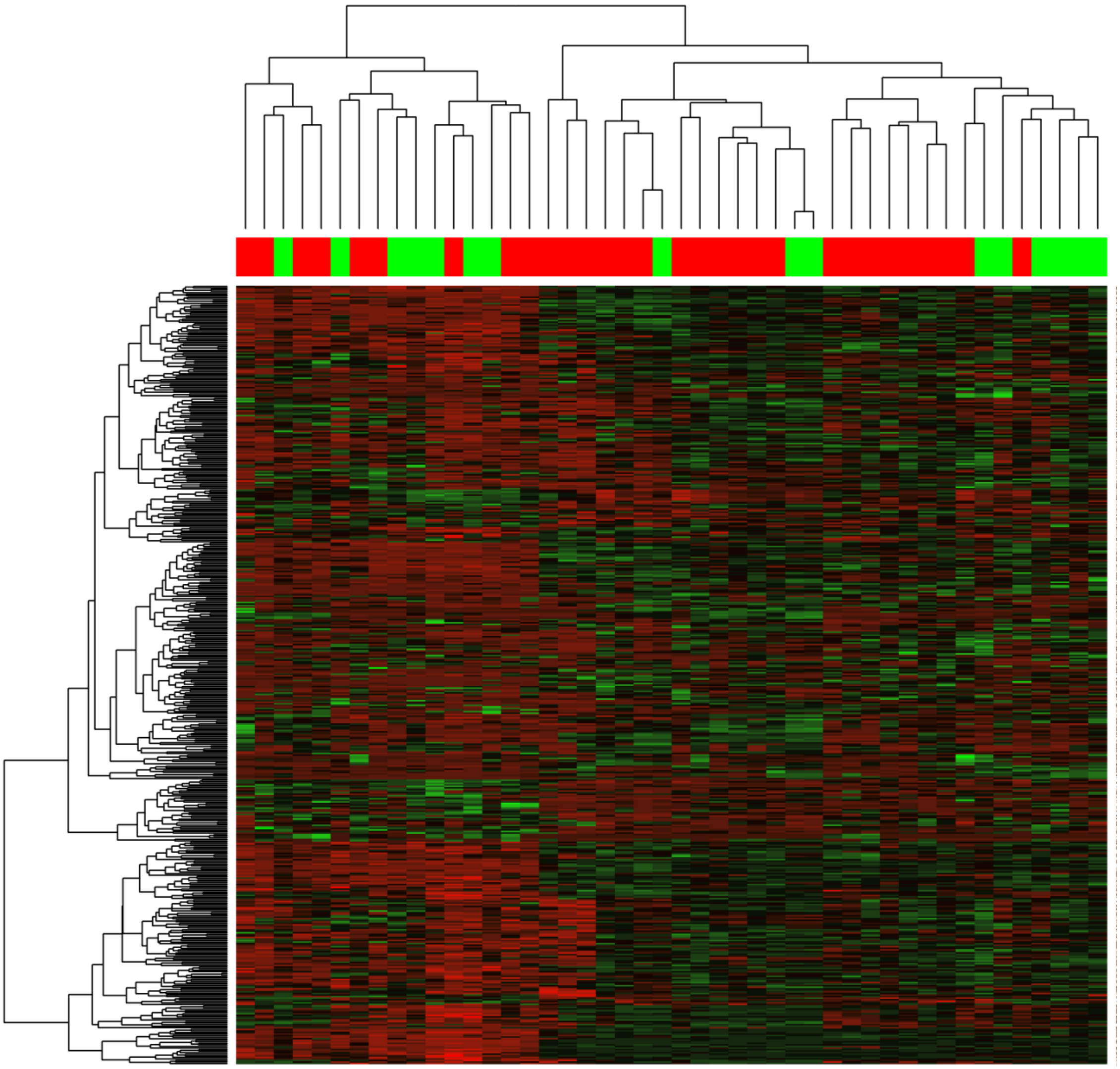


Figure 3.10: Hierarchical clustering of the 500 most variable probes

Clustering of the 46 CS samples according to the 500 most variable probes shows no clustering of *IDH* MUT (red) or WT (green).

3.2.6.2 Supervised analysis

Following on from the unsupervised analysis, investigating differential expression between MUT and WT groups reveals no probe as significantly differentially expressed after p-value adjustment for multiple testing and only two probes with an absolute log fold change (logFC) greater than two. In addition, clustering of the cohort based on the 500 probes with either the lowest p-value or largest absolute logFC reveals no distinct grouping of samples by mutation status (**Figure 3.11**). This further indicates that the hypermethylation phenotype observed in CS does not seem to be correlated with significant changes in gene expression. Possible explanations for this observation will be discussed at the end of this chapter and in *Chapter 6*.

3.2.6.3 Integration of methylation and gene expression profiles

In an attempt to assess whether the observed methylation changes could be linked to differential expression, albeit non-significant, I performed further filtering of both datasets.

The set of 3,057 differentially methylated probes was narrowed down to those annotated to promoter regions (TSS1500 and TSS200); the resulting gene list was further filtered to retain only promoters for which probes were either all hypermethylated or all hypomethylated. This left 525 genes in the methylation dataset, all of which had hypermethylated promoters in the *IDH* MUT cohort relative to WT.

The gene expression set was restricted to contain only genes with either only one probe, or multiple probes that showed the same directionality in logFC: a positive logFC indicates higher expression in the MUT cohort, while a negative logFC shows the reverse. This filtering step yielded an expression dataset of 16,342 genes.

The 525 genes in the methylation dataset and the 16,342 from the gene expression data overlap by 387 unique genes. Of those, only 213 (55%) displayed downregulation of gene expression (logFC < 0) in the MUT samples, while all 387 had hypermethylated promoters. Thus just over half of genes with hypermethylated promoters display any level of transcriptional downregulation, making the functional consequences of 2-HG-induced hypermethylation in CS tumours unclear.

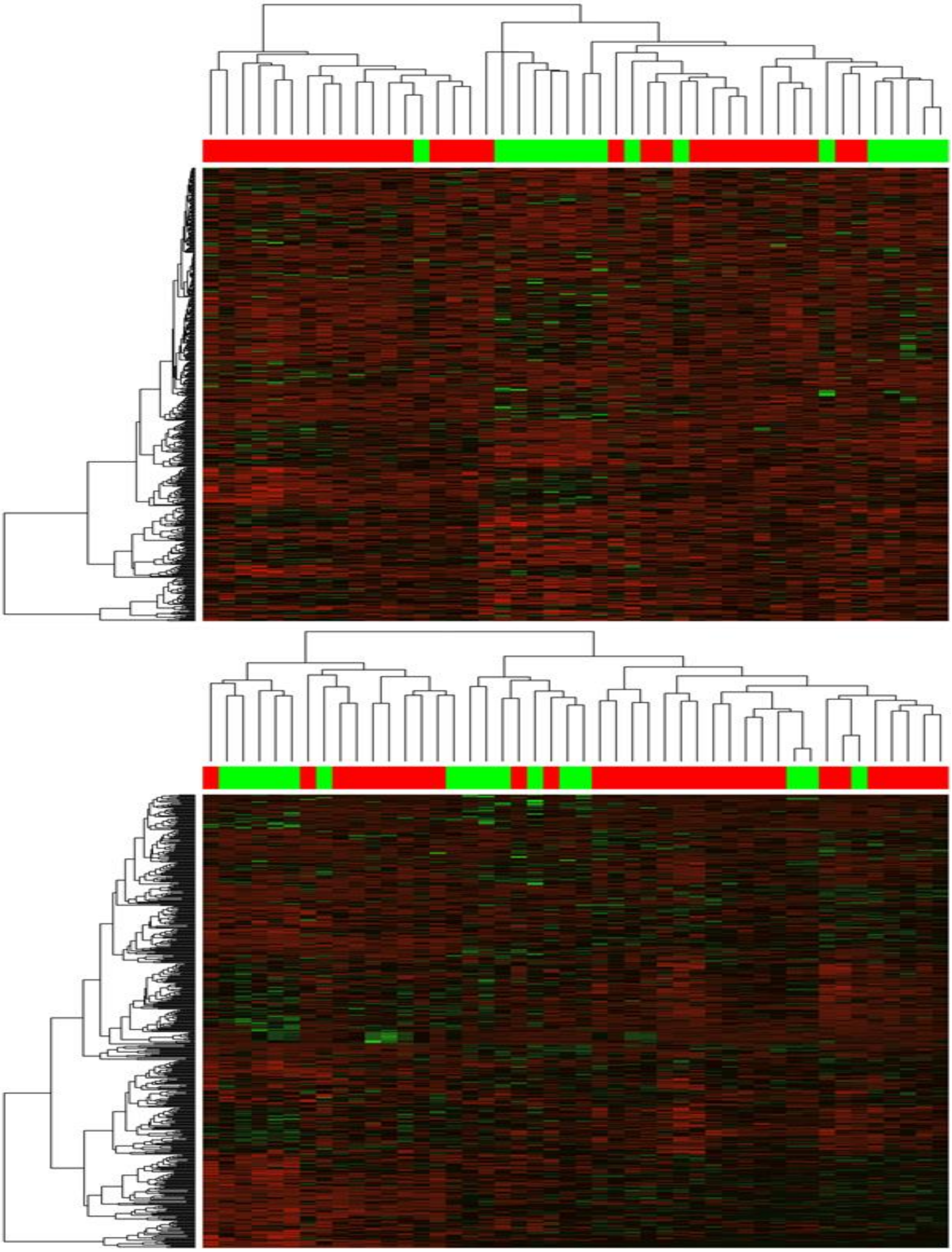


Figure 3.11: Hierarchical clustering of CS samples by p-value and logFC

The top 500 probes as determined by either p-value (top) or logFC (bottom) were used to cluster the 46 CS samples. The samples did not separate into distinct groups based on *IDH* mutation status (MUT = red, WT = green).

3.3 Copy number variation in chondrosarcoma

3.3.1 Using methylation arrays for assessment of copy number

In addition to providing genome-wide coverage of DNA methylation, the 450K array can be used for CNV analysis, as recently described by Feber et al[72]. Some of the results described below were used in the validation of that method. Briefly, the 450K arrays assess the methylation level at a particular locus as the ratio of the signal intensity from the methylated probe to the total intensity, i.e. the sum of methylated and unmethylated signal. It follows that one can use the total intensity measured, after within- and between-array normalisation, as representative of the copy number state. This method has been integrated within the existing ChAMP pipeline[71] and is described in more detail in *Chapter 2*.

3.3.2 Objectives and samples

I proceeded to investigate genomic alterations in chondrosarcoma by combining the test and replication sets of samples from the methylation analysis above in order to potentially identify novel CNVs in CS. Although genomic variations have been reported for CS[61], few of those were found at high-frequency (≥ 20%), and the use of 450K arrays for CNV analysis, with their probe distribution being markedly different from common SNP arrays, may additionally uncover novel sites of alteration.

A total of 51 CS samples were used in this analysis, corresponding to 29 MUT and 22 WT, as well as three pooled blood samples used as a reference. Finally, ten of the tumour samples, six MUT and four WT, were also processed on the Illumina CytoSNP-12 array for validation purposes.

3.3.3 Analysis workflow

ChAMP was run as recommended with a detection p-value filter of 0.01 and a bead cut-off of 0.05. The output of the CNV module of the pipeline was then uploaded onto the GISTIC[73] server to calculate group frequencies of alterations as well as attributing p-values to these alterations, based on background levels of CNVs and probe distribution. In GISTIC, the amplification/deletion threshold was set at 0.3, with a minimum probe number set to four, a confidence level of 0.95 and p-value threshold of 0.05.

3.3.4 Method validation

In order to validate the use of 450K arrays for CNV estimation, ten of the CS samples were also run on a standard SNP array. The raw intensity data was processed similarly to that from the methylation arrays and run through the CNV module of the ChAMP pipeline and the GISTIC algorithm.

As shown in **Figure 3.12**, both methods identified similar alterations overall. However, a noteworthy distinction can be made in these similarities between large (≥ 10 Mb) and focal (≤ 1 Mb) alterations (**Table 3.1**), as detailed below.

3.3.4.1 Large alterations

In investigating large alterations, the vast majority of both gains (82.6%) and losses (75.2%) identified by the 450K array were also found when running the sample on an Illumina SNP array, demonstrating the robustness of this method to identify genomic alterations of 10 Mb or more.

3.3.4.2 Focal alterations

When comparing the list of small genomic changes identified by the two methods, however, the percentages of overlaps are noticeably lower with only 6.7% of gains and 11.1% of losses identified by the 450K and found in the SNP list. As shown in

Table 3.1, this is most likely due to the limited number of probes (< 300,000) and differing distribution on the SNP array. Around half of the focal alterations that were not identified by the SNP array could not have been due to an insufficient number of probes in the region, with the threshold for the minimum number of probes in GISTIC set to four. **Figure 3.13** displays the genomic distribution of probes on the 450K and SNP arrays and illustrates noticeable variation in the probe frequency particularly in gene-poor regions, where the SNP array typically contains fewer probes.

3.3.4.2.1 Comparison with exome sequencing

These ten samples were also profiled for CNV by exome sequencing by Flanagan et al.[61], and, although the gains and losses identified overlapped those found with the 450K by only 13 and 4, respectively, the focal changes were better represented on the SNP chip with 32 of the 78 (41%) alterations also identified by the SNP array. However, the much smaller number of focal alterations identified with exome sequencing means no general conclusions about method agreement can be drawn from this comparison.

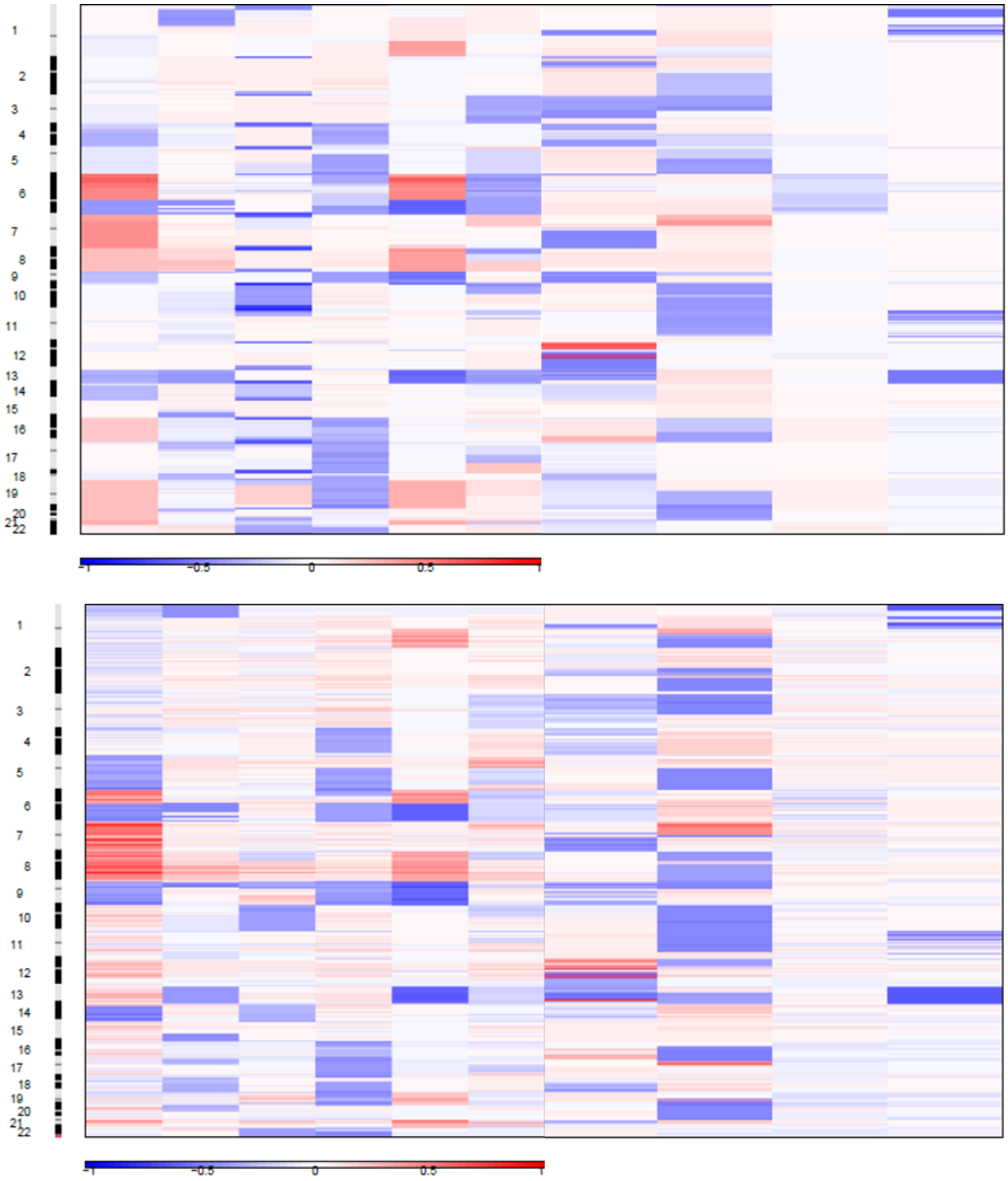


Figure 3.12: CNVs in CS as determined by methylation and SNP arrays

Output from GISTIC for alterations determined by 450K array (top) and Illumina SNP array (bottom). Each column is a sample (1-6: MUT; 7-10: WT), ordered by genomic location.

Large CNVs (≥10 Mb)	#450K Gains	#SNP Gains	#Overlap	%Overlap	#450K Losses	#SNP Losses	#Overlap	%Overlap
MUT 1	14	39	8	57.14	37	40	20	54.05
MUT 2	59	61	52	88.14	23	25	23	100.00
MUT 3	33	57	29	87.88	9	29	7	77.78
MUT 4	54	57	50	92.59	28	25	28	100.00
MUT 5	28	53	27	96.43	39	16	15	38.46
MUT 6	36	50	30	83.33	23	40	21	91.30
WT 1	36	47	33	91.67	26	41	20	76.92
WT 2	40	55	32	80.00	27	26	20	74.07
WT 3	16	45	9	56.25	22	38	13	59.09
WT 4	40	57	37	92.50	5	13	4	80.00
MEAN	35.6	52.1	30.7	**82.59**	23.9	29.3	17.1	**75.17**
Small CNVs (≤1 Mb)	**#450K Gains**	**#SNP Gains**	**#Overlap**	**%Overlap**	**#450K Losses**	**#SNP Losses**	**#Overlap**	**%Overlap**
MUT 1	41	126	1	2.44	32	119	3	9.38
MUT 2	308	19	2	0.65	71	36	5	7.04
MUT 3	54	43	0	0.00	96	43	9	9.38
MUT 4	197	30	5	2.54	74	27	8	10.81
MUT 5	127	15	0	0.00	70	16	4	5.71
MUT 6	86	69	2	2.33	34	37	1	2.94
WT 1	120	50	15	12.50	85	64	12	14.12
WT 2	68	40	1	1.47	39	35	4	10.26
WT 3	7	47	2	28.57	27	83	5	18.52
WT 4	36	28	6	16.67	56	49	13	23.21
MEAN	104.4	46.7	3.4	**6.72**	58.4	50.9	6.4	**11.14**
SNP array markers in Small CNVs (≤1 Mb)	**#450K Gains not in SNP Gains**	**#with < 4 SNP markers**	**%with < 4 SNP markers**		**#450K Losses not in SNP Losses**	**#with < 4 SNP markers**	**%with < 4 SNP markers**	
MUT 1	40	26	65.00		29	16	55.17	
MUT 2	306	199	64.80		66	23	34.85	
MUT 3	54	31	56.60		87	25	28.74	
MUT 4	192	124	63.83		66	26	39.39	
MUT 5	127	78	61.11		66	33	50.00	
MUT 6	84	67	79.76		33	10	30.30	
WT 1	105	64	60.58		73	28	38.36	
WT 2	67	51	76.12		35	20	57.14	
WT 3	5	2	40.00		22	11	50.00	
WT 4	30	18	52.00		43	27	62.79	
MEAN	101	66	**61.98**		52	21.9	**44.67**	

Table 3.1: Comparison of identified CNVs as a function of alteration size

Gains and losses of copy number in samples profiled with both array types (450K and SNP) were overlapped for either large (≥ 10 Mb) or focal (≤ 1 Mb) CNVs.

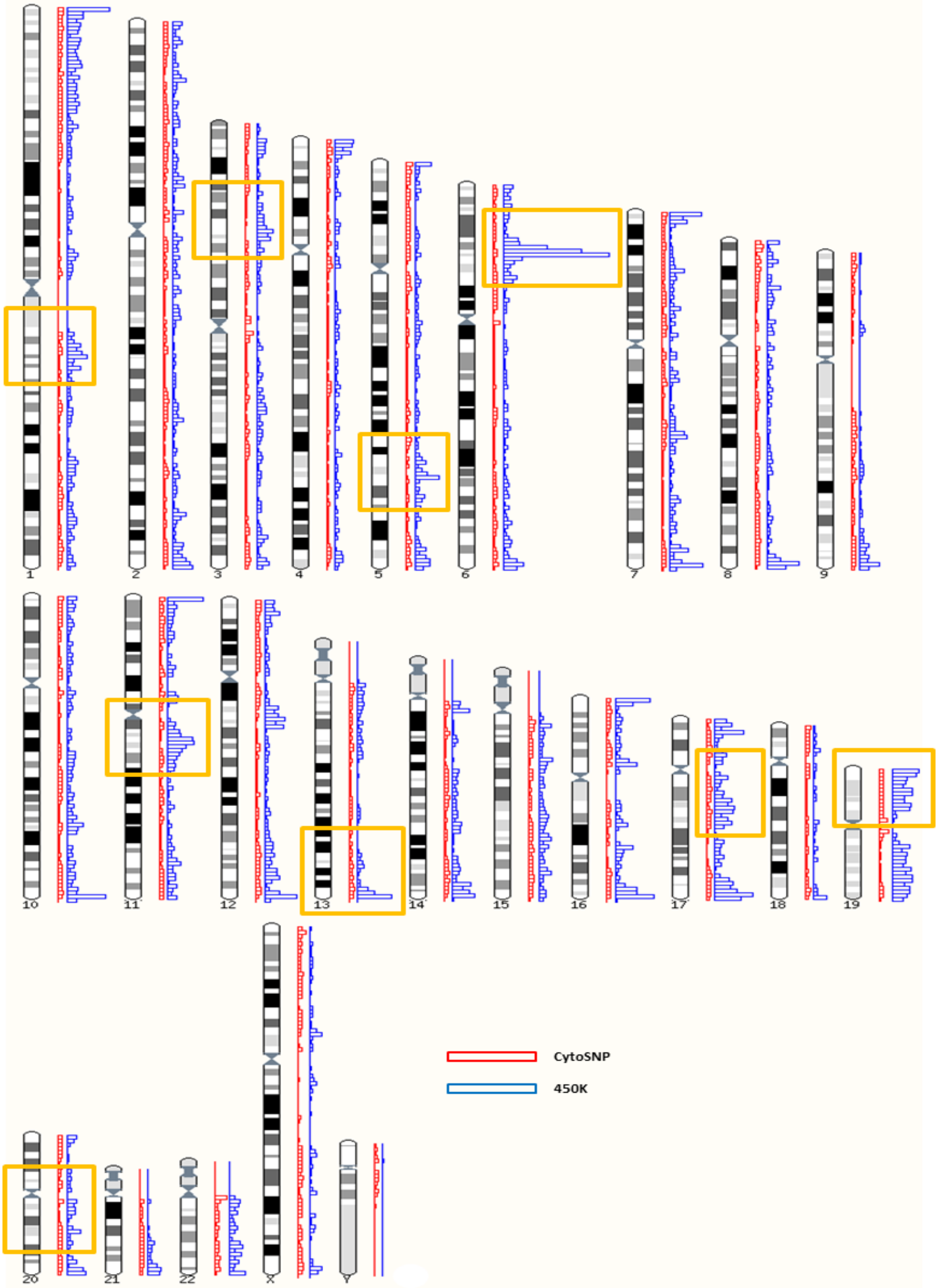


Figure 3.13: Genomic distribution of array probes

Frequency distribution of probes on the CytoSNP (red) and 450K (blue) arrays. Orange boxes indicate example regions over which 450K probes are present at higher frequency and would allow better detection of smaller alterations than the CytoSNP.

3.3.5 Identification of novel alterations in CS

Due to the scarcity of CS, little is known to date about copy number variation associated with the disease, with only ten studies resulting from a literature search on the topic[61,62,84-91]. I combined the genomic locations of CNVs derived in these studies from array-comparative genomic hybridisation (array-CGH), SNP arrays, and exome sequencing into a list of known CS CNVs.

In parallel, after running GISTIC on the 51 CS samples in my cohort, I calculated the frequency of alteration of identified regions, counting only samples with an alteration amplitude ≥ 0.3. This list of regions was then compared to that obtained from the literature to find 23 copy number gains and five losses overlapped. The differences and overlaps between CNVs previously identified and those observed with the 450K are displayed in **Figure 3.14**.

Of particular interest were those recurrent CNVs (frequency > 20%) identified here with the 450K array that had not been previously reported for central CS: a total of 16 gains and three losses. The copy number losses annotate mainly to genes that are commonly structurally variable, such as major histocompatibility complex genes (*HLA-DQA1, HLA-DRB1, HLA-DRB5, HLA-DRB6*) and beta defensins (*DEFB125-129, DEFB132*). However, one noteworthy exception is the osteoblast specific factor, periostin (*POSTN*). This gene has been reported as over-expressed in the stroma of bone metastases of breast cancer[92] and is thought to aid the infiltration of tumour cells in this matrix-rich environment. Its loss in CS could serve the reverse purpose and facilitate the initial detachment of tumour cells for metastasis. Some of the novel gains are also annotated to genes involved in tumour progression and metastasis. For example, hyaluronan synthase 1 (*HAS1*) provides a matrix through which cells can migrate, while matrix metallopeptidase 11

(*MMP11*) is involved in the degradation of extracellular matrix, one of the first steps in metastasis of tumour cells. The tumour necrosis factor receptor TNFRSF10C, also linked to copy number gain, functions as an antagonistic receptor and prevents cytokine-induced apoptosis. Finally, the deubiquitinating enzyme USP17L2 prevents cell cycle arrest through the removal of ubiquitin marks from CDC25A, thus preventing its degradation in response to DNA damage.

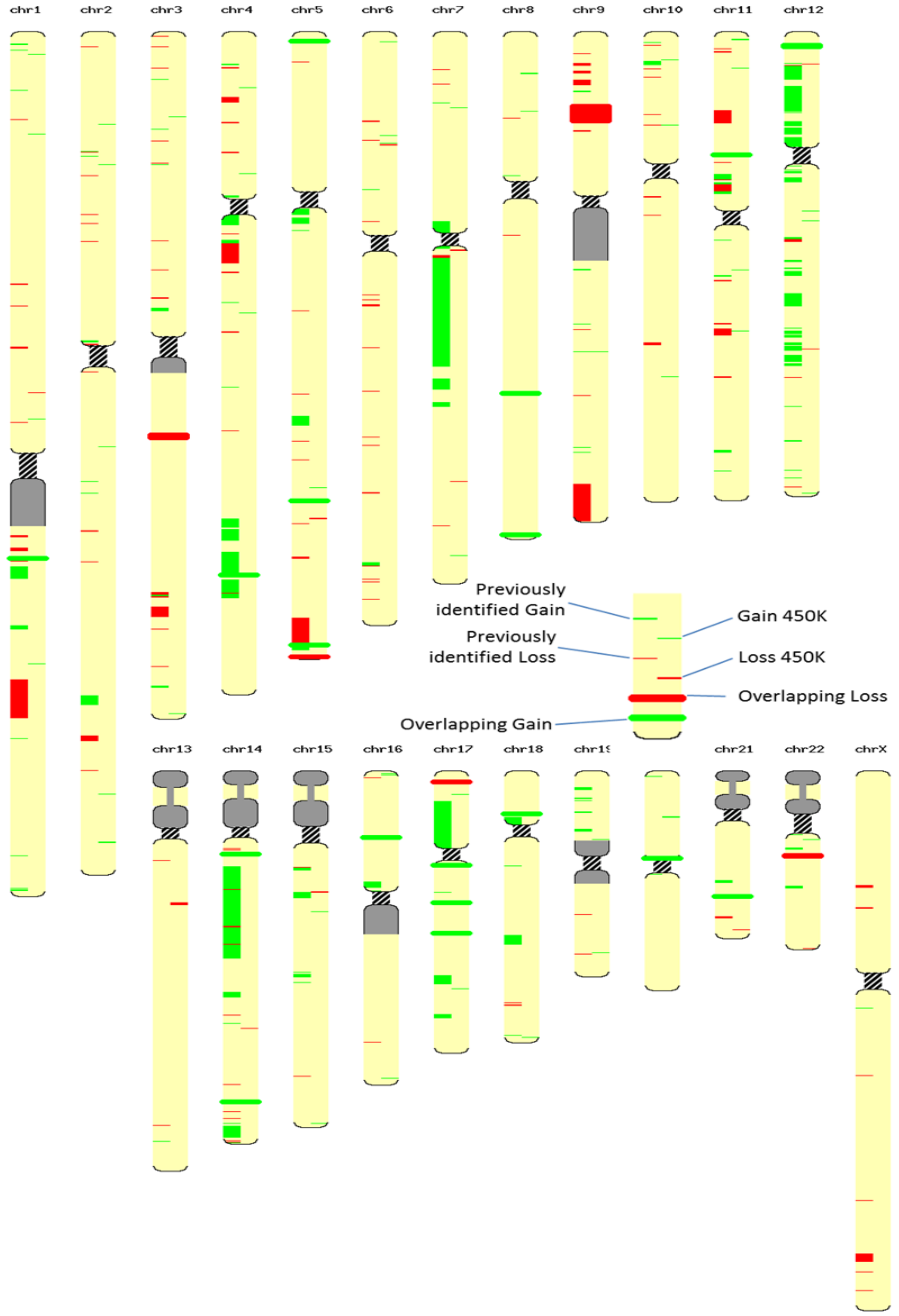


Figure 3.14: Chondrosarcoma CNVs

Display programme written by Roach et al.[93] Copy number gains and losses previously identified are displayed on the left of each chromosome, in green and red, respectively. Those identified with the 450K are on the right. Bars across the entire width of the chromosome represent CNVs present in both sets.

3.4 Conclusions and future work

Mutations in the *IDH* genes are frequent in central CS but the DNA methylation profile of these mutant tumours had not been previously characterised. Using genome-wide methylation arrays, I show here that the hypermethylation phenotype associated with *IDH* gain-of-function mutations in other cancer types[24,29,31] is also present in CS and is enriched for functionally relevant regions such as CpG islands and gene promoters.

This was validated using both the well-established pyrosequencing technique and the novel RainDrop-BSseq based on parallel amplification method followed by next-generation sequencing, and replicated in an independent cohort.

Although the identified methylation signature replicated well across technological platforms and sample cohorts, additional sources of potential variation should be taken into account in future studies. Firstly, while all *IDH1* and *IDH2* mutant samples were grouped together under the umbrella of '*IDH* mutant' in this analysis (albeit with a vast majority of *IDH1* mutants), more subtlety in the observed methylation phenotype might have been obtained if these had been analysed as separate groups, with further subdivision by mutation site: mutations at R140 and R172 in IDH2 have indeed been shown to be associated with markedly different outcomes in AML patients[94]. Secondly, the supervised analysis shown in **Figure 3.4** reveals a gradient in the degree to which *IDH* mutant samples were hypermethylated, with two distinct clusters of mutant samples, one with intermediate to high methylation and the other with high methylation. This could be indicative of patient heterogeneity, with each patient sample showing differing mutational loads with regards to *IDH*. In addition to this difference in mutational load across tumours, tumour heterogeneity within each tumour should also be considered as the fraction of tumour analysed could itself contain a larger number of cells affected by an *IDH* mutation, thus biasing the measured methylation.

Gene expression analysis of the full CS cohort provided intriguing results. Despite the observed widespread DNA hypermethylation, no significantly differentially expressed genes were identified. Moreover, only half of the genes with hypermethylated promoters showed any sign of downregulated expression, suggesting no correlation between those two events, contrary to the established paradigm of transcriptional regulation by DNA methylation. Although some of the other studies investigating *IDH*-linked hypermethylation had reported gene expression changes, these occurred at noticeably fewer genes than expected; in glioma, for example, of the 2,611 genes with significant differential methylation in the promoter region only 429 were both hypermethylated at the promoter and downregulated, while 176 were hypomethylated and upregulated[29]. A possible explanation for these inconsistencies could be that DNA methylation acts as a locking mechanism for prior changes in chromatin[95], and that gene expression changes do not follow but precede the establishment of hypermethylation; with the desired expression pattern in place, the tumour cell would conserve energy by preventing demethylation and thus any dynamic modulation of the epigenetic profile in subsequent cell cycles. It is noteworthy, though, that no difference in survival has been observed between patients with *IDH* mutant CS and those with tumours wild-type for *IDH*[63], suggesting any growth advantage provided by an *IDH* mutation does not translate into more virulent tumour development, at least in chondrosarcoma. However, these mutations occur early in tumourigenesis[63], so alternative mechanisms conferring equivalent growth advantages could have been developed by their *IDH* wild-type counterparts.

Finally, using a recently developed method, I additionally extracted, and validated, CNV information from the methylation arrays and identified a number of novel and recurrent CS alterations in genes involved in tumour progression and metastasis, providing intriguing avenues for future research.

4 META-ANALYSIS OF *IDH*-MUTANT CANCERS

Most of the results presented in this chapter were published[1] and the corresponding abstract and title pages can be found in the *Appendices*. The modelling analysis of the interaction described in this chapter (4.5.3) is based on advice from Ha Phuong Nguyen and Dr. Tracey Barrett.

4.1 Introduction

The *IDH* mutations discussed in the last chapter have previously been studied in the context of other cancer types and are present in ~70% of low-grade gliomas[30] (LGGs), ~10% of acute myeloid leukaemias[30] (AMLs) and cholangiocarcinomas[31] (CCs), as well as at much lower frequencies in other malignancies. In these also, the early[64] mutations in *IDH* genes have been shown to induce a DNA hypermethylation[24,29] phenotype focused particularly on CpG islands. This suggests a common mechanism linking the IDH enzyme's gain of function with a reduction in demethylation, potentially through inhibition of the TET dioxygeneases[96]. Although this inhibition is most likely to be the link between mutations in IDH and DNA hypermethylation in these four malignancies, it should be noted that the same accumulation of 2-HG also affects other dioxygenases, such as histone demethylases and proxyl hydroxylases, which could be contributing to the observed phenotype. With the aim of assessing whether the DNA hypermethylation affects shared pathways and/or tissue-specific processes in each tumour type, I performed a meta-analysis of the CS data described in the previous chapter with publically available datasets from LGG, AML and CC (*Chapter 2*).

4.2 DNA methylation data processing

The AML sample set (n = 398) used in this study was analysed with the HpaII tiny fragment enrichment by ligation-mediated PCR[97] (HELP) assay. It is a restriction enzyme-based method, and, in this particular case, was designed to target CpG sites in gene promoters. The samples from LGG (n = 81), CC (n = 50), as well as the CS data were all processed on the 450K array, which covers not only the gene promoter regions but also gene bodies, untranslated regions and some intergenic loci. This enabled me to analyse both LGG and CC datasets using the same statistical tests and filters as those employed in the CS supervised analysis, as previously detailed (Wilcoxon p-value ≤ 0.001, $|\Delta\beta| \geq 0.35$). At the time, no more suitable alternatives for the AML dataset were available, so in order to facilitate the inclusion of this cancer type in the meta-analysis, I restricted the 450K data to information from probes annotated to CpG islands or shores within promoter regions, thus matching the coverage provided by the HELP assay in terms of functional genomic regions. The comparative analysis was then carried out at the gene level using the pathway analysis software IPA (Ingenuity Systems, *Chapter 2*) and custom R scripts on genes found differentially methylated in at least one of CS, LGG, CC and AML (640, 1,028, 169, and 48 genes, respectively).

4.3 Pathway analysis

A preliminary analysis of overlapping genes in the four sets revealed that CS and LGG had 188 genes in common (random resampling p-value ≤ 10^{-5}), CS, LGG, and CC overlapped by 16 (p-value ≤ 10^{-6}), but no gene appeared in all four cancer types (**Figure 4.1**).

Although no genes were found to be commonly affected in all cancers, the same cellular processes could be targeted through differential methylation of different genes within the same pathway. Using IPA, I thus analysed the four gene sets for shared pathways: molecules from the dataset that were linked to a pathway as annotated in the IPA database were considered for the analysis and given a p-value corresponding to the probability that the association between the genes in this list and each pathway is due to chance alone, based on Fisher's exact test with a further adjustment for multiple testing (Benjamini-Hochberg[70]). Although none of the annotated pathways reached statistical significance, I identified one process that was heavily targeted in all four cancer types: retinoic acid receptor (RAR) activation (**Figure 4.2**).

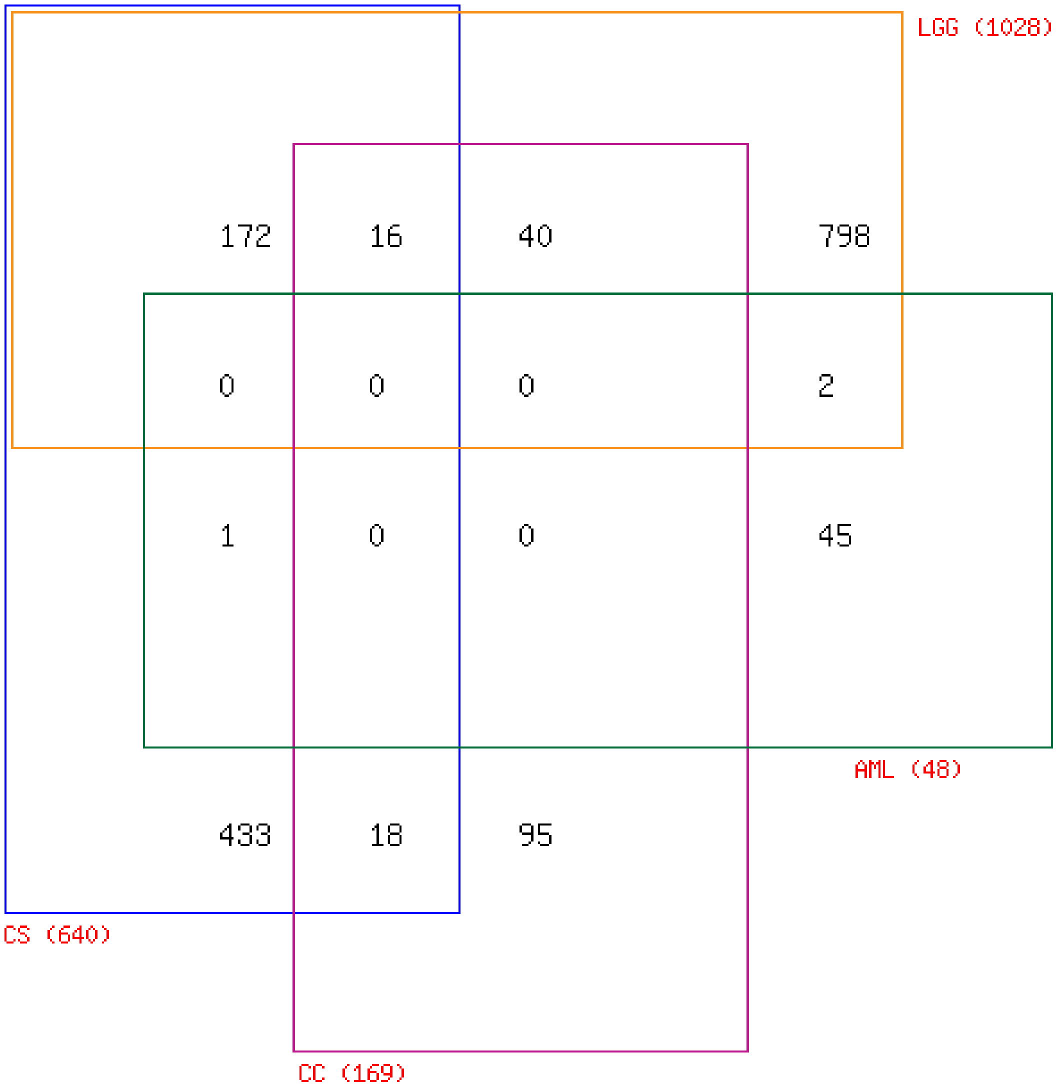


Figure 4.1: Overlap of significantly differentially methylated genes from four cancer types

Individual pairs of cancer types have small overlaps in the affected genes but no single gene was differentially methylated between MUT and WT in all four tumour types.

The RAR activation pathway has been studied in the context of many malignancies: retinoids perform a variety of crucial functions involved for instance in vision, cell proliferation and differentiation, neural and immune functions and their signalling is often compromised early in carcinogenesis[98]; moreover, retinoids are known as potent growth inhibitors in various cancers, including skin, bladder, kidney, prostate, and breast[99]. As shown in the pathway diagram, five genes from CC, two from AML, 17 from LGG and 14 genes from CS were differentially methylated in the RAR activation pathway. For example, retinol-binding protein 1 (RBP1), the carrier protein responsible for the transport of retinol, is present in the gene sets from CS, LGG and CC. This gene has been shown recently to become hypermethylated following the knock-in of a mutant *IDH1* gene into a cancer cell line[100]. In the CS cohort, it displays significant hypermethylation in the promoter region and its gene expression is downregulated, as displayed in **Figure 4.3**.

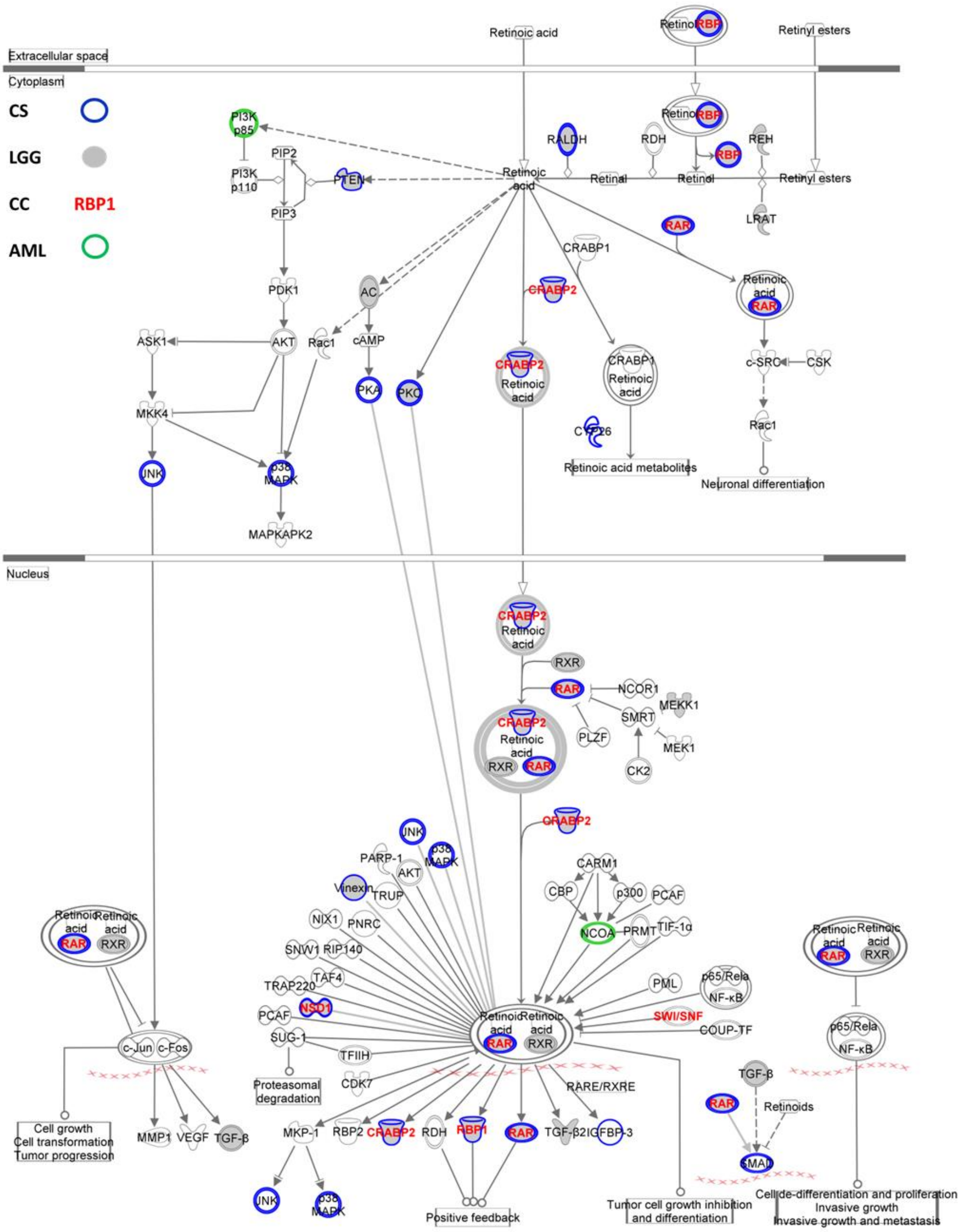


Figure 4.2: The RAR activation pathway in *IDH*-mutant cancers

Schematic diagram of the RAR activation pathway produced with IPA. Molecules are represented as nodes, and the biological relationship between two nodes is represented as an edge (line). All edges are supported by at least one reference from the literature, from a textbook, or from canonical information stored in the Ingenuity Knowledge Base. Nodes are displayed using various shapes that represent the functional class of the gene product. Edges are displayed with various labels that describe the nature of the relationship between the nodes (for example, P for phosphorylation, T for transcription).Original figure from Guilhamon et al.[1]

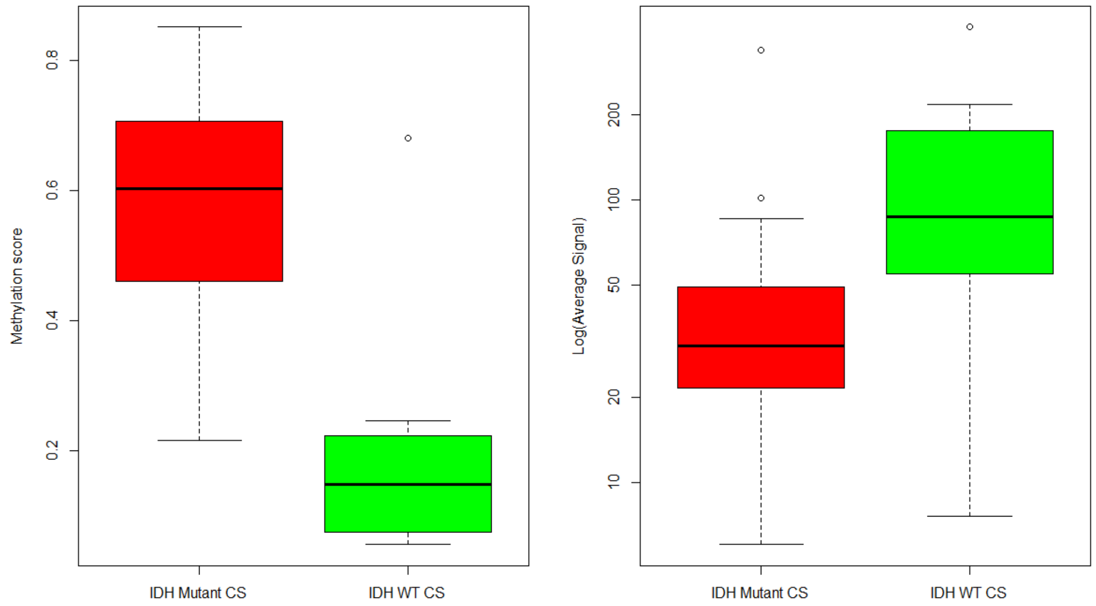


Figure 4.3: RBP1 methylation and gene expression in CS

The RBP1 promoter is significantly hypermethylated (p-value = 6.9 x 10^{-6}) in CS harbouring an *IDH* mutation compared to wild-type samples (left panel). RBP1 gene expression is also significantly downregulated (p-value = 0.018) in mutant samples. Original figure from Guilhamon et al.[1]

Aside from this general cancer pathway, however, the other pathways identified by IPA as associated with the input gene sets were more directly related to each individual tissue of origin than to general tumour development: in CS, for example, affected processes included function of osteoblasts, osteoclasts and chondrocytes; in LGG, axonal guidance; Myc signalling in AML, and in CC, circadian rhythm. These initial results suggested that the observed hypermethylation phenotype might be following a tissue-specific pattern.

In order to further test this hypothesis, I then exclusively analysed genes found differentially methylated in only one cancer type for affected functions and correlations to disease. Using the CS-only genes, the most significant functional category was tissue development (right-tailed Fisher Exact Test p-value = 5.44 x 10^{-5} - 4.76 x 10^{-2}; number of genes n = 46) and it included development of connective tissue in its top functions. In LGG, neurological disease and psychological disorders were the diseases most significantly associated with the gene set (p-value = 2.04 x 10^{-5} - 4.38 x 10^{-2}), while nervous system development and function appeared as the most significant category (p-value = 9.43 x 10^{-4} - 4.38 x 10^{-2}, n = 12), with proliferation of neuronal cells and extension of neurites and axons as its top functions. Similarly in the epithelial disease CC, hair and skin development, including proliferation of epithelial cells, was the most significant functional category (p-value = 2.85 x 10^{-3} - 2.63 x 10^{-2}), as was haematological disease in AML (p-value = 2.34 x 10^{-3} - 4.67 x 10^{-3}).

4.4 Motifs for a TET DNA-binding partner

In the four cancer types I analysed, *IDH*-mutant tumours display widespread DNA hypermethylation, increases in 2-HG, accompanied by disruption of TET function, at least in LGG and AML. The loci affected by the hypermethylation, however, appear to be tissue-specific, as shown by the gene and pathway analysis above. These seemingly contrary observations of a common mechanism but different target sites point to a mechanism of demethylation in which the function of the TET proteins is regulated in a tissue-specific manner. This hypothesis is in fact aligned with the concept of DNA methylation as a tissue-specific signature: if the methylation pattern in healthy cells varies from tissue to tissue, it is not aberrant that the enzymes responsible for its regulation might themselves be controlled in a tissue-specific fashion. I therefore hypothesised that this regulation could be achieved through interaction with a DNA-binding partner that could itself be under transcriptional or post-transcriptional regulation.

To investigate this possibility, I analysed the 100 bp sequences surrounding *IDH*-related MVPs in CS, LGG and CC for common binding motifs using the multiple expectation maximisation for motif elicitation (MEME)[79] suite of tools. The motif 5'-CDGGRA-3' was significantly enriched in the input sequences (MEME p-value = 10^{-3}, discriminative regular expression motif elicitation (DREME) p-value = 7×10^{-70}). I then proceeded to scan 1 kb windows around the MVPs for the presence of this motif, as DNA methylation is generally considered to be tightly correlated over such a distance[101,102], and with no prior knowledge of the mechanics of the interaction between TET proteins and the sought DNA-binding partner, any likely distance of interaction needed to be assessed. The identified motif was present in 93% of sequences tested (8,008/8,582).

With the aid of another tool from the MEME suite, TOMTOM, I compared the motif to those used by known DNA-binding factors for similarity, and identified the early B-cell factor 1 (EBF1) binding motif (**Figure 4.4**) as a significant match (p-value = 0.0025).

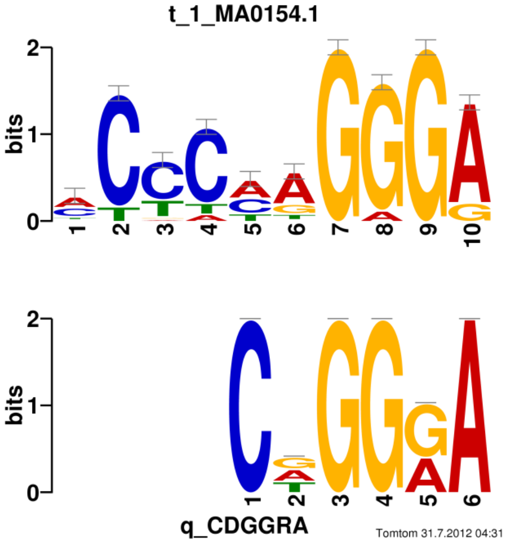


Figure 4.4: EBF1 binding site motif is enriched around MVPs

Motif logo matching between CDGGRA (bottom) and the consensus EBF1 motif (top) as determined by TOMTOM. The offset of the sequence relative to the known motif was used in conjunction with the nucleotide frequencies in each motif to determine the significance of the match. Original figure from Guilhamon et al.[1]

4.5 EBF1: a likely candidate?

Using motif discovery tools facilitated the selection of a possible candidate for the role of DNA-binding partner of the TET enzymes, EBF1. This role is supported by previous findings indicating a function for EBF1 in transcriptional regulation: it has been linked to the induction of *CD79a* promoter demethylation during B-cell differentiation[103], and binding of EBF1 has also been correlated with histone modifications associated with poised chromatin and transcriptional activation[104]. However, before proceeding to experimentally test the interaction of these proteins, I wanted to use available data to ascertain whether EBF1 fulfilled essential requirements to perform that function: 1) is it expressed in the CS samples under study, irrespective of *IDH* mutation status? 2) Is it enriched for (epi)genomic locations where it would be needed to perform its proposed function? 3) Is EBF1 structurally capable of interacting with the TET enzymes in a functional way?

4.5.1 *EBF1* gene expression in CS

The expression of *EBF1* in CS was assessed on 19 MUT and 13 WT samples (**Figure 4.5**) using gene expression microarrays and found to be similar in the two groups, with no significant difference observed (Wilcoxon p-value=0.34) .

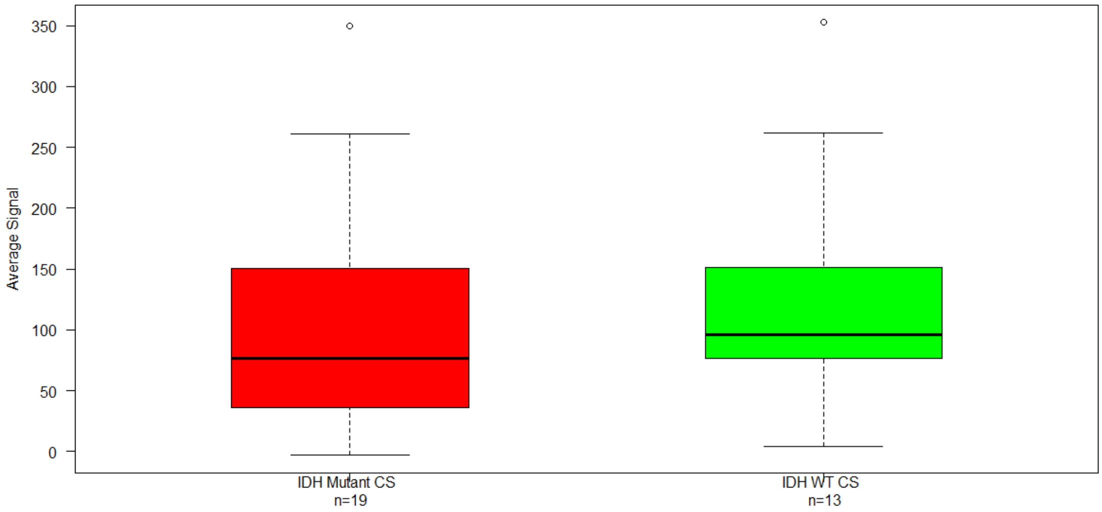


Figure 4.5: EBF1 gene expression in *IDH*-mutant and wild-type CS

Gene expression microarray signal in the mutant and wild-type groups showing no statistically significant difference (p=value = 0.34) in *EBF1* expression. Original figure from Guilhamon et al.[1]

4.5.2 EBF1 binding sites in the genome

The data used here corresponds to chromatin immunoprecipitation experiments followed by next-generation sequencing (ChIP-seq), and was produced by two working groups of the ENCODE consortium and downloaded from UCSC genome browser in the form of ChIP-seq peak coordinates[105]; the data was generated in GM12878 cells, a lymphoblastoid cell.

Each of the ENCODE working groups produced an independent dataset from which peaks were called. Although both identified a similar number of binding sites/peaks (36,140 and 33,410 genome-wide), the overlap between them was poor (56%). In order to only analyse regions with a high likelihood of being true positives, I created a consensus peak list, keeping only those peaks in each dataset that overlapped with a peak from the other. This process yielded a final dataset of 21,113 peaks.

In order to assess enrichment of those peaks for specific genomic and epigenomic features, coordinates for these features were taken from the Ensembl database[106]; shores were defined as 2,000 bp upstream and downstream of each CpG Island, and shelves as the next 2,000 bp from the shore, so that each CpG Island has two shores and two shelves. This replicates the feature definition used on the 450K array[8]. The effective (i.e. mappable) hg19 genome was used as background (2,451,960,000bp).

Enrichment percentages were calculated using the following formula:

$(x/y) - (T/N)$, where:

x = Total bp in peaks in FeatureX

y= Total bp in EBF1 peaks

T= Total bp in background in FeatureX

N= Total bp in background

The enrichments and depletions shown in the waterfall plots below (**Figure 4.6**) were all calculated as being extremely significant by a hypergeometric test (p-value $\leq 10^{-8}$). A noteworthy caveat to this analysis is that the ideal background to calculate significance would have been the 'Input' data for the corresponding experiment. However, firstly the input tracks were not available for these experiments; secondly, it would have been impossible to determine an accurate input consensus sequence for the two datasets; thirdly, the effective genome can be considered an acceptable background in this case, considering ChIP-seq is a genome-wide method.

This analysis clearly shows that EBF1 binding sites are significantly enriched in regions of high CpG density (CpG islands and shores by 6.1% and 6.6%, respectively) and in promoter regions (12.9%), supporting EBF1 as a potential regulator of demethylation.

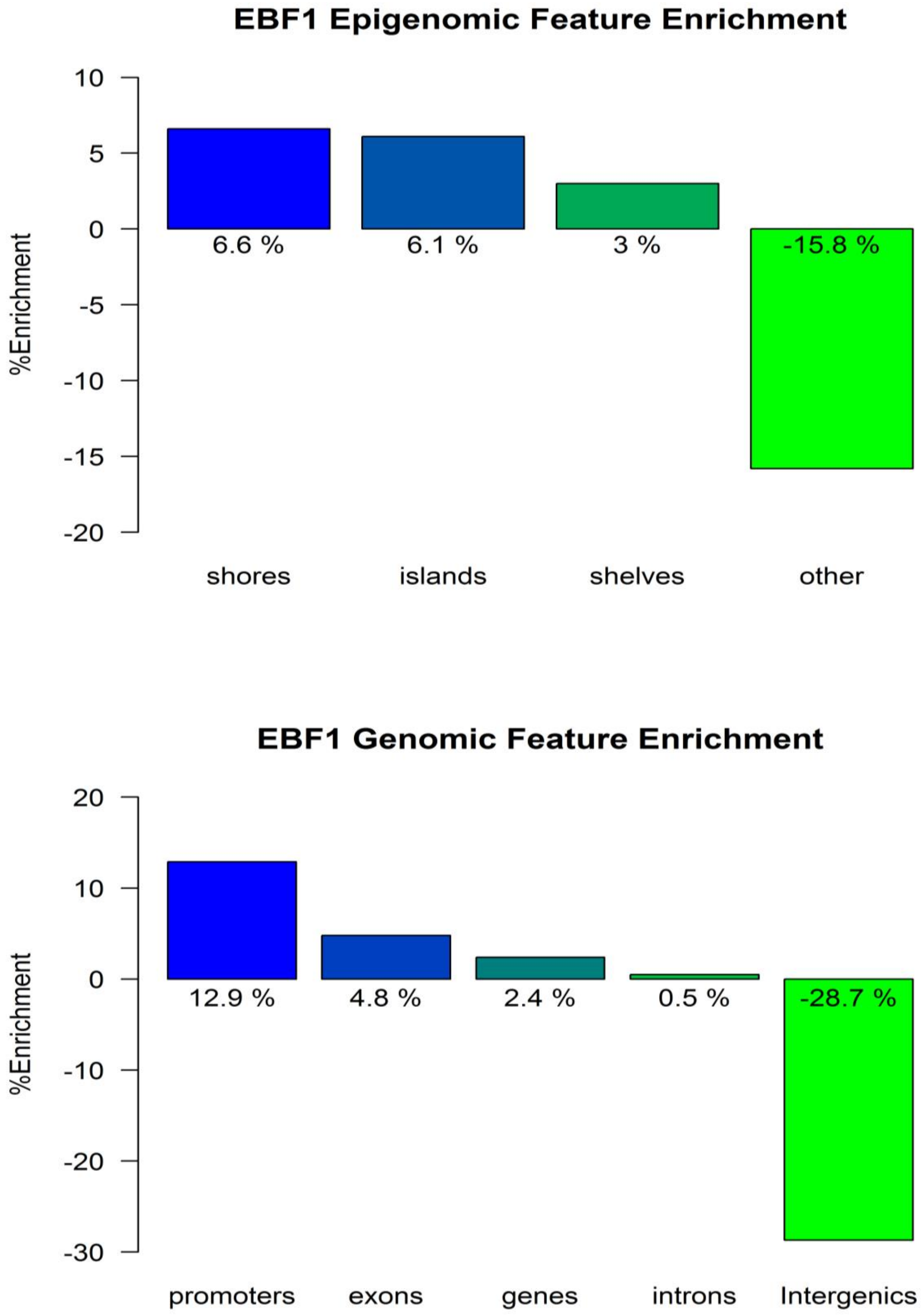


Figure 4.6: Enrichment of EBF1 binding sites in genomic (bottom) and epigenomic (top) features.

EBF1 binding sites are significantly overrepresented in CpG shores and islands and gene promoter regions, and depleted in intergenic regions.

4.5.3 EBF1 and TET2: structure and interaction modelling

The crystal structures of both EBF1 and TET2 with DNA have been solved[19,107,108], enabling modelling analysis of their interaction.

TET2 contains a C-terminal catalytic domain with a double stranded beta helix core (DSBH), which itself carries a cysteine-rich domain at its N-terminus. EBF1 functions as a dimer, and each of the monomers is composed of the following regions: a DNA binding domain (DBD) at the N-terminus forming a symmetric clamp over the binding site; an Immunoglobulin Plexins Transcription factors-like/Transcription factor Immunoglobulin (IPT/TIG) domain thought to be involved in protein-protein interaction; a Helix Loop Helix (HLH) domain responsible for dimerization.

In silico methods can be applied to predict ways in which EBF1 and TET2 might interact. The ZDOCK[109] server, for instance, was designed for protein docking modelling and using as input the Protein Data Bank (PDB) representations of the solved structures for two proteins outputs likely interaction models using an energy-based scoring function. This scoring function is based on 1) knowledge-based potential, or the likelihood of the assessed model being real based on known protein structures and interactions in the PDB, 2) shape complementarity, and 3) electrostatics.

Only one representation of TET2 exists in the PDB at the time of writing, that of human TET2 bound to DNA (PDB ID: 4NM6), but the structure of EBF1 has been solved in both human and mouse; in the case of the former, domains were resolved individually, whereas the mouse Ebf1 structure includes all domains of interest: DBD, IPT/TIG, and HLH. The human and mouse proteins share 100% sequence homology, so I used the available mouse structure (PDB ID: 3MLP) to test the interaction with TET2, as it provided the most complete representation of functional EBF1 domains.

The ZDOCK server returned the interaction model pictured below (**Figure 4.7**). It displays the Ebf1 DBD bound to the DNA and connected to the IPT/TIG and HLH domains, which are predicted and shown as interacting with the TET2 DSBH domain. Collaborators in the Barrett laboratory are pursuing this line of inquiry but current models certainly support a possible interaction between TET2 and EBF1.

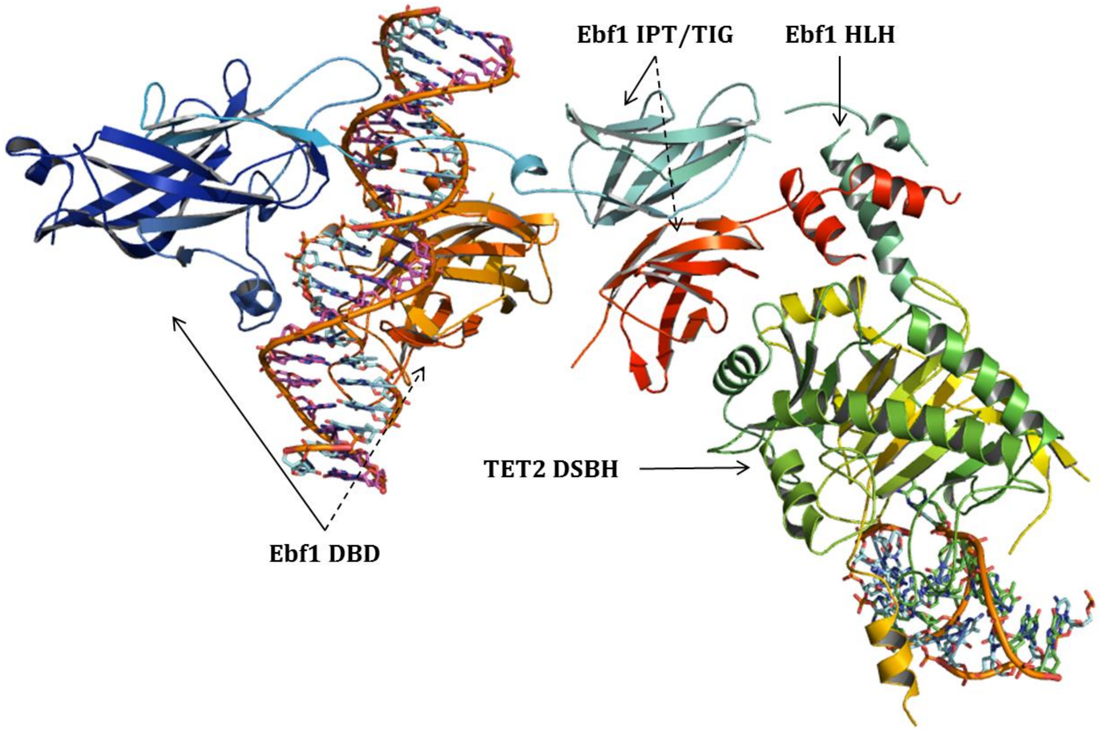


Figure 4.7: Interaction prediction for Ebf1 and TET2

Proteins Ebf1 (left) and TET2 (right) interacting via contact of the TET2 DSBH domain and Ebf1 IPT/TIG and HLH domains. Model created using the ZDOCK server and first described by Dr. Tracey Barrett et al.

4.6 EBF1, a novel interaction partner for TET2

In order to validate the predicted co-localisation and interaction between EBF1 and TET2, I performed two sets of experiments in the SW1353 CS cell line: ChIP-qPCR and co-immunoprecipiation (co-IP) followed by western blotting.

4.6.1 EBF1 and TET2 ChIP

ChIP-qPCR was performed on three loci, using antibodies against TET2 and EBF1. The target sites, located in the vicinity of *CCND2*, *FABP3*, and *FBRSL1*, were selected based on the following criteria (**Table 4.1**): they were hypermethylated in the MUT samples of the CS cohort relative to WT; they showed elevated methylation levels in the SW1353 cell line (β > 0.9); there was at least one predicted binding site for EBF1 within 50 bp, either upstream or downstream.

Targets	MUT CS β-value	WT CS β-value	p-value (BH)	SW1353 β-value
CCND2	0.6238	0.0185	1.92 x 10-4	0.9764
FABP3	0.6331	0.0227	3.23 x 10-4	0.9882
FBRSL1	0.6555	0.0276	1.49 x 10-4	0.984

Table 4.1: Characteristics of selected ChIP targets

In addition, I selected a negative control locus in an intergenic region to normalise the measure enrichments, based on its high methylation levels in CS samples, irrespective of the mutation status and the absence of any predicted EBF1 binding sites in the surrounding sequence.

The enrichments were calculated by the ΔΔCt method, normalising to both the mock IgG IP control and the negative control region where no binding of either protein was expected. As shown in the bar chart below (**Figure 4.8**), all three target sites were enriched for both TET2 and EBF1 binding, with fold enrichments ranging from 4 to 104. Moreover, the TET2:EBF1 ratio of these fold enrichments were similar for the three loci at 9.8, 6.2 and 11.6, further supporting the genomic co-localisation of the two proteins.

A ChIP experiment, however, is performed on a population of cells; thus, although these results support co-localisation of EBF1 and TET2, they do not necessarily suggest temporal co-localisation: in the cell population, cells at different stages in their cycle could have either TET2 or EBF1 bound at these sites without them being present there at the same time. Moreover, this type of experiment functions through immunoprecipitation of a DNA fragment of several hundred base pairs, making it possible for both proteins to be bound within the tested region but not interacting in a manner that would be functionally relevant.

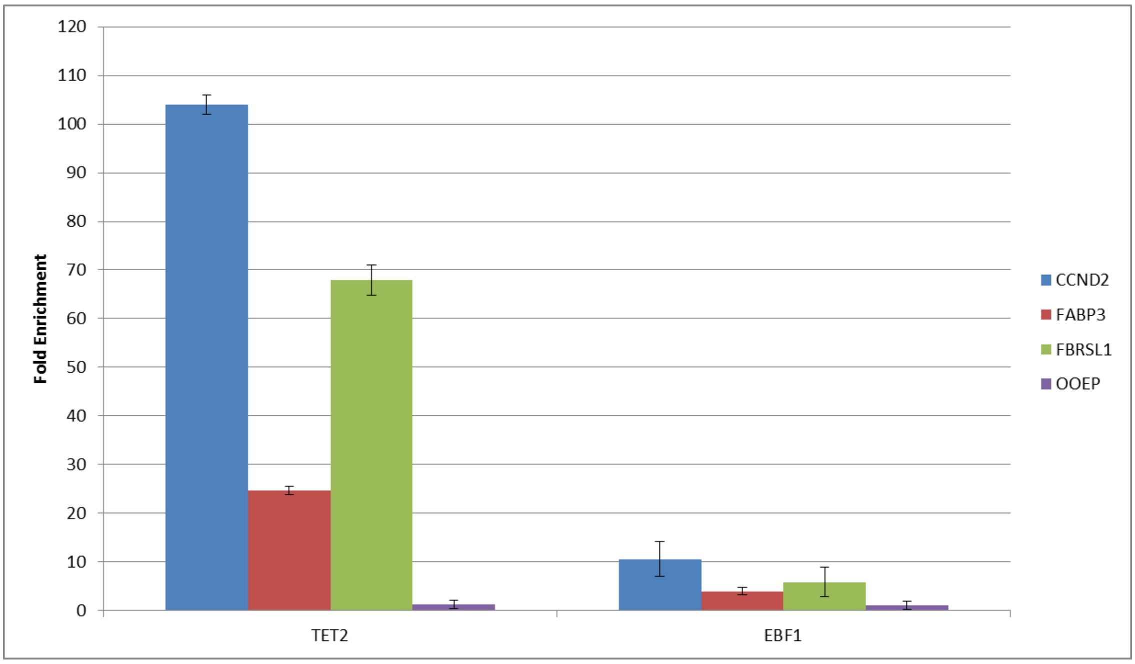


Figure 4.8: ChIP-qPCR for TET2 and EBF1 at three target loci

ChIP-qPCR analysis indicates proportional presence of both TET2 and EBF1 at three loci. Fold enrichments were calculated by normalising to the mock IgG IP control. *OOEP* corresponds to the negative control region. Error bars are based on standard errors of the mean.

4.6.2 Co-IP and western blot of TET2 and EBF1

To alleviate these concerns and demonstrate interaction of TET2 and EBF1, I carried out a co-IP experiment using an antibody targeting TET2 (**Figure 4.9**). After precipitation of TET2 and its interacting proteins, I performed a western blot on the isolated fraction with an antibody against EBF1, and detected a strong EBF1 band, demonstrating the endogenous interaction of TET2 and EBF1 in SW1353 cells.

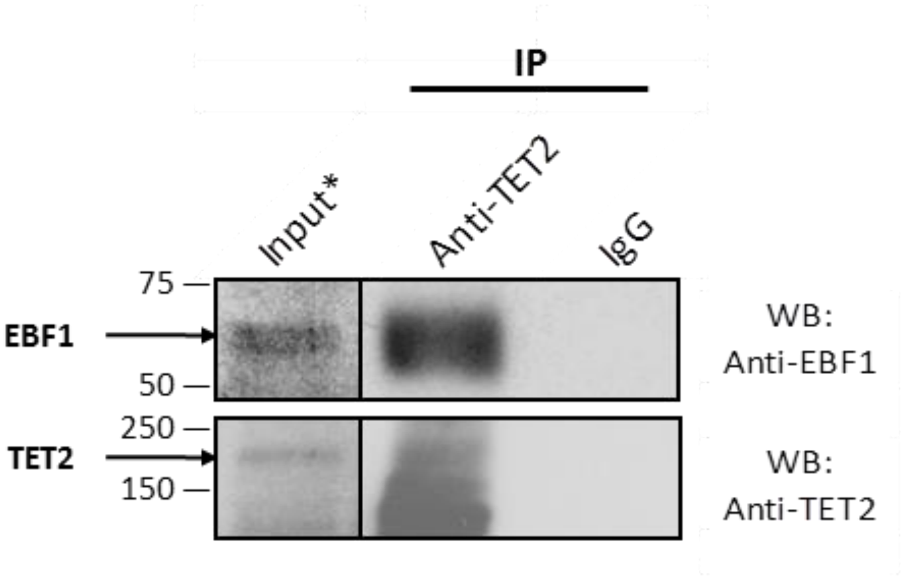


Figure 4.9: Co-IP followed by western blotting supports EBF1 as a TET2 interactor

Interaction of endogenous EBF1 and TET2 in SW1353 cells. The asterisk (*) indicates a longer exposure for the input lysates (5% of total).

4.7 Conclusions and future work

Observing a shared hypermethylation phenotype in CS, AML, LGG and CC tumours with *IDH* mutations suggested that common biological processes might be affected in these malignancies, and the meta-analysis presented here did indeed identify the RAR activation pathway as independently targeted in all four cancer types, with one of its main components, RBP1, being hypermethylated and downregulated in mutant CS. Further gene-level analysis, however, revealed that most of the differentially methylated sites in each neoplasm were tissue-specific, and those only affected in one cancer type are involved in pathways particular to their tissue of origin.

Based on their structure, it is likely that the TET family of dioxygenases, although able to distinguish methylated from unmethylated CpGs with the CXXC domain in the case of TET1 and TET3, would require a binding partner to select loci for demethylation in a tissue-specific manner. Searching for common binding motifs around sites hypermethylated in *IDH* mutant tumours unveiled EBF1 as a likely candidate for this role: its previously identified links to promoter demethylation, constant expression levels in CS tumours irrespective of *IDH* status, and ability to functionally interact with the TET proteins support this conclusion. The ChIP-qPCR and co-IP experiments demonstrate co-localisation and interaction of EBF1 with TET2. It should be noted that I also investigated TET1 and TET3 for potential interaction with EBF1 but TET1 was not expressed in the assayed samples and although TET3 was expressed, no enrichment for the protein was found at any of the hypermethylated sites I investigated. As TET1 and TET3 both contain a CXXC domain known to preferentially bind to unmethylated CpGs while TET2 has undergone an evolutionary loss of that domain, it is conceivable that TET1 and TET3 use a different targeting mechanism (*Chapter 6*).

Although demethylation can be simply and passively achieved though lack of methylation maintenance during replication, for example in the maternal genome after fertilization[110], certain patterns of demethylation have been observed that can only be explained through an active process as in human monocytes, postmitotic cells, differentiating into dendritic cells[111]. The biochemical role of the TET family in active demethylation has been extensively discussed over the past few years, but to date no mechanism has been presented to explain targeted DNA demethylation. The interaction of the TET2 demethylase with the transcription factor EBF1 to affect tissue-specific CpGs presented here constitutes the first supporting evidence for a targeted demethylation pathway.

Future experiments should focus on elucidating the composition of the TET2-EBF1 complex: the experiments described here are unable to discern whether the interaction is direct or mediated by one or more other proteins. Furthermore, additional factors might be involved in the complex, contributing to its function but not involved in the interaction of these two proteins. Mass spectrometry analysis of the TET2 interactome should provide answers to these questions but is at present hindered by the quality of antibodies available.

Collaborators Ha Phuong Nguyen and Dr. Tracey Barrett are additionally developing an assay to test the specific interaction between the truncated catalytic domain of TET2 and the IPT/TIG domain of EBF1, which should provide novel insights into the mechanisms of this demethylation pathway.

5 PDXs AS A DISCOVERY TOOL FOR SARCOMA EPIGENOMICS

Most of the results presented in this chapter have been submitted for publication. The corresponding abstract and title pages can be found in the *Appendices*. DNA from the colon cancer xenografts was provided by the OncoTrack Consortium, and the 450K arrays for these samples were processed by Dr. Lee M. Butcher. Methylation data for the corresponding patient tumours was directly provided by the OncoTrack Consortium. The methylation data from head and neck cancer samples mentioned in the analysis was provided by Dr. Matthias Lechner. The full analysis script can be found in the *Appendices*.

5.1 Introduction

Xenografting of human tumours into mice or rats has been performed since the late 1960s[112], but it was not until the advent of immunodeficient mouse strains (e.g. severe combined immunodeficiency, SCID) in the mid-1980s that the practice became widespread in basic research and preclinical studies[113]. These new models of disease brought with them new hopes of therapeutic advances but have also displayed a number of noteworthy limitations[113]. Firstly, both the surrounding stroma and the blood vessels recruited to the growing tumour during angiogenesis effectively incorporate murine cells into the transplanted tumour. Secondly, placing the xenograft orthotopically is technically challenging, hence most are grown subcutaneously, effectively eliminating the possibility of replicating metastatic disease. Despite these limitations, patient-derived tumour xenografts (PDXs) have proven extremely accurate at predicting drug response in various cancer types[114], and have been used in numerous preclinical studies[115].

Osteosarcoma (OS) is the most common form of primary bone cancer, yet remains incredibly rare with an age-standardised incidence in the UK of 8 and 6 per million in males and females, respectively[52]. Thus, one of the major issues with the study of rare cancers such as OS is the scarcity of primary samples to analyse. This highlights the need for an accurate model of the disease and patient-derived tumour xenografts have been shown in multiple cancer types to better represent the genetic and gene-expression characteristics of tumours than *in vitro* cell lines[116]. Moreover, because OS presents most often in adolescents and young adults, who are less likely to enrol into clinical trials[117], patient recruitment can often take several years, thus enhancing the inherent jeopardy in drug selection for these trials. With this in mind, *in vivo* tumour models that most accurately replicate the patient's condition are a crucial factor in experimental pharmacology.

PDXs constitute one such model that is widely used in preclinical research[118], and OncoTrack, the largest European public-private biomarker consortium which aims to develop novel biomarkers for targeted therapy[119], generated PDXs that were included here as an additional tumour type and an example of a common cancer (colon cancer). Despite the popularity of PDXs, only a few systematic studies have compared their fidelity to the original tumours from which they were derived. Nonetheless, the findings have been encouraging: in pancreatic cancer for instance, gene expression patterns were faithfully retained in PDXs and the majority of the observed changes were associated with pathways reflecting the microenvironment [120]. To my knowledge however, only one study has assessed genome-wide DNA methylation changes in head and neck squamous cell carcinomas using the less powerful Infinium 27K Beadchip, which found no statistically significant changes[121].

To address this gap in our current knowledge, I have carried out a comprehensive assessment of the suitability of PDXs for cancer epigenomics. The assessment included methylome analysis using array- and sequencing-based technologies of primary and secondary PDXs derived from rare (OS) and common (colon cancer) cancers as well as computational simulations.

5.2 Comparison of osteosarcoma PDXs and patient tumours

To investigate the methylation changes linked to deriving xenografts from patient tumours, I subcutaneously inserted osteosarcoma fragments from two patients in the flanks of SCID mice, and grew them over two generations according to the scheme described in **Figure 5.1**. The two patient tumours came from patients of a similar age (13 and 14), both diagnosed with high-grade osteosarcoma in the left distal femur and both taken post-chemotherapy. The tumour labelled as T2P was less cellular than T1P however, perhaps explaining the lower initial yield of successful xenografts. The six mice described in this figure correspond to six of a total of 14 animals used: an initial set of six were given subcutaneous injections of cell suspensions but these led to no tumour growth. In addition, two mice died in the course of the experiment with solid tumour fragments: one in the first generation, after engraftment of fragments from patient tumour T2P, and the other in the second generation after engraftment of fragments from T1X1.

A final sample set consisting of two patient tumours (T1P and T2P), four first generation PDXs, and three second generation PDXs were available for methylation analysis on the Illumina Infinium 450K Beadchips[8].

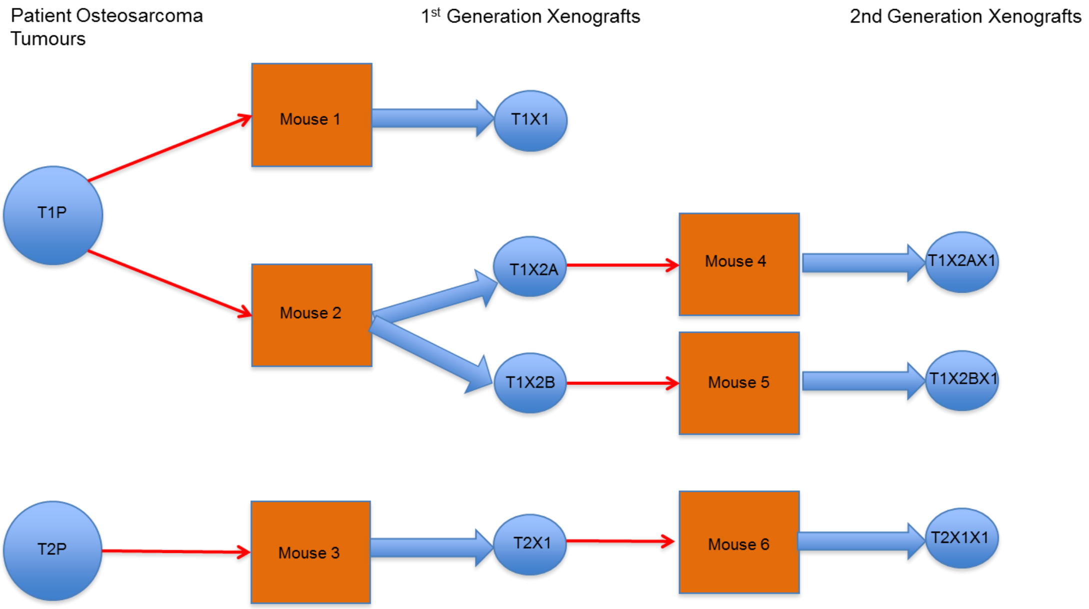


Figure 5.1: Osteosarcoma PDX derivation scheme

A single fragment from each patient tumour, approximately 1 mm in diameter, was inserted subcutaneously into each flank of a SCID mouse. Patient tumour 1 (T1P) gave rise to three first generation PDXs and two second generation PDXs, while Patient tumour 2 (T2P) was used to produce one PDX at each generation.

A major concern with analyses of human tumours grown in mice is the potential for signal contamination by host DNA from tumour vascularisation during its development or from the surrounding stroma when extracting the tumour. In order to eliminate these confounders in this methylation analysis, an additional mouse-only sample was processed on the 450K array and the 45,934 probes passing quality control were removed from downstream analysis. The raw data for all samples was subsequently processed through the ChAMP analysis pipeline[71] (*Chapter 2*) to produce a final dataset of 9 samples and 463,558 probes.

The distributions of methylation at the genome-wide and feature-specific levels for each sample are shown in **Figure 5.2**. Although methylation levels appear remarkably consistent within each tumour set, and in line with expected feature-specific values (e.g. low methylation at CpG islands), there is a slight increase in methylation levels across all features between the two patient tumours and their derivatives.

Specifically assessing methylation differences at each probe between a PDX and its original patient tumour further supports the maintenance of most of the methylome in tumour xenografts: **Figure 5.3a** shows that only a small fraction of the assessed CpG sites display large changes in methylation. From previous work[49], we know that 95% of fully unmethylated probes display β-values ≤ 0.31, while fully methylated probes have β-values ≥ 0.82; thus a Δβ threshold of 0.51 can be used as the minimum change expected for a CpG to be observed as going from fully unmethylated to methylated or vice-versa ("reversed methylation"). Using this threshold in the comparisons of PDXs and patient tumours, as shown in **Figure 5.3b**, an average of only 0.85% of probes in the T1 set (n=5) and 6.35% in the T2 set (n=2) are measured as reversing their methylation status, leading to inaccurate results if using the PDX as a proxy for the patient tumour.

In order to verify that the reversed methylation observed here was not a technical artefact, I performed the same analysis on technical replicates of pooled whole blood samples: three of the replicates were compared to one of the blood samples and only two, two, and zero probes with reversed methylation were observed in each of the respective comparisons. This demonstrates that, although small, the changes observed in the patient tumour - PDX comparisons are true biological shifts rather than technical artefacts.

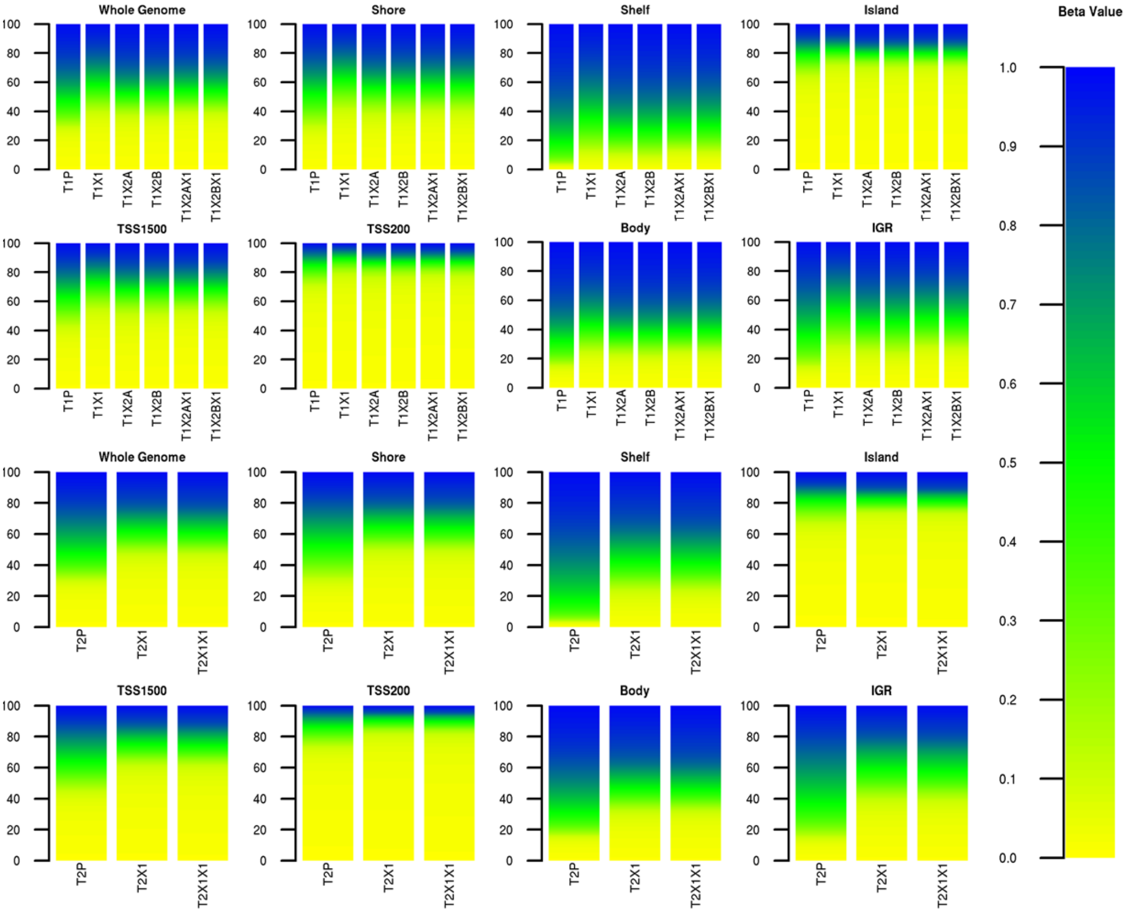
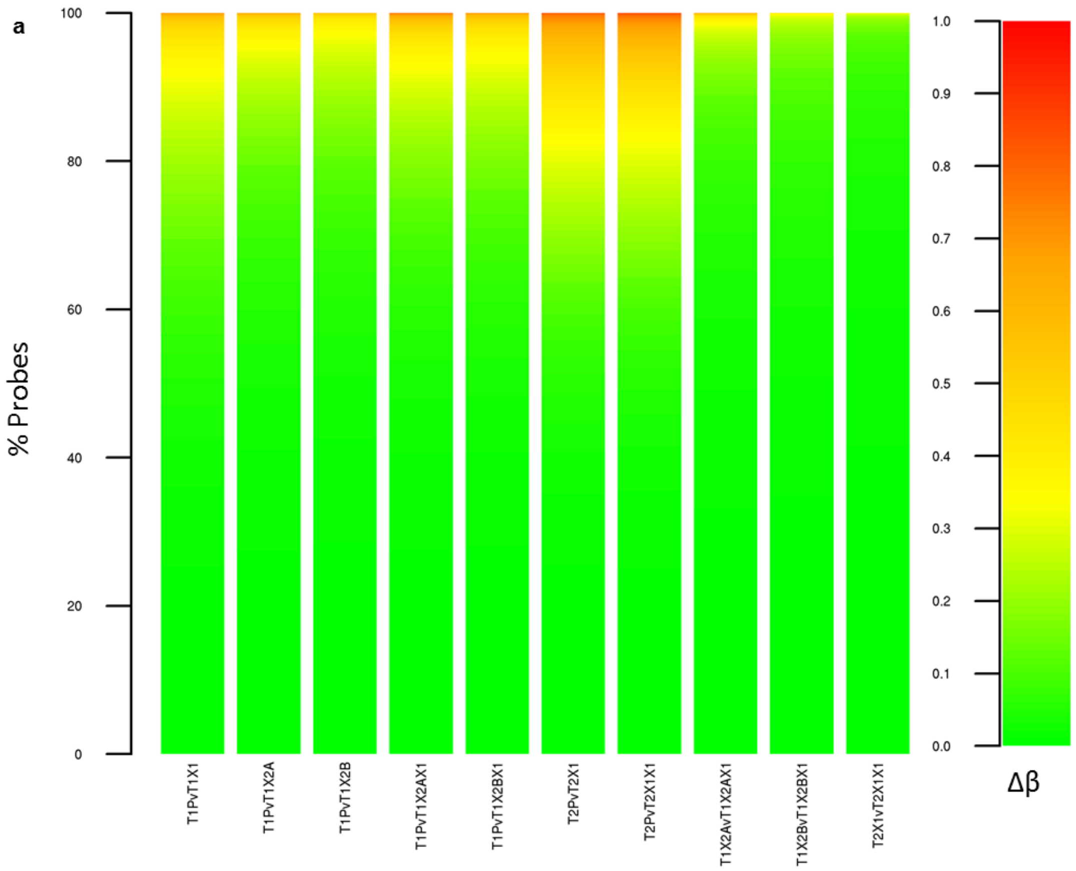


Figure 5.2: DNA methylation distribution by feature

For each feature, in each sample, the β-values are binned into 1% methylation increments (described by the colour scale), and the percentage of probes at each methylation level is shown in the individual plots. The top and bottom eight plots correspond to the T1 and T2 sets, respectively. TSS= Transcription Start Site, IGR= Intergenic Region, Whole Genome= all probes.



b

Comparison	T1P vs T1X1	T1P vs T1X2A	T1P vs T1X2B	T1P vs T1X2AX1	T1P vs T1X2BX1
#CpG(Δβ≥0.51)	5272	2398	797	6994	4153
%CpG(Δβ≥0.51)	1.14	0.52	0.17	1.51	0.90
Comparison	**T2P vs T2X1**	**T2P vs T2X1X1**	**T1X2A vs T1X2AX1**	**T1X2B vs T1X2BX1**	**T2X1 vs T2X1X1**
#CpG(Δß=0.51)	28939	29959	944	32	24
%CpG(Δß=0.51)	6.24	6.46	0.20	0.01	0.01

Figure 5.3: Assessment of methylation changes in OS PDXs

a) For each PDX, the absolute difference (β_{Patient}- $\beta_{\text{Xenograft}}$) is calculated at each probe and binned into 1% methylation difference increments (described by the colour scale); the percentage of probes showing each methylation difference level is shown in the individual plots.

b) Number and percentage of probes in each comparison changing by 0.51 or more, corresponding to all probes going from fully unmethylated to fully methylated and vice-versa.

5.3 Comparison of osteosarcoma PDXs across generations

Interestingly, although each set of PDXs displays this shift with xenografting, a constant profile is then maintained within a xenograft lineage: T1X2A, T1X2B and their 2nd generation tumours all displayed consistent levels across features (**Figure 5.2**), as did T2X1 and T2X1X1, demonstrating that although the change in host is linked to a slight increase in methylation levels, subsequent xenografting is not accompanied by additional changes. This is confirmed by the vastly reduced number of reversed methylation events observed between first and second generation PDXs as opposed to those identified within the first generation; **Figure 5.3** reveals that an average of only 0.07% (n=3) of CpG sites see their methylation scores increase or decrease by over 0.51 after the first generation. This result suggests either an initial reaction to the new host, that is then preserved in further generations (as the mice used were isogenic), or a loss of tumour heterogeneity as only a fragment of the initial patient sample was used for xenografting, or a combination of these two factors. The fact that loss of heterogeneity would be expected to persist in further generations as only a fragment of the grown tumour is transplanted at each passage, and that signal from new host stromal cells and vascularisation affect gene expression in specific pathways (such as extra cellular matrix formation)[120] suggests that the observed epigenetic change is due primarily to implantation of the tumour into a new host.

5.4 Validation with MeDIP-seq

In addition to the methylation arrays, the osteosarcoma PDXs and patient samples were analysed by Methylated DNA Immunoprecipitation followed by low-coverage next-generation sequencing (MeDIP-seq[42]). Alignment, filtering of reads, and calling of differentially methylated regions (DMRs) were performed using the MeDUSA pipeline[81]. In order to minimise read contamination by mouse DNA, the fastq files were aligned separately to the human and mouse genomes and those reads aligning only to mouse were removed from downstream analysis.

In order to ensure that this approach yielded the most accurate set of human reads, I also filtered the reads using the Xenome[122] protocol and compared the outputs (**Figure 5.4**).

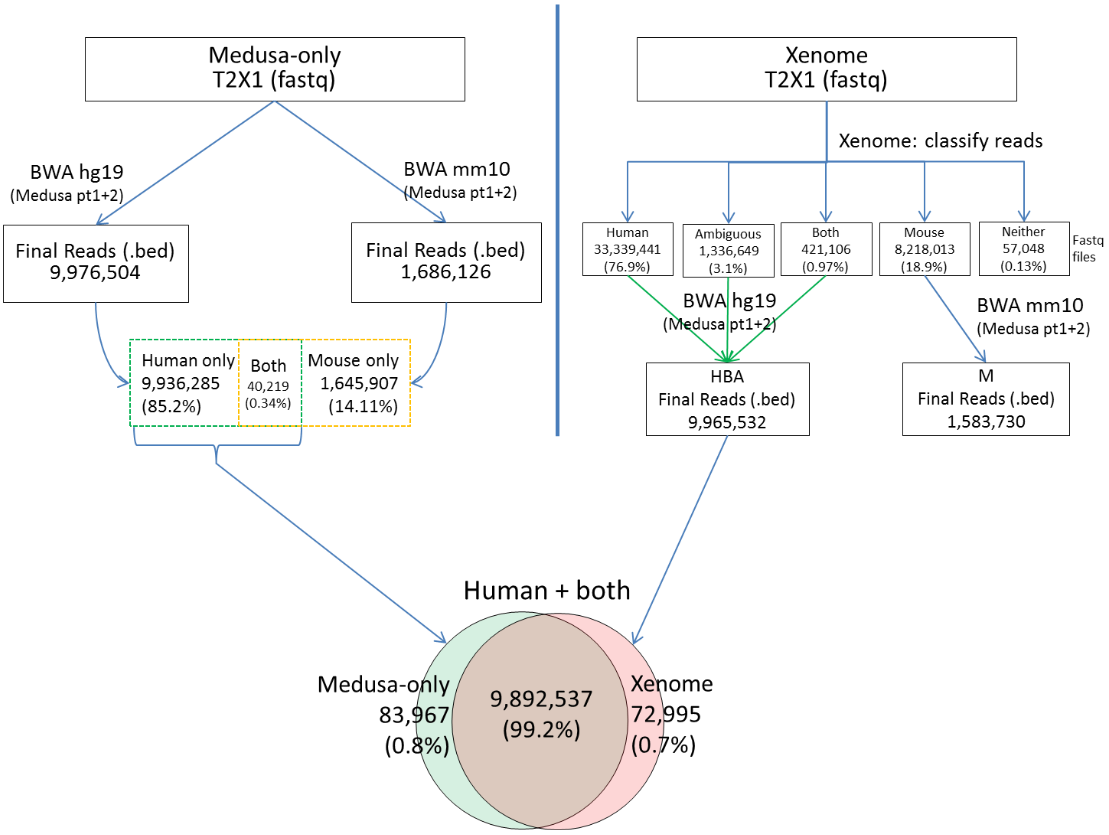


Figure 5.4: Comparison of read filtering methods

Reads from sample T2X1 were filtered using either the Xenome protocol or simple alignments to hg19 and mm10 and exclusion of mouse-only reads. Retaining only reads that mapped to either human-only or both human and mouse, the overlap between the two methods was over 99%. This figure is based on the numbers from the T2X1 analysis but this was performed for all samples.

Briefly, the Xenome method involves first indexing the graft and host genomes, here human (hg19) and mouse (mm10), respectively. The sample reads are then classified according to matches to the indexed genomes into the following five categories: *host*, *graft*, *both*, *ambiguous*, *neither*. As recommended by the authors, I included reads binned into either *graft*, *both*, or *neither*, and obtained an almost complete overlap (98.5 - 99.3%) with the read set obtained by simple alignment to mm10 and hg19 followed by selection of reads mapping to human only or both human and mouse (**Table 5.1**). Therefore, this latter, simpler method is preferable to the Xenome protocol, at least in the case of this dataset, as it produces a nearly identical filtered read set with fewer steps and a much smaller computational load: the processing load is reduced by avoiding the genome indexing and read classifying steps prior to alignment, while data storage is also diminished by circumventing the need to store the indexed host and graft genomes. The final read counts aligning to human, mouse or both as determined by the Medusa-only method are shown in **Table 5.2**.

SAMPLE	T1P		T1X1		T1X2A	
	Count	Percent	Count	Percent	Count	Percent
Final Overlap	2,216,467	99.0	2,328,746	98.5	3,105,555	98.6
SAMPLE	T1X2AX1		T1X2B		T1X2BX1	
	Count	Percent	Count	Percent	Count	Percent
Final Overlap	4,240,738	98.5	3,119,375	98.8	2,844,791	98.7
SAMPLE	T2P		T2X1		T2X1X1	
	Count	Percent	Count	Percent	Count	Percent
Final Overlap	12,103,693	99.3	9,892,537	99.2	14,750,660	99.3

Table 5.1: Overlaps in final read sets between filtering methods

The Medusa-only and Xenome protocols described above were applied separately to each sample. The overlaps in final read sets are shown in this table. They range from 98.5-99.3%.

Interestingly, a small proportion of reads from the patient tumours align only to the mouse genome (0.5% and 0.35% in T1P and T2P, respectively), as shown in **Table 5.2.** These samples were stored separately from the fragments inserted into the mice and at no time came intentionally in contact with murine DNA. Therefore, I investigated two possible causes for this observation: these samples were accidentally contaminated with mouse DNA during the MeDIP portion of the experiment when all samples were processed together; or these reads are technical artefacts.

Sample	T1P		T1X1		T1X2A	
	Count	Percent	Count	Percent	Count	Percent
Human	2,232,237	99	2,351,595	62.97	3,132,476	68.64
Mouse	11,189	0.5	1,357,335	36.34	1,397,076	30.61
Both	5,690	0.25	12,906	0.35	17,095	0.37
Sample	T1X2AX1		T1X2B		T1X2BX1	
	Count	Percent	Count	Percent	Count	Percent
Human	4,292,507	86.27	3,139,160	63.93	2,870,080	75.98
Mouse	655,435	13.17	1,733,618	35.31	882,557	23.36
Both	13,811	0.28	18,677	0.38	12,512	0.33
Sample	T2P		T2X1		T2X1X1	
	Count	Percent	Count	Percent	Count	Percent
Human	12,154,767	99.11	9,936,285	85.2	14,793,302	85.16
Mouse	43,526	0.35	1,645,907	14.11	2,454,625	14.13
Both	32,518	0.27	40,219	0.34	61,171	0.35

Table 5.2: Final MeDIP-seq read counts for OS samples

The fastq files were aligned to both Human (hg19) and Mouse (mm10) genomes. Only reads that aligned to either Human-only or both Human and Mouse were retained for downstream analysis.

The possibility of contamination during processing was assessed by performing the same alignment procedure on data produced from MeDIP-seq experiments conducted by another person (Dr. Matthias Lechner) in a project not involving any mouse DNA at any stage. Data from two human head and neck cancer samples was used for this test, and similar proportions of these reads aligned to mouse only: 0.37% and 0.49%. This suggests that contamination is unlikely to be responsible for these misaligning reads. Further support for this conclusion was provided by the genomic distribution of the mouse-only reads from the OS tumours: 78.2% and 74.2% of those sequences from T1P and T2P, respectively, mapped to repeat regions as annotated by the mouse UCSC repeat masker (mm10), with proportions of repeat types vastly differing from a random distribution; for example, over 30%

of those mouse-only reads mapping to repeats in the OS patient samples corresponded to rRNA (ribosomal RNA) repeats, while only 5% would be expected from a random distribution of contaminating reads. These findings strongly suggest that the small proportion of sequencing reads from the patient samples that aligned only to mouse are the result of a technical artefact and do not represent a biological contamination.

Of the total 1,095 intra-tumour MeDIP-seq DMRs identified across all seven patient tumour/xenograft comparisons, 48 overlapped with 450K probes that showed a reversed methylation profile in the same comparison. The seemingly low overlap between the two methods is due to the bias of MeDIP-seq towards regions of high CpG density[42], such as CpG islands, whereas the sites identified by the 450K as displaying reversed methylation were enriched for intergenic regions with low CpG density, as detailed in section 5.5. Despite this, at those 48 overlapping loci, the directionality of methylation change between patient tumour and xenograft was concordant between the two methods, with the same 22 gains and 26 losses of methylation identified in the PDXs.

Similarly, in an inter-tumour comparison, when assessing the ability of a PDX to substitute for its matched patient tumour in an inter-tumour comparison (i.e. T1P vs T2P), 450K and MeDIP-seq both identified similar trends (**Figure 5.5** and **Figure 5.6**): for each technology, the differences between the patient tumours T1P and T2P were assessed to act as a reference set; each PDX was then compared to the unmatched-patient to see if the same differential methylation was captured. MeDIP-seq showed similar levels of concordance in the comparisons as the methylation array, with the exception of two of the hypomethylation sets (T1PvT2X1 and T1PvT2X1X1) that displayed lower levels of concordance (22.4%

and 17.6%, respectively) in the MeDIP-seq data (**Figure 5.6**). These, however, represent only small absolute differences in concordance (66 and 70 DMRs of the T1P vs T2P comparison were not identified in T1P vs T2X1 and T1P vs T2X1X1, respectively) due to the overall low number of hypomethylated DMRs detected between the two patient tumours (n=85) as compared to hypermethylated (n=1,980).

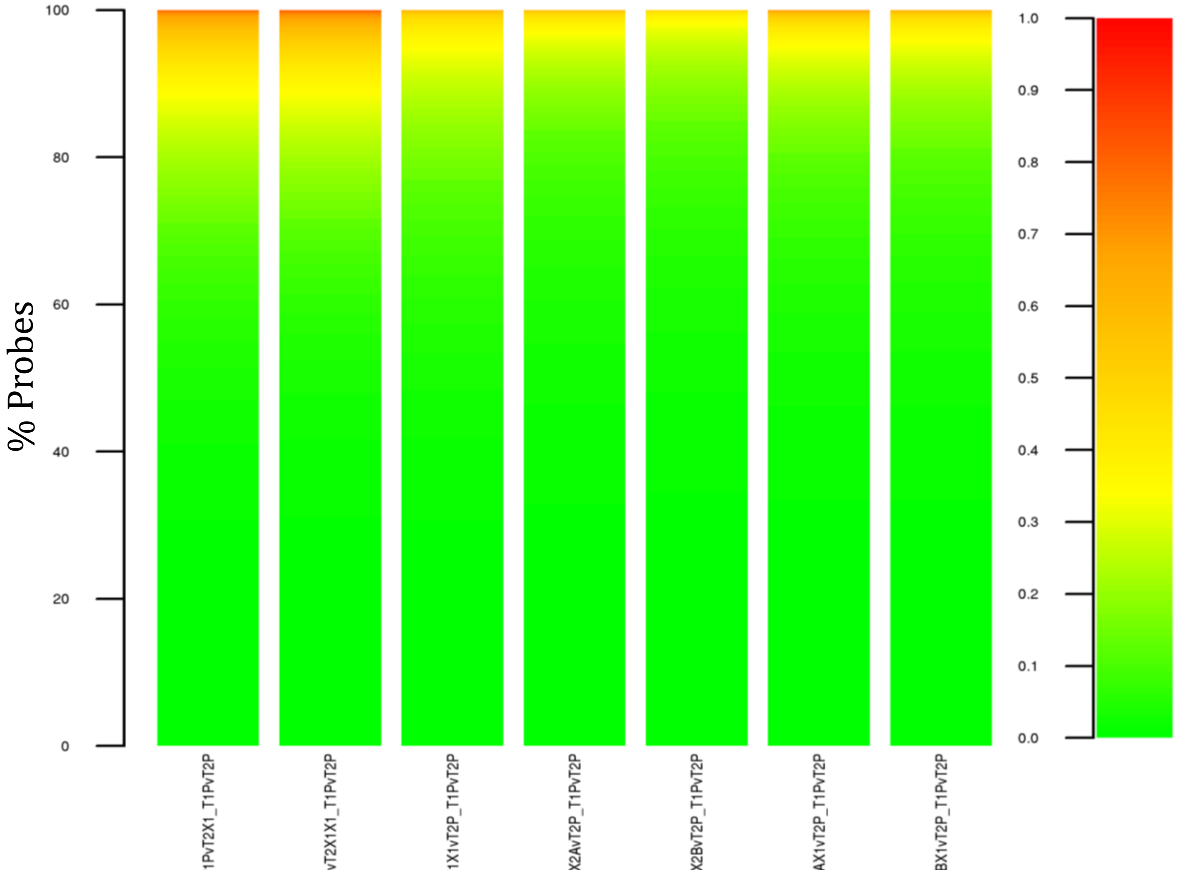


Figure 5.5: PDXs as substitutes for patient tumours: 450K

The absolute difference in β value between the two OS patient tumours is calculated at each probe. The absolute difference between each PDX and the patient tumour from the other tumour set is then assessed, and a ΔΔβ for those two differences is calculated and plotted as in Figure 3. A result close to zero indicates concordance between the two measurements at a given CpG site.

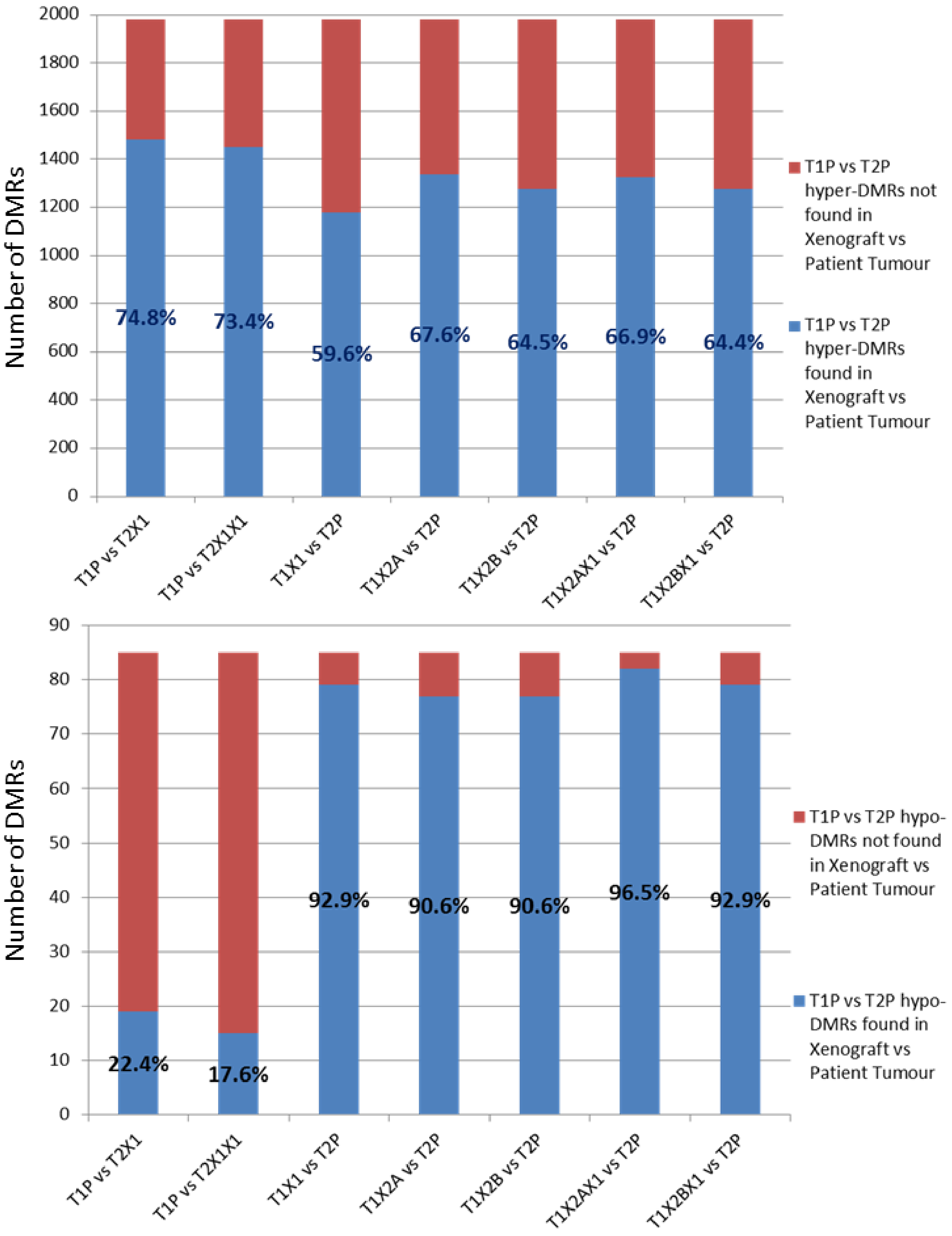
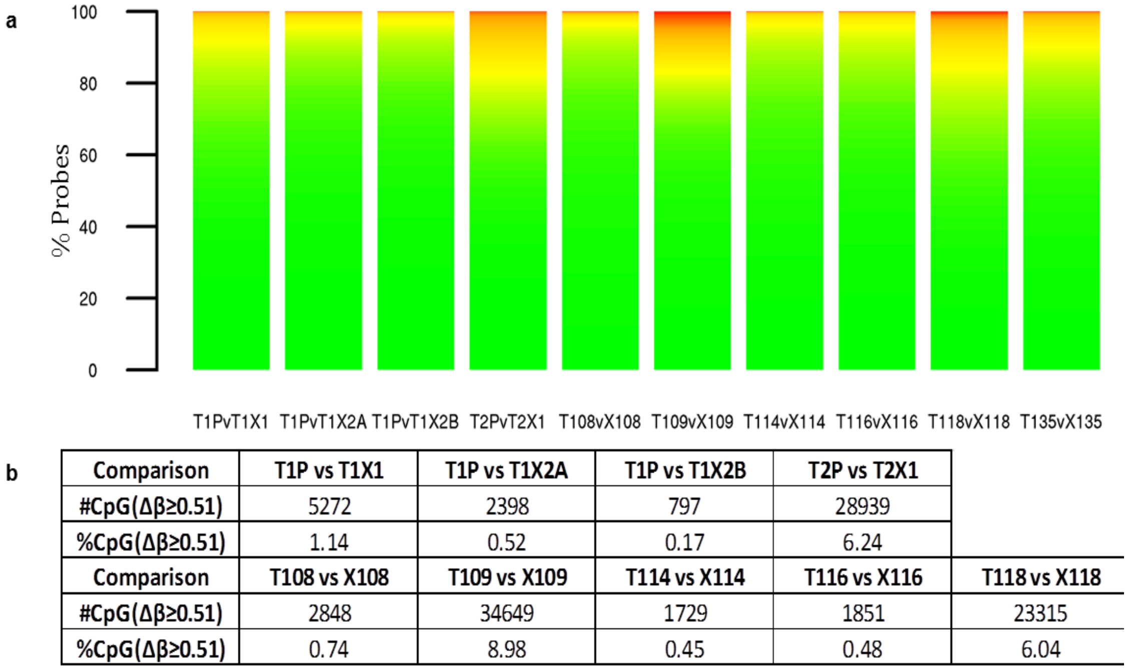


Figure 5.6: PDXs as substitutes for patient tumours: MeDIP-seq

Analogous to the process described above but with MeDIP-seq, the number of DMRs between the two patient tumours that can be recapitulated between a PDX and the patient tumour are shown, for both hyper- and hypo-DMRs

5.5 Methylome changes in colon cancer and osteosarcoma PDXs

In order to further investigate those few CpG sites with changing methylation levels after xenografting, an additional set of six patient tumour/xenograft colon cancer pairs from the OncoTrack consortium were assessed using Illumina 450K arrays and processed with the R package ChAMP. Grouping these with the first generation PDXs derived from OS tumours yields a final cohort of ten sample pairs **(Figure 5.7)**. Using the same Δβ threshold of 0.51 as for the OS samples, a similarly low number of probes were identified as changing with xenografting in the first generation, with an average of 3.18% (n=6).



Comparison	T1P vs T1X1	T1P vs T1X2A	T1P vs T1X2B	T2P vs T2X1	
#CpG(Δβ≥0.51)	5272	2398	797	28939	
%CpG(Δβ≥0.51)	1.14	0.52	0.17	6.24	
Comparison	T108 vs X108	T109 vs X109	T114 vs X114	T116 vs X116	T118 vs X118
#CpG(Δβ≥0.51)	2848	34649	1729	1851	23315
%CpG(Δβ≥0.51)	0.74	8.98	0.45	0.48	6.04

Figure 5.7: Assessment of methylation changes in OS and colon cancer PDXs

a) For each PDX, at each probe, the absolute difference ($\beta_{Patient}$- $\beta_{Xenograft}$) is calculated and binned into 1% methylation difference increments (described by the colour scale); the percentage of probes showing each methylation difference level is shown in the individual plots.
b) Number and percentage of probes in each comparison changing by 0.51 or more, corresponding to all probes going from fully unmethylated to fully methylated and vice-versa.

To assess whether changes in methylation could be generalised to any tumour undergoing this procedure or whether they are tumour or tumour type-specific, the overlap in these changing probes within- as well as between- tumour types was evaluated. Statistically significant overlaps were found within each tumour type, with 236 probes changing in all first generation OS PDXs and five probes in colon cancer PDXs (random resampling p-value<10^{-4}); however, gene ontology tools (GREAT[123], Panther[124], DAVID[125]) did not reveal any particular functional links between these changing sites and no overlap was found between the two tumour types. This suggests that the changes in methylation observed with xenografting are unlikely to be due to a systematic reaction to the xenografting procedure but rather point to tumour-specificity.

Finally, I assessed whether these methylation changes were more likely to occur in certain genomic and/or epigenomic features. As shown in **Figure 5.8**, these probes are depleted for promoter regions and CpG islands, but enriched for intergenic regions, particularly those with low CpG density (p-value <10^{-4}).

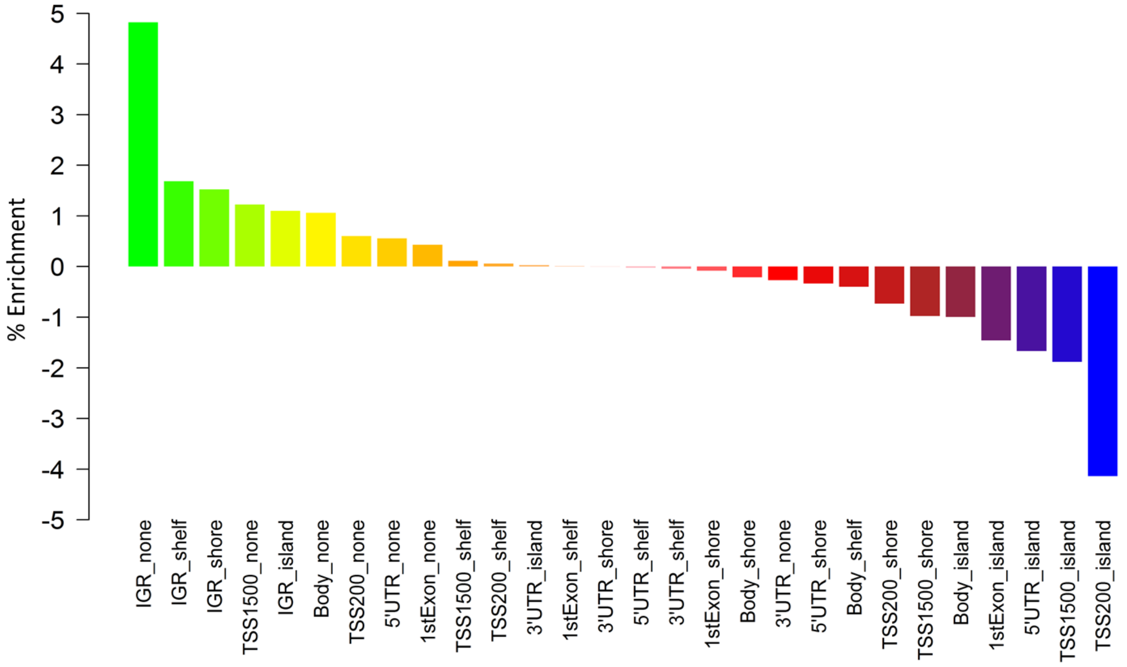


Figure 5.8: Enrichment of (epi)genomic regions with changing methylation status after xenografting

Each probe on the 450K array is annotated to a genomic (TSS1500, Body, 3'UTR…) and epigenomic (island, shore, shelf, none) region. These were combined for each probe to form a unique (epi)genomic annotation and enrichments were calculated using a random resampling strategy. TSS=Transcription start site, IGR = intergenic region

In the OS cohort, one of the patient tumours produced three first-generation PDXs, grown in two animals. Two of the PDXs (T1X2A and T1X2B) were harvested from the same mouse, one from each of the flanks. Despite the limited sample size, this set-up provides novel and important insights into the potential tumour-specificity of the observed changes in methylation. The results displayed in **Table 5.3** reveal that over 86% of probes changing in T1X2B also underwent major changes in T1X2A, and over 64% of changes were common between all three PDXs originating from T1P. These overlaps, much higher than those observed within or across tumour types further confirm tumour specificity of the observed methylation changes that accompany xenografting.

#CpG Overlap		692	
%CpG Overlap		86.83	
Comparison	**T1P vs T1X1**	**T1P vs T1X2A**	**T1P vs T1X2B**
#CpG(Δβ≥0.51)	5272	2398	797
%CpG(Δβ≥0.51)	1.14	0.52	0.17
#CpG Overlap	515		
%CpG Overlap	64.62		

Table 5.3: Overlap of changing CpG sites between PDXs originating from the same patient tumour

T1X2A and T1X2B were grown from T1P in two flanks of the same mouse. T1X1 was grown from T1P in a different animal. Overlap percentages were calculated based on the number of changing sites in T1X2B, the PDX with the fewest changes. Over 86% of probes changing in T1X2B also underwent major changes in T1X2A, and over 64% of changes were common between all three PDXs originating from T1P.

5.6 Practical implications for the use of PDXs in epigenetic studies

With a mean percentage of 2.7% (n = 11,110) of CpG sites undergoing major methylation shifts in first generation xenografts, PDXs appear to be more than adequate proxies for patient samples in methylation studies. However, the tumour-specific nature of these methylation changes implies that no accurate prediction as to which 2.7% of the measured methylation scores will be affected can reasonably be made beyond a general statement concerning enrichment in intergenic regions. In order to aid in the design of future studies, I devised a model to test how many 450K arrays should be run when comparing two groups of samples in order to minimise the effects of these tumour-specific xenografting-linked methylation changes. From a total of 2,000 datasets from Marmal-aid[126], a 450K data repository, I selected *n* (5 ≤ n ≤ 50) random samples. A total of 11,110 random β-values in each sample were then increased or decreased by 0.51 (5,555 of each). I subsequently compared the original *n* samples from Marmal-aid to their modified counterparts and assessed the number of sites that appeared to be significantly differentially methylated between the groups (**Figure 5.9**), as determined by a Wilcoxon rank-sum test with a non-adjusted p-value threshold of 0.05.

This analysis revealed that the maximum number of probes significantly differentially methylated between the groups was eight, and if using 15 or more samples in each group, the xenografting-associated methylation changes might only significantly affect the differences between groups at two loci. This further demonstrates the suitability of tumour xenografts for methylome analysis.

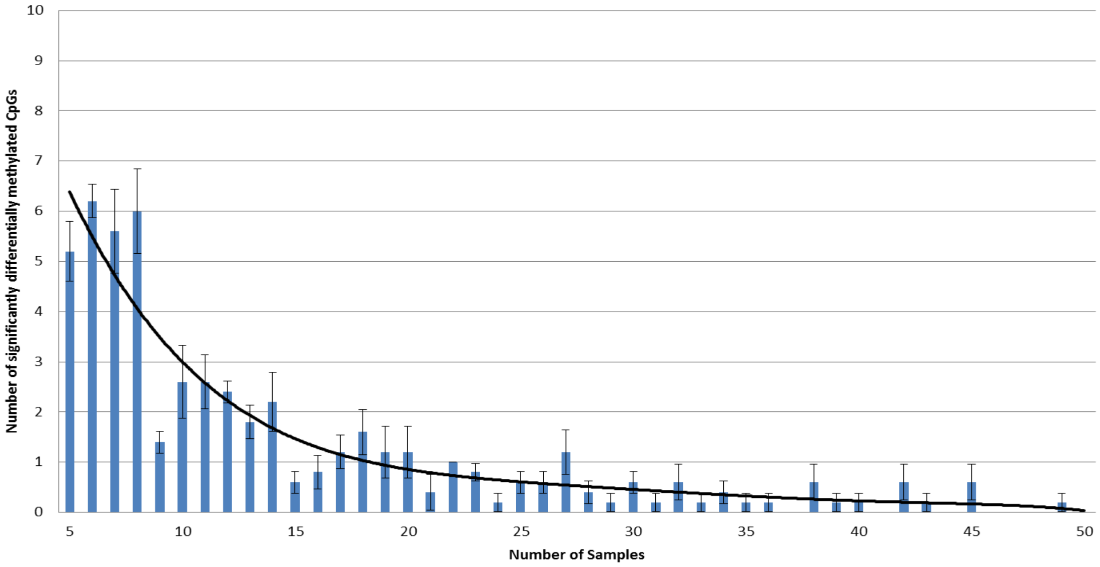


Figure 5.9: Model of the effect of PDX-associated methylation changes

For sample numbers *n* from 5-50, *n* random samples were randomly selected from 2,000 Marmal-aid datasets. Each sample was modified at 11,110 probes by β=0.51 and a Wilcoxon rank-sum test run between the original n samples and the modified versions. The number of significantly differentially methylated probes (p-value ≤ 0.05) for each *n* is plotted against *n*.

5.7 Conclusions and future work

Xenografts of patient tumours are frequently used in both basic research and preclinical drug development as a model of the original malignancy[115]. They represent its morphology, heterogeneity, and development more accurately than can be achieved today from a tumour-derived cell line[116] and can prove invaluable in providing additional sample material for the investigation of rare cancer types as well as a much needed filtering step in experimental pharmacology.

However, our understanding of the genomic, epigenomic, and gene expression shifts that might occur when transplanting a tumour into a new host are poorly understood. Indeed, the xenografting procedure implies drastic changes in stroma, vascularisation, and heterogeneity for the tumour but surprisingly few studies have investigated how well these PDXs replicate the molecular make-up of the patient's disease. Only one systematic genomic profiling of patient tumours and PDXs is available in the literature[127], while the epigenome was also explored only once previously (but with now-outdated technology) finding no significant differences in the xenograft's methylome[121].

With the advent of clinical developments based on epigenetics such as biomarkers for cancer[128] and inhibitors affecting epigenetic modifiers[129], understanding and predicting how PDXs can epigenomically differ from patient tumours is of paramount importance and thus the aim of the study presented here was to provide an initial assessment of these variations.

Using both rare (OS) and common (colon cancer) cancer types, I identified any large changes in methylation across the genome through two generations of PDXs. These were investigated with both microarrays and sequencing-based technologies and revealed that an average of 2.7% of the methylome displays a reversed methylation profile after xenografting, while almost no further changes (<1%) are observed in subsequent generations. The observed changes are tumour-specific and enriched for intergenic regions.

In addition, a model is provided here to guide other researchers in their use of PDXs in epigenetic studies. The 2.7% of CpG sites with reversed methylation cannot be predicted *a priori* due to tumour-specificity, but full confidence in the capacity of a group of PDXs to replicate the methylation profile of the patient tumours can be achieved using 15 or more samples in each group thus diluting the statistical effect that xenografting-associated changes might have on the data.

Future studies should aim to understand the root cause of the xenografting-associated changes to inform the creation of ever more accurate proxies for patient samples as it is becoming increasingly apparent that the ability to replicate the full heterogeneity of a solid tumour will lead to the development of more targeted therapies[130].

Finally, technological advances may also provide alternative solutions to the limitations of existing tumour models such as classical cell lines and xenografts. For example, the development of three dimensional (3D) tumour models[131] may provide an adequate middle ground between the over-simplified 2D tumour cell lines, and the inherently complex *in vivo* xenografts by accurately replicating vascularisation and stromal environment, while avoiding contamination by host DNA and the effects of unorthotopic xenografting.

6 Discussion

6.1 The IDH idea

6.1.1 Consequences of *IDH* mutations

Recurrent *IDH1/2* gain-of-function mutations have now been identified in a number of cancer types, and their impact on the methylation profile of affected cells has become quite clear: instead of the interconversion of isocitrate and α-KG, mutant IDH uses up α-KG to produce the oncometabolite 2-HG, which in turn competitively inhibits a number of α-KG-dependent enzymes, including the TET family of demethylases. This results in a genome-wide DNA hypermethylation phenotype. As described in *Chapter 3*, over 50% of CSs harbour a mutation in *IDH1* or *IDH2* and through the use of the established 450K array and a novel microdroplet PCR-based assay I showed that 3,057 CpG sites are significantly hypermethylated in mutant CS relative to wild-type. These are enriched for island and promoter regions, thus replicating the overall CpG island methylator phenotype observed in glioma and AML.

The functional consequences of this hypermethylation, however, are less clear. The simple expectation would be that hypermethylated gene promoters correlate with downregulation of the corresponding transcript's expression. However, as mentioned in *Chapter 3*, no significant differential expression was identified in *IDH*-mutant CS relative to WT, and no correlation of directionality of change between promoter methylation and gene expression could even be established, with only half of the genes with hypermethylated promoters displaying any sign of downregulation. I have already discussed a possible explanation for these observations, involving a locking mechanism for prior chromatin changes, which could fit the currently accepted paradigm of coupled promoter methylation/gene downregulation. However, a different, and perhaps simpler, way of explaining these apparent inconsistencies is to consider that the promoter-gene association that is generally assumed might be incorrect, or at least incomplete. Indeed, a recent study in the relationship between DNA methylation and gene expression in human fibroblasts concluded that "the location of CpG probes with respect to the gene provides relatively little information about the sign of the correlation"[132], and the authors suggested that histone marks in the vicinity of 5mC were much more indicative of correlation between a particular methylated CpG and a gene. Moreover, others have shown that only 3.3% of intertumour variation in gene expression could be attributed to promoter methylation[133] and that methylation of distal enhancers is often better correlated to gene expression. With the ongoing improvement of both our awareness of distal regulatory regions and the technology required to probe the physical plasticity of the genome (e.g. chromosome conformation capture) it would be beneficial to incorporate the possibility of long-range transcriptional regulation by DNA methylation into the routine investigation of methylation profiles and their consequences.

Finally, further consideration should be given to the confounding impact of variations in mutational load across patient tumours as well as that of the heterogeneity of individual tumours. Subdividing sample cohorts into more precise, genetically similar groups could reveal associations with transcriptional regulation that are not apparent when using the more inclusive criterion of the mere presence of a particular mutation in a given sample.

6.1.2 Hypermethylation and the Krebs cycle

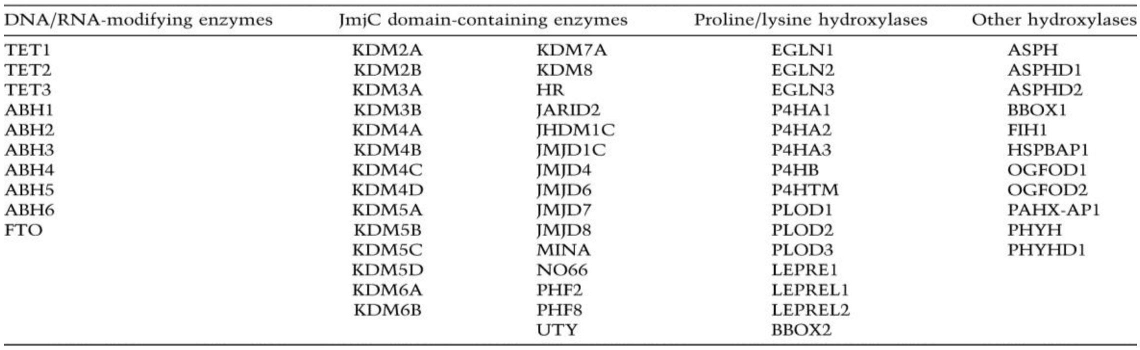
Interestingly, it has recently come to light that the IDH isoforms are not the only members of the Krebs cycle for which mutations are associated with genome-wide methylation changes. In particular, various studies have investigated mutations in succinate dehydrogenase (SDH), the catalysing enzyme for the oxidation of succinate to fumarate. In paragangliomas and pheochromocytomas, for example, two separate studies identified a hypermethylation phenotype associated with mutations in *SDH*[134,135]. SDH is composed of four subunits (SDHA/B/C/D), and the mutations linked to DNA hypermethylation occur mainly in *SDHB*. This is further supported by the fact that cells from *Sdhb*-deficient mice displayed higher levels of 5mC and lower 5hmC than their wild-type counterparts, as well as increased histone methylation[135], thus closely replicating the phenotype observed in *IDH* mutant cells. In fact, the mechanism linking *SDH* mutations to hypermethylation also resembles that of *IDH* as succinate has been shown to competitively inhibit α-KG-dependent dioxygenases including the TET hydroxylases[136]. Where these processes differ is in the type of mutation they originate from: while the IDH proteins gain a new function through mutation of particular residues, SDH is inactivated, and the causal mutations are widespread[135]. An alternative mechanism of SDH inactivation that bypasses the need for a mutation was recently reported[137] and involves the accumulation of 5mC in the promoter region of *SDHC* and the reduced expression of the corresponding transcript.

When *IDH* mutations were initially being investigated for apparent links to hypermethylation, a strong supporting argument for the TET-inhibition mechanism was that mutations in *IDH* and *TET* were mutually exclusive in AML, but were both associated with DNA hypermethylation, suggesting they might be part of the same pathway. Similarly, in the study of paragangliomas[135], only one sample was wild-type for *SDH* but still displayed the same hypermethylation profile; exome sequencing revealed mutations in yet another Krebs cycle protein, fumarate hydratase (FH). This enzyme catalyses the reversible conversion of fumarate to malate, and accumulation of fumarate has also been shown to inhibit α-KG-dependent enzymes, much like 2-HG and succinate.

Attempts at mitigating some of the functional effects of SDH loss have already been made with demethylating 5-aza-2'-deoxycytidine treatment reversing the increased migratory ability of chromaffin cells *in vitro* induced by *Sdhb* loss[135], although it could be expected that being affected by loss-of-function mutations would make these new epigenetic effectors less attractive targets for therapy than IDH.

6.1.3 Beyond DNA methylation

Accumulation of 2-HG does not solely affect the TET enzymes but has, in theory, the potential to inhibit all proteins that depend on α-KG for their function, representing a set of over 70 putative targets, as shown in **Table 6.1**. In practice, however, not all of these enzymes are equally affected by 2-HG, and their sensitivity to the accumulation of this metabolite depends on their affinity for it[138]. The most sensitive happen to be the KDM family of histone demethylases, including those responsible for the demethylation of H3K9 and H3K36 (KDM4A, KDM4C, and KDM2A), giving mutant IDH the potential for major effects on multiple facets of the epigenome.



DNA/RNA-modifying enzymes	JmjC domain-containing enzymes		Proline/lysine hydroxylases	Other hydroxylases
TET1	KDM2A	KDM7A	EGLN1	ASPH
TET2	KDM2B	KDM8	EGLN2	ASPHD1
TET3	KDM3A	HR	EGLN3	ASPHD2
ABH1	KDM3B	JARID2	P4HA1	BBOX1
ABH2	KDM4A	JHDM1C	P4HA2	FIH1
ABH3	KDM4B	JMJD1C	P4HA3	HSPBAP1
ABH4	KDM4C	JMJD4	P4HB	OGFOD1
ABH5	KDM4D	JMJD6	P4HTM	OGFOD2
ABH6	KDM5A	JMJD7	PLOD1	PAHX-AP1
FTO	KDM5B	JMJD8	PLOD2	PHYH
	KDM5C	MINA	PLOD3	PHYHD1
	KDM5D	NO66	LEPRE1	
	KDM6A	PHF2	LEPREL1	
	KDM6B	PHF8	LEPREL2	
		UTY	BBOX2	

Table 6.1: Known and putative α-KG-dependent enzymes

List of proteins with a dependence on α-KG, that could be affected by increased cellular levels of 2-HG. Original table from Losman et al.[139]

6.1.4 Harnessing IDH

Considering their presence in a portion of the patient population for a number of malignancies and their appearance early in tumourigenesis, it would have been plausible for *IDH* mutations to be suitable biomarkers for survival or tumour development. However, they do not appear to have the same prognostic impact across all malignancies. While they have been reported to have a positive effect on the survival of glioblastoma patients, with the median survival increasing from 15 months for those producing only wild-type IDH to 31 months if a mutant allele is present[140], it is still unclear whether this difference is actually related to the *IDH* mutation status or is in fact the consequence of other biological differences between those tumours appearing as primary malignancies (no mutations in *IDH*) and those developing from low-grade gliomas (*IDH1* mutation frequency ~70%)[139]. Reports on prognostic significance differ in AML too, from those showing no difference[141], to those claiming increased[142] or decreased[143] risk of relapse. Finally in CS, no significant difference in survival was identified[63].

While IDH itself seems to have limited use as a biomarker, its product D-2-HG is more promising as intracellular 2-HG levels are directly correlated with the presence of mutant IDH in glioma[27], AML[144], and CS[1]. This is particularly relevant for AML, as 2-HG is excreted from cells and its levels can be assessed in serum[144], making it potentially a useful tool for AML diagnosis. Even solid and less accessible tumours such as gliomas could benefit from 2-HG as a biomarker, as reports suggest that magnetic resonance spectroscopy could be used to detect regions of high 2-HG levels[145], which could potentially aid in surgery preparation or for a non-invasive initial diagnosis. Finally, 2-HG levels have already been shown to decrease with reduced AML burden and increase again with relapse[146], indicating they could be used to monitor the effects of therapy in the case of tumours carrying mutant IDH.

Aside from the potential use of 2-HG as a biomarker, the existence of mutant IDH and its presence in multiple tumour types are already being harnessed by the pharmaceutical industry. The methylation and gene expression data presented in *Chapter 3* were shared with Agios, the company responsible for the development of the inhibitors AG-120 and AG-221. These molecules specifically target mutant forms of IDH1 and IDH2, respectively, and are currently in Phase 1 clinical trials[129,147].

Another way in which our understanding of mutant IDH and its consequences on the cancer epigenome can be utilised is by providing insights into the mechanisms that these mutations affect, specifically DNA demethylation, as discussed in the following section.

6.2 DNA demethylation

6.2.1 A novel mechanism for targeted demethylation

With an apparently similar epigenetic phenotype, linked to the same gain-of-function mutations emerging in multiple cancer types, a comparison of the hypermethylated sites in each cancer was the next logical step. The dataset I generated for CS adequately complemented those already existing for AML and LGG as the only large AML dataset available in the public domain at the time was generated using the HELP assay, hindering a direct comparison with the 450K-generated LGG data. Fortuitously, an additional dataset was generated for yet another cancer type, cholangiocarcinoma (CC), while this work was ongoing and could be incorporated into the meta-analysis, as discussed in *Chapter 4*.

Through a gene and pathway analysis, I showed that although all four cancer types targeted the major cancer pathway involved with the downstream response to retinoic acid receptor activation, there was overall little overlap between any two cancer types in the genes affected by the hypermethylation, and in fact no overlap at all between the four. Thus multiple cancer types, presenting the same mutation, the same consequential increase in cellular 2-HG levels and the same inhibition of TET function, all displayed an overall hypermethylation of the genome, but at different CpG sites. This finding was not entirely surprising as the methylome is known to be tissue-specific, so the enzymes regulating it, such as the TET demethylases, could logically be expected to be themselves under some kind of tissue-specific regulation. This could conceivably be partly achieved through the rearrangement of chromatin to make target regions more or less accessible to TET proteins and CpG sites within open chromatin regions that should remain methylated could theoretically be protected from demethylation by methyl-binding proteins. This seems an unlikely possibility, however, considering the energy expenditure that would be required to constantly maintain in their current state the 80% of genomic CpGs that are normally methylated. Hence the most likely mechanism for tissue-specific regulation of TET-mediated demethylation is through the targeting of TET itself. Both TET1 and TET3 contain a CXXC domain that selectively binds unmethylated CpGs, while TET2 interacts with IDAX (a CXXC domain-containing protein) to compensate for the evolutionary loss of its own CXXC domain[148]. The exact role of these domains remains unclear, although they appear to regulate the degradation of TET[148] . They have in any case not been shown to discriminate between different methylated CpGs and are thus unlikely candidates for the tissue-specific regulation of TET function.

The meta-analysis mentioned above offered a unique opportunity to identify this potential DNA-binding partner for TET as the sites identified were hypermethylated due to the inhibition of TET by 2-HG, and as such should have in their vicinity a sequence recognised by that partner. Using motif analysis, ChIP-qPCR and a co-immunoprecipitation experiment followed by western blotting, I identified EBF1 as a candidate binding partner for TET2 in the first targeted demethylation pathway to be described.

Although EBF1 was indeed expressed in the CS tumour samples and cell lines, was shown through modelling analysis to have a structure compatible with interaction with TET2, and recognises a sequence present around the majority of the hypermethylated sites, future experiments should focus on the following questions to fully clarify the mechanism of TET targeting (**Figure 6.1**). Firstly, the demonstrated interaction only proves TET2 and EBF1 co-precipitate as part of a complex, but cannot distinguish between a direct and indirect interaction; moreover, other proteins potentially involved in the complex could play a crucial role in the targeting of TET2 or in supporting its interaction with EBF1 and should be investigated. Mass spectrometry of the EBF1 and TET2 interactomes should shed light on these questions but will require the availability of more specific antibodies. Secondly, EBF1 is unlikely to be the sole protein with this function and is most likely to be itself under post-transcriptional control so it can target TET2 to a subset of CpG sites depending on cellular requirements. Other targeting partners for TET2 should be investigated, and the resolution of its interactome mentioned above should yield useful candidates.

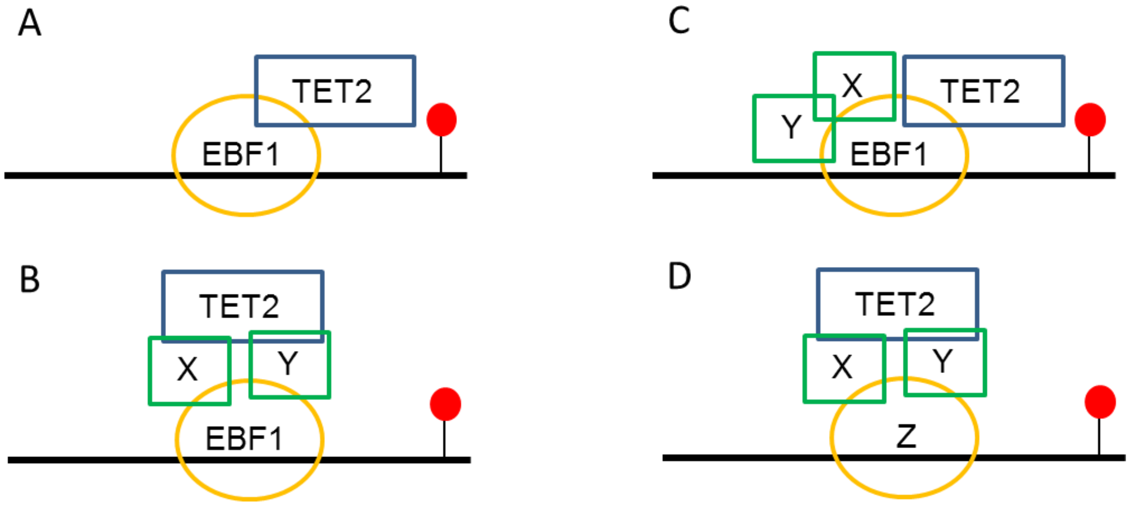


Figure 6.1: Interaction models for targeted demethylation

A) Simple model of EBF1 targeting TET2 to a methylated CpG for demethylation
B) EBF1 and TET2 might not interact directly; other factors might be required
C) Even if the interaction is direct, other factors might be involved
D) EBF1 is unlikely to be the sole protein with this targeting function

6.2.2 Outlook on targeted demethylation therapy

Various attempts have been made in the past to manipulate CpG methylation in both global and targeted ways, with limited success in vivo. DNMT inhibitors (DNMTi), for example, have been used to achieve whole-genome demethylation: the nucleoside analogues, 5-azacytidine and 5-aza-2'-deoxycytidine are FDA-approved in the treatment of myeloid malignancies. The nitrogen on the 5-position of their pyrimidine ring disrupts the interaction between DNA and DNMT and traps DNMTs for proteosomal degradation[14], leading to loss of methylation in daughter cells after replication. However, these nucleoside analogues can be easily hydrolysed in aqueous solution and deaminated by cytidine deaminase, making them unstable and limiting their clinical application. Other cytidine analogues with improved stability, such as zebularine[149], were also investigated but they displayed inefficient metabolic activation or inconsistent clinical efficacy[150]. Non-nucleoside inhibitors have also been developed in the hope that because they do not need to intercalate in the DNA they might be associated with lower toxicity, but they have shown limited hypomethylation activity in living cells[14].

Targeted methylation to repress the expression of specific genes has been attempted using various mechanisms: vander Gun et al.[151] tried to silence epithelial cell adhesion molecule (EpCAM) expression using a mutant version of the CpG-specific DNA methyltransferase M.SssI coupled to a triple-helix-forming oligonucleotide specifically designed for the EpCAM gene. Although they achieved methylation of the targeted CpG, it did not have any effect on the promoter activity. Li et al.[152] fused the catalytic domains of mouse Dnmt3a and Dnmt3b to an engineered zinc finger domain and showed targeted DNA methylation and repression of specific genes involved in Herpes Simplex Virus type 1 (HSV-1) infection in cell culture. Although successful, the authors acknowledged that methylating a specific promoter in culture on a virus genome is significantly different from achieving the same in mammalian cells and *in vivo*.

Targeted demethylation, however, had not been attempted until recently even though it presents distinct advantages over methylation: only one allele of a tumour suppressor gene needs to be demethylated to regain function in the cell, even if the other allele is mutated, while oncogenes will not be repressed unless both alleles are methylated. Moreover, as ~80% of CpG sites in the human genome are methylated, it is much easier to measure and validate a loss than a gain of methylation.

With the advent of novel genomic editing technologies such as transcription activator-like effectors (TALEs) and clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9, targeted alterations in the methylome have recently become more accessible, with the use of DNMT and TET domains already demonstrated in the case of TALEs[153]. Fusions of catalytically inactive Cas9 (dCas9) to transcriptional activation or repression domains (VP64 and KRAB, respectively) have already been shown to successfully affect the expression of human genes[154], and the combination of dCas9 with DNMT and TET domains is currently under development by various groups.

The authors of the study describing the TALE-TET construct raised a major concern regarding the off-target effects they observed. In addition to those due to actual unintended binding of the TALE, they noticed nonspecific demethylation caused by proteins acting from solution. The better understanding of targeted demethylation *in vivo* provided by the work described here could thus improve the specificity of such demethylating tools, by fusing a TET binding partner (or even only the domain interacting with TET) such as EBF1 to the TALE/CRISPR in order to recruit endogenous TET to the chosen site and avoid unintended demethylation from an overexpressed TET in solution.

It is important to note that the work presented here was only possible due to the availability of the Stanmore Musculoskeletal Biobank developed by Prof. Adrienne Flanagan and colleagues. The fundamental issue with investigating rare malignancies is indeed the scarcity of patient samples. This is particularly true in the case of epigenetic studies: higher amounts of starting material are often required in high-throughput technologies compared to standard genomic analyses, for example due to the necessary initial bisulfite conversion and subsequent clean-up steps as in the case of the 450K array, although protocols relying on low cell numbers have been developed[155] and might contribute to alleviating this issue. Targeted approaches also suffer from our current inability to *in vitro* amplify epigenetically-modified DNA (e.g. 5mC, 5hmC) making the finite nature of patient DNA a permanent concern and placing a constant emphasis on experimental optimisation.

Moreover, even when a biobank is available for use in basic research, the low patient numbers remain an issue when attempting to translate a discovery to the clinic. Taking OS as an example, it most often presents in the younger population, the least likely to participate in clinical trials[117], hence patient recruitment can often take several years. This in turn requires an efficient pre-clinical screening of candidate therapeutic agents, which is only possible with an accurate model of the disease in question that can faithfully replicate the *in vivo* conditions of a patient tumour. The use of PDXs, as well as alternative models (e.g. 3D cell culture), for both fundamental research of epigenetic cancer mechanisms and the screening of pre-clinical compounds targeting the epigenome, is discussed in the following section.

6.3 PDXs in epigenetic studies

6.3.1 Tumour cell lines

An inconsistency has emerged between the usefulness of cell lines to further our understanding of cancer biology and the inability to translate that understanding into the clinic. Cell lines certainly present a number of useful characteristics: they are easy to grow in vast numbers to provide large amounts of genomic material; they represent homogeneous populations making results easily reproducible within a cell line; they are easy to manipulate (e.g. genetically, metabolically); and they can faithfully replicate individual aspects of a disease's biology, such as a particular pathway or the interaction of certain proteins, making them ideally suited to the investigation of well-contained mechanisms.

However, those same characteristics can also be drawbacks to using cell lines as representative of patient tumours. The homogeneity of a cell population, for instance, means cultured cells cannot reproduce the heterogeneity of a solid tumour. Furthermore, cell lines, even when subsequently engrafted and grown *in vivo*, have not evolved within the natural tumour environment, and they will have undergone genetic and epigenetic changes to respond to a different set of pressures than those experienced by patient tumour cells[156]. Finally, evidence suggests that a cell line derived from a patient tumour will show a greater divergence in gene expression from that tumour than a xenograft would[156].

6.3.2 Genomics of PDXs

As mentioned above, PDXs have proven particularly apt at replicating the genomic attributes of patient tumours. Most of the evidence for this comes from gene expression experiments. For example, genome-wide gene expression profiling of non-small-cell lung cancer PDXs showed that for 17 patient tumour/PDX pairs, the correlation coefficients ranged from 0.78 to 0.95, with ten of those displaying correlations above 0.9[157]. The gene expression in PDXs for pancreatic cancer was also deemed adequate enough to use some of the PDXs in the corresponding cancer genome sequencing initiative[158]. These advantages over cell lines remain, even when the cells are subsequently returned to an *in vivo* environment, grown into xenografts[156] and compared to PDXs of a similar age, demonstrating that it is the very act of *in vitro* culture that is primarily responsible for the genomic divergence of cell lines.

The only study to date to have investigated the genetic drift associated with xenografting of tumours into mice showed that all CNVs are maintained in PDXs, and that while xenografts do initially present a small number of single nucleotide variants (~4,300), the vast majority of changes that accumulate over time occur in non-coding parts of the genome[127].

6.3.3 PDXs in preclinical drug screening

PDXs have been increasingly used to screen drugs before advancing them to human trials (**Table 6.2**) and xenografts' responses to various compounds have been shown to correlate reasonably well with that of patients. In a study of 15 colorectal cancer PDXs, for example, the xenografts tested for 5-fluorouracil, oxaliplatin or irinotecan exhibited a response concordant with their corresponding patients[159]. In addition, a separate study showed that colorectal PDX models could predict with 90% accuracy the response and resistance to the EGFR inhibitor cetuximab[160].

Tumour model	Approved agent tested	Investigational agent
Pancreatic ductal adenocarcinoma	Gemcitabine, erlotinib	Temsirolimus, saracatinib, bosutinib, MK-1775, IPI-504
NSCLC	Etoposide, carboplatin, gemcitabine, paclitaxel, vinorelbine, cetuximab, erlotinib, docetaxel, docetaxel–vinorelbine, docetaxel–gemcitabine, docetaxel–cisplatin, cisplatin	Sagopilone, diaziquone, pazelliptine, retelliptine
Melanoma	Actinomycin-D, carmustine, doxorubicin bleomycin, cisplatin melphalan, mitomycin-C, vinblastine, cyclophosphamide, ifosfamide, lomustine, 5-FU, methotrexate, etoposide, paclitaxel, vindesine, temozolomide	NA
RCC	Sorafenib, sunitinib	NA
Breast cancer	Doxorubicin, cyclophosphamide, docetaxel, trastuzumab, ifosfamide, cisplatin, capecitabine	Degarelix
HNSCC	Cisplatin, cetuximab	Diaziquone, pazelliptine, retelliptine
GBM	Bevacizumab	NA
Prostate cancer	Bicalutamide	NA
Ovarian cancer	5-FU, cyclophosphamide, doxorubicin, methotrexate, hexamethylmelamine, cisplatin	NA
HCC	5-FU, oxaliplatin, doxorubicin, cisplatin, estradiol, progesterone, dihydrotestosterone	Gefitinib, seocalcitol, brivanib
Abbreviations: 5-FU, 5-fluorouracil; GBM, glioblastoma multiforme; HCC, hepatocellular carcinoma; HNSCC, head and neck squamous-cell carcinoma; NA, not applicable; NSCLC, non-small-cell lung cancer; RCC, renal cell carcinoma.		

Table 6.2: Preclinical drug screening conducted in PDXs

A number of novel and approved drugs have been tested in PDXs from a variety of cancer types. Original table from Tentler et al.[116]

6.3.4 PDXs as accurate epigenomic models of patient tumours

Considering our growing understanding of the role of DNA methylation in tumour development, and the advent of global and targeted epigenetic therapies, it has become necessary to assess whether the epigenome is as stably maintained in PDXs as the genome has been shown to be, so these same tumour models can be used for both basic epigenetic research and preclinical screening of new compounds.

As described in *Chapter 5*, I used both microarray and sequencing-based methods to assess genome-wide variation in methylation between PDXs and the patient tumours they were derived from. This was done on both common and rare cancer types, colon cancer and osteosarcoma, respectively. On average, only 2.7% of assayed CpG sites saw their methylation shift from methylated to unmethylated or vice-versa with xenografting and these variations were tumour-specific. Interestingly, these few changes occurred mainly in intergenic regions, much like the genomic variation mentioned above[127]. However, while single nucleotide variants were found to accumulate with successive xenografting passages in that study[127], almost no further methylation changes were observed in OS and colon cancer samples (< 1%). One of the explanations for the accumulation of genetic variants offered by the authors is 'population bottlenecking' by which the repeated cell population reduction with successive passages may arbitrarily select for passenger mutations. A similar mechanism in the epigenome is possible and two generations of PDXs might not be sufficient to observe it.

Due to the tumour-specificity of the observed changes, no general prediction could be extrapolated regarding which CpG sites are most likely to be affected by xenografting of each individual tumour, Instead, I proposed a model for future studies to estimate the number of individual patient tumours that would need to be xenografted to statistically dilute down the xenografting-associated methylation changes and extract meaningful group characteristics with confidence.

6.3.5 What future for PDXs?

With their accurate representation of the genome and methylome, PDXs will surely continue to be essential tools in both fundamental research and preclinical screening in the near future. However, they still present certain drawbacks that should be considered.

For example, in the case of heterogeneous tumours, the xenografts of a single tumour fragment will be unable to fully replicate the entire tumour and the effects of companion cells with a different genetic or epigenetic make-up in tumour development or drug response might be incorrectly ignored by investigators. In addition, placing a human tumour within a murine environment, surrounded and supported by mouse tissue might have adverse effects on both transplantation success rate, and accurate tumour development; this issue is already being addressed through the co-transplantation of the tumour fragment with stroma from the tumour's original microenvironment[161]. Finally, maintaining live libraries of PDXs for large-scale studies is prohibitively expensive for smaller research groups, and often requires country-specific training and purpose-built facilities, making them a tool unavailable to many researchers.

The constant improvements in sequencing technology might ultimately make the use of xenografts for expansion of patient sample material in basic research redundant. Whole-genome bisulfite sequencing (WGBS), for example, is currently too expensive for routine use, and methylation profiling projects often require large amounts of starting material to use in multiple test and validation experiments. However, considering the cost evolution of sequencing a human genome from ~$3 billion for the first to $1,000 with Illumina's latest technological offering (HiSeq X Ten), it is not inconceivable that WGBS might soon be achievable at an affordable cost.

Finally, the development of 3D tissue culture models that combine the advantages of both PDXs and cell lines while removing some of the concerns associated with xenografts, such as microenvironment and cost, could provide an alternative to PDXs, including in preclinical drug screening[162]. With the development of (epi)genomic editing tools, these models could be further improved to fully recapitulate the heterogeneity of the original malignancy. This would provide the additional advantage of enhancing predictions of tumour evolution as well as drug response due to their close replication of the stresses and pressures the original tumour's cells might be experiencing.

7 REFERENCES

1 Guilhamon, P. *et al.*, Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2, *Nat Commun*,**4**, 2166,(2013)

2 Eccleston, A., DeWitt, N., Gunter, C., Marte, B. & Nath, D., Epigenetics, *Nature*,**447**, 395-395,(2007)

3 Lister, R. *et al.*, Human DNA methylomes at base resolution show widespread epigenomic differences, *Nature*,**462**, 315-322,(2009)

4 Illingworth, R. S. & Bird, A. P., CpG islands--'a rough guide', *FEBS letters*,**583**, 1713-1720,(2009)

5 Gardiner-Garden, M. & Frommer, M., CpG islands in vertebrate genomes, *J Mol Biol*,**196**, 261-282,(1987)

6 Deaton, A. M. & Bird, A., CpG islands and the regulation of transcription, *Genes Dev*,**25**, 1010-1022,(2011)

7 Irizarry, R. A. *et al.*, The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores, *Nat Genet*,**41**, 178-186,(2009)

8 Bibikova, M. *et al.*, High density DNA methylation array with single CpG site resolution, *Genomics*,**98**, 288-295,(2011)

9 Wachter, E. *et al.*, Synthetic CpG islands reveal DNA sequence determinants of chromatin structure, *eLife*,**3**,(2014)

10 Carninci, P. *et al.*, Genome-wide analysis of mammalian promoter architecture and evolution, *Nature Genetics*,**38**, 626-635,(2006)

11 Ramirez-Carrozzi, V. R. *et al.*, A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling, *Cell*,**138**, 114-128,(2009)

12 Birney, E. *et al.*, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*,**447**, 799-816,(2007)

13 Sado, T. *et al.*, X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation, *Dev Biol*,**225**, 294-303,(2000)

14 Yang, X., Lay, F., Han, H. & Jones, P. A., Targeting DNA methylation for epigenetic therapy, *Trends Pharmacol Sci*,**31**, 536-546,(2010)

15 Sasaki, H. & Matsui, Y., Epigenetic events in mammalian germ-cell development: reprogramming and beyond, *Nature reviews. Genetics*,**9**, 129-140,(2008)

16 Ito, S. *et al.*, Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification, *Nature*,**466**, 1129-1133,(2010)

17 He, Y. F. *et al.*, Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA, *Science*,**333**, 1303-1307,(2011)

18 Koh, K. P. *et al.*, Tet1 and tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells, *Cell Stem Cell*,**8**, 200-213,(2011)

19 Hu, L. *et al.*, Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation, *Cell*,**155**, 1545-1555,(2013)

20 Wu, H. & Zhang, Y., Reversing DNA methylation: mechanisms, genomics, and biological functions, *Cell*,**156**, 45-68,(2014)

21 Kohli, R. M. & Zhang, Y., TET enzymes, TDG and the dynamics of DNA demethylation, *Nature*,**502**, 472-479,(2013)

22 Xu, L. *et al.*, A Chemical Probe Targets DNA 5-Formylcytosine Sites and Inhibits TDG Excision, Polymerases Bypass, and Gene Expression, *Chemical science*,**5**, 567-574,(2014)

23 Jones, P. A. & Baylin, S. B., The fundamental role of epigenetic events in cancer, *Nat Rev Genet*,**3**, 415-428,(2002)

24 Figueroa, M. E. *et al.*, Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation, *Cancer Cell*,**18**, 553-567,(2010)

25 Combs, S. E. *et al.*, Prognostic significance of IDH-1 and MGMT in patients with glioblastoma: one step forward, and one step back?, *Radiat Oncol*,**6**, 115,(2011)

26 Parsons, D. W. *et al.*, An Integrated Genomic Analysis of Human Glioblastoma Multiforme, *Science*,**321**, 1807-1812,(2008)

27 Dang, L. *et al.*, Cancer-associated IDH1 mutations produce 2-hydroxyglutarate, *Nature*,**462**, 739-744,(2009)

28 Xu, W. *et al.*, Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of alpha-ketoglutarate-dependent dioxygenases, *Cancer Cell*,**19**, 17-30,(2011)

29 Turcan, S. *et al.*, IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype, *Nature*,**483**, 479-483,(2012)

30 Dang, L., Jin, S. & Su, S. M., IDH mutations in glioma and acute myeloid leukemia, *Trends Mol Med*,**16**, 387-397,(2010)

31 Wang, P. *et al.*, Mutations in isocitrate dehydrogenase 1 and 2 occur frequently in intrahepatic cholangiocarcinomas and share hypermethylation targets with glioblastomas, *Oncogene*,**32**, 3091-3100,(2012)

32 Kurek, K. C. *et al.*, R132C IDH1 Mutations Are Found in Spindle Cell Hemangiomas and Not in Other Vascular Tumors or Malformations, *The American Journal of Pathology*,**182**, 1494-1500,(2013)

33 Reitman, Z. J., Parsons, D. W. & Yan, H., IDH1 and IDH2: not your typical oncogenes, *Cancer Cell*,**17**, 215-216,(2010)

34 Dedeurwaerder, S. *et al.*, A comprehensive overview of Infinium HumanMethylation450 data processing, *Brief Bioinform*,(2013)

35 Langaee, T. & Ronaghi, M., Genetic variation analyses by Pyrosequencing, *Mutat Res*,**573**, 96-102,(2005)

36 Margulies, M. *et al.*, Genome sequencing in microfabricated high-density picolitre reactors, *Nature*,**437**, 376-380,(2005)

37 Schatz, P., Dietrich, D. & Schuster, M., Rapid analysis of CpG methylation patterns using RNase T1 cleavage and MALDI-TOF, *Nucleic Acids Research*,**32**, e167,(2004)

38 Mirmohammadsadegh, A. *et al.*, Epigenetic silencing of the PTEN gene in melanoma, *Cancer research*,**66**, 6546-6552,(2006)

39 Hung, C. C. *et al.*, Quantitative and qualitative analyses of the SNRPN gene using real-time PCR with melting curve analysis, *The Journal of molecular diagnostics : JMD*,**13**, 609-613,(2011)

40 Tost, J. & Gut, I. G., DNA methylation analysis by pyrosequencing, *Nat Protoc*,**2**, 2265-2275,(2007)

41 Taiwo, O. *et al.*, Methylome analysis using MeDIP-seq with low DNA concentrations, *Nature protocols*,**7**, 617-636,(2012)

42 Butcher, L. M. & Beck, S., AutoMeDIP-seq: A high-throughput, whole genome, DNA methylation assay, *Methods*,**52**, 223-231,(2010)

43 Beck, S., Taking the measure of the methylome, *Nat Biotechnol*,**28**, 1026-1028,(2010)

44 Down, T. A. *et al.*, A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis, *Nat Biotechnol*,**26**, 779-785,(2008)

45 Li, N. *et al.*, Whole genome DNA methylation analysis based on high throughput sequencing technology, *Methods*,**52**, 203-212,(2010)

46 Feber, A. *et al.*, Comparative methylome analysis of benign and malignant peripheral nerve sheath tumours, *Genome Res*,**21**, 515-524,(2011)

47 Komori, H. K. *et al.*, Application of microdroplet PCR for large-scale targeted bisulfite sequencing, *Genome research*,**21**, 1738-1745,(2011)

48 Herrmann, A. *et al.*, Pipeline for large-scale microdroplet bisulfite PCR-based sequencing allows the tracking of hepitype evolution in tumors, *PLoS One*,**6**, e21332,(2011)

49 Paul, D. S. *et al.*, Assessment of RainDrop BS-seq as a method for large-scale, targeted bisulfite sequencing, *Epigenetics*,**9**, 678-684,(2014)

50 Denk, F. & McMahon, S. B., Chronic pain: emerging evidence for the involvement of epigenetics, *Neuron*,**73**, 435-444,(2012)

51 Bleyer, A., O'Leary, M., Barr, R. & Ries, L., Cancer Epidemiology in Older Adolescents and Young Adults 15 to 29 Years of Age, Including SEER Incidence and Survival: 1975-2000, *National Cancer Institute, NIH Pub. No. 06-5767*,(2006)

52 McNally, R. J. *et al.*, Small-area analyses of bone cancer diagnosed in Great Britain provide clues to aetiology, *BMC Cancer*,**12**, 270,(2012)

53 Ottaviani, G. & Jaffe, N. in *Pediatric and Adolescent Osteosarcoma* Vol. 152 (ed Norman Jaffe) 15-32 (Springer, 2010).

54 Martin, J. W., Squire, J. A. & Zielenska, M., The genetics of osteosarcoma, *Sarcoma*,**2012**, 627254,(2012)

55 Cleton-Jansen, A. M. in *Bone and Soft Tissue Pathology* Vol. 11 (ed Elsevier) 390-399 (Current Diagnostic Pathology, 2005).

56 Kresse, S. H. *et al.*, Integrative analysis reveals relationships of genetic and epigenetic alterations in osteosarcoma, *PLoS One*,**7**, e48262,(2012)

57 Kansara, M. & Thomas, D. M., Molecular pathogenesis of osteosarcoma, *DNA Cell Biol*,**26**, 1-18,(2007)

58 Sadikovic, B. *et al.*, Identification of interactive networks of gene expression associated with osteosarcoma oncogenesis by integrated molecular profiling, *Hum Mol Genet*,**18**, 1962-1975,(2009)

59 Biermann, J. S. *et al.*, Bone Cancer, *Journal of the National Comprehensive Cancer Network*,**8**, 688-712,(2010)

60 Sakamoto, A. *et al.*, Dedifferentiated chondrosarcoma with leukocytosis and elevation of serum G-CSF. A case report, *World J Surg Oncol*,**4**, 37,(2006)

61 Tarpey, P. S. *et al.*, Frequent mutation of the major cartilage collagen gene COL2A1 in chondrosarcoma, *Nature Genetics*,**45**, 923-926,(2013)

62 Hallor, K. H. *et al.*, Genomic profiling of chondrosarcoma: chromosomal patterns in central and peripheral tumors, *Clinical cancer research : an official journal of the American Association for Cancer Research*,**15**, 2685-2694,(2009)

63 Amary, M. F. *et al.*, IDH1 and IDH2 mutations are frequent events in central chondrosarcoma and central and periosteal chondromas but not in other mesenchymal tumours, *J Pathol*,**224**, 334-343,(2011)

64 Amary, M. F. *et al.*, Ollier disease and Maffucci syndrome are caused by somatic mosaic mutations of IDH1 and IDH2, *Nature genetics*,**43**, 1262-1265,(2011)

65 Bolstad, B. M., preprocessCore: A collection of pre-processing functions, *R package version 1.18.0*

66 Yang, Y. H., marray: Exploratory analysis for two-color spotted microarray data, *R package version 1.34.0*,(2009)

67 Day, A., heatmap.plus: Heatmap with more sensible behavior, *R package version 1.3*,(2007)

68 Warnes, G. R., gtools: Various R programming tools, *R package version 2.7.0*,(2012)

69 Teschendorff, A. E. *et al.*, An epigenetic signature in peripheral blood predicts active ovarian cancer, *PLoS One*,**4**, e8274,(2009)

70 Benjamini, Y. & Hochberg, Y., Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing, *J Roy Stat Soc B Met*,**57**, 289-300,(1995)

71 Morris, T. J. *et al.*, ChAMP: 450k Chip Analysis Methylation Pipeline, *Bioinformatics*,**30**, 428-430,(2014)

72 Feber, A. *et al.*, Using high-density DNA methylation arrays to profile copy number alterations, *Genome Biol*,**15**, R30,(2014)

73 Mermel, C. H. *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)

74 Krueger, F., Kreck, B., Franke, A. & Andrews, S. R., DNA methylome analysis using short bisulfite sequencing data, *Nat Methods*,**9**, 145-151,(2012)

75 Krueger, F. & Andrews, S. R., Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications, *Bioinformatics*,**27**, 1571-1572,(2011)

76 Olshen, V. E. S. a. A., DNAcopy: DNA copy number data analysis, *R package version 1.36.0*,(2007)

77 Smyth, G. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Gentleman R *et al.*) 397-420 (Springer, 2005).

78 Ritchie, M. E., Dunning, M. J., Smith, M. L., Shi, W. & Lynch, A. G., BeadArray expression analysis using bioconductor, *PLoS Comput Biol*,**7**, e1002276,(2011)

79 Bailey, T. L. *et al.*, MEME SUITE: tools for motif discovery and searching, *Nucleic acids research*,**37**, W202-208,(2009)

80 Teschendorff, A. E. *et al.*, A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data, *Bioinformatics*,**29**, 189-196,(2013)

81 Wilson, G. A. *et al.*, Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers, *GigaScience*,**1**, 3,(2012)

82 Portela, A. & Esteller, M., Epigenetic modifications and human disease, *Nat Biotechnol*,**28**, 1057-1068,(2010)

83 Dunning, M., illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4), *R package version 1.22.1*,(2014)

84 Larramendy, M. L. *et al.*, Gains, losses, and amplifications of DNA sequences evaluated by comparative genomic hybridization in chondrosarcomas, *The American Journal of Pathology*,**150**, 685-691,(1997)

85 Larramendy, M. L. *et al.*, Clinical significance of genetic imbalances revealed by comparative genomic hybridization in chondrosarcomas, *Human pathology*,**30**, 1247-1253,(1999)

86 Bovee, J. V. *et al.*, Molecular genetic characterization of both components of a dedifferentiated chondrosarcoma, with implications for its histogenesis, *The Journal of pathology*,**189**, 454-462,(1999)

87 Katoh, M., FNBP2 gene on human chromosome 1q32.1 encodes ARHGAP family protein with FCH, FBH, RhoGAP and SH3 domains, *Int J Mol Med*,**11**, 791-797,(2003)

88 Schorle, C. M. *et al.*, Comparative analysis of imbalances in genomic DNA and mRNA expression levels in chondrosarcoma-derived cell line FSCP-1, *Int J Oncol*,**25**, 1651-1660,(2004)

89 Pansuriya, T. C. *et al.*, Genome-wide analysis of Ollier disease: Is it all in the genes?, *Orphanet J Rare Dis*,**6**, 2,(2011)

90 Rozeman, L. B. *et al.*, Array-comparative genomic hybridization of central chondrosarcoma: identification of ribosomal protein S6 and cyclin-dependent kinase 4 as candidate target genes for genomic aberrations, *Cancer*,**107**, 380-388,(2006)

91 Pansuriya, T. C. *et al.*, Maffucci syndrome: A genome-wide analysis using high resolution single nucleotide polymorphism and expression arrays on four cases, *Genes, chromosomes & cancer*,**50**, 673-679,(2011)

92 Contie, S., Voorzanger-Rousselot, N., Litvin, J., Clezardin, P. & Garnero, P., Increased expression and serum levels of the stromal cell-secreted protein periostin in breast cancer bone metastases, *International journal of cancer. Journal international du cancer*,**128**, 352-360,(2011)

93 Roach, J. C. *et al.*, Analysis of genetic inheritance in a family quartet by whole-genome sequencing, *Science*,**328**, 636-639,(2010)

94 Green, C. L. *et al.*, The prognostic significance of IDH2 mutations in AML depends on the location of the mutation, *Blood*,**118**, 409-412,(2011)

95 Siegfried, Z. & Cedar, H., DNA methylation: a molecular lock, *Current biology : CB*,**7**, R305-307,(1997)

96 Ye, D., Xiong, Y. & Guan, K. L., The mechanisms of IDH mutations in tumorigenesis, *Cell research*,**22**, 1102-1104,(2012)

97 Khulan, B. *et al.*, Comparative isoschizomer profiling of cytosine methylation: the HELP assay, *Genome research*,**16**, 1046-1055,(2006)

98 Tang, X. H. & Gudas, L. J., Retinoids, retinoic acid receptors, and cancer, *Annu Rev Pathol*,**6**, 345-364,(2011)

99 Soprano, K. J. & Soprano, D. R., Retinoic acid receptors and cancer, *J Nutr*,**132**, 3809S-3813S,(2002)

100 Duncan, C. G. *et al.*, A heterozygous IDH1R132H/WT mutation induces genome-wide alterations in DNA methylation, *Genome research*,**22**, 2339-2355,(2012)

101 Eckhardt, F. *et al.*, DNA methylation profiling of human chromosomes 6, 20 and 22, *Nature genetics*,**38**, 1378-1385,(2006)

102 Li, Y. *et al.*, The DNA methylome of human peripheral blood mononuclear cells, *PLoS Biol*,**8**, e1000533,(2010)

103 Maier, H. *et al.*, Early B cell factor cooperates with Runx1 and mediates epigenetic changes associated with mb-1 transcription, *Nat Immunol*,**5**, 1069-1077,(2004)

104 Treiber, T. *et al.*, Early B cell factor 1 regulates B cell gene networks by activation, repression, and transcription- independent poising of chromatin, *Immunity*,**32**, 714-725,(2010)

105 Rosenbloom, K. R. *et al.*, ENCODE data in the UCSC Genome Browser: year 5 update, *Nucleic Acids Res*,**41**, D56-63,(2013)

106 Flicek, P. *et al.*, Ensembl 2014, *Nucleic Acids Res*,**42**, D749-755,(2014)

107 Siponen, M. I. *et al.*, Structural determination of functional domains in early B-cell factor (EBF) family of transcription factors reveals similarities to Rel

DNA-binding proteins and a novel dimerization motif, *The Journal of biological chemistry*,**285**, 25875-25879,(2010)

108 Treiber, N., Treiber, T., Zocher, G. & Grosschedl, R., Structure of an Ebf1:DNA complex reveals unusual DNA recognition and structural homology with Rel proteins, *Genes Dev*,**24**, 2270-2275,(2010)

109 Pierce, B. G. *et al.*, ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers, *Bioinformatics*,**30**, 1771-1773,(2014)

110 Auclair, G. & Weber, M., Mechanisms of DNA methylation and demethylation in mammals, *Biochimie*,**94**, 2202-2211,(2012)

111 Klug, M. *et al.*, Active DNA demethylation in human postmitotic cells correlates with activating histone modifications, but not transcription levels, *Genome Biol*,**11**, R63,(2010)

112 Rygaard, J. & Povlsen, C. O., Heterotransplantation of a human malignant tumour to "Nude" mice, *Acta pathologica et microbiologica Scandinavica*,**77**, 758-760,(1969)

113 Sausville, E. A. & Burger, A. M., Contributions of human tumor xenografts to anticancer drug development, *Cancer Res*,**66**, 3351-3354, discussion 3354,(2006)

114 Fiebig, H. H., Maier, A. & Burger, A. M., Clonogenic assay with established human tumour xenografts: correlation of in vitro to in vivo activity as a basis for anticancer drug discovery, *Eur J Cancer*,**40**, 802-820,(2004)

115 Jin, K. *et al.*, Patient-derived human tumour tissue xenografts in immunodeficient mice: a systematic review, *Clin Transl Oncol*,**12**, 473-480,(2010)

116 Tentler, J. J. *et al.*, Patient-derived tumour xenografts as models for oncology drug development, *Nature reviews. Clinical oncology*,**9**, 338-350,(2012)

117 Janeway, K. A. & Walkley, C. R., Modeling human osteosarcoma in the mouse: From bedside to bench, *Bone*,**47**, 859-865,(2010)

118 Siolas, D. & Hannon, G. J., Patient-derived tumor xenografts: transforming clinical samples into mouse models, *Cancer Res*,**73**, 5315-5319,(2013)

119 Henderson, D. *et al.*, OncoTrack: Personalized medicine approaches for colon cancer driven by genomics and systems biology, *Biotechnology Journal*,(2014, In Press)

120 Martinez-Garcia, R. *et al.*, Transcriptional dissection of pancreatic tumors engrafted in mice, *Genome Med*,**6**, 27,(2014)

121 Hennessey, P. T. *et al.*, Promoter methylation in head and neck squamous cell carcinoma cell lines is significantly different than methylation in primary tumors and xenografts, *PLoS One*,**6**, e20584,(2011)

122 Conway, T. *et al.*, Xenome--a tool for classifying reads from xenograft samples, *Bioinformatics*,**28**, i172-i178,(2012)

123 McLean, C. Y. *et al.*, GREAT improves functional interpretation of cis-regulatory regions, *Nature biotechnology*,**28**, 495-501,(2010)

124 Thomas, P. D. *et al.*, PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification, *Nucleic acids research*,**31**, 334-341,(2003)

125 Huang da, W., Sherman, B. T. & Lempicki, R. A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc*,**4**, 44-57,(2009)

126 Lowe, R. & Rakyan, V. K., Marmal-aid--a database for Infinium HumanMethylation450, *BMC Bioinformatics*,**14**, 359,(2013)

127 Li, S. *et al.*, Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts, *Cell Rep*,**4**, 1116-1130,(2013)

128 Levenson, V. V. & Melnikov, A. A., DNA methylation as clinically useful biomarkers-light at the end of the tunnel, *Pharmaceuticals*,**5**, 94-113,(2012)

129 Wang, F. *et al.*, Targeted inhibition of mutant IDH2 in leukemia cells induces cellular differentiation, *Science*,**340**, 622-626,(2013)

130 Gerlinger, M. *et al.*, Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing, *Nat Genet*,**46**, 225-233,(2014)

131 Nyga, A., Cheema, U. & Loizidou, M., 3D tumour models: novel in vitro approaches to cancer studies, *Journal of cell communication and signaling*,**5**, 239-248,(2011)

132 Wagner, J. R. *et al.*, The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts, *Genome Biol*,**15**, R37,(2014)

133 Aran, D. & Hellman, A., DNA methylation of transcriptional enhancers and cancer predisposition, *Cell*,**154**, 11-13,(2013)

134 Killian, J. K. *et al.*, Succinate dehydrogenase mutation underlies global epigenomic divergence in gastrointestinal stromal tumor, *Cancer Discov*,**3**, 648-657,(2013)

135 Letouze, E. *et al.*, SDH mutations establish a hypermethylator phenotype in paraganglioma, *Cancer Cell*,**23**, 739-752,(2013)

136 Xiao, M. *et al.*, Inhibition of alpha-KG-dependent histone and DNA demethylases by fumarate and succinate that are accumulated in mutations of FH and SDH tumor suppressors, *Genes Dev*,**26**, 1326-1338,(2012)

137 Haller, F. *et al.*, Aberrant DNA hypermethylation of SDHC: a novel mechanism of tumor development in Carney triad, *Endocrine-Related Cancer*,**21**, 567-577,(2014)

138 Chowdhury, R. *et al.*, The oncometabolite 2-hydroxyglutarate inhibits histone lysine demethylases, *EMBO Rep*,**12**, 463-469,(2011)

139 Losman, J. A. & Kaelin, W. G., Jr., What a difference a hydroxyl makes: mutant IDH, (R)-2-hydroxyglutarate, and cancer, *Genes Dev*,**27**, 836-852,(2013)

140 Yan, H. *et al.*, IDH1 and IDH2 mutations in gliomas, *The New England journal of medicine*,**360**, 765-773,(2009)

141 Abdel-Wahab, O., Patel, J. & Levine, R. L., Clinical implications of novel mutations in epigenetic modifiers in AML, *Hematol Oncol Clin North Am*,**25**, 1119-1133,(2011)

142 Rakheja, D., Konoplev, S., Medeiros, L. J. & Chen, W., IDH mutations in acute myeloid leukemia, *Human pathology*,**43**, 1541-1551,(2012)

143 Zhou, K. G. *et al.*, Potential application of IDH1 and IDH2 mutations as prognostic indicators in non-promyelocytic acute myeloid leukemia: a meta-analysis, *Leuk Lymphoma*,**53**, 2423-2429,(2012)

144 Gross, S. *et al.*, Cancer-associated metabolite 2-hydroxyglutarate accumulates in acute myelogenous leukemia with isocitrate dehydrogenase 1 and 2 mutations, *J Exp Med*,**207**, 339-344,(2010)

145 Aydin, K., Ozmen, M., Tatli, B. & Sencer, S., Single-voxel MR spectroscopy and diffusion-weighted MRI in two patients with l-2-hydroxyglutaric aciduria, *Pediatr Radiol*,**33**, 872-876,(2003)

146 Kotz, J., Oncometabolite takedown, *Science-Business eXchange*,**14**,(2013)

147 Rohle, D. *et al.*, An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells, *Science*,**340**, 626-630,(2013)

148 Ko, M. *et al.*, Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX, *Nature*,**497**, 122-126,(2013)

149 Yoo, C. B., Cheng, J. C. & Jones, P. A., Zebularine: a new drug for epigenetic therapy, *Biochem Soc Trans*,**32**, 910-912,(2004)

150 Kratzke, R. A. *et al.*, Response to the methylation inhibitor dihydro-5-azacytidine in mesothelioma is not associated with methylation of p16(INK4a) - Results of cancer and leukemia group B 159904, *Journal of Thoracic Oncology*,**3**, 417-421,(2008)

151 van der Gun, B. T. *et al.*, Targeted DNA methylation by a DNA methyltransferase coupled to a triple helix forming oligonucleotide to down-regulate the epithelial cell adhesion molecule, *Bioconjug Chem*,**21**, 1239-1245,(2010)

152 Li, F. *et al.*, Chimeric DNA methyltransferases target DNA methylation to specific DNA sequences and repress expression of target genes, *Nucleic acids research*,**35**, 100-112,(2007)

153 Maeder, M. L. *et al.*, Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins, *Nat Biotechnol*,**31**, 1137-1142,(2013)

154 Sander, J. D. & Joung, J. K., CRISPR-Cas systems for editing, regulating and targeting genomes, *Nat Biotechnol*,**32**, 347-355,(2014)

155 Smallwood, S. A. *et al.*, Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity, *Nat Methods*,**11**, 817-820,(2014)

156 Daniel, V. C. *et al.*, A primary xenograft model of small-cell lung cancer reveals irreversible changes in gene expression imposed by culture in vitro, *Cancer research*,**69**, 3364-3373,(2009)

157 Fichtner, I. *et al.*, Establishment of patient-derived non-small cell lung cancer xenografts as models for the identification of predictive biomarkers,

Clinical cancer research : an official journal of the American Association for Cancer Research,**14**, 6456-6468,(2008)

158 Jones, S. *et al.*, Core signaling pathways in human pancreatic cancers revealed by global genomic analyses, *Science*,**321**, 1801-1806,(2008)

159 Fichtner, I. *et al.*, Anticancer drug response and expression of molecular markers in early-passage xenotransplanted colon carcinomas, *Eur J Cancer*,**40**, 298-307,(2004)

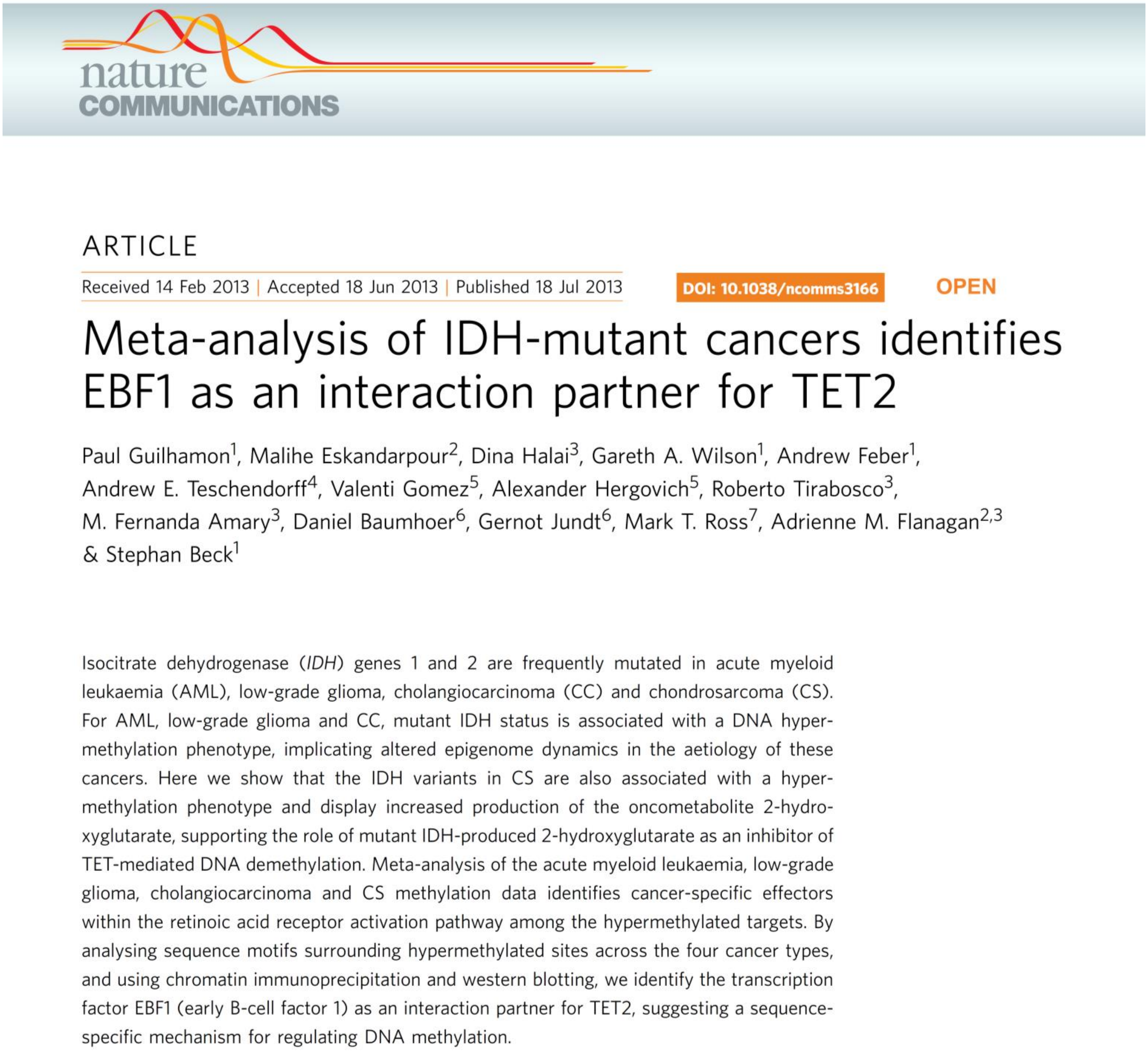
160 Krumbach, R. *et al.*, Primary resistance to cetuximab in a panel of patient-derived tumour xenograft models: activation of MET as one mechanism for drug resistance, *Eur J Cancer*,**47**, 1231-1243,(2011)

161 Richmond, A. & Su, Y., Mouse xenograft models vs GEM models for human cancer therapeutics, *Dis Model Mech*,**1**, 78-82,(2008)

162 Hongisto, V. *et al.*, High-throughput 3D screening reveals differences in drug sensitivities between culture models of JIMT1 breast cancer cells, *PLoS One*,**8**, e77232,(2013)

8 Appendices

Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2, Guilhamon et al., Nature Communications (2013)



nature COMMUNICATIONS

ARTICLE

Received 14 Feb 2013 | Accepted 18 Jun 2013 | Published 18 Jul 2013

DOI: 10.1038/ncomms3166

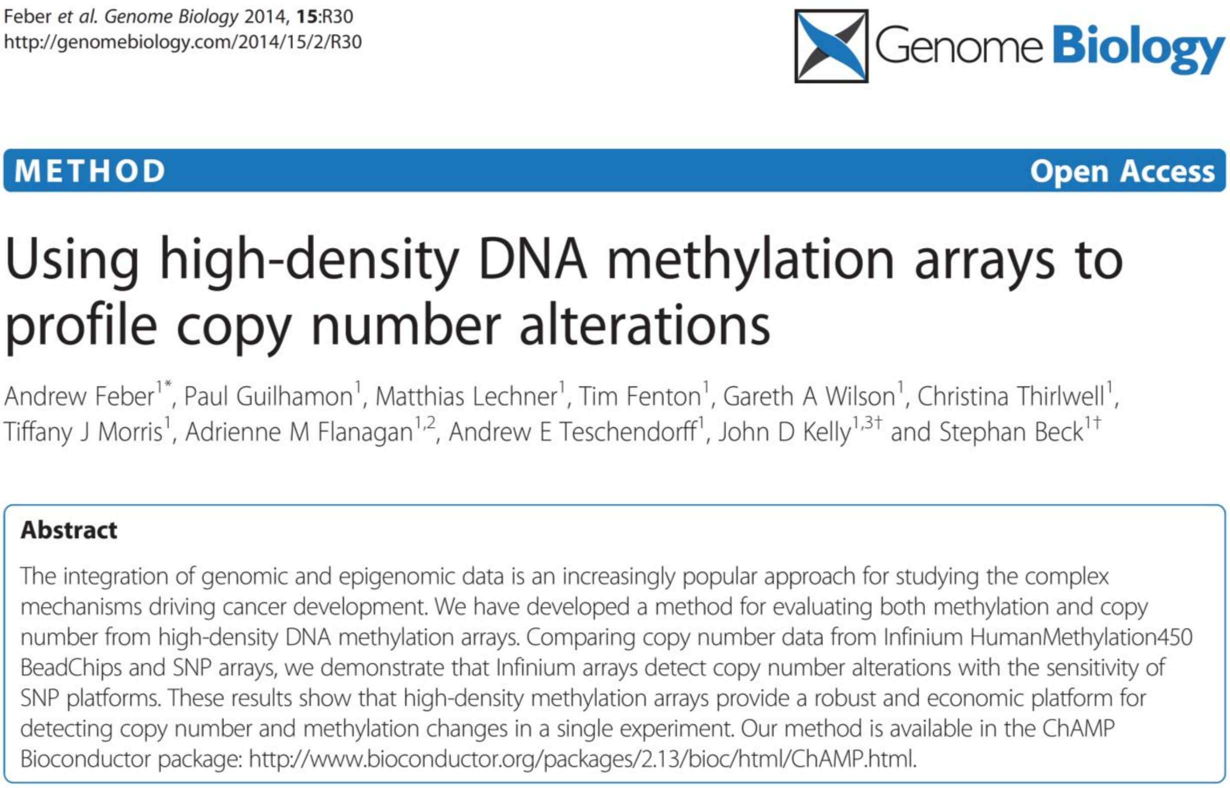
OPEN

Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2

Paul Guilhamon[1], Malihe Eskandarpour[2], Dina Halai[3], Gareth A. Wilson[1], Andrew Feber[1], Andrew E. Teschendorff[4], Valenti Gomez[5], Alexander Hergovich[5], Roberto Tirabosco[3], M. Fernanda Amary[3], Daniel Baumhoer[6], Gernot Jundt[6], Mark T. Ross[7], Adrienne M. Flanagan[2,3] & Stephan Beck[1]

Isocitrate dehydrogenase (*IDH*) genes 1 and 2 are frequently mutated in acute myeloid leukaemia (AML), low-grade glioma, cholangiocarcinoma (CC) and chondrosarcoma (CS). For AML, low-grade glioma and CC, mutant IDH status is associated with a DNA hypermethylation phenotype, implicating altered epigenome dynamics in the aetiology of these cancers. Here we show that the IDH variants in CS are also associated with a hypermethylation phenotype and display increased production of the oncometabolite 2-hydroxyglutarate, supporting the role of mutant IDH-produced 2-hydroxyglutarate as an inhibitor of TET-mediated DNA demethylation. Meta-analysis of the acute myeloid leukaemia, low-grade glioma, cholangiocarcinoma and CS methylation data identifies cancer-specific effectors within the retinoic acid receptor activation pathway among the hypermethylated targets. By analysing sequence motifs surrounding hypermethylated sites across the four cancer types, and using chromatin immunoprecipitation and western blotting, we identify the transcription factor EBF1 (early B-cell factor 1) as an interaction partner for TET2, suggesting a sequence-specific mechanism for regulating DNA methylation.

Using high-density DNA methylation arrays to profile copy number alterations, Feber et al., Genome Biology (2014)



Feber *et al. Genome Biology* 2014, **15**:R30
http://genomebiology.com/2014/15/2/R30

Genome **Biology**

METHOD **Open Access**

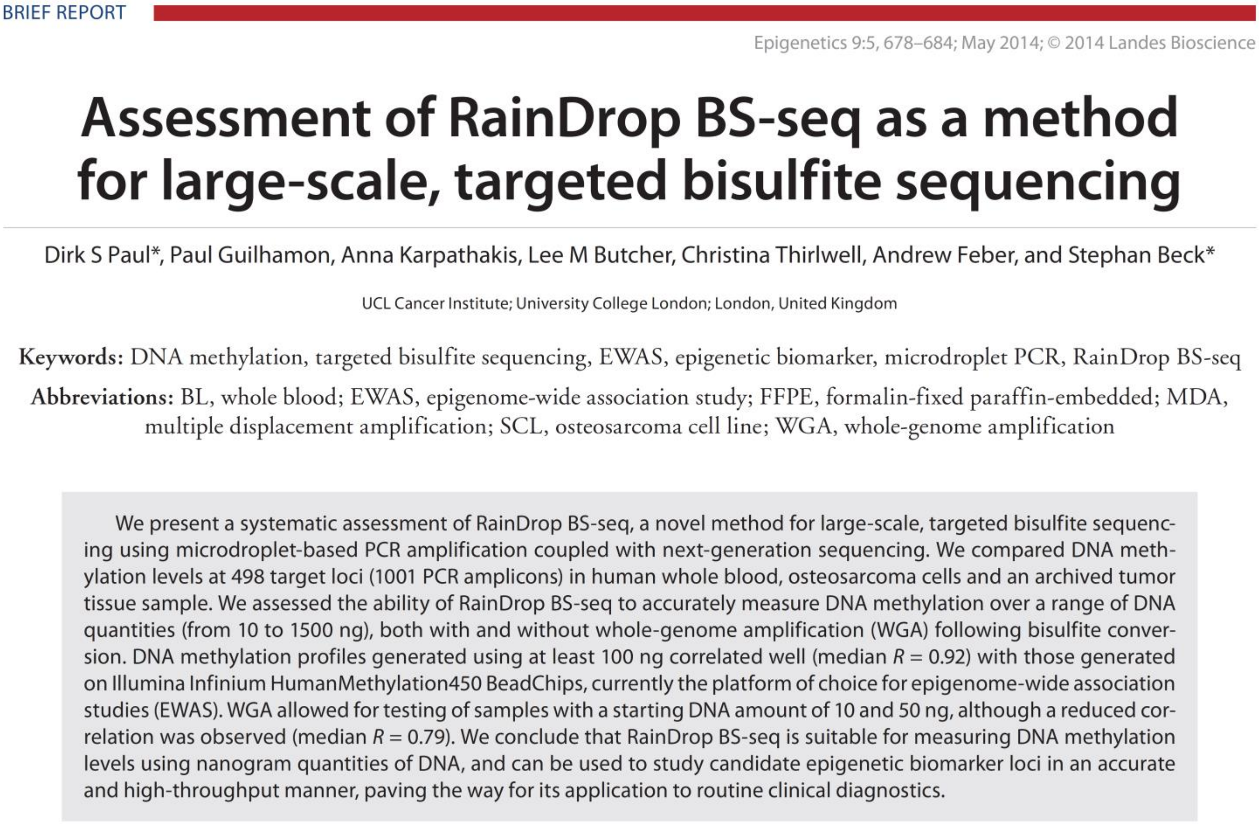
Using high-density DNA methylation arrays to profile copy number alterations

Andrew Feber[1*], Paul Guilhamon[1], Matthias Lechner[1], Tim Fenton[1], Gareth A Wilson[1], Christina Thirlwell[1], Tiffany J Morris[1], Adrienne M Flanagan[1,2], Andrew E Teschendorff[1], John D Kelly[1,3†] and Stephan Beck[1†]

Abstract

The integration of genomic and epigenomic data is an increasingly popular approach for studying the complex mechanisms driving cancer development. We have developed a method for evaluating both methylation and copy number from high-density DNA methylation arrays. Comparing copy number data from Infinium HumanMethylation450 BeadChips and SNP arrays, we demonstrate that Infinium arrays detect copy number alterations with the sensitivity of SNP platforms. These results show that high-density methylation arrays provide a robust and economic platform for detecting copy number and methylation changes in a single experiment. Our method is available in the ChAMP Bioconductor package: http://www.bioconductor.org/packages/2.13/bioc/html/ChAMP.html.

Assessment of RainDrop BS-seq as a method for large-scale, targeted bisulfite sequencing, Paul et al., Epigenetics (2014)



BRIEF REPORT

Epigenetics 9:5, 678–684; May 2014; © 2014 Landes Bioscience

Assessment of RainDrop BS-seq as a method for large-scale, targeted bisulfite sequencing

Dirk S Paul*, Paul Guilhamon, Anna Karpathakis, Lee M Butcher, Christina Thirlwell, Andrew Feber, and Stephan Beck*

UCL Cancer Institute; University College London; London, United Kingdom

Keywords: DNA methylation, targeted bisulfite sequencing, EWAS, epigenetic biomarker, microdroplet PCR, RainDrop BS-seq

Abbreviations: BL, whole blood; EWAS, epigenome-wide association study; FFPE, formalin-fixed paraffin-embedded; MDA, multiple displacement amplification; SCL, osteosarcoma cell line; WGA, whole-genome amplification

We present a systematic assessment of RainDrop BS-seq, a novel method for large-scale, targeted bisulfite sequencing using microdroplet-based PCR amplification coupled with next-generation sequencing. We compared DNA methylation levels at 498 target loci (1001 PCR amplicons) in human whole blood, osteosarcoma cells and an archived tumor tissue sample. We assessed the ability of RainDrop BS-seq to accurately measure DNA methylation over a range of DNA quantities (from 10 to 1500 ng), both with and without whole-genome amplification (WGA) following bisulfite conversion. DNA methylation profiles generated using at least 100 ng correlated well (median *R* = 0.92) with those generated on Illumina Infinium HumanMethylation450 BeadChips, currently the platform of choice for epigenome-wide association studies (EWAS). WGA allowed for testing of samples with a starting DNA amount of 10 and 50 ng, although a reduced correlation was observed (median *R* = 0.79). We conclude that RainDrop BS-seq is suitable for measuring DNA methylation levels using nanogram quantities of DNA, and can be used to study candidate epigenetic biomarker loci in an accurate and high-throughput manner, paving the way for its application to routine clinical diagnostics.

Assessment of tumour xenografts as a discovery tool for cancer epigenomics, Guilhamon et al., Genome Medicine (Under Review)

Paul Guilhamon[1], Lee M. Butcher[1], Nadege Presneau[2,3], Gareth A. Wilson[1,4], Andrew Feber[1], Dirk S. Paul[1], Moritz Schütte[5], Johannes Haybaeck[6], Ulrich Keilholz[7], Jens Hoffman[8], Mark T. Ross[9], Adrienne M. Flanagan[2,10], Stephan Beck[1*]

[1]Medical Genomics, UCL Cancer Institute, University College London, London, UK

[2]Genetics and Cell Biology of Sarcoma, UCL Cancer Institute, University College London, London, UK

[3]Department of Biomedical Sciences, University of Westminster, London, UK

[4]Translational Cancer Therapeutics Laboratory, CR-UK London Research Institute, London, UK

[5]Alacris Theranostics GmbH, Berlin, DE

[6]Institute of Pathology, Medical University of Graz, Graz, Austria

[7]Department of Hematology and Medical Oncology, Charité Comprehensive Cancer Center, Berlin, Germany

[8]EPO-Berlin-Buch GmbH, Berlin, DE

[9]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, UK

[10]Department of Histopathology, Royal National Orthopaedic Hospital NHS Trust, Stanmore, Middlesex, UK

Abstract

Background: The use of tumour xenografts is a well-established research tool in cancer genomics but has not yet been comprehensively evaluated for cancer epigenomics. Methods: In this study, we assessed the suitability of patient-derived tumour xenografts (PDXs) for methylome analysis using Infinium 450K Beadchips and MeDIP-seq.

Results: Controlled for confounding host (mouse) sequences, comparison of primary PDXs and matching patient tumours in a rare (osteosarcoma) and common (colon) cancer revealed that an average 2.7% of the assayed CpG sites undergo major ($\Delta\beta \geq 0.51$) methylation changes in a cancer-specific manner as a result of the xenografting procedure. No significant subsequent methylation changes were observed after a second round of xenografting between primary and secondary PDXs. Based on computational simulation using publically available methylation data, we additionally show that future studies comparing two groups of PDXs should use 15 or more samples in each group to minimise the impact of xenografting-associated changes in methylation on comparison results.

Conclusions: Our results from rare and common cancers indicate that PDXs are a suitable discovery tool for cancer epigenomics and we provide guidance on how to overcome the observed limitations.

MUCOSAL TRANSCRIPTOMICS IMPLICATES UNDER EXPRESSION OF *BRINP3* IN THE PATHOGENESIS OF ULCERATIVE COLITIS, SMITH ET AL., INFLAMMATORY BOWEL DISEASES (ACCEPTED)

Philip J Smith[1,2], Adam P Levine[1], Jenny Dunne[1], Paul Guilhamon[3], Mark Turmaine[4], Gavin W Sewell[1], Nuala R O'Shea[1,2], Roser Vega[2], Jennifer C Paterson[5], Dahmane Oukrif[5], Stephan Beck[3], Stuart L Bloom[2], Marco Novelli[5], Manuel Rodriguez-Justo[5], Andrew M Smith[1], Anthony W Segal[1]

[1]Centre for Molecular Medicine, Division of Medicine, University College London, London, WC1E 6JF, United Kingdom

[2]Department of Gastroenterology, University College London Hospitals NHS Foundation Trust, London, United Kingdom

[3]Medical Genomics, UCL Cancer Institute, University College London, London, United Kingdom

[4]Division of Bioscience, University College London, London, WC1N 1BT, United Kingdom

[5]Advanced Diagnostics, University College London, London, United Kingdom

Abstract

Background and aims: Mucosal abnormalities are potentially important in the primary pathogenesis of ulcerative colitis (UC). We investigated the mucosal transcriptomic expression profiles of biopsies from patients with UC and healthy controls (HC), taken from macroscopically non-inflamed tissue from the terminal ileum and three colonic locations with the objective of identifying abnormal molecules that might be involved in disease development.

Methods: Whole-genome transcriptional analysis was performed on intestinal biopsies taken from 24 UC, 26 HC and 14 patients with Crohn's disease. Differential gene expression analysis was performed at each tissue location separately and results were then meta-analysed using Fisher's method. Significantly differentially expressed genes were validated using qPCR. Gene location within the colon was determined using immunohistochemistry, subcellular fractionation, electron and confocal microscopy. DNA methylation was quantified by pyrosequencing.

Results: Seven probes were abnormally expressed throughout the colon in UC patients with Family with sequence similarity member 5 C (FAM5C) being the most significantly underexpressed. Attenuated expression of FAM5C in UC was independent of inflammation, unrelated to phenotype or treatment, and remained low at rebiopsy approximately 23 months later. FAM5C is localised to the brush border of the colonic epithelium and expression is influenced by DNA methylation within its promoter.

Conclusion: Genome-wide expression analysis of non-inflamed mucosal biopsies from UC patients identified FAM5C as significantly under-expressed throughout the colon in a major sub-set of patients with UC. Low levels of this gene could predispose to or contribute to the maintenance of the characteristic mucosal inflammation seen in this condition.

OXBS-450K: A METHOD FOR ANALYSING HYDROXYMETHYLATION USING 450K BEADCHIPS, STEWART ET AL., METHODS (ACCEPTED)

Sabrina K. Stewart[1], Tiffany J. Morris[1], Paul Guilhamon[1], Harry Bulstrode[2], Martin Bachman[3], Shankar Balasubramanian[3], Stephan Beck[1]

[1]Department of Cancer Biology, UCL Cancer Institute, University College London, London WC1E 6BT

[2]Scottish Centre for Regenerative Medicine, Edinburgh, EH16 4UU

[3]Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, UK

Abstract

DNA methylation analysis has become an integral part of biomedical research. For high-throughput applications such as epigenome-wide association studies, the Infinium HumanMethylation450 (450K) BeadChip is currently the platform of choice. However, BeadChip processing relies on traditional bisulfite (BS) based protocols which cannot discriminate between 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC). Here, we report the adaptation of the recently developed oxidative bisulfite (oxBS) chemistry to specifically detect both 5mC and 5hmC in a single workflow using 450K BeadChips, termed oxBS-450K. Supported by validation using mass spectrometry and pyrosequencing, we demonstrate reproducible (R2 > 0.99) detection of 5hmC in human brain tissue using the optimised oxBS-450K protocol described here.

Comparative methylome analysis identifies new tumour subtypes and biomarkers for transformation of nephrogenic rests into Wilms tumour, Charlton et al., (In Revision)

Jocelyn Charlton[1], Richard D. Williams[1], Neil J. Sebire[1], Sergey Popov[2], Gordan Vujanic[4], Tasnim Chagtai[1], Mariana Maschietto[1], Marisa Alcaide-German[1], Tiffany Morris[3], Lee M. Butcher[3], Paul Guilhamon[3], Stephan Beck[3], Kathy Pritchard-Jones[1]

[1] UCL Institute of Child Health, University College London, 30 Guilford Street, London, WC1N 1EH

[2] The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey, SM2 5NG

[3] UCL Cancer Institute, University College London, 72 Huntley Street, London, WC1E 6BT

[4] Department of Pathology, Cardiff University School of Medicine, Heath Park, Cardiff, CF14 4XN

Abstract

Wilms tumours (WT), characterised by loss of imprinting at 11p15 and a paucity of recurrent mutations, are frequently found in association with nephrogenic rests (NR), lesions reminiscent of embryonic kidney. To test if aberrant DNA methylation is implicated in tumourigenesis, we performed methylome analysis on 20 micro-dissected matched trios (WT, NR and normal kidney (NK)). NR/NK comparison revealed 629 differentially methylated regions (DMRs): 55% were hypermethylated, enriched for domains that are bivalent in embryonic stem cells and for genes expressed during development (P=2.49x10-5). NR/WT comparison revealed two WT subgroups; group-2 WT and NR were epigenetically indistinguishable whereas group-1 WT showed hypomethylation of renal development genes, hypermethylation of known and potential new WT tumour suppressor genes CASP8, H19, MIR195, RB1 and TSPAN32, and included all bilateral cases (P=0.032). This suggests that methylation analysis could aid treatment planning in bilateral disease and that some WT may be candidates for epigenetic-modifier therapy.

METHYLATION ANALYSIS SCRIPT

```
setwd("~/Experiments/450K/FINAL 08.06.12/Wilcox 22.06.12")
source("R:/R/450K scripts/Meth.Wil.R")

##detpval=0.01,adj pval=0.001, deltamedbeta =0.35: without FFPE and FF replicates (mut7+8) or
odd4 : 12+15
data<-read.table("R:/Experiments/450K/FINAL
08.06.12/data.betas.detpval.txt",header=TRUE,sep="\t")
data.min4<-data[,-c(22:25,38:39,46:47)]
data.noreps<-data.min4[,-c(36,37,50,51)]

Methyl.W(data.noreps)
"R:/Experiments/450K/FINAL 08.06.12/Wilcox 22.06.12/minFFreps/"
#472655 probes after det pval


##median difference threshold selection
norm.csdata1<-
read.csv("minFFreps/full_dataset_det.pval.filter_nonorm.0.01.csv",header=T,sep=",")
norm.csdata2<-norm.csdata1[complete.cases(norm.csdata1),]

psort5<-as.matrix(norm.csdata2[,2:28])

psort5.2<-as.matrix(psort5)
psort6<-as.matrix(psort5[,1:27])
wtmed<-as.matrix(c(1:nrow(psort6)))
colnames(wtmed)<-"WT.median"
mutmed<-as.matrix(c(1:nrow(psort6)))
colnames(mutmed)<-"MUT.median"
med.diff<-as.matrix(c(1:nrow(psort6)))
colnames(med.diff)<-"MUTmed-WTmed"

for (x in 1:nrow(psort6)) wtmed[x]<-((median(psort6[x,1:12])))
for (x in 1:nrow(psort6)) mutmed[x]<-((median(psort6[x,13:27])))
for (x in 1:nrow(psort6)) med.diff[x]<-(mutmed[x]-wtmed[x])

psort6.5<-cbind(norm.csdata2,wtmed,mutmed,med.diff)

par(mfrow=c(1,1))
###random selection of 8175 median differences
samp3<-replicate(1000,sample(psort6.5[,31],8175,replace=T),simplify="matrix")

###plot of top8175median differences
norm.cs_topinfo<-read.csv("minFFreps/topinfo.0.001_8175.csv",header=TRUE,sep=",")

png("minFFreps/threshold.png")
par(new=F)
br2<-seq(-1,1,by=0.05)
hist(samp3,freq=F,col=NULL,border="red",xlim=c(-
1,1),main="",xlab=NULL,breaks=br2,ylim=c(0,8))
par(new=T)
hist(norm.cs_topinfo[,6],freq=F,col=NULL,border="blue",main=NULL,xlab= "Delta
Median(Beta)",ylab=NULL,xlim=c(-1,1),breaks=br2,yaxt='n',ylim=c(0,8))
dev.off()

###calculate percentage cut off break
a<-hist(samp3,plot=F,breaks=br2)
b<-cbind(matrix(a$breaks),matrix(c(999999,a$counts)))
```

```
a2<-hist(norm.cs_topinfo[,6],plot=F,breaks=br2)
b2<-cbind(matrix(a2$breaks),matrix(c(999999,a2$counts)))

###Find point in histogram where less than 2% of random values end up by chance-->gives <2%
chance of false positive: 0.02*8175000=163,500
###-->35%

##Figures
###adj-p.values:
all.pvals<-read.table("minFFreps/data_adjp.values.txt",header=TRUE,sep="\t")
br2=seq(0,1,by=0.05)
p_adjusted<-all.pvals[,30]
png("minFFreps//pval_distrib.png")
hist(p_adjusted,col="gray",main="Distribution of p-values (All Data)",xlab="p-value",breaks=br2)
dev.off()

###median difference:
topinfo<-read.csv("minFFreps/topinfo0.001_0.35_3057.modif.csv",header=TRUE,sep=",")
br4=seq(-1,1,by=0.1)

png("minFFreps/med.diff_distrib.png")
hist(topinfo[,6],breaks=br4,labels=c("","","","","",7,8,"","","","","","",925,1381,614,116,6,"",""),col="gr
ey",xaxt="n",ylim=c(0,1500),xlab="<--Hypometh in Mutant    MUT-WT median difference
Hypermeth in Mutant-->",main="")
axis(1,at=c(seq(from=-1,to=1,by=0.2)))
dev.off()

png("minFFreps/med.diff_distrib_04.06.13.png",height=9,width=10,units="in",res=600)
hist(topinfo[,6],breaks=br4,col="grey",xaxt="n",ylim=c(0,1500),xlab="MUT-WT median
difference",main="",cex.lab=1.3,cex.axis=1.3)
axis(1,at=c(seq(from=-1,to=1,by=0.2)),cex.axis=1.3)
text(x=c(-0.45,-0.35,0.35,0.45,0.55,0.65,0.75),
y=c(7,8,925,1381,614,116,6),labels=c(7,8,925,1381,614,116,6),cex=1.2,pos=3)
dev.off()

###islands,shores, shelves:
island<-as.matrix(topinfo[,15])
summary(island)
plot(topinfo[,15],col="blue",main="Observed",ylab="Frequency")

####calculate percentage of each type in top list:
island2<-as.matrix(summary(topinfo[,15]))
row.names=1
tempI<-as.matrix(c(1:nrow(island2)))
for (x in 1:nrow(island2)) tempI[x]<-((island2[x]/nrow(island))*100)
my.islands<-cbind(island2,tempI,sum(tempI))
colnames(my.islands)<-c("My.Total","My.Percentage","Check")

####Extract from full list:
fullislands<-read.table("minFFreps/fullset.islands.txt",header=T,sep="\t")
names(fullislands)
fullislands2<-as.vector(fullislands[,2])

####compare expected to observed:
source("R:/R/rand.R")
com.islands<-as.data.frame(cbind(Random.isl<-rand(fullislands2,3057,1000),my.islands))
attach(com.islands)

com.islands2<-as.data.frame(c(1:nrow(com.islands)))
for (x in 1:nrow(com.islands)) com.islands2[x]<-com.islands["My.Percentage"]-
com.islands["Percentage"]
com.islands3<-cbind(com.islands,com.islands2[,1])
colnames(com.islands3)<-c(colnames(com.islands),"Percentage Enrichment")
com.islands4<-com.islands3[order(com.islands3[,7],decreasing=TRUE),]


png("minFFreps/islands_enrichment_04.06.13.png",height=9,width=7,units="in",res=600)
barplot(com.islands4[,7],main="",col=c("red","yellow","green","blue"),
names.arg=rownames(com.islands4),ylab="%Enrichment",yaxt="n",ylim=c(-
25,25),cex.lab=1.3,cex.names=1.3)
axis(2,at=c(seq(from=-25,to=25,by=10)),cex.axis=1.3)
dev.off()

###refgene groups
norm.group<-read.csv("minFFreps/refgene groups_top3057.csv",header=F,sep=",")


####calculate percentage of each type of refgene group in top list:
norm.group6<-as.matrix(summary(norm.group[,1]))

row.names=1

tempG<-as.matrix(c(1:nrow(norm.group6)))
for (x in 1:nrow(norm.group6)) tempG[x]<-((norm.group6[x]/nrow(norm.group))*100)

my.grps<-cbind(norm.group6,tempG,sum(tempG))
colnames(my.grps)<-c("My.Total","My.Percentage","Check")

####Calculate what percentage of each refgene group would be expected by random chance in a
sample of the same size as top list:

fullgroups<-read.csv("minFFreps/refgene groups_fullset.csv",header=F,sep=",")
names(fullgroups)

fullgroups2<-as.vector(fullgroups[,1])


####compare expected to observed:

com.grps<-as.data.frame(cbind(Random<-rand(fullgroups2,3057,1000),my.grps))
attach(com.grps)

com.grps2<-as.data.frame(c(1:nrow(com.grps)))

for (x in 1:nrow(com.grps)) com.grps2[x]<-com.grps["My.Percentage"]-com.grps["Percentage"]
com.grps3<-cbind(com.grps,com.grps2[,1])
colnames(com.grps3)<-c(colnames(com.grps),"Percentage Enrichment")
com.grps4<-com.grps3[order(com.grps3[,7],decreasing=TRUE),]

png("minFFreps/refgene
groups_enrichment_04.06.13.png",height=9,width=10,units="in",res=600)
barplot(com.grps4[,7],main="",col=c("red","orange","yellow","light green","dark
green","blue","purple"), names.arg=rownames(com.grps4),ylab="%Enrichment",ylim=c(-
10,10),cex.axis=1.3,cex.lab=1.3,cex.names=1.3)
dev.off()

###heatmap supervised analysis
clin.info<-read.table("minFFreps/clin.info3.1.txt",header=T,sep="\t")
clin.info<-read.table("minFFreps/clin.info3.1_21.08.12.txt",header=T,sep="\t")
clin.info3<-clin.info[-c(11,12,19,23,21,29),]

#calculate correlation btwn IDH status and 2HG
```

```
idh.cor<-c(rep(1,12),rep(2,15))
hg.cor<-clin.info3[,10]
cor.test(idh.cor,hg.cor,method="spearman") #rho=0.8422435  p-value = 3.618e-08

#calculate correlation btwn cluster and 2HG
cluster.f2<-function(Cluster){
 if (Cluster == "Cluster 1") 1
 else if (Cluster == "Cluster 2") 2
 else if (Cluster == "Cluster 3") 3
}
cluster.cor <- sapply(clin.info3$Cluster, cluster.f2)
cor.test(cluster.cor,hg.cor,method="spearman") #rho=0.7878581  p-value = 1.071e-06

#calculate correlation btwn IDH status and Age
age.cor<-clin.info3[,6]
cor.test(idh.cor,age.cor,method="spearman") #rho=0.4073254  p-value = 0.03496

#calculate correlation btwn IDH status and sex
sex.f2<-function(Sex) {if (Sex == "F") 2 else 1 }
sex.cor <- sapply(clin.info3$Sex, sex.f2)

cor.test(idh.cor,sex.cor,method="spearman")#rho=-0.16855  p-value = 0.4007

#calculate correlation btwn IDH status and Grade
grade.f2<-function(Grade){
 if (Grade == "Grade 1") 1
 else if (Grade == "Grade 2") 2
 else if (Grade == "Grade 3") 3
 else if (Grade == "Dediff") 4
}
grade.cor <- sapply(clin.info3$Grade, grade.f2)

cor.test(idh.cor,grade.cor,method="spearman")#rho=-0.2551171  p-value = 0.199

library(heatmap.plus)
library(gtools)
source("R:/R/heatmap3_040613.R")

#associate colours with different aspects of clin.info

group.f <- function(Group) {if (Group == "WT") "green" else "red" }
grp.col3 <- sapply(clin.info3$Group, group.f)

sex.f<-function(Sex) {if (Sex == "F") "grey" else "black" }
sex.col3 <- sapply(clin.info3$Sex, sex.f)

grade.f<-function(Grade){
 if (Grade == "Grade 1") "yellow"
 else if (Grade == "Grade 2") "orange"
 else if (Grade == "Grade 3") "orange3"
 else if (Grade == "Dediff") "brown"
}
grade.col3 <- sapply(clin.info3$Grade, grade.f)

age.f<-function(Age){
 if (Age < 50) "yellow"
 else if (Age >= 50 & Age<60) "light green"
 else if (Age >= 60 & Age<70) "light blue"
 else if (Age >= 70 & Age<85) "tomato"
 else if (Age == "NA") "black"
}
```

```
age.col3 <- sapply(clin.info3$Age, age.f)

hg.f<-function(ngpml.2HG){
  if (ngpml.2HG <1) "black"
  else if (ngpml.2HG < 1000 & ngpml.2HG >1 ) "yellow"
  else if (ngpml.2HG >= 1000 & ngpml.2HG < 5000 ) "green"
  else if (ngpml.2HG >= 5000 & ngpml.2HG < 10000 ) "blue"
  else if (ngpml.2HG >= 10000 & ngpml.2HG) "red"

}
ngpml.2HG.col3 <- sapply(clin.info3$ngpml.2HG, hg.f)


cluster.f<-function(Cluster){
  if (Cluster == "Cluster 1") "light blue"
  else if (Cluster == "Cluster 2") "light blue4"
  else if (Cluster == "Cluster 3") "dark blue"
}
cluster.col3 <- sapply(clin.info3$Cluster, cluster.f)

all.colors3 <- cbind(grp.col3,sex.col3,grade.col3,age.col3,ngpml.2HG.col3,cluster.col3)
colnames(all.colors3)<-c("IDH","Sex","Grade","Age","2HG","Cluster")

#plot heatmap

clab3<-all.colors3


main_title=""
par(cex.main=1)

Lab.palette2 <- colorRampPalette(c("yellow","light blue","blue"), space = "Lab")

#function arguments below:
#top.cs2=needs to be a matrix with rows corresponding to probes and columns to samples; all
numeric: do not include a column with targetIDs
#hclustfun: as far as I know only hclust works here, but you can try others
#labCol: you need to change the name in there from top.cs2 to your data name
#NumColSideColors: spaces out the colour bars on top of the heatmap based on the number of bars
you are placing there: eg if you have 6 variables you want to plot, NumColSideColors=6

top.noreps<-read.csv("minFFreps/topinfo_and_betas0.001_0.35_3057.csv",header=T,sep=",")
top.noreps2<-as.matrix(top.noreps[,2:28])

#renaming samples
colnames(top.noreps)
clin.info3[,2]
comp.names<-cbind(colnames(top.noreps2[,1:27]),as.data.frame(clin.info3[,2]))

top.noreps3<-top.noreps[,1:28]
colnames(top.noreps3)=c("TargetID",as.character(clin.info3[,1]))
top.noreps4<-as.matrix(top.noreps3[,2:28])
colnames(top.noreps4)


source("R:/R/heatmap3_040613.R")

png("minFFreps/supervised_heatmap_3057_04.06.13.png",height=9,width=10,units="in",res=600)

heatmap.3(top.noreps4,na.rm = TRUE, scale="none",hclustfun=hclust, dendrogram="column",
margins=c(14,18), Rowv=FALSE, Colv=TRUE, ColSideColors=clab3, symbreaks=FALSE, key=TRUE,
symkey=FALSE, density.info="none", trace="none", main=main_title,
labCol=colnames(top.noreps4),cexRow=0.1, col=Lab.palette2(20), NumColSideColors=7,
KeyValueName="Beta Value",cexCol=1.1)

legend("bottomleft",legend=c("IDH","Mutant","Wild-Type","","Sex","F","M","","Grade","Grade
1","Grade 2","Grade 3","Dediff","","Age","<50","51-60","61-70",">71","","2-
HG(ng/ml)","<1,000","1,001-5,000","5,001-10,000",">10,000","N/A","","Cluster","1: Low","2:
Intermediate","3: High"),
fill=c("white","red","green","white","white","grey","black","white","white","yellow","orange","orang
e3","brown","white","white","yellow","light green","light
blue","tomato","white","white","yellow","green","blue","red","black","white","white","light
blue","light blue4","dark blue"), border=FALSE, bty="n", y.intersp = 0.85, cex=1.1)

dev.off()

##Consensus clustering
source("http://bioconductor.org/biocLite.R")
biocLite("ConsensusClusterPlus")
library(ConsensusClusterPlus)


#
# cs.var<-read.csv("R:/Experiments/450K/FINAL
08.06.12/Wilcox_min4/full_dataset_det.pval.filter_nonorm.0.01.csv",header=TRUE,sep=",")
# cs.var2<-cs.var[complete.cases(cs.var),]


cs.var2<-norm.csdata1[complete.cases(norm.csdata1),]
mad.m<-matrix(1:nrow(cs.var2))
colnames(mad.m)<-"MAD"
mad.m<-apply(cs.var2[,2:28],1,mad)


cs.mad3<-cbind(cs.var2,mad.m)
cs.mad4<-cs.mad3[order(cs.mad3[,29],decreasing=TRUE),]


mad150<-cs.mad4[1:150,1:28]
# write.table(mad100,"R:/Experiments/450K/FINAL
08.06.12/Wilcox_min4/mad100.data.txt",col.names=T,sep="\t",row.names=F)
mad150.noIDs<-as.matrix(mad150[,2:28])

##ConsensusClusterPlusHighRes is in sep R script
results.mad150.km<-
ConsensusClusterPlusHighRes(mad150.noIDs,clusterAlg="km",maxK=9,reps=500,distance="euclid
ean",plot="png",title="minFFreps/150km.consensus_31.05.13_7/",tmyPal=Lab.palette2(20))

results.mad150.km[[2]][[2]]$order
results.mad150.km[[4]][["consensusClass"]]

# mad300<-cs.mad4[1:300,1:28]
# # write.table(mad100,"R:/Experiments/450K/FINAL
08.06.12/Wilcox_min4/mad100.data.txt",col.names=T,sep="\t",row.names=F)
# mad300.noIDs<-as.matrix(mad300[,2:28])
#
#
# results.mad300.km<-
ConsensusClusterPlus(mad300.noIDs,clusterAlg="km",maxK=9,reps=500,distance="euclidean",plot
="pdf",title="minFFreps/300km.consensus/",tmyPal=Lab.palette2(20))
#
# results.mad300.km[[4]][["consensusClass"]] #2mutants miscluster
```

```
#
# results.mad300.km[[3]][["consensusClass"]] #1 Wt cluster, 2 mutant clusters, 1 mut sample misclusters

# mad500<-cs.mad4[1:500,1:28]
# # write.table(mad100,"R:/Experiments/450K/FINAL 08.06.12/Wilcox_min4/mad100.data.txt",col.names=T,sep="\t",row.names=F)
# mad500.noIDs<-as.matrix(mad500[,2:28])
#
#
# results.mad500.km<-ConsensusClusterPlus(mad500.noIDs,clusterAlg="km",maxK=9,reps=500,distance="euclidean",plot="pdf",title="minFFreps/500km.consensus/",tmyPal=Lab.palette2(20))
#
# results.mad500.km[[4]][["consensusClass"]] #2mutants miscluster
#
# results.mad500.km[[3]][["consensusClass"]] #1 Wt cluster, 2 mutant clusters, 1 mut sample misclusters


##plot heatmap with samples in order of consensus cluster: need to reorder the data and the clin.info and the colors

mad150.noIDs2<-mad150.noIDs
colnames(mad150.noIDs2)=c(as.character(clin.info3[,1]))

top.cs2.cons.o<-mad150.noIDs2[,c(results.mad150.km[[4]][[2]]$order)]
clin.info.cons.o<-clin.info3[c(results.mad150.km[[4]][[2]]$order),]

grp.col.o <- sapply(clin.info.cons.o$Group, group.f)
sex.col.o <- sapply(clin.info.cons.o$Sex, sex.f)
grade.col.o <- sapply(clin.info.cons.o$Grade, grade.f)
age.col.o <- sapply(clin.info.cons.o$Age, age.f)
ngpml.2HG.col.o <- sapply(clin.info.cons.o$ngpml.2HG, hg.f)

all.colors.o <- cbind(grp.col.o,sex.col.o,grade.col.o,age.col.o,ngpml.2HG.col.o)
colnames(all.colors.o)<-c("IDH","Sex","Grade","Age","HG")

clab.o <- all.colors.o

main_title=""
par(cex.main=1)

Lab.palette2 <- colorRampPalette(c("yellow","light blue","blue"), space = "Lab")
##make the same plot again to get the color key for consensus
#save as png
png(file="minFFreps/consensus key.png",height=9,width=10,units="in",res=600)
heatmap.3(top.cs2.cons.o,na.rm = TRUE, scale="none",hclustfun=hclust, dendrogram=NULL, margins=c(14,18), Rowv=FALSE, Colv=FALSE, ColSideColors=clab.o, symbreaks=FALSE, key=TRUE, symkey=FALSE, density.info="none", trace="none", main=main_title, labCol=colnames(top.cs2.cons.o),cexRow=0.1, col=Lab.palette2(20), NumColSideColors=5, KeyValueName="Consensus")

legend("left",legend=c("IDH","Mutant","Wild-Type","","Sex","F","M","","Grade","Grade 1","Grade 2","Grade 3","Dediff","","Age","<50","51-60","61-70",">71","","2-HG(ng/ml)","<1,000","1,001-5,000","5,001-10,000",">10,000","N/A"), fill=c("white","red","green","white","white","grey","black","white","white","yellow","orange","orange3","brown","white","white","yellow","light green","light blue","tomato","white","white","yellow","green","blue","red","black"), border=FALSE, bty="n", y.intersp = 0.9, cex=0.7)
```

```
dev.off()

##calculate mean beta value for each cluster in to 3057 probes
top3057<-read.csv("R:/Experiments/450K/FINAL
08.06.12/Wilcox_min4/minFFreps/topinfo_and_betas0.001_0.35_3057.csv",sep=",",header=T)
top3057_2<-top3057[,2:28]

clin.info<-read.table("R:/Experiments/450K/FINAL 08.06.12/clin.info3.1.txt",header=T,sep="\t")
clin.info[,2]
clin.info3<-clin.info[-c(11,12,19,23,21,29),]

high.clust<-top3057_2[,c(27,13,15,22,16,18,14,25)]
int.clust<-top3057_2[,-c(1:12,27,13,15,22,16,18,14,25)]
low.clust<-top3057_2[,c(1:12)]

high.clust2<-stack(high.clust)$values
median(high.clust2) #0.7515689
mean(high.clust2) #0.725014

int.clust2<-stack(int.clust)$values
median(int.clust2) #0.5477938
mean(int.clust2) #0.5159579

low.clust2<-stack(low.clust)$values
median(low.clust2) #0.1594047
mean(low.clust2) #0.2072573


########################Automated analysis

Methyl.W<-function(dataset,
          dataset.layout=c(1,2,3),
          det.pval.filter=c(FALSE,TRUE),
          det.pval.filterlevel=0.01,
          quant.norm=c(FALSE,TRUE),
          total_group1_samples=0,
          total_group2_samples=0,
          folder="C:\\",
          p.val=0,
          group1.name="Group1",
          group2.name="Group2",
          med.diff.threshold=0){

  dataset<-dataset[complete.cases(dataset),]


  cat("\n","Dataset layout. Choose one of the following:","\n","\n","1)TargetIDs, then beta values for
each sample","\n","\n","2)TargetIDs, then, for each sample: 1 column beta, 1 column det. p-val(in
that order)","\n","\n","3)TargetIDs, then, for each sample: 1 column det. p-val, 1 column Signal A, 1
column signal B (in that order)","\n") # prompt
  dataset.layout<-scan(n=1,what=character())


  cat("\n","Name of sample group 1?","\n") # prompt
  group1.name<-scan(n=1,what=character())

  cat("\n","How many samples in group 1?","\n") # prompt
  total_group1_samples<-scan(n=1)

  cat("\n","Name of sample group 2?","\n") # prompt
  group2.name<-scan(n=1,what=character())
```

```
 cat("\n","How many samples in group 2?","\n") # prompt
 total_group2_samples<-scan(n=1)

 total_number_samples<-total_group1_samples+total_group2_samples

 cat("\n","Folder location for output files","\n","(Eg:","T:\\Paul\\450K\\450Kauto\\)","\n") #
prompt
 folder<-scan(n=1,what=character())

 if (dataset.layout==1)  {                                    #######just betas
  if(ncol(dataset)>((total_number_samples)+1)|ncol(dataset)<((total_number_samples)+1))
stop("Layout different from layout 1 requirements")
  det.pval.filter==FALSE
  det.pval.filterlevel==0.01
  quant.norm==FALSE
  dataset<-dataset[complete.cases(dataset),]
  print("Dimensions of dataset (no det pval filter or normalisation): ")
  print(dim(dataset))

 }

 else if (dataset.layout==2){                                   ####beta,det.pval
  if(ncol(dataset)>((2*total_number_samples)+1)|ncol(dataset)<((2*total_number_samples)+1))
stop("Layout different from layout 2 requirements")
  quant.norm==FALSE
  cat("\n","Apply a Detection p-value filter? (TRUE/FALSE)","\n") # prompt
  det.pval.filter<-scan(n=1,what=logical())

  if (det.pval.filter==FALSE) {
   det.pval.filterlevel==0.01
   print("The Detection pval filter is OFF & No quantile Normalisation possible")
   dataset<-dataset[,c(1,seq(from=2,to=ncol(dataset),by=2))]
   dataset<-dataset[complete.cases(dataset),]
   print("Dimensions of dataset (no det pval filter or normalisation): ")
   print(dim(dataset))
   write.csv (dataset, file=paste(folder,"full_dataset_nofilter_nonorm.",".csv",sep=""), row.names=F,
quote=FALSE)
  }

  else if (det.pval.filter==TRUE) {
   cat("\n","What Detection p-value threshold?","\n") # prompt
   det.pval.filterlevel<-scan(n=1)
   print("The Detection pval filter is ON & No quantile Normalisation possible")
   data.pvals<-dataset[,c(1,seq(from=3,to=ncol(dataset),by=2))]
   data.betas<-dataset[,c(1,seq(from=2,to=ncol(dataset),by=2))]
   data.test<-data.pvals[rowSums(data.pvals[,c(2:ncol(data.pvals))]>det.pval.filterlevel)<1,]
   idx1<-match(data.test[,1],data.betas[,1])
   dataset<-data.betas[idx1,]
   dataset<-dataset[complete.cases(dataset),]
   print("Dimensions of dataset after det pval filter (no normalisation): ")
   print(dim(dataset))
   write.csv (dataset,
file=paste(folder,"full_dataset_det.pval.filter_nonorm.",det.pval.filterlevel,".csv",sep=""),
row.names=F, quote=FALSE)

  }
 }

 else if (dataset.layout==3) {                                  ####det.pval,A,B
```

```
  if(ncol(dataset)>((3*total_number_samples)+1)|ncol(dataset)<((3*total_number_samples)+1))
stop("Layout different from layout 3 requirements")
  cat("\n","Apply a Detection p-value filter? (TRUE/FALSE)","\n") # prompt
  det.pval.filter<-scan(n=1,what=logical())

  if (det.pval.filter==FALSE) {
   det.pval.filterlevel==0.01
   cat("\n","Perform quantile normalisation of the data? (TRUE/FALSE)","\n") # prompt
   quant.norm<-scan(n=1,what=logical())

   if (quant.norm==FALSE) {
    print("The Detection pval filter is OFF & Quantile Normalisation is OFF")
    betas.fromsignals<-
matrix(c(1:(total_number_samples*(nrow(dataset)))),byrow=TRUE,nrow=nrow(dataset))
    print(dim(betas.fromsignals))
    for (x in 1:nrow(dataset))  for (y in seq(from=3, to=ncol(dataset),by=3))
betas.fromsignals[x,y/3]<-(dataset[x,y+1]/(dataset[x,y]+dataset[x,y+1]))

    betas.and.TIDs<-cbind(dataset[,1],as.data.frame(betas.fromsignals))
    names.list1<-c(colnames(dataset[,c(1,seq(from=2, to=ncol(dataset),by=3))]))
    names.df1<-as.data.frame(matrix(1:length(names.list1),byrow=TRUE,nrow=1))
    for (x in 1:length(names.list1)) names.df1[x]<-strsplit(names.list1[x],".Detection.Pval")
    colnames(betas.and.TIDs)<-names.df1
    dataset<-betas.and.TIDs
    dataset<-dataset[complete.cases(dataset),]
    print("Dimensions of dataset (no det pval filter or normalisation): ")
    print(dim(dataset))
    write.csv (dataset, file=paste(folder,"full_dataset_nofilter_nonorm.",".csv",sep=""),
row.names=F, quote=FALSE)
   }

   else if (quant.norm==TRUE) {
    print("The Detection pval filter is OFF & Quantile Normalisation is ON")
    data.sigA<-as.matrix(dataset[,c(seq(from=3,to=ncol(dataset),by=3))])
    data.sigB<-as.matrix(dataset[,c(seq(from=4,to=ncol(dataset),by=3))])
    library(preprocessCore)
    sigA.qnorm<-normalize.quantiles(data.sigA,copy=F)
    sigB.qnorm<-normalize.quantiles(data.sigB,copy=F)

    betas.fromsignals<-
matrix(c(1:(total_number_samples*(nrow(dataset)))),byrow=TRUE,nrow=nrow(dataset))
    for (x in 1:nrow(betas.fromsignals))  for (y in 1:ncol(betas.fromsignals))
betas.fromsignals[x,y]<-(data.sigB[x,y]/(data.sigA[x,y]+data.sigB[x,y]))

    betas.and.TIDs<-cbind(dataset[,1],as.data.frame(betas.fromsignals))
    names.list1<-c(colnames(dataset[,c(1,seq(from=2, to=ncol(dataset),by=3))]))
    names.df1<-as.data.frame(matrix(1:length(names.list1),byrow=TRUE,nrow=1))
    for (x in 1:length(names.list1)) names.df1[x]<-strsplit(names.list1[x],".Detection.Pval")
    colnames(betas.and.TIDs)<-names.df1
    dataset<-betas.and.TIDs
    dataset<-dataset[complete.cases(dataset),]
    print("Dimensions of dataset after normalisation (no det pval filter): ")
    print(dim(dataset))
    write.csv (dataset, file=paste(folder,"full_dataset_nofilter_quant.norm.",".csv",sep=""),
row.names=F, quote=FALSE)
   }
  }

  else if (det.pval.filter==TRUE) {
   cat("\n","What Detection p-value threshold?","\n") # prompt
   det.pval.filterlevel<-scan(n=1)
```

```
    cat("\n","Perform quantile normalisation of the data? (TRUE/FALSE)","\n") # prompt
    quant.norm<-scan(n=1,what=logical())

    if(quant.norm==FALSE) {
      print("The Detection pval filter is ON & Quantile Normalisation is OFF")
      data.pvals<-dataset[,c(1,seq(from=2,to=ncol(dataset),by=3))]
      data.test<-data.pvals[rowSums(data.pvals[,c(2:ncol(data.pvals))]>det.pval.filterlevel)<1,]

      betas.fromsignals<-
matrix(c(1:(total_number_samples*(nrow(dataset)))),byrow=TRUE,nrow=nrow(dataset))
      for (x in 1:nrow(dataset))  for (y in seq(from=3, to=ncol(dataset),by=3))
betas.fromsignals[x,y/3]<-(dataset[x,y+1]/(dataset[x,y]+dataset[x,y+1]))

      betas.and.TIDs<-cbind(dataset[,1],as.data.frame(betas.fromsignals))
      names.list1<-c(colnames(dataset[,c(1,seq(from=2, to=ncol(dataset),by=3))]))
      names.df1<-as.data.frame(matrix(1:length(names.list1),byrow=TRUE,nrow=1))
      for (x in 1:length(names.list1)) names.df1[x]<-strsplit(names.list1[x],".Detection.Pval")
      colnames(betas.and.TIDs)<-names.df1
      idx1<-match(data.test[,1],betas.and.TIDs[,1])
      dataset<-betas.and.TIDs[idx1,]
      dataset<-dataset[complete.cases(dataset),]
      print("Dimensions of dataset after det pval filter (no normalisation): ")
      print(dim(dataset))
      write.csv (dataset,
file=paste(folder,"full_dataset_det.pval.filter_nonorm.",det.pval.filterlevel,".csv",sep=""),
row.names=F, quote=FALSE)
    }

    else if (quant.norm==TRUE) {
      print("The Detection pval filter is ON & Quantile Normalisation is ON")
      data.pvals<-dataset[,c(1,seq(from=2,to=ncol(dataset),by=3))]
      data.test<-data.pvals[rowSums(data.pvals[,c(2:ncol(data.pvals))]>det.pval.filterlevel)<1,]

      data.sigA<-as.matrix(dataset[,c(seq(from=3,to=ncol(dataset),by=3))])
      data.sigB<-as.matrix(dataset[,c(seq(from=4,to=ncol(dataset),by=3))])
      library(preprocessCore)
      sigA.qnorm<-normalize.quantiles(data.sigA,copy=F)
      sigB.qnorm<-normalize.quantiles(data.sigB,copy=F)
      print("Data normalisation complete")
      betas.fromsignals<-
matrix(c(1:(total_number_samples*(nrow(dataset)))),byrow=TRUE,nrow=nrow(dataset))
      for (x in 1:nrow(betas.fromsignals))  for (y in 1:ncol(betas.fromsignals))
betas.fromsignals[x,y]<-(data.sigB[x,y]/(data.sigA[x,y]+data.sigB[x,y]))
      print("beta-value calculations complete")
      betas.and.TIDs<-cbind(dataset[,1],as.data.frame(betas.fromsignals))

      names.list1<-c(colnames(dataset[,c(1,seq(from=2, to=ncol(dataset),by=3))]))
      names.df1<-as.data.frame(matrix(1:length(names.list1),byrow=TRUE,nrow=1))
      for (x in 1:length(names.list1)) names.df1[x]<-strsplit(names.list1[x],".Detection.Pval")
      colnames(betas.and.TIDs)<-names.df1

      idx1<-match(data.test[,1],betas.and.TIDs[,1])
      dataset<-betas.and.TIDs[idx1,]
      dataset<-dataset[complete.cases(dataset),]
      print("Dimensions of dataset after det pval filter and normalisation: ")
      print(dim(dataset))
      write.csv (dataset,
file=paste(folder,"full_dataset_det.pval.filter_quant.norm.",det.pval.filterlevel,".csv",sep=""),
row.names=F, quote=FALSE)
    }
```

```
  }
 }


 cat("\n","Adjusted p-value threshold after t-test?","\n") # prompt
 p.val<-scan(n=1)

 cat("\n","Median difference threshold?","\n") # prompt
 med.diff.threshold<-scan(n=1)

 ##1)STATS

 ####a)t.tests + p.value adjustment: Filter 1


 norm.mcs<-as.matrix(dataset[1:nrow(dataset),2:c(total_number_samples+1)])
 print("Dimensions of starting dataset without TargetID column are:")
 print(dim(norm.mcs))
 f<-c(rep(0,total_group1_samples),rep(1,total_group2_samples))
 pv = rep(0,nrow(norm.mcs))
 for (i in 1:nrow(norm.mcs)) { pv[i]=wilcox.test(norm.mcs[i,] ~ f)$p.value }

 pv2<-as.matrix(pv)

 #performs t.tests on each row

 norm.csdatatt<-cbind(dataset,pv2)
 print("Dimensions of dataset after wilcoxon test:")
 print(dim(norm.csdatatt))
 #combines norm.csdata dataframe with the stats for each row

 norm.csp<-
cbind(norm.csdatatt,p.adjust(norm.csdatatt[,c(total_number_samples+2)],method="BH"))
 #multiple test correction (adjusts the p values)
 print("Dimensions of dataset after p.value adjustment:")
 print(dim(norm.csp))
 write.table (norm.csp, file=paste(folder,"data_adjp.values.txt",sep=""), sep="\t", col.names=T,
row.names=F, quote=FALSE)


 norm.psort3<-norm.csp[order(norm.csp[,c(total_number_samples+3)]),]
 norm.psort4<-norm.psort3[norm.psort3[,c(total_number_samples+3)]<=p.val,]
 attach(norm.psort4,warn.conflicts=FALSE)
 print("Dimensions of dataset after p.value threshold selection:")
 print(dim(norm.psort4))

 write.table(norm.psort4,
file=paste(folder,"data_adjp.values_",p.val,"_",nrow(norm.psort4),".txt",sep=""), sep="\t",
col.names=T, row.names=F, quote=FALSE)

 ####b)Means, Medians
 norm.psort5.2<-as.matrix(norm.psort4)
 norm.psort6<-as.matrix(norm.psort4[,2:c(total_number_samples+3)])

 group1.med<-as.matrix(c(1:nrow(norm.psort6)))
 colnames(group1.med)<-group1.name
 group2.med<-as.matrix(c(1:nrow(norm.psort6)))
 colnames(group2.med)<-group2.name
 med.diff<-as.matrix(c(1:nrow(norm.psort6)))
 colnames(med.diff)=paste("med.diff.",group2.name,".minus.",group1.name,sep="")
 for (x in 1:nrow(norm.psort6)) group1.med[x]<-
((median(norm.psort6[x,2:c(total_group1_samples+1)])))
 for (x in 1:nrow(norm.psort6)) group2.med[x]<-
((median(norm.psort6[x,c(total_group1_samples+2):c(total_number_samples+1)])))
 for (x in 1:nrow(norm.psort6)) med.diff[x]<-(group2.med[x]-group1.med[x])

 norm.psort7<-data.frame(norm.psort5.2,group1.med,group2.med,med.diff)

 norm.psort8<-norm.psort7[order(norm.psort7[,1]),]
 attach(norm.psort8,warn.conflicts=FALSE)
 print(names(norm.psort8))


 ##2) GENETIC INFO + FILTER 2

 norm.gen.info<-
read.csv("T:\\Infinium_Methylation_Data_Repository\\InfiniumManifest\\genomic_info2.csv",hea
der=T,sep=",")
 norm.gen.info2<-norm.gen.info[order(norm.gen.info[,1]),]
 norm.idx<-match(norm.psort8[,1],norm.gen.info2[,1])
 norm.gen.info3<-norm.gen.info2[norm.idx,]
 norm.cs_topinfo_and_betas<-cbind(norm.psort8,norm.gen.info3)
 write.csv (norm.cs_topinfo_and_betas,
file=paste(folder,"data_betas_and_gen.info.",p.val,"_",nrow(norm.psort4),".csv",sep=""),
row.names=F, quote=FALSE)


 norm.cs_topinfo<-
norm.cs_topinfo_and_betas[,c(1,c(total_number_samples+2):c(total_number_samples+22))]
 write.csv (norm.cs_topinfo, file=paste(folder,"topinfo.",p.val,"_",nrow(norm.psort4),".csv",sep=""),
row.names=F, quote=FALSE)

 norm.cs_topinfo2<-norm.cs_topinfo[rev(order(abs(norm.cs_topinfo[,6]))),]

 norm.cs_topinfo3<-norm.cs_topinfo2[abs(norm.cs_topinfo2[,6])>=med.diff.threshold,]
 write.csv (norm.cs_topinfo3,
file=paste(folder,"topinfo",p.val,"_",med.diff.threshold,"_",nrow(norm.cs_topinfo3),".csv",sep=""),
row.names=F, quote=FALSE)


 norm.cs_topinfo4<-
norm.cs_topinfo_and_betas[abs(norm.cs_topinfo_and_betas[,c(total_number_samples+6)])>=med.di
ff.threshold,]
 write.csv (norm.cs_topinfo4,
file=paste(folder,"topinfo_and_betas",p.val,"_",med.diff.threshold,"_",nrow(norm.cs_topinfo4),".csv",
sep=""), row.names=F, quote=FALSE)

 norm.genes<-norm.cs_topinfo3[,"UCSC_REFGENE_NAME"]
 norm.genes2<-as.vector(norm.genes)

 write.csv (norm.genes2, file=paste(folder,p.val,"_",med.diff.threshold,"_genes",".csv",sep=""),
row.names=F, quote=FALSE)

 norm.genes3<-as.matrix(unique(unlist(strsplit(norm.genes2, "\\;"))))
 write.csv (norm.genes3,
file=paste(folder,p.val,"_",med.diff.threshold,"_genes_nodups",".csv",sep=""), row.names=F,
quote=FALSE)

 print("Number of selected probes after p-value and median difference filters:")
 print(dim(norm.cs_topinfo3))
}
```

INTEGRATED METHYLATION AND GENE EXPRESSION ANALYSIS SCRIPT

```
#data for 46 CS samples processed on Illumina HT12v4 arrays, exported from GenomeStudio:
probe level info, no normalisation or bg correction
#subsequent analysis follows method described by Ritchie et al.(2011)
setwd("~/Experiments/GEM/CS/CS_110814")
library (limma)

cs<-read.ilmn(files="allcs_probe_raw.txt",ctrlfiles="allcs_controls.txt", probeid
="ProbeID",annotation ="TargetID",other.columns=c("Detection Pval","Avg_ NBEADS"))
dim(cs)
cs$targets
cs$E[1:5 , ]
table (cs$genes$Status)

proportion<-propexpr(cs)
proportion
t.test(proportion[1:30],proportion[31:46])
#same proportion of probes expressed in both sample grps:
#
# Welch Two Sample t-test
#
# data:  proportion[1:30] and proportion[31:46]
# t = -1.5732, df = 37.143, p-value = 0.1242
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   -0.030152424  0.003792949
# sample estimates:
#   mean of x mean of y
# 0.4343614 0.4475411

#normalisation
cs.norm<-neqc(cs)
dim(cs.norm)
par(mfrow=c(3,1))
boxplot(log2(cs$E[cs$gene$Status=="regular",]),
range=0,las=2,xlab="",ylab=expression(log[2](intensity)),main="Regular probes")

boxplot(log2(cs$E[cs$gene$Status=="NEGATIVE",]),
range=0,las=2,xlab="",ylab=expression(log[2](intensity)),main="Negative Control probes probes")
boxplot(cs.norm$E, range=0, ylab=expression(log[2](intensity)),las=2,xlab=22,main="Regular
probes, NEQC normalized")

##MDS
#by idh status
par(mfrow=c(1,1))
plotMDS(cs.norm$E,col=c(rep("red",30),rep("green",16)))
png("gx_mds.png",height=9,width=10,units="in",res=600)
plotMDS(cs.norm$E,col=c(rep("red",30),rep("green",16)))
dev.off()
#by batch
par(mfrow=c(1,1))
png("gx_mds_batch.png",height=9,width=10,units="in",res=600)
plotMDS(cs.norm$E,col=c(rep("blue",16),rep("orange",14),rep("blue",6),rep("orange",10)))
dev.off()
##filtering based on probe annotation (bad, good, perfect)
library(illuminaHumanv4.db)
illuminaHumanv4()
```

```
ids<-as.character(rownames(cs.norm))
ids2<-unlist(mget(ids,revmap(illuminaHumanv4ARRAYADDRESS),ifnotfound=NA)) ##converts illuminaHumanv4 illumina IDs with rpefix ILMN to numeric ArrayAddressIDs  found in cs.norm
qual<-unlist(mget(ids2,illuminaHumanv4PROBEQUALITY,ifnotfound=NA)) ##queries probe quality of new IDs, gives result as a list, which is then converted to a vector
table(qual)
AveSignal=rowMeans(cs.norm$E)
boxplot(AveSignal~qual)
rem<-qual=="No match"|qual=="Bad"
cs.norm.filt<-cs.norm[!rem,]
dim(cs.norm)
dim(cs.norm.filt)


##cluster samples based on highly variable probes
IQR<-apply(cs.norm.filt$E,1,IQR,na.rm=TRUE)
topVar<-order(IQR,decreasing=TRUE)[1:500]
d<-dist(t(cs.norm.filt$E[topVar,]))
plot(hclust(d),main="Cluster on 500 most variable probes")

##make heatmap to show differences btwn grps
Lab.palette2 <- colorRampPalette(c("red","black","green"), space = "Lab")

png("gx_heatmap.png",height=9,width=10,units="in",res=600)
heatmap(cs.norm.filt$E[topVar,],keep.dendro=T,margins=c(10,5),ColSideColors=c(rep("red",30),rep("green",16)),col=Lab.palette2(20))
dev.off()

##Differential Expression Analysis
#####summarise values rom rep arrays +set up contrast btwn diff samples +assess array quality
rna<-factor(c(rep("MUT",30),rep("WT",16)))
design<-model.matrix(~0+rna)
colnames(design)<-levels(rna)
aw<-arrayWeights(cs.norm.filt,design)
aw
fit<-lmFit(cs.norm.filt,design,weights=aw)
contrasts<-makeContrasts(MUT-WT,levels=design)
contr.fit<-eBayes(contrasts.fit(fit,contrasts))
topTable(contr.fit,coef=1)
par(mfrow=c(1,2))
volcanoplot(contr.fit, main="MUT-WT")
smoothScatter(contr.fit$Amean,contr.fit$coef,xlab="average intensity",ylab="log-ratio")
abline(h=0,col=2,lty=2)

#####Annotation
library(illuminaHumanv4.db)
ids <- as.character(rownames(contr.fit$genes))
ids2 <- unlist(mget(ids, revmap(illuminaHumanv4ARRAYADDRESS), ifnotfound=NA))
chr <- mget(ids2, illuminaHumanv4CHR, ifnotfound = NA)
cytoband<- mget(ids2, illuminaHumanv4MAP, ifnotfound = NA)
entrezid <- mget(ids2, illuminaHumanv4ENTREZID, ifnotfound = NA)
symbol <- mget(ids2, illuminaHumanv4SYMBOL, ifnotfound = NA)
genename <- mget(ids2, illuminaHumanv4GENENAME, ifnotfound = NA)
anno <- data.frame(Ill_ID = ids2, Chr = as.character(chr),Cytoband=as.character(cytoband),EntrezID=as.numeric(entrezid),Symbol=as.character(symbol),Name=as.character(genename))
contr.fit$genes <- anno
topTable(contr.fit,sort.by="none")
#write.fit(contr.fit, file ="cs_results_110814.txt",adjust="BH",F.adjust="BH",sep="\t")

##heatmaps
```

```
Lab.palette2 <- colorRampPalette(c("red","black","green"), space = "Lab")
a<-topTable(contr.fit,sort.by="none",n=Inf)
a2<-a[order(a$P.Value,decreasing=F),]
an<-rownames(a2[1:500,])
png("gx_heatmap_top500pvalue.png",height=9,width=10,units="in",res=600)
heatmap(cs.norm.filt$E[an,],keep.dendro=T,margins=c(10,5),ColSideColors=c(rep("red",30),rep("g
reen",16)),col=Lab.palette2(20))
dev.off()
a3<-a[order(abs(a$logFC),decreasing=T),]
am<-rownames(a3[1:500,])
png("gx_heatmap_top500abslogFC.png",height=9,width=10,units="in",res=600)
heatmap(cs.norm.filt$E[am,],keep.dendro=T,margins=c(10,5),ColSideColors=c(rep("red",30),rep("g
reen",16)),col=Lab.palette2(20))
dev.off()

#Write full results table
res<-topTable(contr.fit,sort.by="p",n=Inf)
write.table(res,"cs_results_110814.txt",sep="\t",quote=F, row.names=F)
#gx with  p-value<=0.05 (non-adjusted) #1646 probes
res_0.05<-res[which(res$P.Value<=0.05),]
#gx with logFC >abs(2) (no p-value cut-off) #2 probes
res_fc2<-res[which(abs(res$logFC)>=2),]
#gx with logFC >abs(2) (p-value<=0.05 (non-adjusted)) #same 2 probes as above
res_fc2_0.05<-res[which(abs(res$logFC)>=2 & res$P.Value<=0.05),]
#gx with logFC >abs(1.5)(no p-value cut-off) #5 probes
res_fc15<-res[which(abs(res$logFC)>=1.5),]
#gx with logFC >abs(1.5) (p-value<=0.05 (non-adjusted))
res_fc15_0.05<-res[which(abs(res$logFC)>=1.5 & res$P.Value<=0.05),] # same 5 probes as above

#############Integrate with Methylation
setwd("~/Experiments/GEM/CS/CS_110814")
#logFC explained: if +ve, then higher expression in mut; if -ve, then higher expression in wt

#split gx and me by unique gene name
me<-read.delim("Meth_3057.txt",sep="\t",header=T)
me.l<-split(me,me$UCSC_REFGENE_NAME,drop=TRUE)
gx<-read.delim("cs_results_110814.txt",sep="\t",header=T)
#gx.l<-split(gx,gx$Symbol,drop=TRUE) too memory intensive
#instead copied the cs_results_110814.txt file on the server and ran ####awk '{print > $5".txt"}'
cs_results_110814.txt#### to split the file by Symbol into multiple files each named after one gene;
these were then copied back onto the desktop in the folder gx_split_bygene

setwd("~/Experiments/GEM/CS/CS_110814/gx_split_bygene/")
a<-list.files()
gx.l<-lapply(a, read.delim, sep = "\t",header=FALSE)
names(gx.l)<-gsub(".txt","",a)
gx.l<-lapply(gx.l,setNames,nm=names(gx))

#in genes in me, remove probes not in TSS region, and remove any genes with no probes after that
filter
keep_TSS<-function(data) {
  data<-data[which(data$UCSC_REFGENE_GROUP=="TSS1500" |
data$UCSC_REFGENE_GROUP=="TSS200"),]
}
me2.l<-lapply(me.l,keep_TSS)
to_remove<-vector()
for (i in 1:length(me2.l)){
  if (nrow(me2.l[[i]])==0) to_remove<-c(to_remove,names(me2.l)[i])
}
isNameInIndex <- names(me2.l) %in% to_remove
me3.l<-me2.l[!isNameInIndex]
```

```
#keep in me genes only those with either one probe or multiple probes that agree within each gene
#all the ones left are hypermethylated in MUT so all agree

#keep in gx only those with either one probe or multiple probes that agree within the gene
options(warn=2)#transforms warnings into errors so stops loop at problematic row# use warn=0
tor eturn to default
gx.l<-gx.l[-3546] #contained the cs_results_110814 file in there
to_remove2<-vector()
for (i in 1:length(gx.l)){
  if (nrow(gx.l[[i]])>1) {
    tot<-nrow(gx.l[[i]])
    pos<-nrow(gx.l[[i]][which(gx.l[[i]]$logFC > 0),])
    neg<-nrow(gx.l[[i]][which(gx.l[[i]]$logFC < 0),])
    if (tot!=pos & tot!=neg) to_remove2<-c(to_remove2,names(gx.l)[i])
  }
}
isNameInIndex2 <- names(gx.l) %in% to_remove2
gx2.l<-gx.l[!isNameInIndex2]

#make list for gx and me with only genes contained in both
#525 unique gene names in me; 16342 unique gene names in gx
b<-intersect(unique(names(me3.l)),unique(names(gx2.l)))
length(b)  #387
me4.l<-me3.l[which(names(me3.l) %in% b)]
gx4.l<-gx2.l[which(names(gx2.l) %in% b)]
setwd("~/Experiments/GEM/CS/CS_110814")
save(me4.l, file="me4.RData")
save(gx4.l,file="gx4.RData")

#single data frame for gx
res<-data.frame(rep(0,length(b)),rep(0,length(b)))
names(res)<-c("Gene","AVGlogFC")
for (i in 1:length(gx4.l)){
  res[i,1]<-names(gx4.l[i])
  res[i,2]<-mean(gx4.l[[i]]$logFC)
}
write.table(res,"results_387genes_probemin1.txt",row.names=FALSE,quote=FALSE,sep="\t")

#how many go each way?
nrow(res[which(res[,2]<0),]) #213 which represents 55% of the 387 genes....

#number of probes in each me gene?
a<-vector()
for (i in 1:length(me3.l)){a<-c(a,(nrow(me3.l[[i]])))}

#repeat analysis above with only me genes with num probes>=n
#######n>=2
to_remove3<-vector()
for (i in 1:length(me3.l)){
  if (nrow(me3.l[[i]])<2) to_remove3<-c(to_remove3,names(me3.l)[i])
}
isNameInIndex3 <- names(me3.l) %in% to_remove3
me4.l<-me3.l[!isNameInIndex3]

#228 unique gene names in me; 16342 unique gene names in gx
b2<-intersect(unique(names(me4.l)),unique(names(gx2.l)))
length(b2)  #171

me4.l<-me4.l[which(names(me4.l) %in% b2)]
gx4.l<-gx2.l[which(names(gx2.l) %in% b2)]
```

```
#single data frame for gx
res2<-data.frame(rep(0,length(b2)),rep(0,length(b2)))
names(res2)<-c("Gene","AVGlogFC")

for (i in 1:length(gx4.l)){
  res2[i,1]<-names(gx4.l[i])
  res2[i,2]<-mean(gx4.l[[i]]$logFC)
}

#how many go each way?
nrow(res2[which(res2[,2]<0),]) #102 which represents 59.6% of the 171 genes....
write.table(res2,"results_171genes_probemin2.txt",row.names=FALSE,quote=FALSE,sep="\t")

#######n>=3
to_remove3<-vector()
for (i in 1:length(me3.l)){
  if (nrow(me3.l[[i]])<3) to_remove3<-c(to_remove3,names(me3.l)[i])
}
isNameInIndex3 <- names(me3.l) %in% to_remove3
me4.l<-me3.l[!isNameInIndex3]

#113 unique gene names in me; 16342 unique gene names in gx
b2<-intersect(unique(names(me4.l)),unique(names(gx2.l)))
length(b2)  #89

me4.l<-me4.l[which(names(me4.l) %in% b2)]
gx4.l<-gx2.l[which(names(gx2.l) %in% b2)]

#single data frame for gx
res2<-data.frame(rep(0,length(b2)),rep(0,length(b2)))
names(res2)<-c("Gene","AVGlogFC")

for (i in 1:length(gx4.l)){
  res2[i,1]<-names(gx4.l[i])
  res2[i,2]<-mean(gx4.l[[i]]$logFC)
}

#how many go each way?
nrow(res2[which(res2[,2]<0),]) #55 which represents 61.8% of the 89 genes....
write.table(res2,"results_89genes_probemin3.txt",row.names=FALSE,quote=FALSE,sep="\t")


#######n>=5
to_remove3<-vector()
for (i in 1:length(me3.l)){
  if (nrow(me3.l[[i]])<5) to_remove3<-c(to_remove3,names(me3.l)[i])
}
isNameInIndex3 <- names(me3.l) %in% to_remove3
me4.l<-me3.l[!isNameInIndex3]

#32 unique gene names in me; 16342 unique gene names in gx
b2<-intersect(unique(names(me4.l)),unique(names(gx2.l)))
length(b2)  #25

me4.l<-me4.l[which(names(me4.l) %in% b2)]
gx4.l<-gx2.l[which(names(gx2.l) %in% b2)]

#single data frame for gx
res2<-data.frame(rep(0,length(b2)),rep(0,length(b2)))
names(res2)<-c("Gene","AVGlogFC")
```

```
for (i in 1:length(gx4.l)){
  res2[i,1]<-names(gx4.l[i])
  res2[i,2]<-mean(gx4.l[[i]]$logFC)
}

#how many go each way?
nrow(res2[which(res2[,2]<0),]) #16 which represents 64% of the 25 genes....
write.table(res2,"results_25genes_probemin5.txt",row.names=FALSE,quote=FALSE,sep="\t")
```

CNV ANALYSIS SCRIPT

~~~~
#450K CNV: All samples
Ran modified ChAMP CNA on all MUTs (test+rep) and all WTs (test+rep) + prepared files for  and ran GISTIC
##Run modified ChAMP CNA
The script of champ CNA function was adapted by Andy to use a reference control group; need more than one sample in reference grp so actually using 'Blood' grp as reference and 'reference' grp as blood sample grp.
###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples"
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples")
library(ChAMP)
myLoad=champ.load(methValue="B",QCimages=TRUE,filterXY=FALSE,filterDetP=TRUE,filterBeads=TRUE,beadCutoff=0.05,detPcut=0.01)
boxplot(myLoad$intensity) #intensities look evenly distributed: no batch correction
source("R:/R/CNV/champ_cna_modif_240414.R")
```
The output can be found in resultsChamp/CNA
##Prep files for  and run GISTIC 2.0
Take all the output files for each sample group and stick them together;write file without header;also prepare marker files for hg19
###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples"
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples")
all<-list.files("resultsChamp/CNA",pattern=".txt",full.names=TRUE)
mut<-all[grep("CNA/MUT",all)]
wt<-all[grep("CNA/WT",all)]
mut_full<-do.call("rbind", lapply(mut, read.delim, header = FALSE))
mut_full<-mut_full[-which(mut_full[,1]=="ID"),]
write.table(mut_full,"all_MUT_for_gistic.txt",col.names=F,sep="\t",quote=F,row.names=FALSE)
wt_full<-do.call("rbind", lapply(wt, read.delim, header = FALSE))
wt_full<-wt_full[-which(wt_full[,1]=="ID"),]
write.table(wt_full,"all_WT_for_gistic.txt",col.names=F,sep="\t",quote=F,row.names=FALSE)
#marker files for GISTIC;remove unknown chromosomes from hg19 version
hg19<-read.delim("markers_file_hg19.txt",sep="\t",header=F)
hg19b<-hg19[-which(hg19[,2]==""),]
hg19b<-droplevels(hg19b)
write.table(hg19b,"markers_file_hg19.txt",sep="\t",quote=FALSE,col.names=FALSE,row.names=FALSE)
```
###Run GISTIC 2.0 for all samples separated into MUT and WT
Ran it twice(once for mut once for wt) at http://genepattern.broadinstitute.org/ with the following settings
- refgene file: Human Hg19
- seg file: all_MUT_for_gistic.txt OR all_WT_for_gistic.txt
- markers file: markers_file_hg19.txt
- array list file: none
- cnv file: none
- gene gistic: yes
- amp/del threshold: 0.3
- join segment file: 4
- qv thresh: 0.05
- removeX: yes
- cap val: 1.5
- confidence level: 0.95
- run broad analysis: yes
~~~~

- broad length cutoff: 0.98
- max sample segs: 10000
- arm peel: yes
- output prefix:allmut OR allWT

Output is saved within "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples/GISTIC_output"

###Run GISTIC 2.0 for all samples combined into one group
Used same settings as above but the seg file is all_MUTandWT_for_gistic.txt

Output is saved within "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples_mixed/GISTIC_output"

###Run GISTIC 2.0 for all 450K samples that overlap with a CytoSNP sample
Run twice (once for mut, once for wt) and used same settings as above with the all_MUT_for_gistic.txt or all_WT_for_gistic.txt seg file
Use array list files arraylistfile_MUT_450matchingcCyto and arraylistfile_WT_450matchingcCyto

Output is saved within "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_450samples_matching_cytosnp/GISTIC_output"

#CytoSNP CNV: all samples
Run DnaCopy and GISTIC on cytosnp data (adapted from script by A.Feber);these are hg19 build

##In R:
working directory:"~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytosnp"

~~~~
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytosnp")
cyto<-read.table("Chondro_R_vlaue.txt", header=TRUE, sep="\t")

#change sample names to match 450k
name_trans<-read.delim("cytosnp_450k_names.txt",sep="\t",header=F)
temp<-seq(1:11)
for (i in 1:11){
  temp[i]<-
as.character(name_trans[which(paste("X",name_trans[,1],".R",sep="")==names(cyto[i+3])),2])
}

new_names<-c(names(cyto[1:3]),temp,"BC")
names(cyto)<-new_names

names<-names(cyto)

#Quantile normalise
library(preprocessCore)

tmp<-as.matrix(cyto[,4:15])
tmpqn<-normalize.quantiles(tmp)
cytoqn<-cbind(cyto[,1:3],tmpqn)
names(cytoqn)<-names

#Calculate Log2
cytoqnlog<-log2(cytoqn[,4:15])
cytoqnlog<-cbind(cyto[,1:3], cytoqnlog)

#quantile norm log ratio
cytoqnlogratio<-vector()

for (i in 4:15){
  test<-cytoqnlog[,i]-cytoqnlog[,15]
~~~~

~~~
  cytoqnlogratio<-cbind(cytoqnlogratio,test)
}
cytoqnlogratio<-cbind(cytoqn[,1:3],cytoqnlogratio)
colnames(cytoqnlogratio)<-colnames(cytoqn)

#Replcases Chr X adn Y with 23 and 24
levels(cytoqnlogratio$Chr)[levels(cytoqnlogratio$Chr)=='X']='23'
levels(cytoqnlogratio$Chr)[levels(cytoqnlogratio$Chr)=='Y']='24'

#converts factors to numeric intergers
ymp<-as.numeric(levels(cytoqnlogratio$Chr))[cytoqnlogratio$Chr]

#Runs CNA
library(DNAcopy)

CNA.object <- CNA(cbind(cytoqnlogratio[,4:14]), ymp, cytoqnlogratio$Position ,data.type =
"logratio", sampleid = names[4:14])

smoothed.CNA.object <- smooth.CNA(CNA.object)
segment.smoothed.CNA.object <- segment(smoothed.CNA.object, verbose = 1)

seg<-print(segment.smoothed.CNA.object)
table_name<-"ALL_cyto_qn.txt"


#remove sample that wasnt run on 450k WT 13 and reorder table with muts and WT
seg2<-seg[-which(seg[,1]=="WT_13_FF_13394"),]

write.table(seg2,table_name, sep="\t", col.names=F, row.names=F, quote=FALSE)

#make markers file for gistic
markers<-cyto[,1:3]
write.table(markers,"markers_file_cyto_hg19.txt",sep="\t",quote=FALSE,col.names=FALSE,row.names=FALSE)

```

##Run GISTIC 2.0

Use same settings as above
Out put is saved within "~/Experiments/Other
Experiments/CNV/20140815_CS_CNV/CNV_cytosnp/GISTIC_output"




#450K vs CytoSNP on matched samples

##Preparing data
###In R:

working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV"

```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV")

cyto<-read.delim("CNV_cytosnp/ALL_cyto_qn.txt",sep="\t",header=F)
~~~

~~~~
k450mut<-read.delim("CNV_450samples_matching_cytosnp/all_MUT_for_gistic.txt",sep="\t",header=F)
k450wt<-read.delim("CNV_450samples_matching_cytosnp/all_WT_for_gistic.txt",sep="\t",header=F)
k450<-rbind(k450mut[which(gsub("._qn", "", k450mut[,1]) %in% cyto[,1]),],k450wt[which(gsub("._qn", "", k450wt[,1]) %in% cyto[,1]),])


#calculate segment lengths
seg_length<-function(df){
 for (i in 1:nrow(df)){
   df[i,7]<-df[i,4]-df[i,3]
 }
 return(df)
}

cyto<-seg_length(cyto)
k450<-seg_length(k450)
k450[,1]<-gsub("._qn", "", k450[,1])


cyto_amp<-cyto[which(cyto[,6]>0),]
cyto_amp$ALTID<-paste("chr",cyto_amp[,2],"_",cyto_amp[,3],"_",cyto_amp[,4],sep="")
cyto_amp<-split(cyto_amp,cyto_amp[,1])

cyto_del<-cyto[which(cyto[,6]<0),]
cyto_del$ALTID<-paste("chr",cyto_del[,2],"_",cyto_del[,3],"_",cyto_del[,4],sep="")
cyto_del<-split(cyto_del,cyto_del[,1])



k450_amp<-k450[which(k450[,6]>0),]
k450_amp$ALTID<-paste("chr",k450_amp[,2],"_",k450_amp[,3],"_",k450_amp[,4],sep="")
k450_amp<-split(k450_amp,k450_amp[,1])

k450_del<-k450[which(k450[,6]<0),]
k450_del$ALTID<-paste("chr",k450_del[,2],"_",k450_del[,3],"_",k450_del[,4],sep="")
k450_del<-split(k450_del,k450_del[,1])


par(new=F)
hist(log10(k450[,7]),col=NULL,border="orange",xlim=c(0,9),main="",xlab=NULL,ylim=c(0,1000))
par(new=T)
hist(log10(cyto[,7]),col=NULL,border="black",xlim=c(0,9),main=NULL,xlab="log10(CNV size)",ylim=c(0,1000),ylab=NULL,yaxt='n')


```

##Comparing Cytosnp and 450 on matched samples:LARGE ALTERATIONS (>=10Mb)

###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV"
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV")

#select large alterations (>10Mb)
cyto_large<-cyto[which(cyto[,7]>=10000000),]
k450_large<-k450[which(k450[,7]>=10000000),]

#Find overlap: prepare files for bedtools in galaxy
~~~~

~~~~
gal450<-
data.frame(paste("chr",k450_large[,2],sep=""),k450_large[,3],k450_large[,4],k450_large[,1])
names(gal450)<-c("chr","start","stop","SegmentID")
gal450_l<-list()
count=1
for (i in unique(gal450$SegmentID)){
  gal450_l[[count]]<-subset(gal450,gal450$SegmentID==i)
  gal450_l[[count]]$SegmentID<-paste(i,"_",seq(1,nrow(gal450_l[[count]]),1),sep="")
  count=count+1
}
names(gal450_l)<-unique(gal450$SegmentID)

lapply(names(gal450_l),function (x) write.table(format(gal450_l[[x]],scientific=FALSE),
file=paste("CNV_cytoVS450/",x,"_coords_withID_450_forgalaxay.txt",sep=""),sep="\t",col.names=F,
row.names=F,quote=F )   )

galcyto<-data.frame(paste("chr",cyto_large[,2],sep=""),cyto_large[,3],cyto_large[,4],cyto_large[,1])
names(galcyto)<-c("chr","start","stop","SegmentID")
galcyto_l<-list()
count=1
for (i in unique(galcyto$SegmentID)){
  galcyto_l[[count]]<-subset(galcyto,galcyto$SegmentID==i)
  galcyto_l[[count]]$SegmentID<-paste(i,"_",seq(1,nrow(galcyto_l[[count]]),1),sep="")
  count=count+1
}
names(galcyto_l)<-unique(galcyto$SegmentID)

lapply(names(galcyto_l),function (x) write.table(format(galcyto_l[[x]],scientific=FALSE),
file=paste("CNV_cytoVS450/",x,"_coords_withID_cyto_forgalaxay.txt",sep=""),sep="\t",col.names=F,
row.names=F,quote=F )   )

```
~~~~

###In Galaxy:

The files coords_withID_450_forgalaxay.txt and coords_withID_cyto_forgalaxay.txt for each matched sample pair were uploaded to Galaxy(http://bifx-core.bio.ed.ac.uk:8080/galaxy/)as interval format to assess overlaps using the tool in: 'Operate on Genomic Intervals-->Join', with the following settings:
- Join: coords_withID_450_forgalaxay.txt
- with: coords_withID_cyto_forgalaxay.txt
- with min overlap: 1bp
- Return: Only records that are joined (INNER JOIN)

The output file was downloaded to R:/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVS450/xxxx_OVERLAP.txt

###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVS450"

~~~~
```{r,eval=F}

setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVS450")
temp=list.files(pattern="*_OVERLAP")
sample.l<-sapply(temp,function(x)
read.delim(x,header=FALSE,sep="\t",row.names=NULL),simplify=FALSE)

results<-matrix(rep(0,50),nrow=10)
names<-names(sample.l)
for (i in 1:length(names)){
~~~~

~~~~
  results[i,1]<-gsub("_OVERLAP.txt", "", names[i])
  results[i,2]<-nrow(gal450[which(gal450$SegmentID==results[i,1]),])
  results[i,3]<-nrow(galcyto[which(galcyto$SegmentID==results[i,1]),])
  results[i,4]<-nrow(unique(sample.l[[i]][4]))
  results[i,5]<-100*(as.numeric(results[i,4])/as.numeric(results[i,2]))
}
colnames(results)<-
c("SampleID","N_large_alt_in_450","N_large_alt_in_cyto","N_large_alt_overlap_in_cyto","Percentage_overlap_in_cytosnp")


#split amp and del


create_ALTID<-function(df){
  df$ALTID_k450<-paste(df[,1],"_",df[,2],"_",df[,3],sep="")
  df$ALTID_cyto<-paste(df[,5],"_",df[,6],"_",df[,7],sep="")
  return(df)
}

sample.l<-lapply(sample.l,create_ALTID)

results_amp<-matrix(rep(0,50),nrow=10)
names<-names(sample.l)
for (i in 1:length(names)){
  results_amp[i,1]<-gsub("_OVERLAP.txt", "", names[i])
  results_amp[i,2]<-nrow(k450_amp[[i]][which(k450_amp[[i]][,7]>=10000000),])
  results_amp[i,3]<-nrow(cyto_amp[[i]][which(cyto_amp[[i]][,7]>=10000000),])
  results_amp[i,4]<-length(unique(sample.l[[i]][which(sample.l[[i]][,9] %in% k450_amp[[i]][,8] &
sample.l[[i]][,10] %in% cyto_amp[[i]][,8]),4]))
  results_amp[i,5]<-100*(as.numeric(results_amp[i,4])/as.numeric(results_amp[i,2]))
}
colnames(results_amp)<-
c("SampleID","N_large_amp_in_450","N_large_amp_in_cyto","N_large_amp_overlap_in_cyto","Percentage_overlap_in_cytosnp")

results_del<-matrix(rep(0,50),nrow=10)
names<-names(sample.l)
for (i in 1:length(names)){
  results_del[i,1]<-gsub("_OVERLAP.txt", "", names[i])
  results_del[i,2]<-nrow(k450_del[[i]][which(k450_del[[i]][,7]>=10000000),])
  results_del[i,3]<-nrow(cyto_del[[i]][which(cyto_del[[i]][,7]>=10000000),])
  results_del[i,4]<-length(unique(sample.l[[i]][which(sample.l[[i]][,9] %in% k450_del[[i]][,8] &
sample.l[[i]][,10] %in% cyto_del[[i]][,8]),4]))
  results_del[i,5]<-100*(as.numeric(results_del[i,4])/as.numeric(results_del[i,2]))
}
colnames(results_del)<-
c("SampleID","N_large_del_in_450","N_large_del_in_cyto","N_large_del_overlap_in_cyto","Percentage_overlap_in_cytosnp")

```


##Comparing Cytosnp and 450 on matched samples:SMALL ALTERATIONS (<=1Mb)

###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV"
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV")

#select small alterations (<=1Mb)
~~~~

~~~~
cyto_small<-cyto[which(cyto[,7]<=1000000),]
k450_small<-k450[which(k450[,7]<=1000000),]

#Find overlap: prepare files for bedtools in galaxy

gal450small<-
data.frame(paste("chr",k450_small[,2],sep=""),k450_small[,3],k450_small[,4],k450_small[,1])
names(gal450small)<-c("chr","start","stop","SegmentID")
gal450small_l<-list()
count=1
for (i in unique(gal450small$SegmentID)){
  gal450small_l[[count]]<-subset(gal450small,gal450small$SegmentID==i)
  gal450small_l[[count]]$SegmentID<-paste(i,"_",seq(1,nrow(gal450small_l[[count]]),1),sep="")
  count=count+1
}
names(gal450small_l)<-unique(gal450small$SegmentID)

lapply(names(gal450small_l),function (x) write.table(format(gal450small_l[[x]],scientific=FALSE),
file=paste("CNV_cytoVS450/",x,"_coords_withID_450_forgalaxay_SMALL_ALT.txt",sep=""),sep="\t",
col.names=F,row.names=F,quote=F )   )

galcytosmall<-
data.frame(paste("chr",cyto_small[,2],sep=""),cyto_small[,3],cyto_small[,4],cyto_small[,1])
names(galcytosmall)<-c("chr","start","stop","SegmentID")
galcytosmall_l<-list()
count=1
for (i in unique(galcytosmall$SegmentID)){
  galcytosmall_l[[count]]<-subset(galcytosmall,galcytosmall$SegmentID==i)
  galcytosmall_l[[count]]$SegmentID<-paste(i,"_",seq(1,nrow(galcytosmall_l[[count]]),1),sep="")
  count=count+1
}
names(galcytosmall_l)<-unique(galcytosmall$SegmentID)

lapply(names(galcytosmall_l),function (x) write.table(format(galcytosmall_l[[x]],scientific=FALSE),
file=paste("CNV_cytoVS450/",x,"_coords_withID_cyto_forgalaxay_SMALL_ALT.txt",sep=""),sep="\t",
col.names=F,row.names=F,quote=F )   )

```
~~~~

###In Galaxy:
The files coords_withID_450_forgalaxay_SMALL_ALT.txt and coords_withID_cyto_forgalaxay_SMALL_ALT.txt for each matched sample pair were uploaded to Galaxy(http://bifx-core.bio.ed.ac.uk:8080/galaxy/)as interval format to assess overlaps using the tool in: 'Operate on Genomic Intervals-->Join', with the following settings:
- Join: coords_withID_450_forgalaxay_SMALL_ALT.txt
- with: coords_withID_cyto_forgalaxay_SMALL_ALT.txt
- with min overlap: 1bp
- Return: Only records that are joined (INNER JOIN)

The output file was downloaded to R:\Experiments\Other Experiments\CNV\20140815_CS_CNV\CNV_cytoVS450\small\xxxx_SMALL_ALT_OVERLAP.txt

###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVS450/small"

~~~~
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVS450/small")
temp=list.files(pattern="*_SMALL_ALT_OVERLAP")
sample_small.l<-sapply(temp,function(x)
read.delim(x,header=FALSE,sep="\t",row.names=NULL),simplify=FALSE)
~~~~

```
results_small<-matrix(rep(0,50),nrow=10)
names<-names(sample_small.l)
for (i in 1:length(names)){
  results_small[i,1]<-gsub("_SMALL_ALT_OVERLAP.txt", "", names[i])
  results_small[i,2]<-nrow(gal450small[which(gal450small$SegmentID==results_small[i,1]),])
  results_small[i,3]<-nrow(galcytosmall[which(galcytosmall$SegmentID==results_small[i,1]),])
  results_small[i,4]<-nrow(unique(sample_small.l[[i]][4]))
  results_small[i,5]<-100*(as.numeric(results_small[i,4])/as.numeric(results_small[i,2]))
}
colnames(results_small)<-
c("SampleID","N_small_alt_in_450","N_small_alt_in_cyto","N_small_alt_overlap_in_cyto","Percentage_overlap_in_cytosnp")


#split amp and del


create_ALTID<-function(df){
  df$ALTID_k450<-paste(df[,1],"_",df[,2],"_",df[,3],sep="")
  df$ALTID_cyto<-paste(df[,5],"_",df[,6],"_",df[,7],sep="")
  return(df)
}

sample_small.l<-lapply(sample_small.l,create_ALTID)

results_small_amp<-matrix(rep(0,50),nrow=10)
names<-names(sample_small.l)
for (i in 1:length(names)){
  results_small_amp[i,1]<-gsub("_SMALL_ALT_OVERLAP.txt", "", names[i])
  results_small_amp[i,2]<-nrow(k450_amp[[i]][which(k450_amp[[i]][,7]<=1000000),])
  results_small_amp[i,3]<-nrow(cyto_amp[[i]][which(cyto_amp[[i]][,7]<=1000000),])
  results_small_amp[i,4]<-length(unique(sample_small.l[[i]][which(sample_small.l[[i]][,9] %in%
k450_amp[[i]][,8] & sample_small.l[[i]][,10] %in% cyto_amp[[i]][,8]),4]))
  results_small_amp[i,5]<-
100*(as.numeric(results_small_amp[i,4])/as.numeric(results_small_amp[i,2]))
}
colnames(results_small_amp)<-
c("SampleID","N_small_amp_in_450","N_small_amp_in_cyto","N_small_amp_overlap_in_cyto","Percentage_overlap_in_cytosnp")

results_small_del<-matrix(rep(0,50),nrow=10)
names<-names(sample_small.l)
for (i in 1:length(names)){
  results_small_del[i,1]<-gsub("_SMALL_ALT_OVERLAP.txt", "", names[i])
  results_small_del[i,2]<-nrow(k450_del[[i]][which(k450_del[[i]][,7]<=1000000),])
  results_small_del[i,3]<-nrow(cyto_del[[i]][which(cyto_del[[i]][,7]<=1000000),])
  results_small_del[i,4]<-length(unique(sample_small.l[[i]][which(sample_small.l[[i]][,9] %in%
k450_del[[i]][,8] & sample_small.l[[i]][,10] %in% cyto_del[[i]][,8]),4]))
  results_small_del[i,5]<-
100*(as.numeric(results_small_del[i,4])/as.numeric(results_small_del[i,2]))
}
colnames(results_small_del)<-
c("SampleID","N_small_del_in_450","N_small_del_in_cyto","N_small_del_overlap_in_cyto","Percentage_overlap_in_cytosnp")

```

much lower overlap: check how many markers in cytosnp in 450k regions where no overlap

###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVS450/small"

~~~
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVS450/small")
`%ni%` <- Negate(`%in%`)

names<-names(sample_small.l)
nool_amp<-list()
nool_del<list()
for (i in 1:length(names)){
 nool_amp[[i]]<-k450_amp[[i]][which(k450_amp[[i]][,7]<=1000000 & k450_amp[[i]][,8] %ni%
sample_small.l[[i]][,9]),]
 nool_del[[i]]<-k450_del[[i]][which(k450_del[[i]][,7]<=1000000 & k450_del[[i]][,8] %ni%
sample_small.l[[i]][,9]),]
}

markers<-read.delim("R:/Experiments/Other
Experiments/CNV/20140815_CS_CNV/CNV_cytosnp/markers_file_cyto_hg19.txt",sep="\t",header=
F)

cyto_mark_num<-function(df){
 for (i in 1:nrow(df)){
   df$num_cyto_markers[i]<-nrow(markers[which(markers[,3]>=df[i,3] & markers[,3]<=df[i,4]),])
 }
 return(df)
}

nool_amp2<-lapply(nool_amp,cyto_mark_num)
nool_del2<-lapply(nool_del,cyto_mark_num)

nool_amp_results<-matrix(rep(0,40),nrow=10)
for (i in 1:length(names)){
 nool_amp_results[i,1]<-gsub("_SMALL_ALT_OVERLAP.txt", "", names[i])
 nool_amp_results[i,2]<-nrow(nool_amp2[[i]])
 nool_amp_results[i,3]<-nrow(nool_amp2[[i]][which(nool_amp2[[i]][,9]<4),])
 nool_amp_results[i,4]<-
100*(as.numeric(nool_amp_results[i,3])/as.numeric(nool_amp_results[i,2]))
}
colnames(nool_amp_results)<-
c("SampleID","N_small_amp_nooverlap","N_small_amp_with_lessthan4cytomarkers","Percentage_s
mall_amp_with_lessthan4cytomarkers")


nool_del_results<-matrix(rep(0,40),nrow=10)
for (i in 1:length(names)){
 nool_del_results[i,1]<-gsub("_SMALL_ALT_OVERLAP.txt", "", names[i])
 nool_del_results[i,2]<-nrow(nool_del2[[i]])
 nool_del_results[i,3]<-nrow(nool_del2[[i]][which(nool_del2[[i]][,9]<4),])
 nool_del_results[i,4]<-100*(as.numeric(nool_del_results[i,3])/as.numeric(nool_del_results[i,2]))
}
colnames(nool_del_results)<-
c("SampleID","N_small_del_nooverlap","N_small_del_with_lessthan4cytomarkers","Percentage_smal
l_del_with_lessthan4cytomarkers")

```
~~~

##Writing the comparisons to file

~~~
###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVS450"

```{r,eval=F}

setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVS450")

write.table(results_amp,"Large_amp_10mb_comp.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)
write.table(results_del,"Large_del_10mb_comp.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)
write.table(results_small_amp,"Small_amp_1mb_comp.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)
write.table(results_small_del,"Small_del_1mb_comp.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)

write.table(nool_amp_results,"Small_amp_nool_cytomarkers.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)
write.table(nool_del_results,"Small_del_nool_cytomarkers.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)

```



#Exome vs CytoSNP
The 10 samples run on CytoSNP were also processed by Tarpey et al. by exome and/or SNP 6.0 array
Very low overlap between 450K and Exome, so what's the overlap between exome and snp?


##Preparing Data:

###In R:
working directory:"~/Experiments/Other Experiments/CNV/20140815_CS_CNV"

```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV")

cyto<-read.delim("CNV_cytosnp/ALL_cyto_qn.txt",sep="\t",header=F)

#read in exome samples from amf paper (Tarpey et al 2013)
amf<-read.delim("CNV_allsamples_mixed/sup_table7_Tarpey2013.txt",sep="\t",header=T)

amf<-amf[,c(5,2,3,4,1,1)]

#keep only the amf samples that overlap with cytosnp
equiv_ID<-read.delim("CNV_allsamples_mixed/Tarpey_PG_ID_overlaps.txt",sep="\t",header=T)

amf2<-amf[which(amf$Sample %in% equiv_ID$Tarpey_ID),]


#rename the amf samples with the cytosnp sample names

amf3<-merge(amf2,equiv_ID, by.x="Sample", by.y="Tarpey_ID")
amf3[,1]<-amf3[,7]
amf3<-amf3[,-7]

#calculate segment lengths
seg_length<-function(df){
~~~

~~~
  for (i in 1:nrow(df)){
    df[i,7]<-df[i,4]-df[i,3]
  }
  return(df)
}

cyto<-seg_length(cyto)
amf3<-seg_length(amf3)

cyto_amp<-cyto[which(cyto[,6]>0),]
cyto_amp$ALTID<-paste("chr",cyto_amp[,2],"_",cyto_amp[,3],"_",cyto_amp[,4],sep="")
cyto_amp<-split(cyto_amp,cyto_amp[,1])

cyto_del<-cyto[which(cyto[,6]<0),]
cyto_del$ALTID<-paste("chr",cyto_del[,2],"_",cyto_del[,3],"_",cyto_del[,4],sep="")
cyto_del<-split(cyto_del,cyto_del[,1])

amf_amp<-amf3[which(amf3[,6]=="Amp"),]
amf_amp$ALTID<-paste("chr",amf_amp[,2],"_",amf_amp[,3],"_",amf_amp[,4],sep="")
amf_amp<-split(amf_amp,amf_amp[,1])

amf_del<-amf3[which(amf3[,6]=="Del"),]
amf_del$ALTID<-paste("chr",amf_del[,2],"_",amf_del[,3],"_",amf_del[,4],sep="")
amf_del<-split(amf_del,amf_del[,1])


```


##Comparing Exome and CytoSNP on matched samples: Large Alterations (>=10Mb):


###In R:
working directory: “~/Experiments/Other Experiments/CNV/20140815_CS_CNV”
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV")

#select large alterations (>10Mb)
cyto_large<-cyto[which(cyto[,7]>=10000000),]
amf_large<-amf3[which(amf3[,7]>=10000000),]

#prepare files for galaxy:cyto
galcyto<-data.frame(paste("chr",cyto_large[,2],sep=""),cyto_large[,3],cyto_large[,4],cyto_large[,1])
names(galcyto)<-c("chr","start","stop","SegmentID")
galcyto_l<-list()
count=1
for (i in unique(galcyto$SegmentID)){
  galcyto_l[[count]]<-subset(galcyto,galcyto$SegmentID==i)
  galcyto_l[[count]]$SegmentID<-paste(i,"_",seq(1,nrow(galcyto_l[[count]]),1),sep="")
  count=count+1
}
names(galcyto_l)<-unique(galcyto$SegmentID)

lapply(names(galcyto_l),function (x) write.table(format(galcyto_l[[x]],scientific=FALSE),
file=paste("CNV_cytoVSamf/",x,"_coords_withID_cyto_forgalaxay.txt",sep=""),sep="\t",col.names=F,
row.names=F,quote=F )   )

#prepare files for galaxy:amf
galamf<-data.frame(paste("chr",amf_large[,2],sep=""),amf_large[,3],amf_large[,4],amf_large[,1])
names(galamf)<-c("chr","start","stop","SegmentID")
galamf_l<-list()
~~~

~~~
count=1
for (i in unique(galamf$SegmentID)){
 galamf_l[[count]]<-subset(galamf,galamf$SegmentID==i)
 galamf_l[[count]]$SegmentID<-paste(i,"_",seq(1,nrow(galamf_l[[count]]),1),sep="")
 count=count+1
}
names(galamf_l)<-unique(galamf$SegmentID)

lapply(names(galamf_l),function (x) write.table(format(galamf_l[[x]],scientific=FALSE),
file=paste("CNV_cytoVSamf/",x,"_coords_withID_amf_forgalaxay.txt",sep=""),sep="\t",col.names=F,r
ow.names=F,quote=F )   )
#only one large alteration in amf
```

###In Galaxy:
The files coords_withID_cyto_forgalaxay.txt and coords_withID_amf_forgalaxay.txt for the matched
sample pair (WT14) were uploaded to Galaxy(http://bifx-core.bio.ed.ac.uk:8080/galaxy/)as
interval format to assess overlaps using the tool in: 'Operate on Genomic Intervals–>Join', with the
following settings:
- Join: coords_withID_cyto_forgalaxay.txt
- with: coords_withID_amf_forgalaxay.txt
- with min overlap: 1bp
- Return: Only records that are joined (INNER JOIN)


The output file was downloaded to "R:/Experiments/Other
Experiments/CNV/20140815_CS_CNV/CNV_cytoVSamf/xxxx_OVERLAP.txt"

###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVSamf"
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVSamf")

temp=list.files(pattern="*_OVERLAP")
sample.l<-sapply(temp,function(x)
read.delim(x,header=FALSE,sep="\t",row.names=NULL),simplify=FALSE)
names(sample.l)<-gsub("_OVERLAP.txt", "", names(sample.l))

results<-matrix(rep(0,50),nrow=10)
names<-names(sample.l)
for (i in 1:length(names)){
 results[i,1]<-names[i]
 results[i,2]<-nrow(galamf[which(galamf$SegmentID==results[i,1]),])
 results[i,3]<-nrow(galcyto[which(galcyto$SegmentID==results[i,1]),])
 results[i,4]<-nrow(unique(sample.l[[i]][4]))
 results[i,5]<-100*(as.numeric(results[i,4])/as.numeric(results[i,2]))
}
colnames(results)<-
c("SampleID","N_large_alt_in_amf","N_large_alt_in_cyto","N_large_alt_overlap_in_cyto","Percentage_
amf_that_overlap_in_cytosnp")



#split amp and del



create_ALTID<-function(df){
 df$ALTID_amf<-paste(df[,5],"_",df[,6],"_",df[,7],sep="")
 df$ALTID_cyto<-paste(df[,1],"_",df[,2],"_",df[,3],sep="")
 return(df)
~~~

~~~~
}

sample.l<-lapply(sample.l,create_ALTID)

results_amp<-matrix(rep(0,50),nrow=10)
names<-names(sample.l)
for (i in 1:length(names)){
  results_amp[i,1]<-names[i]
  results_amp[i,2]<-
nrow(amf_amp[[match(results_amp[i,1],names(amf_amp))]][which(amf_amp[[match(results_amp[i,1],names(amf_amp))]][,7]>=10000000),])
  results_amp[i,3]<-
nrow(cyto_amp[[match(results_amp[i,1],names(cyto_amp))]][which(cyto_amp[[match(results_amp[i,1],names(cyto_amp))]][,7]>=10000000),])
  results_amp[i,4]<-
length(unique(sample.l[[match(results_amp[i,1],names(sample.l))]][which(sample.l[[match(results_amp[i,1],names(sample.l))]][,9] %in% amf_amp[[match(results_amp[i,1],names(amf_amp))]][,8] & sample.l[[match(results_amp[i,1],names(sample.l))]][,10] %in% cyto_amp[[match(results_amp[i,1],names(cyto_amp))]][,8]),4]))
  results_amp[i,5]<-100*(as.numeric(results_amp[i,4])/as.numeric(results_amp[i,2]))
}
colnames(results_amp)<-
c("SampleID","N_large_amp_in_amf","N_large_amp_in_cyto","N_large_amp_overlap_in_cyto","Percentage_overlap_in_cytosnp")

results_del<-matrix(rep(0,50),nrow=10)
names<-names(sample.l)
for (i in 1:length(names)){
  results_del[i,1]<-gsub("_OVERLAP.txt", "", names[i])
  results_del[i,2]<-
nrow(amf_del[[match(results_del[i,1],names(amf_del))]][which(amf_del[[match(results_del[i,1],names(amf_del))]][,7]>=10000000),])
  results_del[i,3]<-
nrow(cyto_del[[match(results_del[i,1],names(cyto_del))]][which(cyto_del[[match(results_del[i,1],names(cyto_del))]][,7]>=10000000),])
  results_del[i,4]<-
length(unique(sample.l[[match(results_del[i,1],names(sample.l))]][which(sample.l[[match(results_del[i,1],names(sample.l))]][,9] %in% amf_del[[match(results_del[i,1],names(amf_del))]][,8] & sample.l[[match(results_del[i,1],names(sample.l))]][,10] %in% cyto_del[[match(results_del[i,1],names(cyto_del))]][,8]),4]))
  results_del[i,5]<-100*(as.numeric(results_del[i,4])/as.numeric(results_del[i,2]))
}
colnames(results_del)<-
c("SampleID","N_large_del_in_amf","N_large_del_in_cyto","N_large_del_overlap_in_cyto","Percentage_overlap_in_cytosnp")


```
~~~~

##Comparing Exome and CytoSNP on matched samples: Small Alterations (<=1Mb):

###In R:
working directory: “~/Experiments/Other Experiments/CNV/20140815_CS_CNV”

~~~~
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV")

#select small alterations (<=1Mb)
cyto_small<-cyto[which(cyto[,7]<=1000000),]
~~~~

```
amf_small<-amf3[which(amf3[,7]<=1000000),]

#prepare files for galaxy:cyto
galcytosmall<-
data.frame(paste("chr",cyto_small[,2],sep=""),cyto_small[,3],cyto_small[,4],cyto_small[,1])
names(galcytosmall)<-c("chr","start","stop","SegmentID")
galcytosmall_l<-list()
count=1
for (i in unique(galcytosmall$SegmentID)){
  galcytosmall_l[[count]]<-subset(galcytosmall,galcytosmall$SegmentID==i)
  galcytosmall_l[[count]]$SegmentID<-paste(i,"_",seq(1,nrow(galcytosmall_l[[count]]),1),sep="")
  count=count+1
}
names(galcytosmall_l)<-unique(galcytosmall$SegmentID)

lapply(names(galcytosmall_l),function (x) write.table(format(galcytosmall_l[[x]],scientific=FALSE),
file=paste("CNV_cytoVSamf/small/",x,"_coords_withID_cyto_forgalaxay_SMALL_ALT.txt",sep=""),se
p="\t",col.names=F,row.names=F,quote=F )   )

#prepare files for galaxy:amf
galamfsmall<-
data.frame(paste("chr",amf_small[,2],sep=""),amf_small[,3],amf_small[,4],amf_small[,1])
names(galamfsmall)<-c("chr","start","stop","SegmentID")
galamfsmall_l<-list()
count=1
for (i in unique(galamfsmall$SegmentID)){
  galamfsmall_l[[count]]<-subset(galamfsmall,galamfsmall$SegmentID==i)
  galamfsmall_l[[count]]$SegmentID<-paste(i,"_",seq(1,nrow(galamfsmall_l[[count]]),1),sep="")
  count=count+1
}
names(galamfsmall_l)<-unique(galamfsmall$SegmentID)

lapply(names(galamfsmall_l),function (x) write.table(format(galamfsmall_l[[x]],scientific=FALSE),
file=paste("CNV_cytoVSamf/small/",x,"_coords_withID_amf_forgalaxay_SMALL_ALT.txt",sep=""),sep
="\t",col.names=F,row.names=F,quote=F )   )

```

###In Galaxy:
The files coords_withID_cyto_forgalaxay_SMALL_ALT.txt and coords_withID_amf_forgalaxay_SMALL_ALT.txt for the matched sample pairs were uploaded to Galaxy(http://bifx-core.bio.ed.ac.uk:8080/galaxy/)as interval format to assess overlaps using the tool in: 'Operate on Genomic Intervals–>Join', with the following settings:
- Join: coords_withID_cyto_forgalaxay_SMALL_ALT.txt
- with: coords_withID_amf_forgalaxay_SMALL_ALT.txt
- with min overlap: 1bp
- Return: Only records that are joined (INNER JOIN)

The output file was downloaded to "R:/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVSamf/small/xxxx_SMALL_ALT_OVERLAP.txt"

###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVSamf/small"

```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVSamf/small")

temp=list.files(pattern="*_OVERLAP")
sample_small.l<-sapply(temp,function(x)
read.delim(x,header=FALSE,sep="\t",row.names=NULL),simplify=FALSE)
```

```
names(sample_small.l)<-gsub("_SMALL_ALT_OVERLAP.txt", "", names(sample_small.l))

results_small<-matrix(rep(0,50),nrow=10)
names<-names(sample_small.l)
for (i in 1:length(names)){
  results_small[i,1]<-names[i]
  results_small[i,2]<-nrow(galamfsmall[which(galamfsmall$SegmentID==results_small[i,1]),])
  results_small[i,3]<-nrow(galcytosmall[which(galcytosmall$SegmentID==results_small[i,1]),])
  results_small[i,4]<-nrow(unique(sample_small.l[[i]][4]))-1
  results_small[i,5]<-100*(as.numeric(results_small[i,4])/as.numeric(results_small[i,2]))
}
colnames(results_small)<-
c("SampleID","N_small_alt_in_amf","N_small_alt_in_cyto","N_small_alt_overlap_in_cyto","Percentage_
amf_that_overlap_in_cytosnp")



#split amp and del

create_ALTID<-function(df){
  df$ALTID_amf<-paste(df[,5],"_",df[,6],"_",df[,7],sep="")
  df$ALTID_cyto<-paste(df[,1],"_",df[,2],"_",df[,3],sep="")
  return(df)
}

sample_small.l<-lapply(sample_small.l,create_ALTID)

results_small_amp<-matrix(rep(0,50),nrow=10)
names<-names(sample_small.l)

for (i in 1:length(names)){
  results_small_amp[i,1]<-names[i]
  results_small_amp[i,2]<-
nrow(amf_amp[[match(results_small_amp[i,1],names(amf_amp))]][which(amf_amp[[match(results
_small_amp[i,1],names(amf_amp))]][,7]<=1000000),])
  results_small_amp[i,3]<-
nrow(cyto_amp[[match(results_small_amp[i,1],names(cyto_amp))]][which(cyto_amp[[match(resul
ts_small_amp[i,1],names(cyto_amp))]][,7]<=1000000),])
  results_small_amp[i,4]<-
length(unique(sample_small.l[[match(results_small_amp[i,1],names(sample_small.l))]][which(sam
ple_small.l[[match(results_small_amp[i,1],names(sample_small.l))]][,9] %in%
amf_amp[[match(results_small_amp[i,1],names(amf_amp))]][,8] &
sample_small.l[[match(results_small_amp[i,1],names(sample_small.l))]][,10] %in%
cyto_amp[[match(results_small_amp[i,1],names(cyto_amp))]][,8]),4]))
  results_small_amp[i,5]<-
100*(as.numeric(results_small_amp[i,4])/as.numeric(results_small_amp[i,2]))
}
colnames(results_small_amp)<-
c("SampleID","N_small_amp_in_amf","N_small_amp_in_cyto","N_small_amp_overlap_in_cyto","Perce
ntage_overlap_in_cytosnp")

results_small_del<-matrix(rep(0,50),nrow=10)
names<-names(sample_small.l)

for (i in 1:length(names)){
  results_small_del[i,1]<-names[i]
  results_small_del[i,2]<-
nrow(amf_del[[match(results_small_del[i,1],names(amf_del))]][which(amf_del[[match(results_sma
ll_del[i,1],names(amf_del))]][,7]<=1000000),])
```

~~~
  results_small_del[i,3]<-nrow(cyto_del[[match(results_small_del[i,1],names(cyto_del))]][which(cyto_del[[match(results_small_del[i,1],names(cyto_del))]][,7]<=1000000),])
  results_small_del[i,4]<-length(unique(sample_small.l[[match(results_small_del[i,1],names(sample_small.l))]][which(sample_small.l[[match(results_small_del[i,1],names(sample_small.l))]][,9] %in% amf_del[[match(results_small_del[i,1],names(amf_del))]][,8] & sample_small.l[[match(results_small_del[i,1],names(sample_small.l))]][,10] %in% cyto_del[[match(results_small_del[i,1],names(cyto_del))]][,8]),4]))
  results_small_del[i,5]<-100*(as.numeric(results_small_del[i,4])/as.numeric(results_small_del[i,2]))
}
colnames(results_small_del)<-c("SampleID","N_small_del_in_amf","N_small_del_in_cyto","N_small_del_overlap_in_cyto","Percentage_overlap_in_cytosnp")

```




##Writing the comparisons to file
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVSamf"

###In R:
working directory: "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVSamf/small"
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_cytoVSamf")

write.table(results_amp,"Large_amp_10mb_comp.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)
write.table(results_del,"Large_del_10mb_comp.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)
write.table(results_small_amp,"Small_amp_1mb_comp.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)
write.table(results_small_del,"Small_del_1mb_comp.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=F)

```
#Novel alterations in central CS
Extracted list of published alterations from output of pubmed search for chondrosarcoma and copy number;
##Compile list of known CNVs
###From Tarpey, 2013
####In R:
working directory:"~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples_mixed"

```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples_mixed")
#load in list of alterations from nat gen paper (49 cs):
amf<-read.delim("sup_table7_Tarpey2013.txt",sep="\t",header=T)
amf_del<-amf[which(amf$Type=="Del"),2:5]
amf_amp<-amf[which(amf$Type=="Amp"),2:5]
amf_del$Chromosome<-paste("chr",amf_del$Chromosome,sep="")
amf_amp$Chromosome<-paste("chr",amf_amp$Chromosome,sep="")

write.table(amf_del,"amf_del_forgalaxy.txt",sep="\t",col.names=F,row.names=F,quote=F)
write.table(amf_amp,"amf_amp_forgalaxy.txt",sep="\t",col.names=F,row.names=F,quote=F)
```
~~~

~~~
####In Linux:
Can't actually do this in galaxy(wanted to merge overlapping regions and get a count of samples contributing to each overlap), need to do this is linux with bedtools;
Copy amf_del and amf_amp to server and do the following:
```{r,eval=F}
cd paulg/CS_CNV/
sortBed -i amf_amp_forgalaxy.txt | bedtools merge -i stdin -nms > amf_amp_MERGED.bed
sortBed -i amf_del_forgalaxy.txt | bedtools merge -i stdin -nms > amf_del_MERGED.bed
```
####Back on local PC in R
working directory:"~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples_mixed"

```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples_mixed")
amf_amp<-read.delim("amf_amp_MERGED.bed",sep="\t",header=F)
amf_del<-read.delim("amf_del_MERGED.bed",sep="\t",header=F)
names(amf_amp)<-c("Chr","Start","End","Samples")
names(amf_del)<-c("Chr","Start","End","Samples")
#files saved to "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis" as Tarpey2013_Gains_GRCh37_hg19_merged and Tarpey2013_Losses_GRCh37_hg19_merged
```
###Combine with Hallor, 2009
Hallor files saved in "~/Experiments/Other Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis" as Hallor2009_Gains_hg19_merged and Hallor2009_Losses_hg19_merged
####In R:
working directory:"~/Experiments/Other Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis")
tarpey_amp<-read.delim("Tarpey2013_Gains_GRCh37_hg19_merged.txt",sep="\t",header=F)
tarpey_del<-read.delim("Tarpey2013_Losses_GRCh37_hg19_merged.txt",sep="\t",header=F)
hallor_amp<-read.delim("Hallor2009_Gains_hg19_merged.txt",sep="\t",header=F)
hallor_del<-read.delim("Hallor2009_Losses_hg19_merged.txt",sep="\t",header=F)
#combine the two sets and write to file:
known_amp<-rbind(tarpey_amp,hallor_amp)
known_del<-rbind(tarpey_del,hallor_del)
write.table(known_amp,"known_Gains_hg19.txt",col.names=F,row.names=F,sep="\t",quote=F)
write.table(known_del,"known_Losses_hg19.txt",col.names=F,row.names=F,sep="\t",quote=F)
```
##Load 450K CNVs as output by GISTIC
###Run ChAMP CNA pipeline on all 450K samples together
####In R:
working directory:"~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples_mixed"

```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples_mixed")
library(ChAMP)
myLoad=champ.load(methValue="B",QCimages=TRUE,filterXY=FALSE,filterDetP=TRUE,filterBeads=TRUE,beadCutoff=0.05,detPcut=0.01)
source("R:/R/CNV/champ_cna_modif_240414.R")
#output in resultsChamp/CNA
```
####In GISTIC:
same settings as for the all samples at the very top of this script but put the allWT and allmut files together into one and change output prefix to all_samples_mixed
####In R:
~~~

~~~~
Load in all lesions conf 0.95 from GISTIC; the original lesions file (without the '_MODIF' suffix) is
slightly modified in excel
working directory:"~/Experiments/Other
Experiments/CNV/20140815_CS_CNV/CNV_allsamples_mixed"

```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/CNV_allsamples_mixed")
gis<-
read.delim("GISTIC_output/all_samples_mixed/all_samples_mixed.all_lesions.conf_95_MODIF.txt",s
ep="\t",header=T)
#51 samples total
#across each row, calc sum of samples with '1' or '2' (==amplitude >0.3), put sum in sep column,
then calculate percentage of samples
gis<-gis[1:113,]
for (i in 1:nrow(gis)) {
  gis$Total_morethan_1[i]<-(sum(gis[i,10:60]==1)+sum(gis[i,10:60]==2))
  gis$Perc_morethan_1[i]<-100*(gis$Total_morethan_1[i]/51)
  gis$Total_at_2[i]<-sum(gis[i,10:60]==2)
  gis$Perc_at_2[i]<-100*(gis$Total_at_2[i]/51)
}
#split into amp and del
gis_amp<-gis[grep("Amplification",gis[,1]),c(1:9,61:64)]
gis_del<-gis[grep("Deletion",gis[,1]),c(1:9,61:64)]
#write to file in "~/Experiments/Other
Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis"
write.table(gis_amp,"R:/Experiments/Other
Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis/k450_Gains_hg19.txt",col.names
=T,row.names=F,sep="\t",quote=F)
write.table(gis_del,"R:/Experiments/Other
Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis/k450_Losses_hg19.txt",col.name
s=T,row.names=F,sep="\t",quote=F)
```
##Compare 450K and previously known CNVs
###In R:
working directory:"~/Experiments/Other
Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis")
known_amp<-read.delim("known_Gains_hg19.txt",header=F,sep="\t")
known_del<-read.delim("known_Losses_hg19.txt",header=F,sep="\t")
gis_amp<-read.delim("k450_Gains_hg19.txt",header=T,sep="\t")
gis_del<-read.delim("k450_Losses_hg19.txt",header=T,sep="\t")
library(IRanges)
library(GenomicRanges)


known_amp.gr <- GRanges(seqnames=known_amp[,1], ranges=IRanges(start=known_amp[,2],
end= known_amp[,3]))
known_del.gr <- GRanges(seqnames=known_del[,1], ranges=IRanges(start=known_del[,2], end=
known_del[,3]))

gis_amp.gr <- GRanges(seqnames=gis_amp[,3], ranges=IRanges(start=gis_amp[,4], end=
gis_amp[,5]))
gis_del.gr <- GRanges(seqnames=gis_del[,3], ranges=IRanges(start=gis_del[,4], end= gis_del[,5]))
#check overlaps btwn gis and known
amp_countOverlap <- countOverlaps(gis_amp.gr, known_amp.gr)
amp_findOverlap <- as.data.frame(findOverlaps(gis_amp.gr, known_amp.gr)) #22
amp_gis_idx<-as.vector(amp_findOverlap[,1])
amp_known_idx<-as.vector(amp_findOverlap[,2])
temp1<-data.frame()
for (i in amp_gis_idx) temp1<-rbind(temp1,gis_amp[i,])
~~~~

~~~
temp2<-data.frame()
for (i in amp_known_idx) temp2<-rbind(temp2,known_amp[i,])
amp_gisknown_overlap<-cbind(temp1,temp2)
write.table(amp_gisknown_overlap,"amp_gisknown_overlap.txt",sep="\t",col.names=T,row.names=F,quote=F)
del_countOverlap <- countOverlaps(gis_del.gr, known_del.gr)
del_findOverlap <- as.data.frame(findOverlaps(gis_del.gr, known_del.gr)) #5
del_gis_idx<-as.vector(del_findOverlap[,1])
del_known_idx<-as.vector(del_findOverlap[,2])
temp1<-data.frame()
for (i in del_gis_idx) temp1<-rbind(temp1,gis_del[i,])
temp2<-data.frame()
for (i in del_known_idx) temp2<-rbind(temp2,known_del[i,])
del_gisknown_overlap<-cbind(temp1,temp2)
write.table(del_gisknown_overlap,"del_gisknown_overlap.txt",sep="\t",col.names=T,row.names=F,quote=F)
#make tables for those that don't overlap from gis
`%ni%` <- Negate(`%in%`)
amp_gis_NOoverlap<-gis_amp[which(rownames(gis_amp) %ni% amp_gis_idx),]
del_gis_NOoverlap<-gis_del[which(rownames(gis_del) %ni% del_gis_idx),]
write.table(amp_gis_NOoverlap,"amp_gis_NOoverlap.txt",sep="\t",col.names=T,row.names=F,quote=F)
write.table(del_gis_NOoverlap,"del_gis_NOoverlap.txt",sep="\t",col.names=T,row.names=F,quote=F)
#make tables for those that don't overlap from amf
`%ni%` <- Negate(`%in%`)
amp_known_NOoverlap<-known_amp[which(rownames(known_amp) %ni% amp_known_idx),]
del_known_NOoverlap<-known_del[which(rownames(known_del) %ni% del_known_idx),]
write.table(amp_known_NOoverlap,"amp_known_NOoverlap.txt",sep="\t",col.names=T,row.names=F,quote=F)
write.table(del_known_NOoverlap,"del_known_NOoverlap.txt",sep="\t",col.names=T,row.names=F,quote=F)
```
##Plot Karyotype
Make three files to use in this utility to plot a karyotype: "http://db.systemsbiology.net/gestalt/cgi-pub/genomeMapBlocks.pl" ; this will plot one set of regions on the left side of the chrom and one set on the right in chosen colours
Info about the utility: "https://groups.google.com/forum/?hl=en#!topic/isb-famgen/SJd6ee0ujyA and here: https://www.biostars.org/p/16738/"
make one file for all known only and one for all gis only and one for those alterations overlapping in both. the first two need 4 columns: chr, start,stop, colour; overlapping file needs 5: chr start stop ID colour
Plot known-only on the left and gis-only on the right and use the following colours:
- green: gains
- red: losses
The third file with the overlapping regions:
- green common gains
- red common losses
###In R:
working directory:"~/Experiments/Other Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis"
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis")
#file for background colour
bg<-read.delim("karyo_background.txt",sep="\t",header=F)
bg$Start<-1
bg<-bg[,c(1,4,2,3)]
bg3<-bg
names(bg3)<-c("Chr","Start","Stop","Colour")
~~~

~~~~
file1a<-amp_known_NOoverlap[,1:3]
file1a$Colour<-"green"
names(file1a)<-c("Chr","Start","Stop","Colour")
file1b<-del_known_NOoverlap[,1:3]
file1b$Colour<-"red"
names(file1b)<-c("Chr","Start","Stop","Colour")
bg1<-bg
names(bg1)<-c("Chr","Start","Stop","Colour")
file1<-rbind(bg1,file1a,file1b)
write.table(file1,"karyo_file1.txt",sep="\t",col.names=F,row.names=F,quote=F)
file2a<-amp_gis_NOoverlap[,3:5]
file2a$Colour<-"green"
names(file2a)<-c("Chr","Start","Stop","Colour")
file2b<-del_gis_NOoverlap[,3:5]
file2b$Colour<-"red"
names(file2b)<-c("Chr","Start","Stop","Colour")
bg2<-bg
names(bg2)<-c("Chr","Start","Stop","Colour")
file2<-rbind(bg2,file2a,file2b)
write.table(file2,"karyo_file2.txt",sep="\t",col.names=F,row.names=F,quote=F)
file3a<-amp_gisknown_overlap[,3:5]
file3a$Colour<-"green"
names(file3a)<-c("Chr","Start","Stop","Colour")
file3b<-del_gisknown_overlap[,3:5]
file3b$Colour<-"red"
names(file3b)<-c("Chr","Start","Stop","Colour")
file3<-rbind(file3a,file3b)
file3$ID<-"test"
file3<-file3[,c(1,2,3,5,4)]
write.table(file3,"karyo_file3.txt",sep="\t",col.names=F,row.names=F,quote=F)
#when plotting, change in the data for gis, at chr20 the deletion starting at 1 to starting at 2
otherwise bg is blue for some reason
```
##Novel recurrent CNVs
Pick non-overlapping alterations from 450k(gis), that have recurrence >20% (20% of samples with
amplitude>0.3) and are larger than 10Kb
###In R:
working directory:"~/Experiments/Other
Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis"
```{r,eval=F}
setwd("~/Experiments/Other Experiments/CNV/20140815_CS_CNV/pubmed_files_for_analysis")
amp<-read.delim("amp_gis_NOoverlap.txt",sep="\t",header=T)
del<-read.delim("del_gis_NOoverlap.txt",sep="\t",header=T)
amp2<-amp[which(amp$Region_length >= 10000 & amp$Perc_morethan_1>20),]
del2<-del[which(del$Region_length >= 10000 & del$Perc_morethan_1>20),]
amp3<-amp2[,c(3:5)]
del3<-del2[,c(3:5)]
write.table(amp3,"amp_novel.txt",sep="\t",col.names=F,row.names=F,quote=F)
write.table(del3,"del_novel.txt",sep="\t",col.names=F,row.names=F,quote=F)
```
~~~~

FULL ANALYSIS SCRIPT

ANALYSIS OF 450K DATA

```
cd /medical_genomics/paulg/xeno_450k
##########RUN CHAMP ON T1vT2
#Open R
setwd("/medical_genomics/paulg/xeno_450k")
library(ChAMP)
myLoad=champ.load(methValue="B",QCimages=TRUE,filterXY=TRUE,filterDetP=TRUE,filterBeads=
TRUE,beadCutoff=0.05,detPcut=0.01)
myNorm=champ.norm(methValue="B",norm="BMIQ",filterXY=TRUE,QCimages=TRUE)
save(myNorm,file="myNorm.RData")
norm<-myNorm$beta
write.table(norm,"all_norm_betas.txt",sep="\t",row.names=TRUE)
#norm<-read.delim("all_norm_betas.txt",sep="\t",row.names=1,header=T)
champ.SVD()
limma=champ.MVP(bedFile=TRUE)
lasso=champ.lasso(fromFile=TRUE, limma=limma,bedFile=TRUE)
champ.CNA()

norm.anno<-merge(norm,limma,by="row.names")
write.table(norm.anno,"all_norm_betas_annotated.txt",sep="\t",row.names=FALSE)
#norm.anno<-read.delim("all_norm_betas_annotated.txt",sep="\t",header=T)
#Close R

##########DISTRIBUTION OF METHYLATION BY FEATURE
#filter the list of normalised  probes to exclude all those with detectable signal in mouse only.
#ran MB1 kidney on same chip as xeno samples; 45934 probes pass det pvalue threshold of 0.01;
need to remove these from analysis (some already removed as in sex chr or snps); probe list
extracted from genomestudio and saved in paulg/xeno_450k/Analysis_mouse_filtered/

#Open R
setwd("/medical_genomics/paulg/xeno_450k")
#norm.anno<-read.delim("all_norm_betas_annotated.txt",sep="\t",header=T)
#ex.probes<-
read.delim("Analysis_mouse_filtered/mouse_probes_to_exclude.txt",sep="\t",header=T)
#names(ex.probes)<-"probeID"
#norm.anno.ex<-norm.anno[!(norm.anno$probeID %in% ex.probes$probeID),]
#write.table(norm.anno.ex,"Analysis_mouse_filtered/filtered_norm_betas_annotated.txt",sep="\t",r
ow.names=FALSE)
norm.anno.ex<-
read.delim("Analysis_mouse_filtered/filtered_norm_betas_annotated.txt",sep="\t",header=T)
source("/medical_genomics/paulg/scripts_misc/color_bar.R") #has function to plot color bar
legend
Lab.palette.1<-colorRampPalette(c("yellow","green","blue"))
options(digits=22)

feat2.l<-list()
		feat2.l$Whole_Genome<-norm.anno.ex
		feat2.l$Shore<-
norm.anno.ex[which(norm.anno.ex$RELATION_TO_UCSC_CPG_ISLAND=="N_Shore" |
norm.anno.ex$RELATION_TO_UCSC_CPG_ISLAND=="S_Shore"),]
		feat2.l$Shelf<-
norm.anno.ex[which(norm.anno.ex$RELATION_TO_UCSC_CPG_ISLAND=="N_Shelf" |
norm.anno.ex$RELATION_TO_UCSC_CPG_ISLAND=="S_Shelf"),]
		feat2.l$Island<-
norm.anno.ex[which(norm.anno.ex$RELATION_TO_UCSC_CPG_ISLAND=="Island"),]
		feat2.l$TSS1500<-norm.anno.ex[which(norm.anno.ex$feature.1=="TSS1500"),]
```

```
	feat2.l$TSS200<-norm.anno.ex[which(norm.anno.ex$feature.1=="TSS200"),]
	feat2.l$UTR3<-norm.anno.ex[which(norm.anno.ex$feature.1=="3'UTR"),]
	feat2.l$UTR5<-norm.anno.ex[which(norm.anno.ex$feature.1=="5'UTR"),]
	feat2.l$Exon_1st<-norm.anno.ex[which(norm.anno.ex$feature.1=="1stExon"),]
	feat2.l$Body<-norm.anno.ex[which(norm.anno.ex$feature.1=="Body"),]
	feat2.l$IGR<-norm.anno.ex[which(norm.anno.ex$feature.1=="IGR"),]
	feat2.l$Enhancer<-norm.anno.ex[which(norm.anno.ex$ENHANCER==TRUE),]
	feat2.l$miRNA<-norm.anno.ex[grep("^MIR",norm.anno.ex$gene.1),]

#plot T1
results_T1.ex<-vector("list",13)
names(results_T1.ex)<-c("Whole
Genome","Shore","Shelf","Island","TSS1500","TSS200","3'UTR","5'UTR","1st
Exon","Body","IGR","Enhancer","miRNA")
results_T1.ex<-lapply(results_T1.ex,function(x)
matrix(rep(0,600),nrow=100,dimnames=list(seq(1,100,1),colnames(norm.anno.ex[,2:7]))))
for (k in 1:13){
	for (i in 2:7) {results_T1.ex[[k]][1,i-1]<-
100*(nrow(feat2.l[[k]][which(feat2.l[[k]][,i]<=0.01),])/nrow(feat2.l[[k]])) }
	for (i in 2:7) for (j in seq(0.02,1,0.01)) {results_T1.ex[[k]][(j*100),i-1]<-
100*(nrow(feat2.l[[k]][which(feat2.l[[k]][,i]>(j-0.01) & feat2.l[[k]][,i]<=j),])/nrow(feat2.l[[k]]))}
	surplus<-vector("list",6)
	for (i in 2:7) {surplus[[i-1]]<-(100-(sum(results_T1.ex[[k]][1:100,i-1])))/100}
	for (i in 2:7) for (j in 1:100) {results_T1.ex[[k]][j,i-1]<-results_T1.ex[[k]][j,i-1]+surplus[[i-
1]]}
	}
temp_T1<-Reduce(cbind,results_T1.ex)
colnames(temp_T1)<-paste(rep(names(results_T1.ex),each=6),"_",colnames(temp_T1),sep="")
write.table(temp_T1,"Analysis_mouse_filtered/Features_T1.txt",sep="\t",col.names=T,row.names=
F)

png(file="Analysis_mouse_filtered/Features_T1.png",height=4,width=4.5,units="in",res=600)
par(mfrow=c(4,4))
for (k in 1:13){
	par(mar=c(1.5,1.6,1,0.5))
	out<-
barplot(results_T1.ex[[k]][,1:6],beside=FALSE,col=Lab.palette.1(100),border=NA,main=names(res
ults_T1.ex[k]),las=1,xaxt='n')
	mtext(c("T1P","T1X1","T1X2A","T1X2B","T1X2AX1","T1X2BX1"),side=1,at=out,cex=1,las
=2)
	}
dev.off()




#plot T2
results_T2.ex<-vector("list",13)
names(results_T2.ex)<-c("Whole
Genome","Shore","Shelf","Island","TSS1500","TSS200","3'UTR","5'UTR","1st
Exon","Body","IGR","Enhancer","miRNA")
results_T2.ex<-lapply(results_T2.ex,function(x)
matrix(rep(0,300),nrow=100,dimnames=list(seq(1,100,1),colnames(norm.anno.ex[,8:10]))))
for (k in 1:13){
	for (i in 8:10) {results_T2.ex[[k]][1,i-7]<-
100*(nrow(feat2.l[[k]][which(feat2.l[[k]][,i]<=0.01),])/nrow(feat2.l[[k]])) }
	for (i in 8:10) for (j in seq(0.02,1,0.01)) {results_T2.ex[[k]][(j*100),i-7]<-
100*(nrow(feat2.l[[k]][which(feat2.l[[k]][,i]>(j-0.01) & feat2.l[[k]][,i]<=j),])/nrow(feat2.l[[k]]))}
	surplus<-vector("list",3)
	for (i in 8:10) {surplus[[i-7]]<-(100-(sum(results_T2.ex[[k]][1:100,i-7])))/100}
```

```
		for (i in 8:10) for (j in 1:100) {results_T2.ex[[k]][j,i-7]<-results_T2.ex[[k]][j,i-
7]+surplus[[i-7]]}
		}

temp_T2<-Reduce(cbind,results_T2.ex)
colnames(temp_T2)<-paste(rep(names(results_T2.ex),each=3),"_",colnames(temp_T2),sep="")
write.table(temp_T2,"Analysis_mouse_filtered/Features_T2.txt",sep="\t",col.names=T,row.names=
F)



png(file="Analysis_mouse_filtered/Features_T2.png",height=4,width=4.5,units="in",res=600)
par(mfrow=c(4,4))
for (k in 1:13){
		par(mar=c(1.5,1.6,1,0.5))
		out<-
barplot(results_T2.ex[[k]][,1:3],beside=FALSE,col=Lab.palette.1(100),border=NA,main=names(res
ults_T2.ex[k]),las=1,xaxt='n')
		mtext(c("T2P","T2X1","T2X1X1"),side=1,at=out,cex=1,las=2)
		}
dev.off()



#plot colour bar legend for both plots
png(file="Analysis_mouse_filtered/Features_legend.png",height=4,width=0.65,units="in",res=600)
par(mar=c(1.5,1.6,1,0.5))
color.bar(Lab.palette.1(100), min=0, max=1,title='Beta Value')
dev.off()


#Close R

########INTRA-TUMOUR DIFFERENCES: DELTA BETA ON P vs X
```

Sample_20746_7	T1P
Sample_20746_A	T1X1
Sample_20747_LA	T1X2A
Sample_20747_RA	T1X2B
Sample_20747_LB	T1X2AX1
Sample_20747_RB	T1X2BX1
Sample_18727	T2P
Sample_18727_A	T2X1
Sample_18727_B	T2X1X1

```
#Open R
setwd("/medical_genomics/paulg/xeno_450k/Analysis_mouse_filtered")
data<-read.delim("filtered_norm_betas_annotated.txt",sep="\t",header=T) #all normalised beta
values minus those that pass det pval filter 0.01 in mouse kidney sample
source("/medical_genomics/paulg/scripts_misc/color_bar.R") #has function to plot color bar
legend
Lab.palette.2<-colorRampPalette(c("green","yellow","orange","red"))
options(digits=22)


dB.l<-list() #create list of delta beta values
		dB.l$T1PvT1X1<-data.frame(data$probeID,abs(data$Sample_20746_7-
data$Sample_20746_A),(data$Sample_20746_7-data$Sample_20746_A))
		dB.l$T1PvT1X2A<-data.frame(data$probeID,abs(data$Sample_20746_7-
data$Sample_20747_LA),(data$Sample_20746_7-data$Sample_20747_LA))
		dB.l$T1PvT1X2B<-data.frame(data$probeID,abs(data$Sample_20746_7-
data$Sample_20747_RA),(data$Sample_20746_7-data$Sample_20747_RA))
```

```
		dB.l$T1PvT1X2AX1<-data.frame(data$probeID,abs(data$Sample_20746_7-
data$Sample_20747_LB),(data$Sample_20746_7-data$Sample_20747_LB))
		dB.l$T1PvT1X2BX1<-data.frame(data$probeID,abs(data$Sample_20746_7-
data$Sample_20747_RB),(data$Sample_20746_7-data$Sample_20747_RB))
		dB.l$T2PvT2X1<-data.frame(data$probeID,abs(data$Sample_18727-
data$Sample_18727_A),(data$Sample_18727-data$Sample_18727_A))
		dB.l$T2PvT2X1X1<-data.frame(data$probeID,abs(data$Sample_18727-
data$Sample_18727_B),(data$Sample_18727-data$Sample_18727_B))
		dB.l$T1X2AvT1X2AX1<-data.frame(data$probeID,abs(data$Sample_20747_LA-
data$Sample_20747_LB),(data$Sample_20747_LA-data$Sample_20747_LB))
		dB.l$T1X2BvT1X2BX1<-data.frame(data$probeID,abs(data$Sample_20747_RA-
data$Sample_20747_RB),(data$Sample_20747_RA-data$Sample_20747_RB))
		dB.l$T2X1vT2X1X1<-data.frame(data$probeID,abs(data$Sample_18727_A-
data$Sample_18727_B),(data$Sample_18727_A-data$Sample_18727_B))

results_dB<-matrix(rep(0,1000),nrow=100,dimnames=list(seq(1,100,1),names(dB.l)))

for (k in 1:10){
		results_dB[1,k]<-100*(nrow(dB.l[[k]][which(dB.l[[k]][,2]<=0.01),])/nrow(dB.l[[k]]))
		for (j in seq(0.02,1,0.01)) {results_dB[(j*100),k]<-
100*(nrow(dB.l[[k]][which(dB.l[[k]][,2]>(j-0.01) & dB.l[[k]][,2]<=j),])/nrow(dB.l[[k]]))}
		surplus<-(100-(sum(results_dB[1:100,k])))
		results_dB[1,k]<-results_dB[1,k]+surplus
}

write.table(results_dB,"dB_table.txt",sep="\t",col.names=NA,row.names=TRUE,quote=FALSE)
png(file="dB.png",height=4,width=4.5,units="in",res=600)
par(mar=c(3.5,2,1.2,4))
out<-
barplot(results_dB,beside=FALSE,col=Lab.palette.2(100),border=NA,main="DeltaBeta",las=1,xaxt='
n')
mtext(names(dB.l),side=1,at=out,cex=1,las=3)
subplot(color.bar(Lab.palette.2(100), min=0, max=1),x=14,y=50,size=c(0.3,3.2))
dev.off()


#Those probes that do change btwn primary and xeno: what genes? what function in human? in
mouse? random or overlap? more than expected by chance?

#Selecting probes that change: set a threshold of 0.51 difference; rationale: from Lee's work we
know fully unmeth can show up as high as 0.31 and fully meth as low as 0.82 so 0.51 difference
should capture all probes that go from unmeth to meth or vice versa.
options(digits=5)
ch.l<-list()
for (k in 1:10) {ch.l[[k]]<-dB.l[[k]][which(dB.l[[k]][,2]>=0.51),]}
names(ch.l)<-names(dB.l)
sapply(names(ch.l),function(x) write.table(ch.l[[x]],
file=paste(x,"_deltabeta0.51",".txt",sep=""),sep="\t",quote=F,row.names=F))


ch.m<-
matrix(rep(0,20),nrow=2,dimnames=list(c("Num_probes_ch>0.51","Percentage"),names(ch.l)))
for (k in 1:10) {
ch.m[1,k]<-nrow(ch.l[[k]])
ch.m[2,k]<-(ch.m[1,k]/463558)*100
}
write.table(ch.m,"dB_changes_0.51.txt",sep="\t",quote=F,col.names=NA)

#what overlaps?
library(made4)
int12<-intersect(ch.l[[1]][,1],ch.l[[2]][,1])
```

```
int34<-intersect(ch.l[[3]][,1],ch.l[[4]][,1])
int1234<-intersect(int12,int34)
int_T1<-intersect(int1234,ch.l[[5]][,1]) #415
int_T2<-intersect(ch.l[[6]][,1],ch.l[[7]][,1]) #25595

#overlap btwn all?
int_T1T2<-intersect(int_T1,int_T2) #171

#significant overlap?
all<-data[,1] #all probe IDs on the array (minus those excluded at the start)

temp<-
matrix(rep(0,30000),nrow=10000,dimnames=list(c(1:10000),c("T1_rand_overlap","T2_rand_overl
ap","T1T2_rand_overlap")))
for (i in 1:10000){
temp1.1<-sample(all,5272)
temp1.2<-sample(all,2398)
temp1.3<-sample(all,797)
temp1.4<-sample(all,6994)
temp1.5<-sample(all,4153)
temp2.1<-sample(all,28939)
temp2.2<-sample(all,29959)
temp3.1<-sample(all,415)
temp3.2<-sample(all,25595)

int1.12<-intersect(temp1.1,temp1.2)
int1.34<-intersect(temp1.3,temp1.4)
int1.1234<-intersect(int1.12,int1.34)
temp[i,1]<-length(intersect(int1.1234,temp1.5))

temp[i,2]<-length(intersect(temp2.1,temp2.2))
temp[i,3]<-length(intersect(temp3.1,temp3.2))

rm(temp1.1,temp1.2,temp1.3,temp1.4,temp1.5,temp2.1,temp2.2,int1.12,int1.34,int1.1234,temp3.1,
temp3.2)
}
max(temp[,1]) #0
max(temp[,2]) #2229
max(temp[,3]) #47
#yes, overlaps are significant for both tumours at empirical pvalue<=10^-4


#what's the det pvalue of those probes that overlap on the MB1 kidney sample?
mb1.det<-read.delim("../MB1_kidney_alldetpvalues.txt",header=T,sep="\t")
int_T1_2<-data.frame(as.factor(int_T1))
names(int_T1_2)<-"TargetID"
int_T2_2<-data.frame(as.factor(int_T2))
names(int_T2_2)<-"TargetID"

int_T1_det<-merge(int_T1_2,mb1.det)
int_T2_det<-merge(int_T2_2,mb1.det)

no clear link between det pvalue and the overlap; seem to be all over the place.

#export table of hg19 coordinates of those overlapping probes in each tumour set, liftover to mm9,
and run in GREAT.
full.anno<-read.delim("../Full_450K_Annotation.txt",sep="\t",header=T,row.names=1)
int_T1.anno<-full.anno[(rownames(full.anno) %in% int_T1_2$TargetID),c(1,2)]
int_T1.anno$chromosome<-paste("chr",int_T1.anno[,1],sep="")
int_T1.anno$stop<-(int_T1.anno$MAPINFO + 1)
int_T1.anno<-int_T1.anno[,c(3,2,4)]
```

```
int_T2.anno<-full.anno[(rownames(full.anno) %in% int_T2_2$TargetID),c(1,2)]
int_T2.anno$chromosome<-paste("chr",int_T2.anno[,1],sep="")
int_T2.anno$stop<-(int_T2.anno$MAPINFO + 1)
int_T2.anno<-int_T2.anno[,c(3,2,4)]
write.table(int_T1.anno,"T1_overlap_changing_probes_hg19coord.bed",sep="\t",col.names=F,row.names=F,quote=F)
write.table(int_T2.anno,"T2_overlap_changing_probes_hg19coord.bed",sep="\t",col.names=F,row.names=F,quote=F)


#Check dB as well for two technical replicates run on same chip to see what differences to expect if sample identical.
#read in BC1-4 from first 450k experiment; pvalue det filter at 0.01 already applied in genome studio
setwd("/medical_genomics/paulg/xeno_450k/")
bc<-read.delim("BC_replicates_betas_detpval0.01.txt",header=T,sep="\t")
Lab.palette.2<-colorRampPalette(c("green","yellow","orange","red"))
source("/medical_genomics/paulg/scripts_misc/color_bar.R") #has function to plot color bar legend
library(Hmisc)

bc.l<-list() #create list of delta beta values
        bc.l$BC1vBC2<-data.frame(bc$TargetID,abs(bc$BC1.AVG_Beta-bc$BC2.AVG_Beta))
        bc.l$BC1vBC3<-data.frame(bc$TargetID,abs(bc$BC1.AVG_Beta-bc$BC3.AVG_Beta))
        bc.l$BC1vBC4<-data.frame(bc$TargetID,abs(bc$BC1.AVG_Beta-bc$BC4.AVG_Beta))



results_bc<-matrix(rep(0,300),nrow=100,dimnames=list(seq(1,100,1),names(bc.l)))

for (k in 1:3){
        results_bc[1,k]<-100*(nrow(bc.l[[k]][which(bc.l[[k]][,2]<=0.01),])/nrow(bc.l[[k]]))
        for (j in seq(0.02,1,0.01)) {results_bc[(j*100),k]<-100*(nrow(bc.l[[k]][which(bc.l[[k]][,2]>(j-0.01) & bc.l[[k]][,2]<=j),])/nrow(bc.l[[k]]))}
        surplus<-(100-(sum(results_bc[1:100,k])))
        results_bc[1,k]<-results_bc[1,k]+surplus
}


png(file="Analysis_mouse_filtered/dB_BCreplicates.png",height=4,width=4.5,units="in",res=600)
par(mar=c(3.5,2,1.2,4))
out<-barplot(results_bc,beside=FALSE,col=Lab.palette.2(100),border=NA,main="DeltaBeta",las=1,xaxt='n')
mtext(names(bc.l),side=1,at=out,cex=1,las=3)
subplot(color.bar(Lab.palette.2(100), min=0, max=1),x=4.2,y=50,size=c(0.3,3.2))
dev.off()

#evidently much less change;Selecting probes that change: set a threshold of 0.5 difference;
bc_ch.l<-list()
for (k in 1:3) {bc_ch.l[[k]]<-bc.l[[k]][which(bc.l[[k]][,2]>=0.5),]}
names(bc_ch.l)<-names(bc.l)
bc_ch.m<-matrix(rep(0,3),nrow=1,dimnames=list(1,names(bc_ch.l)))
for (k in 1:3) {bc_ch.m[1,k]<-nrow(bc_ch.l[[k]])}
#only 0,2 and 2 probes change >=0.5.



##########INTER TUMOUR DIFFERENCES: HOW WELL DO XENOS SUBSTITUTE FOR PRIMARY?
#Compare the differences found between T1P and T2P vs those found between (eg) T1P and T2X1 (delta delta beta)
```

```
#T1PvT2P
#T1PvT2X1
#T1PvT2X1X1
#T1X1vT2P
#T1X2AvT2P
#T1X2BvT2P
#T1X2AX1vT2P
#T1X2BX1vT2P

setwd("/medical_genomics/paulg/xeno_450k/Analysis_mouse_filtered")
ddB.l<-list() #create list of delta delta beta values
        ddB.l$T1PvT2X1_T1PvT2P<-data.frame(data$probeID,abs(abs(data$Sample_20746_7-
data$Sample_18727_A)-abs(data$Sample_20746_7-data$Sample_18727)))
        ddB.l$T1PvT2X1X1_T1PvT2P<-data.frame(data$probeID,abs(abs(data$Sample_20746_7-
data$Sample_18727_B)-abs(data$Sample_20746_7-data$Sample_18727)))
        ddB.l$T1X1vT2P_T1PvT2P<-data.frame(data$probeID,abs(abs(data$Sample_20746_A-
data$Sample_18727)-abs(data$Sample_20746_7-data$Sample_18727)))
        ddB.l$T1X2AvT2P_T1PvT2P<-
data.frame(data$probeID,abs(abs(data$Sample_20747_LA-data$Sample_18727)-
abs(data$Sample_20746_7-data$Sample_18727)))
        ddB.l$T1X2BvT2P_T1PvT2P<-
data.frame(data$probeID,abs(abs(data$Sample_20747_RA-data$Sample_18727)-
abs(data$Sample_20746_7-data$Sample_18727)))
        ddB.l$T1X2AX1vT2P_T1PvT2P<-
data.frame(data$probeID,abs(abs(data$Sample_20747_LB-data$Sample_18727)-
abs(data$Sample_20746_7-data$Sample_18727)))
        ddB.l$T1X2BX1vT2P_T1PvT2P<-
data.frame(data$probeID,abs(abs(data$Sample_20747_RB-data$Sample_18727)-
abs(data$Sample_20746_7-data$Sample_18727)))

results_ddB<-matrix(rep(0,700),nrow=100,dimnames=list(seq(1,100,1),names(ddB.l)))

for (k in 1:7){
        results_ddB[1,k]<-
100*(nrow(ddB.l[[k]][which(ddB.l[[k]][,2]<=0.01),])/nrow(ddB.l[[k]]))
        for (j in seq(0.02,1,0.01)) {results_ddB[(j*100),k]<-
100*(nrow(ddB.l[[k]][which(ddB.l[[k]][,2]>(j-0.01) & ddB.l[[k]][,2]<=j),])/nrow(ddB.l[[k]]))}
        surplus<-(100-(sum(results_ddB[1:100,k])))
        results_ddB[1,k]<-results_ddB[1,k]+surplus
}

write.table(results_ddB,"ddB_table.txt",sep="\t",col.names=NA,row.names=TRUE,quote=FALSE)
png(file="ddB.png",height=4,width=4.5,units="in",res=600)
par(mar=c(3.5,2,1.2,4))
out<-
barplot(results_ddB,beside=FALSE,col=Lab.palette.2(100),border=NA,main="DeltaDeltaBeta",las=1
,xaxt='n')
mtext(names(ddB.l),side=1,at=out,cex=1,las=3)
subplot(color.bar(Lab.palette.2(100), min=0, max=1),x=9.8,y=50,size=c(0.3,3.2))
dev.off()


##########EXCLUDED MOUSE PROBES
#Open R
setwd("/medical_genomics/paulg/xeno_450k")

#full.anno<-read.delim("Full_450K_Annotation.txt",sep="\t",header=T,row.names=1)
#ex.probes<-
read.delim("Analysis_mouse_filtered/mouse_probes_to_exclude.txt",sep="\t",header=T)
#ex.probes.anno<-full.anno[(rownames(full.anno) %in% ex.probes$TargetID),]
```

```
#write.table(ex.probes.anno,"Analysis_mouse_filtered/mouse_probes_to_exclude_annotated.txt",se
p="\t",col.names=NA)
ex.probes.anno<-
read.delim("Analysis_mouse_filtered/mouse_probes_to_exclude_annotated.txt",sep="\t",header=T,r
ow.names=1)
#The coordinates of the probes to exlude were converted to mm9 using liftover and fed into GREAT
for ontology analysis (using the full 450K coords also converted to mm9 as background); results in
R:/Experiments/Xenograft/450K_xeno/Excluded mouse probes/
```

ANALYSIS OF MEDIP-SEQ DATA

```
####################Separating Human and Mouse Reads: Manual method

##########Run Medusa pt1+2
#medusa is run twice, once aligning to human (hg19), once to mouse (mm10); version used is v2.1
#fastq files are located in /medical_genomics/medip_seq/paul_180613/sequence_data

#config file for running medusa on human is
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt12_031213.cfg
#alignments (incl. bed and sorted BAM files) to human are located in
/medical_genomics/paulg/xeno_medip/human_align
#run medusa on human using this command in /medical_genomics/paulg/medusa/v2_1 :
nohup perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt12_031213.cfg -t
Sample_20746_7,Sample_20746_A,Sample_20747_LA,Sample_20747_RA,Sample_20747_LB,Sample_
20747_RB,Sample_18727,Sample_18727_A,Sample_18727_B -c 0 &

#config file for running medusa on mouse is
/medical_genomics/paulg/xeno_medip/configs/medusa_mouse_pt12_031213.cfg
#alignments (incl. bed and sorted BAM files) to mouse are located in
/medical_genomics/paulg/xeno_medip/mouse_align
#run medusa on mouse using this command in /medical_genomics/paulg/medusa/v2_1 :
nohup perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_mouse_pt12_031213.cfg -t
Sample_20746_7,Sample_20746_A,Sample_20747_LA,Sample_20747_RA,Sample_20747_LB,Sample_
20747_RB,Sample_18727,Sample_18727_A,Sample_18727_B -c 0 &

##########Classify reads
#Using the bed files generated by medusa (stored in human_align and mouse_align) categorise read
pairs as human, mouse, both.
cd /medical_genomics/paulg/xeno_medip/human_vs_mouse
#Open R

source("/medical_genomics/paulg/scripts_misc/set_comps_function.R")
set_comps(Sample_list=c("Sample_18727", "Sample_18727_A", "Sample_18727_B",
"Sample_20746_7", "Sample_20746_A", "Sample_20747_LA", "Sample_20747_LB",
"Sample_20747_RA", "Sample_20747_RB"),
hfolder="/medical_genomics/paulg/xeno_medip/human_align/",
mfolder="/medical_genomics/paulg/xeno_medip/mouse_align/",
outfolder="/medical_genomics/paulg/xeno_medip/human_vs_mouse/")

#Close R

#remove quotes from the sample bed files:
sed -i 's/\"//g' *.bed

##########Side Analysis to look at reads from primary human tumour that map only to mouse
#Checked that Matt's HN samples also have read pairs that map only to mouse, in similar
proportions: HN105(54349 mouse only/224172 mouse read pairs), HN96(45100 mouse
only/181766 mouse read pairs)
```

```
#Load /medical_genomics/paulg/xeno_medip/human_vs_mouse/Sample_18727_mouse.bed and Sample_20746_7_mouse.bed into Ensembl to plot distribution across genome: covers whole genome

#Import the repeatmasker file (to /medical_genomics/paulg/xeno_medip/human_vs_mouse/mouseonly_fromhumanonly/) for mouse mm10 from ucsc to compare with list of overlaps; Only keep the chr, start and end and repeat type columns (cut) and remove header line (sed);then do overlaps

cut -f 6,7,8,12 mouse_UCSC_repeatmasker_021213 > mouse_UCSC_repeatmasker_021213_coords.bed
sed -i 1d mouse_UCSC_repeatmasker_021213_coords.bed #1d stands for 1st line, delete

intersectBed -a ../Sample_18727_mouse.bed -b mouse_UCSC_repeatmasker_021213_coords.bed -wa -u -f 0.25|wc -l
#32314 (74.2%)
intersectBed -a ../Sample_20746_7_mouse.bed -b mouse_UCSC_repeatmasker_021213_coords.bed -wa -u -f 0.25|wc -l
#8752 (78.2%)

#Write these to file (with the entry from repeats file for repeat type; will get more entries because if A overlaps with multiple entries in B, will get all of them)
intersectBed -a ../Sample_18727_mouse.bed -b mouse_UCSC_repeatmasker_021213_coords.bed -wa -wb -f 0.25 > Sample_18727_mouse_REPEATS.bed
intersectBed -a ../Sample_20746_7_mouse.bed -b mouse_UCSC_repeatmasker_021213_coords.bed -wa -wb -f 0.25 > Sample_20746_7_mouse_REPEATS.bed

#Get distribution of repeat types
cat Sample_18727_mouse_REPEATS.bed | awk '{print $10}' | sort | uniq -c > Sample_18727_mouse_REPEATS_distrib.text
cat Sample_20746_7_mouse_REPEATS.bed | awk '{print $10}' | sort | uniq -c > Sample_20746_7_mouse_REPEATS_distrib.text

#Conclusion: When trying to identify cell type from mouse reads in xeno samples, use mouse repeat masker on those reads first to make sure reads used are from mouse cells in the human tumour and not from human reads that map to mouse anyway.



####################Separating Human and Mouse Reads: Xenome method

#run from within /medical_genomics/paulg/xeno_medip/xenome/

##########Index human(graft) and mouse(host) genomes

./xenome-1.0.1-r/xenome index --kmer-size 25 --max-memory 25 --num-threads 8 --verbose --prefix ../index/ --graft /local_data/genomic_data/human/GRCh37/ucsc/human_hg19_full.fa --host /local_data/genomic_data/mouse/GRCm38/ucsc_mm10/mm10_main.fa

##########Classify sample reads

./xenome-1.0.1-r/xenome classify --max-memory 25 --num-threads 8 --verbose --prefix ./index/ --pairs --fastq-in /medical_genomics/medip_seq/paul_180613/sequence_data/Sample_18727_R1.fastq --fastq-in /medical_genomics/medip_seq/paul_180613/sequence_data/Sample_18727_R2.fastq --graft-name human --host-name mouse --output-filename-prefix Sample_18727 >Sample_18727_stats.txt

./xenome-1.0.1-r/xenome classify --max-memory 25 --num-threads 8 --verbose --prefix ./index/ --pairs --fastq-in /medical_genomics/medip_seq/paul_180613/sequence_data/Sample_18727_A_R1.fastq --fastq-in
```

```
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_18727_A_R2.fastq --graft-
name human --host-name mouse --output-filename-prefix Sample_18727_A >
Sample_18727_A_stats.txt

./xenome-1.0.1-r/xenome classify --max-memory 25 --num-threads 8 --verbose --prefix ./index/ --
pairs --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_18727_B_R1.fastq --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_18727_B_R2.fastq --graft-
name human --host-name mouse --output-filename-prefix Sample_18727_B >
Sample_18727_B_stats.txt

./xenome-1.0.1-r/xenome classify --max-memory 25 --num-threads 8 --verbose --prefix ./index/ --
pairs --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20746_7_R1.fastq --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20746_7_R2.fastq --graft-
name human --host-name mouse --output-filename-prefix Sample_20746_7 >
Sample_20746_7_stats.txt

./xenome-1.0.1-r/xenome classify --max-memory 25 --num-threads 8 --verbose --prefix ./index/ --
pairs --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20746_A_R1.fastq --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20746_A_R2.fastq --graft-
name human --host-name mouse --output-filename-prefix Sample_20746_A >
Sample_20746_A_stats.txt

./xenome-1.0.1-r/xenome classify --max-memory 25 --num-threads 8 --verbose --prefix ./index/ --
pairs --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20747_LA_R1.fastq --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20747_LA_R2.fastq --graft-
name human --host-name mouse --output-filename-prefix Sample_20747_LA >
Sample_20747_LA_stats.txt

./xenome-1.0.1-r/xenome classify --max-memory 25 --num-threads 8 --verbose --prefix ./index/ --
pairs --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20747_RA_R1.fastq --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20747_RA_R2.fastq --graft-
name human --host-name mouse --output-filename-prefix Sample_20747_RA >
Sample_20747_RA_stats.txt

./xenome-1.0.1-r/xenome classify --max-memory 25 --num-threads 8 --verbose --prefix ./index/ --
pairs --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20747_LB_R1.fastq --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20747_LB_R2.fastq --graft-
name human --host-name mouse --output-filename-prefix Sample_20747_LB >
Sample_20747_LB_stats.txt

./xenome-1.0.1-r/xenome classify --max-memory 25 --num-threads 8 --verbose --prefix ./index/ --
pairs --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20747_RB_R1.fastq --fastq-in
/medical_genomics/medip_seq/paul_180613/sequence_data/Sample_20747_RB_R2.fastq --graft-
name human --host-name mouse --output-filename-prefix Sample_20747_RB >
Sample_20747_RB_stats.txt

##########Run Medusa pt1+2
#The output of xenome is fastq files for human, mouse, both, neither, ambiguous. Need to merge
fastq files for each sample. Xenome recommends using human+both+ambiguous(hba) together and
mouse+both+ambiguous(mba) together. Run hba and mba but also run mouse alone on xeno
samples to get mouse only signature.

#merge fastq files and then move them to merged_fastqs
cd /medical_genomics/paulg/xeno_medip/xenome/
```

```
cat Sample_18727_human_1.fastq Sample_18727_both_1.fastq Sample_18727_ambiguous_1.fastq >
Sample_18727_hba_R1.fastq
cat Sample_18727_human_2.fastq Sample_18727_both_2.fastq Sample_18727_ambiguous_2.fastq >
Sample_18727_hba_R2.fastq
cat Sample_18727_A_human_1.fastq Sample_18727_A_both_1.fastq
Sample_18727_A_ambiguous_1.fastq > Sample_18727_A_hba_R1.fastq
cat Sample_18727_A_human_2.fastq Sample_18727_A_both_2.fastq
Sample_18727_A_ambiguous_2.fastq > Sample_18727_A_hba_R2.fastq
cat Sample_18727_B_human_1.fastq Sample_18727_B_both_1.fastq
Sample_18727_B_ambiguous_1.fastq > Sample_18727_B_hba_R1.fastq
cat Sample_18727_B_human_2.fastq Sample_18727_B_both_2.fastq
Sample_18727_B_ambiguous_2.fastq > Sample_18727_B_hba_R2.fastq
cat Sample_20746_7_human_1.fastq Sample_20746_7_both_1.fastq
Sample_20746_7_ambiguous_1.fastq > Sample_20746_7_hba_R1.fastq
cat Sample_20746_7_human_2.fastq Sample_20746_7_both_2.fastq
Sample_20746_7_ambiguous_2.fastq > Sample_20746_7_hba_R2.fastq
cat Sample_20746_A_human_1.fastq Sample_20746_A_both_1.fastq
Sample_20746_A_ambiguous_1.fastq > Sample_20746_A_hba_R1.fastq
cat Sample_20746_A_human_2.fastq Sample_20746_A_both_2.fastq
Sample_20746_A_ambiguous_2.fastq > Sample_20746_A_hba_R2.fastq
cat Sample_20747_LA_human_1.fastq Sample_20747_LA_both_1.fastq
Sample_20747_LA_ambiguous_1.fastq > Sample_20747_LA_hba_R1.fastq
cat Sample_20747_LA_human_2.fastq Sample_20747_LA_both_2.fastq
Sample_20747_LA_ambiguous_2.fastq > Sample_20747_LA_hba_R2.fastq
cat Sample_20747_LB_human_1.fastq Sample_20747_LB_both_1.fastq
Sample_20747_LB_ambiguous_1.fastq > Sample_20747_LB_hba_R1.fastq
cat Sample_20747_LB_human_2.fastq Sample_20747_LB_both_2.fastq
Sample_20747_LB_ambiguous_2.fastq > Sample_20747_LB_hba_R2.fastq
cat Sample_20747_RA_human_1.fastq Sample_20747_RA_both_1.fastq
Sample_20747_RA_ambiguous_1.fastq > Sample_20747_RA_hba_R1.fastq
cat Sample_20747_RA_human_2.fastq Sample_20747_RA_both_2.fastq
Sample_20747_RA_ambiguous_2.fastq > Sample_20747_RA_hba_R2.fastq
cat Sample_20747_RB_human_1.fastq Sample_20747_RB_both_1.fastq
Sample_20747_RB_ambiguous_1.fastq > Sample_20747_RB_hba_R1.fastq
cat Sample_20747_RB_human_2.fastq Sample_20747_RB_both_2.fastq
Sample_20747_RB_ambiguous_2.fastq > Sample_20747_RB_hba_R2.fastq

cat Sample_18727_mouse_1.fastq Sample_18727_both_1.fastq Sample_18727_ambiguous_1.fastq >
Sample_18727_mba_R1.fastq
cat Sample_18727_mouse_2.fastq Sample_18727_both_2.fastq Sample_18727_ambiguous_2.fastq >
Sample_18727_mba_R2.fastq
cat Sample_18727_A_mouse_1.fastq Sample_18727_A_both_1.fastq
Sample_18727_A_ambiguous_1.fastq > Sample_18727_A_mba_R1.fastq
cat Sample_18727_A_mouse_2.fastq Sample_18727_A_both_2.fastq
Sample_18727_A_ambiguous_2.fastq > Sample_18727_A_mba_R2.fastq
cat Sample_18727_B_mouse_1.fastq Sample_18727_B_both_1.fastq
Sample_18727_B_ambiguous_1.fastq > Sample_18727_B_mba_R1.fastq
cat Sample_18727_B_mouse_2.fastq Sample_18727_B_both_2.fastq
Sample_18727_B_ambiguous_2.fastq > Sample_18727_B_mba_R2.fastq
cat Sample_20746_7_mouse_1.fastq Sample_20746_7_both_1.fastq
Sample_20746_7_ambiguous_1.fastq > Sample_20746_7_mba_R1.fastq
cat Sample_20746_7_mouse_2.fastq Sample_20746_7_both_2.fastq
Sample_20746_7_ambiguous_2.fastq > Sample_20746_7_mba_R2.fastq
cat Sample_20746_A_mouse_1.fastq Sample_20746_A_both_1.fastq
Sample_20746_A_ambiguous_1.fastq > Sample_20746_A_mba_R1.fastq
cat Sample_20746_A_mouse_2.fastq Sample_20746_A_both_2.fastq
Sample_20746_A_ambiguous_2.fastq > Sample_20746_A_mba_R2.fastq
cat Sample_20747_LA_mouse_1.fastq Sample_20747_LA_both_1.fastq
Sample_20747_LA_ambiguous_1.fastq > Sample_20747_LA_mba_R1.fastq
```

```
cat Sample_20747_LA_mouse_2.fastq Sample_20747_LA_both_2.fastq
Sample_20747_LA_ambiguous_2.fastq > Sample_20747_LA_mba_R2.fastq
cat Sample_20747_LB_mouse_1.fastq Sample_20747_LB_both_1.fastq
Sample_20747_LB_ambiguous_1.fastq > Sample_20747_LB_mba_R1.fastq
cat Sample_20747_LB_mouse_2.fastq Sample_20747_LB_both_2.fastq
Sample_20747_LB_ambiguous_2.fastq > Sample_20747_LB_mba_R2.fastq
cat Sample_20747_RA_mouse_1.fastq Sample_20747_RA_both_1.fastq
Sample_20747_RA_ambiguous_1.fastq > Sample_20747_RA_mba_R1.fastq
cat Sample_20747_RA_mouse_2.fastq Sample_20747_RA_both_2.fastq
Sample_20747_RA_ambiguous_2.fastq > Sample_20747_RA_mba_R2.fastq
cat Sample_20747_RB_mouse_1.fastq Sample_20747_RB_both_1.fastq
Sample_20747_RB_ambiguous_1.fastq > Sample_20747_RB_mba_R1.fastq
cat Sample_20747_RB_mouse_2.fastq Sample_20747_RB_both_2.fastq
Sample_20747_RB_ambiguous_2.fastq > Sample_20747_RB_mba_R2.fastq

cp Sample_18727_mouse_1.fastq Sample_18727_mouse_R1.fastq
cp Sample_18727_mouse_2.fastq Sample_18727_mouse_R2.fastq
cp Sample_18727_A_mouse_1.fastq Sample_18727_A_mouse_R1.fastq
cp Sample_18727_A_mouse_2.fastq Sample_18727_A_mouse_R2.fastq
cp Sample_18727_B_mouse_1.fastq Sample_18727_B_mouse_R1.fastq
cp Sample_18727_B_mouse_2.fastq Sample_18727_B_mouse_R2.fastq
cp Sample_20746_7_mouse_1.fastq Sample_20746_7_mouse_R1.fastq
cp Sample_20746_7_mouse_2.fastq Sample_20746_7_mouse_R2.fastq
cp Sample_20746_A_mouse_1.fastq Sample_20746_A_mouse_R1.fastq
cp Sample_20746_A_mouse_2.fastq Sample_20746_A_mouse_R2.fastq
cp Sample_20747_LA_mouse_1.fastq Sample_20747_LA_mouse_R1.fastq
cp Sample_20747_LA_mouse_2.fastq Sample_20747_LA_mouse_R2.fastq
cp Sample_20747_LB_mouse_1.fastq Sample_20747_LB_mouse_R1.fastq
cp Sample_20747_LB_mouse_2.fastq Sample_20747_LB_mouse_R2.fastq
cp Sample_20747_RA_mouse_1.fastq Sample_20747_RA_mouse_R1.fastq
cp Sample_20747_RA_mouse_2.fastq Sample_20747_RA_mouse_R2.fastq
cp Sample_20747_RB_mouse_1.fastq Sample_20747_RB_mouse_R1.fastq
cp Sample_20747_RB_mouse_2.fastq Sample_20747_RB_mouse_R2.fastq

#it seems xenome removed the @ sign at the start of the readIDs in the fastq and the + on the 3rd
line of each read.
sed -i 's/HSQ/@HSQ/g;s/^$/+/g' *.fastq


#medusa is run 3 times, once aligning hba to human (hg19), once mba to mouse (mm10) and once
mouse to mouse; version used is v2.1
#fastq files are located in /medical_genomics/paulg/xeno_medip/xenome/merged_fastqs/

#config file for running medusa on human is
/medical_genomics/paulg/xeno_medip/configs/medusa_xenome_hba_pt12_061213.cfg
#alignments (incl. bed and sorted BAM files) to human are located in
/medical_genomics/paulg/xeno_medip/xenome/hba_align
#run medusa on human using this command in /medical_genomics/paulg/medusa/v2_1 :
nohup perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_xenome_hba_pt12_061213.cfg -t
Sample_20746_7_hba,Sample_20746_A_hba,Sample_20747_LA_hba,Sample_20747_RA_hba,Sample_
20747_LB_hba,Sample_20747_RB_hba,Sample_18727_hba,Sample_18727_A_hba,Sample_18727_B_
hba -c 0 &

#config file for running medusa on mouse is
/medical_genomics/paulg/xeno_medip/configs/medusa_xenome_mba_pt12_061213.cfg
#alignments (incl. bed and sorted BAM files) to mouse are located in
/medical_genomics/paulg/xeno_medip/xenome/mba_align
#run medusa on mouse using this command in /medical_genomics/paulg/medusa/v2_1 :
nohup perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_xenome_mba_pt12_061213.cfg -t
```

```
Sample_20746_7_mba,Sample_20746_A_mba,Sample_20747_LA_mba,Sample_20747_RA_mba,Samp
le_20747_LB_mba,Sample_20747_RB_mba,Sample_18727_mba,Sample_18727_A_mba,Sample_1872
7_B_mba -c 0 &

#config file for running medusa on mouse is
/medical_genomics/paulg/xeno_medip/configs/medusa_xenome_m_pt12_061213.cfg
#alignments (incl. bed and sorted BAM files) to mouse are located in
/medical_genomics/paulg/xeno_medip/xenome/m_align
#run medusa on mouse using this command in /medical_genomics/paulg/medusa/v2_1 :
nohup perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_xenome_m_pt12_061213.cfg -t
Sample_20746_7_mouse,Sample_20746_A_mouse,Sample_20747_LA_mouse,Sample_20747_RA_mo
use,Sample_20747_LB_mouse,Sample_20747_RB_mouse,Sample_18727_mouse,Sample_18727_A_m
ouse,Sample_18727_B_mouse -c 0 &


#for some reason hba for Sample 18727 has not aligned (made error when pasting h ba and a
togeher and then changing headers (inserted 2 @' signs); rerun medusa on just that sample:
#config file for running medusa is
/medical_genomics/paulg/xeno_medip/configs/medusa_xenome_hba_pt12_18727_200114.cfg
#run medusa on human using this command in /medical_genomics/paulg/medusa/v2_1 :
nohup perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_xenome_hba_pt12_18727_200114.cfg -t
Sample_18727_hba -c 0 &



####################Comparing reads obtained from the two methods

#Compare reads from human+both to hba; mouse+both to mba; mouse to m

#Using the bed files generated by medusa (stored in xenome/hba_align, xenome/mba_align,
xenome/m_align, human_align and mouse_align)
cd /medical_genomics/paulg/xeno_medip/Manual_vs_Xenome

#Open R

source("/medical_genomics/paulg/scripts_misc/set_comps_function_2.R")
set_comps_2(Sample_list=c("Sample_18727", "Sample_18727_A", "Sample_18727_B",
"Sample_20746_7", "Sample_20746_A", "Sample_20747_LA", "Sample_20747_LB",
"Sample_20747_RA", "Sample_20747_RB"))

#Close R

#Conclusion: get almost identical results so might as well use simpler manual method


##########DMRs:INDIVIDUAL COMPARISONS (THESE HAVE BEEN RUN WITH PVALUE 0.05
AND 0.1; latest version of config files has 0.05 (and so the files in the DMR folders that don't have
the pvalue within the name have been overwritten with latest version as well (0.05))
Sample_20746_7      T1P
Sample_20746_A      T1X1
Sample_20747_LA     T1X2A
Sample_20747_RA     T1X2B
Sample_20747_LB     T1X2AX1
Sample_20747_RB     T1X2BX1
Sample_18727        T2P
Sample_18727_A      T2X1
Sample_18727_B      T2X1X1
```

```
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1PvT2P_230114.cfg -t
Sample_20746_7 -c Sample_18727
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1PvT2X1_230114.cfg -t
Sample_20746_7 -c Sample_18727_A
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1PvT2X1X1_230114.cfg -t
Sample_20746_7 -c Sample_18727_B
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1X1vT2P_060214.cfg -t
Sample_20746_A -c Sample_18727
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1X2AvT2P_060214.cfg -t
Sample_20747_LA -c Sample_18727
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1X2BvT2P_060214.cfg -t
Sample_20747_RA -c Sample_18727
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1X2AX1vT2P_060214.cfg -t
Sample_20747_LB -c Sample_18727
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1X2BX1vT2P_060214.cfg -t
Sample_20747_RB -c Sample_18727

perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1X1vT2X1_060214.cfg -t
Sample_20746_A -c Sample_18727_A

perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1PvT1X1_060214.cfg -t
Sample_20746_7 -c Sample_20746_A
#THESE HAVE ONLY BEEN RUN AT PVALUE 0.05:
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1PvT1X2A_200214.cfg -t
Sample_20746_7 -c Sample_20747_LA
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1PvT1X2B_200214.cfg -t
Sample_20746_7 -c Sample_20747_RA
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1PvT1X2AX1_200214.cfg -t
Sample_20746_7 -c Sample_20747_LB
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T1PvT1X2BX1_200214.cfg -t
Sample_20746_7 -c Sample_20747_RB
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T2PvT2X1_200214.cfg -t
Sample_18727 -c Sample_18727_A
perl medusa.pl -p
/medical_genomics/paulg/xeno_medip/configs/medusa_human_pt34_T2PvT2X1X1_200214.cfg -t
Sample_18727 -c Sample_18727_B

#these commands list all line counts for the dmr files
find /medical_genomics/paulg/xeno_medip/human_align/ -name
"medips_dmrs_hyperTreatment_p0.05.bed" -exec wc -l {} \;
find /medical_genomics/paulg/xeno_medip/human_align/ -name
"medips_dmrs_hypoTreatment_p0.05.bed" -exec wc -l {} \;

#results from pvalue 0.1
                  hyper/hypo  # overlap with T1PvT2P   % overlap with T1PvT2P
T1PvT2P           2162/88
```

```
T1PvT2X1          2410/70         1623/21               75.1/23.9
T1PvT2X1X1        2501/77         1594/17               73.7/19.3
T1X1vT2P          4386/230   1246/82               57.6/93.2
T1X2AvT2P         5271/200   1423/82               65.8/93.2
T1X2BvT2P         4388/188   1354/80               62.6/90.1
T1X2AX1vT2P            6459/371   1417/85               65.5/96.6
T1X2BX1vT2P            4425/206   1351/82               62.5/93.2

T1X1vT2X1         4395/132   1153/25               53.3/28.4
T1PvT1X1          460/126

#results from pvalue 0.05
                  hyper/hypo  # overlap with T1PvT2P   % overlap with T1PvT2P
T1PvT2P           1980/85
T1PvT2X1          2129/61         1482/19               74.8/22.4
T1PvT2X1X1        2298/66         1453/15               73.4/17.6
T1X1vT2P          4158/223   1180/79               59.6/92.9
T1X2AvT2P         4999/191   1338/77               67.6/90.6
T1X2BvT2P         4126/177   1277/77               64.5/90.6
T1X2AX1vT2P            6074/359   1324/82               66.9/96.5
T1X2BX1vT2P            4203/200   1276/79               64.4/92.9

T1X1vT2X1         3918/119   1078/22               54.4/25.9

T1PvT1X1          373/74
T1PvT1X2A         368/151
T1PvT1X2B         373/68
T1PvT1X2AX1            600/54
T1PvT1X2BX1            313/67
T2PvT2X1          376/795
T2PvT2X1X1        735/1193

###############overlap of dmrs intra-tumour with 450k mvps
#####first assign probeIDs to all DMRs that have a match on the 450k
###create full_450k coordinates table with probeIDs
cd /medical_genomics/xeno_450k
#Open R
anno<-read.delim("Full_450K_Annotation.txt",sep="\t",header=TRUE)
anno2<-data.frame(anno[,1])
anno2$Chr<-paste("chr",anno$CHR,sep="")
anno2$Start<-anno$MAPINFO
anno2$End<-anno$MAPINFO +1
anno3<-anno2[,c(2,3,4,1)]
anno3<-anno3[-which(anno3$Chr=="chr"),]

write.table(anno3,"medip_v_450/coords_450.bed",col.names=F,row.names=F,quote=F,sep="\t")
#Close R

###-wa -wb -loj: use bedtools with these options to report all the original coordinates of A(dmrs)
whether there's an overlap or not and B(450karray coords) if there is an overlap.

cd /medical_genomics/paulg/xeno_450k/medip_v_450

intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X1/medips_dmrs_hyperTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X1_hyperdmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X2A/medips_dmrs_hyperTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X2A_hyperdmr_450.txt
```

```
intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X2B/medips_dmrs_hyperTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X2B_hyperdmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X2AX1/medips_dmrs_hyperTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X2AX1_hyperdmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X2BX1/medips_dmrs_hyperTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X2BX1_hyperdmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T2PvT2X1/medips_dmrs_hyperTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T2PvT2X1_hyperdmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T2PvT2X1X1/medips_dmrs_hyperTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T2PvT2X1X1_hyperdmr_450.txt

intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X1/medips_dmrs_hypoTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X1_hypodmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X2A/medips_dmrs_hypoTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X2A_hypodmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X2B/medips_dmrs_hypoTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X2B_hypodmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X2AX1/medips_dmrs_hypoTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X2AX1_hypodmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T1PvT1X2BX1/medips_dmrs_hypoTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T1PvT1X2BX1_hypodmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T2PvT2X1/medips_dmrs_hypoTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T2PvT2X1_hypodmr_450.txt
intersectBed -a
../../xeno_medip/human_align/dmr_T2PvT2X1X1/medips_dmrs_hypoTreatment_p0.05.bed -b
coords_450.bed -wa -wb -loj > T2PvT2X1X1_hypodmr_450.txt


#####then see how many of intra tumour mvps(large >0.51) are validated
cd /medical_genomics/paulg/xeno_450k
#Open R

setwd("/medical_genomics/paulg/xeno_450k")
temp1=list.files("Analysis_mouse_filtered/",pattern="*deltabeta0.51.txt")
samp1.l<-sapply(temp1,function(x)
read.delim(paste("Analysis_mouse_filtered/",x,sep=""),header=T,sep="\t",row.names=NULL),simpl
ify=FALSE)
#remove 1stv2nd gen comps
samp1.l<-samp1.l[-c(6,7,10)]
#rename objects in list and colnames and label each probe as hyper or hypo in xeno
for (k in 1:7){
names(samp1.l)[k]<-gsub("_deltabeta0.51.txt","",names(samp1.l)[k])
names(samp1.l[[k]])<-c("probeID","ChinXeno","DeltaBeta")
for (i in 1:nrow(samp1.l[[k]])){
if (samp1.l[[k]][i,3]<0) samp1.l[[k]][i,2]<-"hyper"
else if (samp1.l[[k]][i,3]>0) samp1.l[[k]][i,2]<-"hypo"
}
}
```

```
#Do the same for medip results(these were done with prim as treatment and xeno as control, so
hypo and hyper are reversed)
temp2=list.files("medip_v_450/",pattern="*dmr_450.txt")
samp2.l<-sapply(temp2,function(x)
read.delim(paste("medip_v_450/",x,sep=""),header=F,sep="\t",row.names=NULL),simplify=FALSE)
for (k in 1:14){
samp2.l[[k]]<-samp2.l[[k]][-which(samp2.l[[k]][8]=="."),c(4,8)]
}

for (k in c(1,3,5,7,9,11,13)){
samp2.l[[k]]$ChinXeno<-"hypo"
samp2.l[[k]]<-samp2.l[[k]][,c(2,3,1)]
names(samp2.l[[k]])<-c("probeID","ChinXeno_dmr","dmrID")
}

for (k in c(2,4,6,8,10,12)){
samp2.l[[k]]$ChinXeno<-"hyper"
samp2.l[[k]]<-samp2.l[[k]][,c(2,3,1)]
names(samp2.l[[k]])<-c("probeID","ChinXeno_dmr","dmrID")
}

samp3.l<-samp2.l
for (k in c(1,3,5,7,9,11,13)){
samp3.l[[k]]<-rbind(samp2.l[[k]],samp2.l[[k+1]])
}

samp2.l<-samp3.l[c(1,3,5,7,9,11,13)]
names(samp2.l)<-names(samp1.l)

#create matrix of unique counts of probe IDs for each sample comp and each
direction(hyper/hypo)
comps<-matrix(rep(0,28),nrow=4,
        dimnames=list(c("hyper_450","hypo_450","hyper_medip","hypo_medip"),names(samp1.l
)))
for (k in 1:7) {
comps[1,k]<-nrow(samp1.l[[k]][which(samp1.l[[k]][,2]=="hyper"),])
comps[2,k]<-nrow(samp1.l[[k]][which(samp1.l[[k]][,2]=="hypo"),])
comps[3,k]<-nrow(samp2.l[[k]][which(samp2.l[[k]][,2]=="hyper"),])
comps[4,k]<-nrow(samp2.l[[k]][which(samp2.l[[k]][,2]=="hypo"),])
}

write.table(comps,"medip_v_450/hyper_hypo_numbers.txt",col.names=NA,sep="\t",quote=F)
#merge the two sample data frames by probeID
samp12.l<-list()
for (k in 1:7) {
samp12.l[[k]]<-merge(samp1.l[[k]],samp2.l[[k]],by="probeID")
}
names(samp12.l)<-names(samp1.l)

#write counts and percentages of overlaps
comps2<-matrix(rep(0,40),nrow=5,dimnames=list(c("Total probes","Num Same Change in Medip
and 450","Perc same change","Num hyper","Num hypo"),c(names(samp1.l),"Total")))

for (k in 1:7) {
comps2[1,k]<-nrow(samp12.l[[k]])
if (comps2[1,k] != 0) {
comps2[2,k]<-nrow(samp12.l[[k]][which(samp12.l[[k]][2]==samp12.l[[k]][4]),])
comps2[3,k]<-(comps2[2,k]/comps2[1,k])*100
comps2[4,k]<-nrow(samp12.l[[k]][which(samp12.l[[k]][2]=="hyper"),])
comps2[5,k]<-nrow(samp12.l[[k]][which(samp12.l[[k]][2]=="hypo"),])
}
else {
comps2[2,k]<-0
comps2[3,k]<-0
comps2[4,k]<-0
comps2[5,k]<-0
}
}
comps2[1,8]<-sum(comps2[1,c(1:7)])
comps2[2,8]<-sum(comps2[2,c(1:7)])
comps2[3,8]<-(comps2[2,8]/comps2[1,8])*100
comps2[4,8]<-sum(comps2[4,c(1:7)])
comps2[5,8]<-sum(comps2[5,c(1:7)])


write.table(comps2,"medip_v_450/overlap_medip_450_perc.txt",col.names=NA,sep="\t",quote=F)
```

ONCOTRACK 450K

```
####################ANALYSIS OF 450K DATA
cd /medical_genomics/paulg/xeno_450k/Oncotrack

##########RUN CHAMP ON T1vT2
#Results are stored in /medical_genomics/paulg/xeno_450k/resultsChamp
#Open R

setwd("/medical_genomics/paulg/xeno_450k/Oncotrack")
library(ChAMP)
myLoad=champ.load(methValue="B",QCimages=TRUE,filterXY=TRUE,filterDetP=TRUE,filterBeads=
TRUE,beadCutoff=0.05,detPcut=0.01)
myNorm=champ.norm(methValue="B",norm="BMIQ",filterXY=TRUE,QCimages=TRUE)
save(myNorm,file="myNorm.RData")
norm<-myNorm$beta
write.table(norm,"onco_all_norm_betas.                    txt",sep="\t",row.names=TRUE)
#norm<-read.delim("onco_all_norm_betas.txt",sep="\t",row.names=1,header=T)
champ.SVD()
limma=champ.MVP(bedFile=TRUE)
lasso=champ.lasso(fromFile=TRUE, limma=limma,bedFile=TRUE)
champ.CNA()

norm.anno<-merge(norm,limma,by="row.names")
write.table(norm.anno,"onco_all_norm_betas_annotated.txt",sep="\t",row.names=FALSE)
#norm.anno<-read.delim("onco_all_norm_betas_annotated.txt",sep="\t",header=T)

#Close R

#filter the list of normalised  probes to exclude all those with detectable signal in mouse only.
#ran MB1 kidney on same chip as xeno samples; 45934 probes pass det pvalue threshold of 0.01;
need to remove these from analysis (some already removed as in sex chr or snps); probe list
extracted from genomestudio and saved in paulg/xeno_450k/Oncotrack/Analysis_mouse_filtered/

#Open R
setwd("/medical_genomics/paulg/xeno_450k/Oncotrack")

#norm.anno<-read.delim("onco_all_norm_betas_annotated.txt",sep="\t",header=T)
#ex.probes<-
read.delim("Analysis_mouse_filtered/mouse_probes_to_exclude.txt",sep="\t",header=T)
#names(ex.probes)<-"probeID"
#norm.anno.ex<-norm.anno[!(norm.anno$probeID %in% ex.probes$probeID),]
```

```
#write.table(norm.anno.ex,"Analysis_mouse_filtered/onco_filtered_norm_betas_annotated.txt",sep=
"\t",row.names=FALSE)
norm.anno.ex<-
read.delim("Analysis_mouse_filtered/onco_filtered_norm_betas_annotated.txt",sep="\t",header=T)


########INTRA-TUMOUR DIFFERENCES: DELTA BETA ON P vs X

#Open R
setwd("/medical_genomics/paulg/xeno_450k/Oncotrack/Analysis_mouse_filtered")
data<-read.delim("onco_filtered_norm_betas_annotated.txt",sep="\t",header=T) #all normalised
beta values minus those that pass det pval filter 0.01 in mouse kidney sample
source("/medical_genomics/paulg/scripts_misc/color_bar.R") #has function to plot color bar
legend
Lab.palette.2<-colorRampPalette(c("green","yellow","orange","red"))
options(digits=22)


dB.l<-list() #create list of delta beta values
        dB.l$T108vX108<-data.frame(data$probeID,abs(data$T108-data$X108))
        dB.l$T109vX109<-data.frame(data$probeID,abs(data$T109-data$X109))
        dB.l$T114vX114<-data.frame(data$probeID,abs(data$T114-data$X114))
        dB.l$T116vX116<-data.frame(data$probeID,abs(data$T116-data$X116))
        dB.l$T118vX118<-data.frame(data$probeID,abs(data$T118-data$X118))
        dB.l$T135vX135<-data.frame(data$probeID,abs(data$T135-data$X135))

results_dB<-matrix(rep(0,600),nrow=100,dimnames=list(seq(1,100,1),names(dB.l)))

for (k in 1:6){
        results_dB[1,k]<-100*(nrow(dB.l[[k]][which(dB.l[[k]][,2]<=0.01),])/nrow(dB.l[[k]]))
        for (j in seq(0.02,1,0.01)) {results_dB[(j*100),k]<-
100*(nrow(dB.l[[k]][which(dB.l[[k]][,2]>(j-0.01) & dB.l[[k]][,2]<=j),])/nrow(dB.l[[k]]))}
        surplus<-(100-(sum(results_dB[1:100,k])))
        results_dB[1,k]<-results_dB[1,k]+surplus
}

write.table(results_dB,"onco_dB_table.txt",sep="\t",col.names=NA,row.names=TRUE,quote=FALSE)
png(file="dB.png",height=4,width=4.5,units="in",res=600)
par(mar=c(3.5,2,1.2,4))
out<-
barplot(results_dB,beside=FALSE,col=Lab.palette.2(100),border=NA,main="DeltaBeta",las=1,xaxt='
n')
mtext(names(dB.l),side=1,at=out,cex=1,las=3)
subplot(color.bar(Lab.palette.2(100), min=0, max=1),x=9.8,y=50,size=c(0.3,3.2))
dev.off()


#Those probes that do change btwn primary and xeno

#Selecting probes that change: set a threshold of 0.51 difference; rationale: from Lee's work we
know fully unmeth can show up as high as 0.31 and fully meth as low as 0.82 so 0.51 difference
should capture all probes that go from unmeth to meth or vice versa.
options(digits=5)
ch.l<-list()
for (k in 1:6) {ch.l[[k]]<-dB.l[[k]][which(dB.l[[k]][,2]>=0.51),]}
names(ch.l)<-names(dB.l)
sapply(names(ch.l),function(x) write.table(ch.l[[x]],
file=paste(x,"_deltabeta0.51",".txt",sep=""),sep="\t",quote=F,row.names=F))
```

```
ch.m<-matrix(rep(0,12),nrow=2,dimnames=list(c("Num_probes_ch>0.51","Percentage"),names(ch.l)))
for (k in 1:6) {
ch.m[1,k]<-nrow(ch.l[[k]])
ch.m[2,k]<-(ch.m[1,k]/385724)*100
}
write.table(ch.m,"dB_changes_0.51.txt",sep="\t",quote=F,col.names=NA)

#what overlaps?
library(made4)
int12<-intersect(ch.l[[1]][,1],ch.l[[2]][,1])
int34<-intersect(ch.l[[3]][,1],ch.l[[4]][,1])
int1234<-intersect(int12,int34)
int56<-intersect(ch.l[[5]][,1],ch.l[[6]][,1])

#overlap btwn all?
int_onco<-intersect(int1234,int56) #5

#significant overlap?
all<-data[,1] #all probe IDs on the array (minus those excluded at the start)

temp<-matrix(rep(0,10000),nrow=10000,dimnames=list(c(1:10000),c("onco_rand_overlap")))
for (i in 1:10000){
temp1.1<-sample(all,2848)
temp1.2<-sample(all,34649)
temp1.3<-sample(all,1729)
temp1.4<-sample(all,1851)
temp1.5<-sample(all,23315)
temp1.6<-sample(all,9306)

int1.12<-intersect(temp1.1,temp1.2)
int1.34<-intersect(temp1.3,temp1.4)
int1.1234<-intersect(int1.12,int1.34)
int1.56<-intersect(temp1.5,temp1.6)
int1.123456<-intersect(int1.1234,int1.56)
temp[i,1]<-length(int1.123456)

rm(temp1.1,temp1.2,temp1.3,temp1.4,temp1.5,temp1.6,int1.12,int1.34,int1.1234,int1.56,int1.123456)
}
max(temp[,1]) #0

#yes, overlaps are significant at empirical pvalue<=10^-4


################################################Put Onco and 1st gen OS samples together
setwd("/medical_genomics/paulg/xeno_450k")
onco<-read.delim("Oncotrack/Analysis_mouse_filtered/onco_filtered_norm_betas_annotated.txt",sep="\t",header=T)
os<-read.delim("Analysis_mouse_filtered/filtered_norm_betas_annotated.txt",sep="\t",header=T)

source("/medical_genomics/paulg/scripts_misc/color_bar.R") #has function to plot color bar legend
Lab.palette.2<-colorRampPalette(c("green","yellow","orange","red"))
options(digits=22)


dB.l<-list() #create list of delta beta values
```

```
	dB.l$T1PvT1X1<-
data.frame(os$probeID,os$CHR,os$MAPINFO,os$gene.1,os$feature.1,os$RELATION_TO_UCSC_CPG_
ISLAND,os$feat.rel,abs(os$Sample_20746_7-os$Sample_20746_A),(os$Sample_20746_7-
os$Sample_20746_A))
	dB.l$T1PvT1X2A<-
data.frame(os$probeID,os$CHR,os$MAPINFO,os$gene.1,os$feature.1,os$RELATION_TO_UCSC_CPG_
ISLAND,os$feat.rel,abs(os$Sample_20746_7-os$Sample_20747_LA),(os$Sample_20746_7-
os$Sample_20747_LA))
	dB.l$T1PvT1X2B<-
data.frame(os$probeID,os$CHR,os$MAPINFO,os$gene.1,os$feature.1,os$RELATION_TO_UCSC_CPG_
ISLAND,os$feat.rel,abs(os$Sample_20746_7-os$Sample_20747_RA),(os$Sample_20746_7-
os$Sample_20747_RA))
	dB.l$T2PvT2X1<-
data.frame(os$probeID,os$CHR,os$MAPINFO,os$gene.1,os$feature.1,os$RELATION_TO_UCSC_CPG_
ISLAND,os$feat.rel,abs(os$Sample_18727-os$Sample_18727_A),(os$Sample_18727-
os$Sample_18727_A))
	dB.l$T108vX108<-
data.frame(onco$probeID,onco$CHR,onco$MAPINFO,onco$gene.1,onco$feature.1,onco$RELATION
_TO_UCSC_CPG_ISLAND,onco$feat.rel,abs(onco$T108-onco$X108),(onco$T108-onco$X108))
	dB.l$T109vX109<-
data.frame(onco$probeID,onco$CHR,onco$MAPINFO,onco$gene.1,onco$feature.1,onco$RELATION
_TO_UCSC_CPG_ISLAND,onco$feat.rel,abs(onco$T109-onco$X109),(onco$T109-onco$X109))
	dB.l$T114vX114<-
data.frame(onco$probeID,onco$CHR,onco$MAPINFO,onco$gene.1,onco$feature.1,onco$RELATION
_TO_UCSC_CPG_ISLAND,onco$feat.rel,abs(onco$T114-onco$X114),(onco$T114-onco$X114))
	dB.l$T116vX116<-
data.frame(onco$probeID,onco$CHR,onco$MAPINFO,onco$gene.1,onco$feature.1,onco$RELATION
_TO_UCSC_CPG_ISLAND,onco$feat.rel,abs(onco$T116-onco$X116),(onco$T116-onco$X116))
	dB.l$T118vX118<-
data.frame(onco$probeID,onco$CHR,onco$MAPINFO,onco$gene.1,onco$feature.1,onco$RELATION
_TO_UCSC_CPG_ISLAND,onco$feat.rel,abs(onco$T118-onco$X118),(onco$T118-onco$X118))
	dB.l$T135vX135<-
data.frame(onco$probeID,onco$CHR,onco$MAPINFO,onco$gene.1,onco$feature.1,onco$RELATION
_TO_UCSC_CPG_ISLAND,onco$feat.rel,abs(onco$T135-onco$X135),(onco$T135-onco$X135))

results_dB<-matrix(rep(0,1000),nrow=100,dimnames=list(seq(1,100,1),names(dB.l)))

for (k in 1:10){
	results_dB[1,k]<-100*(nrow(dB.l[[k]][which(dB.l[[k]][,8]<=0.01),])/nrow(dB.l[[k]]))
	for (j in seq(0.02,1,0.01)) {results_dB[(j*100),k]<-
100*(nrow(dB.l[[k]][which(dB.l[[k]][,8]>(j-0.01) & dB.l[[k]][,8]<=j),])/nrow(dB.l[[k]]))}
	surplus<-(100-(sum(results_dB[1:100,k])))
	results_dB[1,k]<-results_dB[1,k]+surplus
}

write.table(results_dB,"OS_ONCO/dB_table.txt",sep="\t",col.names=NA,row.names=TRUE,quote=FA
LSE)
png(file="OS_ONCO/dB.png",height=1.6,width=4.5,units="in",res=600)
par(mar=c(1.5,2,1.2,4))
out<-barplot(results_dB,beside=FALSE,col=Lab.palette.2(100),border=NA,main="",las=1,xaxt='n')
mtext(names(dB.l),side=1,at=out,cex=1,las=1)
subplot(color.bar(Lab.palette.2(100), min=0, max=1,title='Beta
Difference'),x=14,y=50,size=c(0.2,1.0))
dev.off()


#Those probes that do change btwn primary and xeno

#Selecting probes that change: set a threshold of 0.51 difference; rationale: from Lee's work we
know fully unmeth can show up as high as 0.31 and fully meth as low as 0.82 so 0.51 difference
should capture all probes that go from unmeth to meth or vice versa.
```

```
options(digits=5)
ch.l<-list()
for (k in 1:10) {ch.l[[k]]<-dB.l[[k]][which(dB.l[[k]][,8]>=0.51),]}
names(ch.l)<-names(dB.l)
sapply(names(ch.l),function(x) write.table(ch.l[[x]],
file=paste("OS_ONCO/",x,"_deltabeta0.51",".txt",sep=""),sep="\t",quote=F,row.names=F))


ch.m<-
matrix(rep(0,40),nrow=4,dimnames=list(c("Num_probes_ch>0.51","Percentage","Percentage
change Hypermethylation","Percentage change Hypomethylation"),names(ch.l)))
for (k in 5:10) {
ch.m[1,k]<-nrow(ch.l[[k]])
ch.m[2,k]<-(ch.m[1,k]/385724)*100
ch.m[3,k]<-(nrow(ch.l[[k]][which(ch.l[[k]][,9]<0),])/nrow(ch.l[[k]]))*100
ch.m[4,k]<-(nrow(ch.l[[k]][which(ch.l[[k]][,9]>0),])/nrow(ch.l[[k]]))*100
}
for (k in 1:4) {
ch.m[1,k]<-nrow(ch.l[[k]])
ch.m[2,k]<-(ch.m[1,k]/463558)*100
ch.m[3,k]<-(nrow(ch.l[[k]][which(ch.l[[k]][,9]<0),])/nrow(ch.l[[k]]))*100
ch.m[4,k]<-(nrow(ch.l[[k]][which(ch.l[[k]][,9]>0),])/nrow(ch.l[[k]]))*100

}
write.table(ch.m,"OS_ONCO/dB_changes_0.51.txt",sep="\t",quote=F,col.names=NA)

####put together data from all those probes that change
probes<-matrix()
for (k in 1:10){probes<-rbind(probes,as.matrix(ch.l[[k]][1]))}
probes<-unique(probes)
probes<-probes[-1,]
probes<-as.factor(probes)
temp_os<-os[which(os$probeID %in% probes),]
temp_onco<-onco[which(onco$probeID %in% probes),]
all<-merge(temp_os,temp_onco,by="probeID")
all<-all[,c(1,3:6,9:10,42:53,18:19,21,25,30,33,37)]

all$Diff_Sample_20746_A<-all$Sample_20746_A-all$Sample_20746_7
all$Diff_Sample_20747_LA<-all$Sample_20747_LA-all$Sample_20746_7
all$Diff_Sample_20747_RA<-all$Sample_20747_RA-all$Sample_20746_7
all$Diff_Sample_18727_A<-all$Sample_18727_A-all$Sample_18727
all$Diff_X108<-all$X108-all$T108
all$Diff_X109<-all$X109-all$T109
all$Diff_X114<-all$X114-all$T114
all$Diff_X116<-all$X116-all$T116
all$Diff_X118<-all$X118-all$T118
all$Diff_X135<-all$X135-all$T135

all$HyperPerc<-rep(0,nrow(all))
all$HypoPerc<-rep(0,nrow(all))
for (i in 1:nrow(all)){
all[i,37]<-10*(length(all[i,which(all[i,27:36]>0)]))
all[i,38]<-10*(length(all[i,which(all[i,27:36]<0)]))
}

write.table(all,"OS_ONCO/all.txt",sep="\t",quote=F,col.names=T,row.names=F)



#all<-read.delim("all.txt",sep="\t",header=T)
anno<-read.delim("Full_450K_Annotation.txt",sep="\t",header=TRUE)
```

```
feat<-as.matrix(all$feat.rel.x)
feat2<-as.matrix(summary(all$feat.rel.x))
tempI<-as.matrix(c(1:nrow(feat2)))
for (x in 1:nrow(feat2)) tempI[x]<-((feat2[x]/nrow(feat))*100)
my.feat<-cbind(feat2,tempI,sum(tempI))
colnames(my.feat)<-c("My.Total","My.Percentage","Check")

fullfeat<-as.vector(anno$feat.rel)
source("/medical_genomics/paulg/scripts_misc/rand.R")

com.feat<-as.data.frame(cbind(Random.feat<-rand(fullfeat,nrow(all),1000),my.feat))
attach(com.feat)

com.feat2<-as.data.frame(c(1:nrow(com.feat)))
for (x in 1:nrow(com.feat)) com.feat2[x]<-com.feat["My.Percentage"]-com.feat["Percentage"]
com.feat3<-cbind(com.feat,com.feat2[,1])
colnames(com.feat3)<-c(colnames(com.feat),"Percentage Enrichment")
com.feat4<-com.feat3[order(com.feat3[,7],decreasing=TRUE),]

Lab.palette.3<-colorRampPalette(c("green","yellow","orange","pink","red","brown","blue"))
png("OS_ONCO/features_enrichment.png",height=6,width=10,units="in",res=600)
par(mar=c(7,2,1.2,4))
out<-barplot(com.feat4[,7],main="",ylab="%Enrichment",yaxt="n",ylim=c(-
5,5),cex.lab=1.3,col=Lab.palette.3(nrow(com.feat4)),border=NA,xaxt="n")
mtext(rownames(com.feat4),side=1,at=out,cex=1,las=2)
axis(2,at=c(seq(from=-5,to=5,by=1)),cex.axis=1.3,las=2)
dev.off()


####Calculate overlaps within and across tumour types
temp<-list.files("OS_ONCO/",pattern="*_deltabeta0.51.txt")
temp.l<-sapply(temp,function(x)
read.delim(paste("OS_ONCO/",x,sep=""),header=T,sep="\t",row.names=NULL),simplify=FALSE)


names(temp.l)<-gsub("_deltabeta0.51.txt","",names(temp.l))

#ONCO:
in1<-intersect(temp.l[[1]][,1],temp.l[[2]][,1])
in2<-intersect(temp.l[[3]][,1],temp.l[[4]][,1])
in3<-intersect(temp.l[[5]][,1],temp.l[[6]][,1])
in12<-intersect(in1,in2)
in123<-intersect(in12,in3)
length(in123) #5
in123.df<-data.frame(in123)
names(in123.df)<-"onco.probeID"
onco_ol<-merge(temp.l[[1]],in123.df)

write.table(onco_ol,"OS_ONCO/onco_overlap_info.txt",sep="\t",col.names=TRUE,row.names=FALSE
,quote=FALSE)

#OS:
in1<-intersect(temp.l[[7]][,1],temp.l[[8]][,1])
in2<-intersect(temp.l[[9]][,1],temp.l[[10]][,1])
in12<-intersect(in1,in2)
length(in12) #236
in12.df<-data.frame(in12)
names(in12.df)<-"os.probeID"
os_ol<-merge(temp.l[[7]],in12.df)
os_ol
```

```
write.table(os_ol,"OS_ONCO/os_overlap_info.txt",sep="\t",col.names=TRUE,row.names=FALSE,quote=FALSE)


#OS AND ONCO
test<-merge(onco_ol,os_ol,by.x="onco.probeID",by.y="os.probeID")
#no overlap
```