Differential DNA Hydroxy/Methylation Analyses for Improving Prognostic Accuracy in Localised PDAC and a Better Understanding of PDAC Progression



Aiyan Yau

Linacre College

University of Oxford

Supervised by

Professor Eric O'Neill

Michaelmas 2023

This thesis is submitted to the Department of Oncology,

University of Oxford, for the degree of

Master of Science by Research in Oncology

Declaration

I hereby declare that this thesis is entirely my own work, and except where otherwise stated, describes my own research.

Aiyan Yau

Linacre College

Acknowledgements

First and foremost, I would like to thank my supervisor, Eric O'Neill and my mentor, Frances Willenbrock who provided immense support and trust in my research. I would not be able to finish my study and thesis without their guidance and encouragement. I would also like to thank my parents, who took such good care of me over the years and for always being here when I need emotional support. I am also really glad to have met my "Basement Angels" (cohort-mates): Minn, Alexa, Kelly, Cedric, Teo and Alex; who are all so diligent and inspiring. Special thanks go to YT Lau, who helped me recover from breakdowns and gave me the strength to follow my own path in life. Lastly, I would like to thank my best friend Cindy, for believing in me and carrying me through the hard times. I would not have made it this far without her.

Table of Contents

Abstract ····· iii
List of Figures and Tables ····· v
List of Acronyms and Symbols ····· vii

Chapter 1: Introduction **1**
1.1 Project Initiative ····· 1
1.2 Hallmarks of PDAC ····· 3
1.2.1 Low Mutational Burden ····· 3
1.2.2 Desmoplastic Stroma and Immunosuppressive Tumour Microenvironment 4
1.3 PDAC Molecular Subtypes ····· 5
1.3.1 Classical vs Squamous (Basal-like) ····· 6
1.4 Epigenomic Profiling in PDAC ····· 7
1.4.1 DNA Hydroxy/Methylation Analysis ····· 10
1.5 Project Aims ····· 11

Chapter 2: Materials and Workflow **12**
2.1 Materials ····· 12
2.2 Workflow ····· 13

Chapter 3: Pairwise Differential Analyses **14**
3.1 Pre-processing of Raw Data ····· 14
3.1.1 Filtering Samples and Probes ····· 14
3.1.2 Normalisation ····· 16
3.1.3 Batch and Positional Effect Correction ····· 18
3.1.4 Constructing Final Matrices ····· 19
3.2 Differential Hydroxymethylation Analyses ····· 20
3.2.1 DhMP ····· 20
3.2.2 DhMR ····· 22
3.3 Differential Methylation Analysis ····· 27
3.3.1 DMP ····· 27
3.3.2 DMR ····· 29

Chapter 4: Scoring **30**
4.1 DhMP ····· 30
4.2 DhMR ····· 31
4.3 DMP ····· 32
4.4 DMR ····· 32

Chapter 5: Testing the Predictive Power of Selected Signatures **33**

5.1 Methodology ····· 33
5.1.1 Unsupervised HCA ····· 33
5.1.2 MCA ····· 34
5.1.3 Optimisation ····· 36
5.2 Results ····· 37
5.2.1. All 92 Variables ····· 37
5.2.2. Individual Analyses for Signatures from the 4 Differential Analyses ····· 39
5.2.3. Final Signature List ····· 40

Chapter 6: Preparation for Validation **42**

6.1 Candidate Genes: Differential Analyses Output ····· 42
6.2 Candidate Genes: PPI and Functional Enrichment Analysis ····· 44
6.2.1 PPI analysis with STRING ····· 44
6.2.2 Investigation of GO_BP Functional Enrichments with Cytoscape ····· 45
6.3 Selection of Genes ····· 48

Chapter 7: In vitro Validation – RT qPCR **49**

7.1 Methodology ····· 49
7.2 Output Interpretation ····· 50
7.3 Results ····· 51

Chapter 8: Discussion **53**

8.1 Project Review ····· 53
8.1.1 Challenges ····· 53
8.1.2 Result summary ····· 55
8.2 Limitations ····· 55
8.3 Implications and Future work ····· 57
8.3.1 Overlap with Classical/Squamous Subtype ····· 57
8.3.2 RT qPCR Follow-up ····· 57
8.4 Conclusion ····· 61

Bibliography **63**

Appendix **71**

Abstract

Despite immense efforts, the prognosis of pancreatic ductal adenocarcinoma (PDAC) remained very poor. The main reasons include the lack of early-stage symptoms and effective screening programmes, leading to delayed clinical presentations, with only 15-20% of patients being diagnosed with localised resectable tumours. However, ~40% of these patients develop distant recurrence within 12 months, suggesting that while clinically identified as localised, a subset of the patients likely had micro-metastatic lesions and neoadjuvant chemotherapies might have been the better option. There are strong demands for improving prognostic accuracy and this project aims to distinguish PDAC patients with metastasis-prone tumours from those with truly localised resectable tumours via DNA methylation (5mC) and hydroxymethylation (5hmC) signatures.

Starting with raw 5hmC & 5mC data, pairwise differential analyses were performed to identify significantly differentially hydroxy/methylated CpGs - D(h)MPs and significantly differentially hydroxy/methylated regions - D(h)MRs. ChAMP detected 3 DhMPs being significantly hypo-hydroxymethylated in patients with distant recurrence, whereas 21 DhMRs were significantly differentially hydroxymethylated. Similarly, 35459 DMPs and 551 DMRs were detected. Scorings were performed for subsequent unsupervised hierarchical clustering analyses to check how well the identified signatures distinguish between the two groups. The results showed that 5(h)mC profiles indeed have the power to separate patients into their prognostic groups. To gain a better understanding of the underlying mechanisms, functional enrichment analysis was performed for the top signatures. As validation, RT qPCR of selected genes in classical & squamous PDAC cell lines were performed; in which 6 of the candidate genes have a

profound transcription pattern, suggesting an overlap between our hydroxy/methylomic-driven stratification and the transcriptomic-driven classical/squamous stratification.

In conclusion, this preliminary study showed that 5(h)mC signatures have the potential to improve prognostic accuracy. Most importantly, this approach can be adapted for other scientific questions, bringing new opportunities to combat PDAC and different diseases.

List of Figures and Tables

Figures:

*Fig. 1.1, 1.3, 1.4, 3.5 & 3.7 were created with https://biorender.com/

Chapter 1

1.1 Distinct Groups of Clinically Localised Patients ····· 2
1.2 Epigenetic Regulation via DNA Hydroxy/Methylation ····· 8
1.3 Cytosine Oxidation Cycle ····· 9
1.4 5mC/5hmC Measurements via BS & oxBS ····· 10

Chapter 3

3.1 pd file ····· 15
3.2 Normalisation ····· 17
3.3 Epic Array Positional Effects ····· 18
3.4 Batch and Positional Effect Correction ····· 19
3.5 DhMP Analysis ····· 20
3.6 DhMP Outputs ····· 21
3.7 DhMR Analysis ····· 24
3.8 Top 70 Genes Mostly Enriched by Significant CpGs ····· 28
3.9 *PCDHGA4* Methylation Status ····· 28

Chapter 4

4.1 Scoring by DhMP Signature Thresholds ····· 30
4.2 Scoring by DhMR Signature Thresholds ····· 31

Chapter 5

5.1 Visualising MCA Results (Example 1) ····· 35
5.2 Visualising MCA Results (Example 2) ····· 36
5.3 Clustered Heatmap [92 Variables] ····· 37
5.4 MCA Factor Map [92 Variables] ····· 38
5.5 Individual Clustered Heatmaps ····· 39
5.6 Clustered Heatmap [68 Variables] ····· 40
5.7 MCA Factor Map [68 Variables] ····· 41

Chapter 6

6.1 Candidate Genes from Previous Analyses ····· 43
6.2 PPI Network ····· 44

6.3 Exploring Functional Enrichment Results ····· 46
6.4 Identification of "Core Genes" ····· 47
6.5 Final Gene List for RT qPCR ····· 48

Chapter 7

7.1 Genes Upregulated in Squamous ····· 52
7.2 Genes Downregulated in Squamous ····· 52

Appendix

A1 All Genes Enriched by CpGs within significant DMRs ····· 71
A2 pvclust Dendrogram [92 Variables] ····· 71
A3 pvclust Dendrograms for different signature categories ····· 72
A4 MCA Factor Maps for different signature categories ····· 73
A5 pvclust Dendrogram [68 Variables] ····· 74

Tables:

Chapter 3

1 DhMR Output ····· 26
2 DMP Output ····· 27
3 DMR Output ····· 29

Appendix

A1 Primer list ····· 75
A2 Relative Fold Changes for genes showing interesting trends ····· 76
A3 ANOVA/Tukey HSD test results ····· 76

List of Acronyms and Symbols

Acronyms

$\Delta\bar{\beta}$-value	Difference between NR & DR_B group average β-values
$\Delta h\bar{\beta}$-value	Difference between NR & DR_B group average hβ-values
$\Delta h\bar{M}$-value	Difference between NR & DR_B group average hM-values
$\Delta\bar{M}$-value	Difference between NR & DR_B group average M-values
5fc	5'formylcytosine
5hmC	5'hydroxymethylcytosine/ DNA Hydroxymethylation
5mC	5'methylcytosine/ DNA Methylation
β-values	5mC/oxBS Data
AREA	Area Under Regression Curve for Particular Region
AU p-values	Approximately Unbiased p-values
AVG	Regional $\Delta h\bar{M}$ Average/ Regional $\Delta\bar{M}$ Average
BO	Biological Process
BMIQ	Beta Mixture Quantile
CAF	Cancer-associated Fibroblasts
CpG	Cytosine-phosphate-Guanine
C_t	Cycle threshold
DhMP	Differentially Hydroxymethylated Positions/Probes
DhMR	Differentially Methylated Regions
DMP	Differentially Methylated Positions/Probes
DMR	Differentially Methylated Regions
DR_B (0-12m)	Distant Recurrence & Both Local+Distant Recurrence (0-12 months)
ECM	Extracellular Matrix
EMT	Epithelial to Mesenchymal Transition
FE	Functional Enrichment Core Genes
GEF	Guanine Nucleotide Exchange Factor
GO	Gene Ontology

GOI	Gene of Interest
GUI	Graphical User Interface
hβ-values	5hmC/BS – oxBS Data
HCA	Hierarchical Clustering Analysis
hM-values	Logit-transformed hβ-values
iCAF	Inflammatory Cancer-associated Fibroblasts
IGR	Intergenic Region
LOESS	Locally Estimated Scatterplot Smoothing
MCA	Multiple Correspondence Analysis
M-values	Logit-transformed β-values
MDSC	Myeloid-derived Suppressor Cells
NR	No Recurrence
OS	Overall Survival
PC	Pancreatic Cancer
PDAC	Pancreatic Ductal Adenocarcinoma
PFS	Progression-free Survival
PPI	Protein-protein Interaction
SVD	Singular Value Decomposition
PSC	Pancreatic Stellate Cells
TAM	Tumour-associated Macrophages
TP	Top 10 Genes Enriched by Significant DMPs
TR	Top 10 Genes Enriched by CpGs of Significant DMRs
TME	Tumour Microenvironment

Gene Symbols

ACTB	Actin Beta
CDC42	Cell Division Cycle 42
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A
CTNNA2	Catenin Alpha 2

DOCK8	Dedicator of Cytokinesis 8
EGFR	Epidermal Growth Factor Receptor
FAS	Fas Cell Surface Death Receptor
GATA6	GATA Binding Protein 6
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase
GDNF	Glial Cell Derived Neurotrophic Factor
HNF1A	HNF1 Homeobox A
HNF4A	Hepatocyte Nuclear Factor 4 Alpha
ITGA4	Integrin Subunit Alpha 4
ISLR2	Immunoglobulin Superfamily Containing Leucine Rich Repeat 2
KRAS	KRAS Proto-Oncogene, GTPase
MYC	MYC Proto-Oncogene, BHLH Transcription Factor
PCDHGA4	Protocadherin Gamma Subfamily A, 4
PD1/PDCD1	Programmed Cell Death 1
PDX1	Pancreatic and Duodenal Homeobox 1
RAP1/RAP1A	RAP1A, Member of RAS Oncogene Family
RAP2/RAP2A	RAP2A, Member of RAS Oncogene Family
RAPGEF6	Rap Guanine Nucleotide Exchange Factor 6
SMAD4	SMAD Family Member 4
SPRR1A	Small Proline Rich Protein 1A
STAT3	Signal Transducer and Activator of Transcription 3
TET2	Tet Methylcytosine Dioxygenase 2
TIM3/HAVCR2	Hepatitis A Virus Cellular Receptor 2
TP53	Tumour Protein P53
TP63	Tumour Protein P63
WASP/WAS	WASP Actin Nucleation Promoting Factor
WIP/ WIPF1	WAS/WASL Interacting Protein Family Member 1

Chapter 1: Introduction

1.1 Project initiative

Pancreatic cancer (PC) is the 12th most common cancer across the globe and has the 4th highest cancer mortality rate in Western countries (Dell'Aquila *et al.*, 2020). In fact, PC incidence is predicted to increase by 6% between 2014 and 2035 in the UK (Smittenaar *et al.*, 2016) and is projected to be causing the 2nd highest number of cancer-related death in the US by 2040 (Rahib *et al.*, 2021). Despite immense efforts, the prognosis of PC patients remained very poor for the past 50 years; with a 5-year survival of 7.3% (Sophie John & John Broggio, 2019). While the pancreas exhibits both endocrine and exocrine functionalities, over 95% of the cells within this heterochrony organ are exocrine cells. This explains partially why most pancreatic cancers develop from the exocrine compartment, with about 90% of PC cases being diagnosed as pancreatic ductal adenocarcinoma (PDAC) – the focus of this study (Orth *et al.*, 2019). One of the greatest challenges in PDAC is the lack of early-stage symptoms, which leads to delayed, hence advance-staged clinical presentation (Wood *et al.*, 2022). This is also why it has been difficult to fully cure and extend the disease free survival for most PDAC patients.

While the conventional therapeutic strategies for PDAC includes surgical resection (for localised resectable tumours), radiotherapy and systemic chemotherapies (most common ones being gemcitabine and FOLFIRINOX); the limited improvement in prognosis indicates that there is still a great demand for new treatments, especially targeted and personalised therapies (Wood *et al.*, 2022). Nevertheless, as with other diseases, the development of new therapies is extremely time-consuming and costly. In

addition, while only about 15-20% of the patients were diagnosed with localised resectable tumours at initial presentation, data showed that more than 40% of these patients in fact suffers a relapse within 12 months despite well-executed resections, with 80% being distant recurrence (Groot *et al.*, 2019; Principe *et al.*, 2021). This suggests that while clinically identified as localised, a subgroup of patients likely had micro-metastatic lesions (as depicted in Fig. 1.1) and neoadjuvant chemotherapies might have been the better option. A refinement of the screening programmes and the discovery of new prognostic markers for PDAC are therefore, as equally crucial. As such, this project aims to study how we can make use of the advancement in epigenomic technologies to construct a framework that can serve as a tool for the identification of stratification signatures which can improve prognostic accuracy, for the extraction of key information that could enhance our understanding of PDAC pathology and highlight candidates for further drug discovery research.

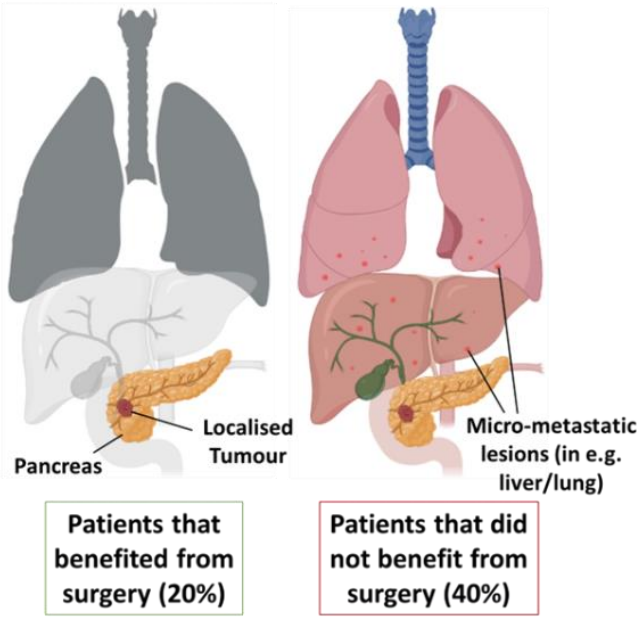


Fig. 1.1 Distinct Groups of Clinically Localised Patients

This is an illustration of how ~40% of the patients that were diagnosed with a localised, resectable tumour might in fact be bearing micro-static lesions in nearby organs, which explains why they might suffer from distant recurrence shortly after surgery.

1.2 Hallmarks of PDAC

Before introducing the actual research, it would be beneficial to keep note of features that are specific to PDAC, on top of most of the general hallmarks of cancer such as enabling replicative immortality, uncontrolled proliferation, apoptotic resistance, and metabolic reprogramming (Hanahan and Weinberg, 2011).

1.2.1 Low Mutational Burden

Firstly, unlike the majority of cancers, most PC/PDAC cases are known to have a relatively low mutational load, in which early research specified recurring mutations mainly occurring within *KRAS*, the proto-oncogene and within tumour suppressive genes: *CDKN2A*, *TP53* and *SMAD4* (Gao *et al.*, 2020). Amongst the 4 most frequently mutated genes, it is known that *KRAS* mutations act as the driver mutation, with more than 90% of the PDAC cases containing at least one mutation within the gene and is essential for the initiation of oncogenesis (Orth *et al.*, 2019). Study involving a comprehensive genetic evaluation by Jones et al. (2008) on the other hand, revealed 63 genetic alterations in pancreatic cancer (mostly point mutations), which are enriched in 12 sets of tumorigenesis-related signalling pathways. KRAS, TGF-β and Wnt/Notch signalling, for example are regulatory pathways that contained at least 1 altered gene in 100% of the advanced PC tumours in Jones' study; with each pathway enriched with 5, 29 and 37 of the genetically altered genes out of the 63 identified respectively (Jones *et al.*, 2008). While the relatively low mutational burden at first may seemed to be a beneficial feature, as with many cancer-related traits, it can also favour tumour growth due to fewer neoantigens available as targets for immune cells and therapeutics. Studies have also described that targeting of the key oncogenes e.g., *KRAS* has been extremely

difficult, with many genetic or small molecule inhibitors (some only targeting an uncommon form of KRAS mutation) still in their early stage of development or clinical trials . Further research is required to study if targeting KRAS on its own is sufficient to bring significant improvement in prognosis and explore the mechanisms behind KRAS inhibition resistance (Qian *et al.*, 2020; He, Liu and Wang, 2022).

1.2.2 Desmoplastic Stroma and Immunosuppressive Tumour Microenvironment

Another significant characteristic of PDAC is the dense desmoplastic stroma which might make up as high as 80% of the tumour volume. It is mainly consisted of endothelial cells, fibroblasts, extracellular matrix (ECM), and different immune populations (Bulle and Lim, 2020). Desmoplasia or in other words, the high level of fibrosis within a tumour, facilitates the build-up of a hypoxic tumour microenvironment (TME) which encourages angiogenesis, tumour invasion and metastasis; as well as transcriptional, epigenetic, and metabolic reprogramming via the activation of oncogenes and silencing of suppressor genes (Bulle and Lim, 2020; Yamasaki, Yanai and Onishi, 2020; Maneshi *et al.*, 2021). Interestingly, hypoxia enhances desmoplasia in return by e.g., activating resident pancreatic stellate cells (PSCs) which is one of the origins of cancer-associated fibroblasts (CAFs) that generate stiff and remodelled e.g., growth-factor (e.g., TGF-β) enriched ECM (Bulle and Lim, 2020; Yamasaki, Yanai and Onishi, 2020).

Unsurprisingly, such fibrotic stromal compartment also demonstrates pro-tumorigenic properties by functioning as a physical and biological barrier, selectively blocking the infiltration of anti-tumour lymphocytes and restricting the proper activation of effector T cell-mediated cytotoxicity, hence turning PDAC into "immune cold tumours"(Huber

et al., 2020; Masugi, 2022). On the other hand, research also showed that the pro-oncogenic inflammatory CAFs (iCAFs) upregulate exhaustion and apoptotic markers such as *TIM3*, *PD1* and *FAS* on T cells, hence contributing to the inactivation and death of cytotoxic T cells (Geng *et al.*, 2021; Narayanan, Vicent and Ponz-Sarvisé, 2021; Koppensteiner *et al.*, 2022). Activated PSCs or CAFs also supports the immunosuppressive TME through the restriction of $CD8^+$ cell migration, STAT3-mediated recruitment of T-reg stimulating myeloid-derived suppressor cells (MDSCs), and M1 to M2 polarisation, fostering a conversion into tumour-associated macrophages (TAM) (Ene-Obong *et al.*, 2013; Mace *et al.*, 2013; Sperb, Tsesmelis and Wirth, 2020; Deng *et al.*, 2022). In conclusion, besides the dense stroma physically prevents proper therapeutic diffusion, the stroma's hypoxic and fibrotic properties and severe disruption in Treg and CD8 activities are also key factors leading to disappointing immunotherapy responses (Daniel *et al.*, 2019; Bulle and Lim, 2020; Masugi, 2022).

1.3 PDAC Molecular Subtypes

While another important feature of PDAC is the intra-/inter-tumour heterogeneity, numerous studies were conducted to classify PDAC into different molecular subtypes and determine their association with distinct prognostic outcomes. As described in Martens et al.'s review (2019), the transcriptomic-driven stratification of PDAC is not only beneficial, but a necessary part of PDAC research as it facilitates the discovery of new drug targets, the identification of biomarkers for improving prognostic accuracy and effective treatment selection, while deepening our understanding of PDAC progression. Over the years, the results from major PDAC molecular subtyping studies including the ones conducted by Collisson et al. (2011); Moffitt et al. (2015) and Bailey et al. (2016), were widely evaluated, for instance, by applying the identified

transcriptomic signatures on varying data sets, or by performing stratification with similar methods then comparing signatures to check for overlapping. Techniques used include gene expression microarray, RNA-sequencing and integrated genomic analysis (+ mutational data) (Martens *et al.*, 2019). Thereafter, it is widely agreed, based on the overlapping signatures and subtyping from independent studies that there are 2 major PDAC subtypes: classical (progenitor) vs squamous (basal-like) (Torres and Grippo, 2018; Martens *et al.*, 2019).

1.3.1 Classical vs Squamous (Basal-like)

As mentioned, one of the most important differences between the 2 subtypes is the distinction between clinical outcomes, with classical tumours being associated with better treatment responses, longer progression-free survival (PFS) and overall survival (OS); whereas patients with squamous tumours were associated with significantly worse outcomes (Bailey *et al.*, 2016; Torres and Grippo, 2018; Malinova *et al.*, 2021). Extensive analyses of the transcriptional signatures such as gene set enrichment analyses did not only serve as further confirmation of the functional relevance in such stratification and association with prognosis, but also provided insight into underpinning mechanisms.

For instance, the classical subtype is enriched with higher levels of differentiation and epithelial cell-fate markers (including critical transcription factors: *GATA6*, *HNF1A* & *HNF4A*); while exhibiting lipogenic metabolic and "normal" stromal signatures (Daemen *et al.*, 2015; Martens *et al.*, 2019; O'Kane *et al.*, 2020). Patients with the classical subtypes also displayed a higher sensitivity with FOLFIRINOX, gemcitabine and EGFR inhibitors like erlotinib (Torres and Grippo, 2018; Xu *et al.*, 2021).

Squamous tumours on the other hand are characterised with stemness (poorly differentiated) and reduced expression & activities of endodermal genes (notably *GATA6*, *HNF1A* & *HNF4A*) (Martens *et al.*, 2019; O'Kane *et al.*, 2020). They are also associated with an "activated" stromal subtype and exhibit a higher dependence on Wnt, ΔNp63 & TGFβ signalling which promotes epithelial to mesenchymal transition (EMT), tumour progression and metastasis (Martens *et al.*, 2019; Mattiolo *et al.*, 2022). Unsurprisingly, squamous tumours have glycolytic metabolic profiles (linked with downregulated *HNF4A*), indicating an outrageously elevated demand in the generation of energy and macromolecules to accommodate the rapid tumour growth (Daemen *et al.*, 2015; Liu *et al.*, 2020). While patients with PDAC of squamous subtype showed limited response to conventional and existing medications, research suggests that they may benefit the most from (neo)adjuvant, combination, and next generation (targeted) therapies (Torres and Grippo, 2018; Xu *et al.*, 2021; Mattiolo *et al.*, 2022).

It is also worth noting that classical and squamous subtypes of similar properties are also found in multiple solid tumours, for example in breast and colorectal tumours, hinting that the dysregulation of particular groups of pathways might be a common theme in multiple cancer types and comparative studies and knowledge integration might be valuable (Xu *et al.*, 2021; Jaaks *et al.*, 2022; Mattiolo *et al.*, 2022).

1.4 Epigenomic Profiling in PDAC

Notably, besides the mutational/transcriptomic-driven PDAC subtyping described in 1.3, the rapid development of epigenetic evaluation techniques (both wet-lab and bioinformatics approaches), along with the surge in the number of research highlighting the significance of epigenetic reprogramming in tumorigenesis, encouraged similar

classification studies that focus on the epigenomic profiles in PDAC patients. Apart from microRNAs working as regulators, as described in Ciernikova et al.'s paper (2020) and illustrated partially in Fig. 1.2, other epigenetic mechanisms include the methylation of histones/DNA which is associated with transcriptional suppression; and the acetylation or hydroxymethylation of histones/DNA which is usually associated with transcriptional activation .

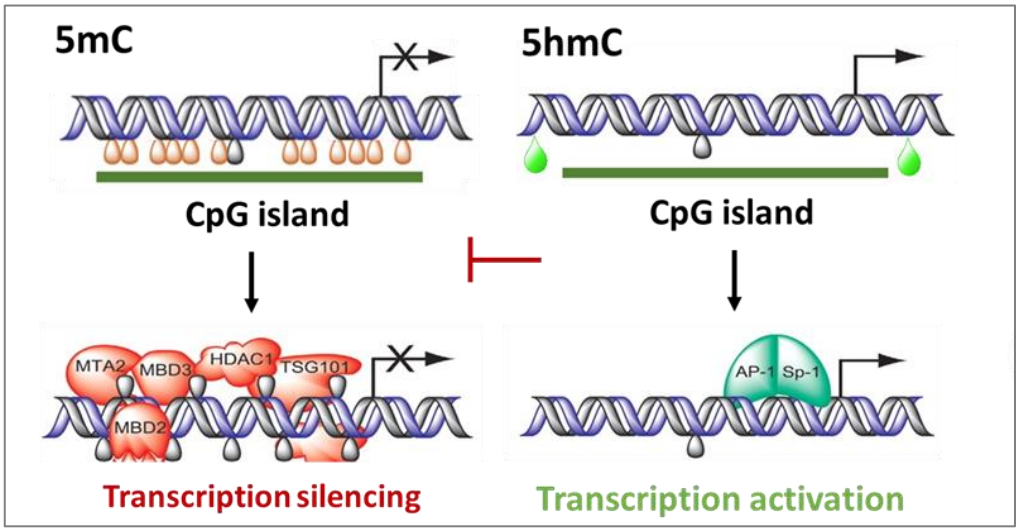


Fig. 1.2 Epigenetic Regulation via DNA Hydroxy/Methylation
As illustrated, DNA hydroxy/methylation cause a cascade of amplifying events like chromatin remodelling, further DNA or histone modifications and the recruitment of appropriate proteins for the formation of transcription activating or silencing complexes respectively (Figure adapted from Brabson et al. (2021)).

It is noteworthy that we decided to set DNA hydroxy/methylation as the focus of this project because a relatively high number of unsupervised studies exploring the distinct DNA methylation profiles in PDAC suggested promising results that are in concordance with previously described transcriptional, metabolic, and prognostic trends (Lomberk et al., 2018; Ciernikova et al., 2020; Eyres et al., 2020). Some of the main findings regarding DNA methylome in varying PDAC subtypes include a significantly elevated levels of methylation observed at endodermal and differentiation markers (e.g., *GATA6* & *PDX1*) in squamous tumours, which is in line with their silencing observed in

previous transcriptomic classification studies (Eyres *et al.*, 2020; Xu *et al.*, 2021; Roalsø *et al.*, 2022). In contrast, significant hypomethylations are observed at oncogenic drivers such as *MYC*, *TP63* & *Wnt*-related genes, which again match the upregulations and downstream activations observed in corresponding expression profiles (Eyres *et al.*, 2020; Xu *et al.*, 2021; Roalsø *et al.*, 2022). The limited number of research in-relation to PDAC hydroxymethylation profiles in the current field on the other hand means that, this is a great opportunity to gain new insights for the understanding of the complex biology behind PDAC oncogenesis and detection of new prognostic biomarkers & therapeutic targets. Moreover, studies revealed that, rather than just being an intermediate in the passive demethylation pathway (Fig. 1.3), 5-hydroxymethylated cytosines (5hmCs) actively prevent the methylation of the CpG islands and maintain transcription activation (e.g., in the brain and embryonic stem cells); signifying another layer of epigenetic regulation that is worth considering (Eyres *et al.*, 2020; Guler *et al.*, 2020). The fact that hydroxy/methylation status can be measured using circulating tumour DNA, also showcased the signatures' great clinical translational potential as non-invasive prognostic markers (Guler *et al.*, 2020; Roalsø *et al.*, 2022).

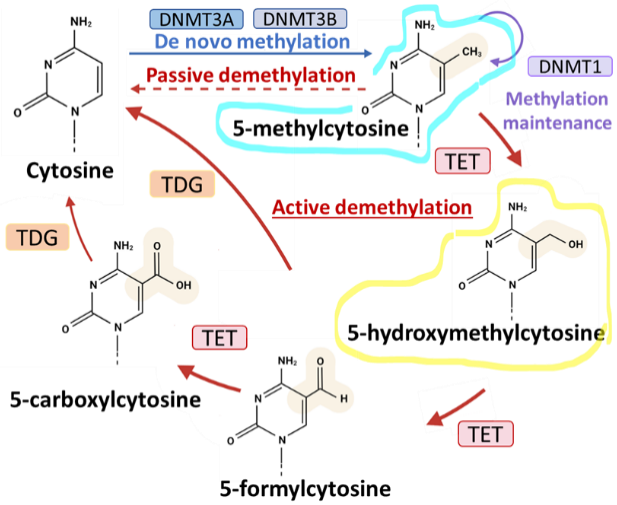


Fig.1.3 Cytosine Oxidation Cycle
Instead of merely being an intermediate, 5hmCs play role in 5mC inhibition and active demethylation, hence help maintaining active transcription.

1.4.1 DNA Hydroxy/Methylation Analysis

As described in Huang and Wang's review (2019), one of the most common ways of measuring DNA methylation is to perform microarray or next generation sequencing after sodium bisulfite treatment because unmethylated cytosines would be converted to uracils and hence sequenced as thymines, whereas the methylated cytosines would remain intact – technique commonly referred to as bisulfite sequencing (BS). This however is not the best approach as it cannot distinguish between 5mCs and 5hmCs. Fortunately, this can be resolved by an extra oxidation step before bisulfite treatment, which would convert the 5hmCs into (5formylcytosine or) 5fCs, which would also be converted into uracils and sequenced as thymines – technique commonly referred to as oxidative bisulfite sequencing (oxBS) (Huang and Wang, 2019; Tiedemann *et al.*, 2021). Therefore, as illustrated in Fig. 1.4, both 5mC and 5hmC levels can be quantified and used in combination for different bioinformatics analysis including differential hydroxy/methylation analyses, just as conventional differential expression analyses (Eyres *et al.*, 2020; Tiedemann *et al.*, 2021).

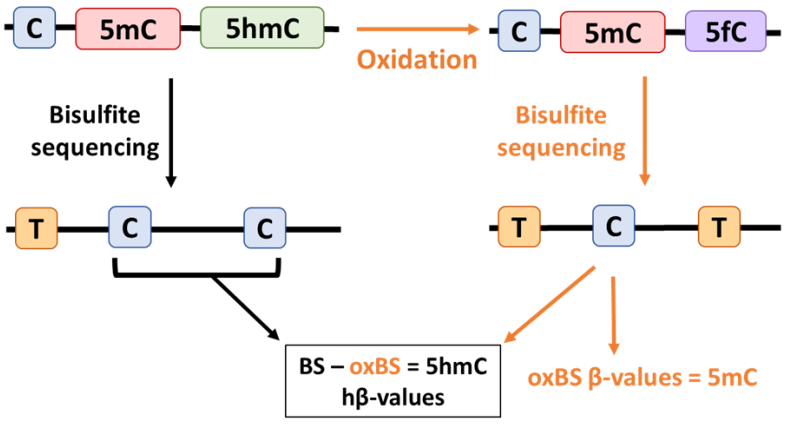


Fig.1.4 5mC/5hmC Measurements via BS & oxBS
Here is a schematic illustration of how bisulfite sequencing and the extra oxidation step in oxBS works to distinguish between different forms of cytosines, allowing the extraction of both β-values and hβ-values for differential analyses.

1.5 Project Aims

In summary, building on such concepts, the ultimate goal of this project is to compare the hydroxy/methylomes in PDAC patients and determine if patients that recurred within 12 months after surgery indeed represent a hidden subtype within the group of patients diagnosed with localised resectable tumours. Next, if any signatures representing the metastatic subtype were found, further evaluation such as functional enrichment analyses and protein-protein interactions would be conducted for a better understanding of the mechanisms fostering such distinct prognostic outcomes. Lastly, after narrowing down candidate genes carefully, RT qPCR would be performed for cell lines of classical and squamous subtypes, in an attempt to test the applicability of the identified signatures in vitro and gain more insight into the underlying biology.

Chapter 2: Materials and Workflow

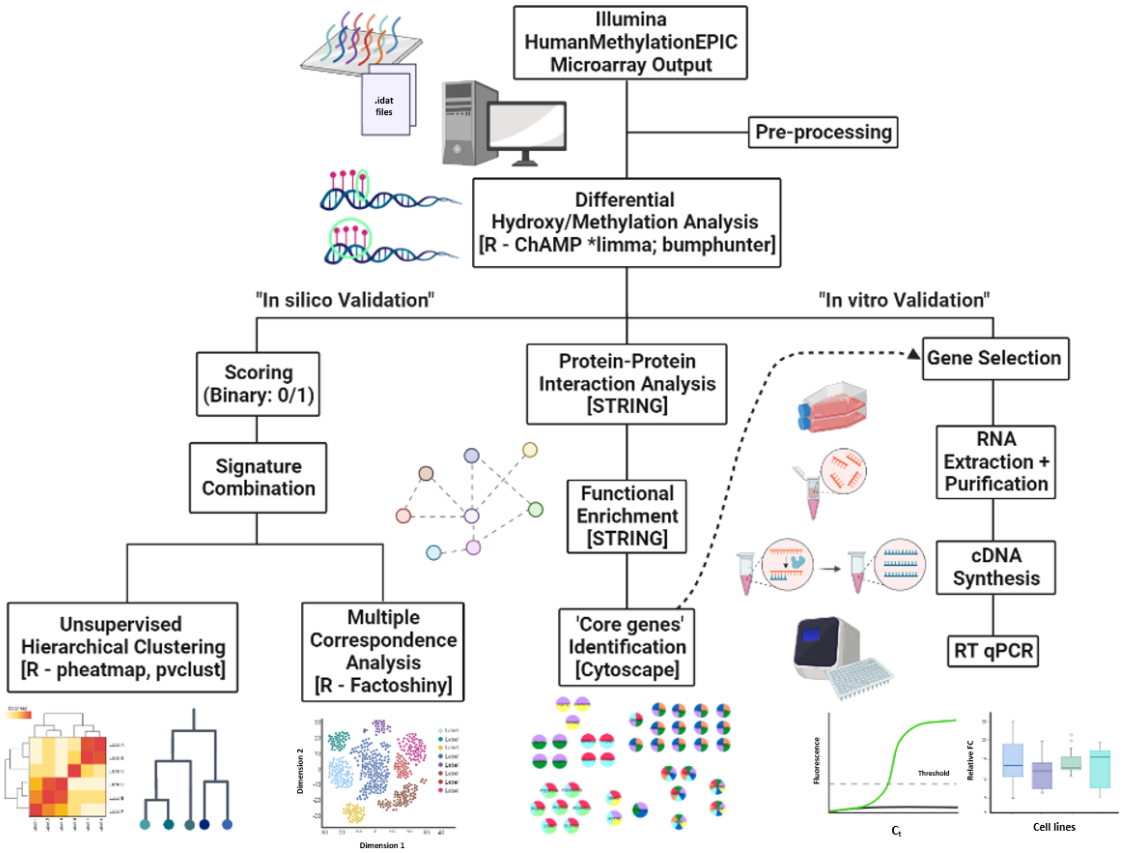
2.1 Materials

In fact, this project was inspired by the work performed by Michael Eyre, a previous DPhil student, in which his work revealed the association between a diminished expression of *TET2*, the 5-methylcytosine hydroxylase, with a significant 5hmC reduction and the epigenetic inactivation of downstream differentiation-promoting genes such as *GATA6*, which leads to the development of a more invasive, squamous-like phenotype in PDAC (Eyres *et al.,* 2020). Michael's research also demonstrated that metformin and vitamin C (ascorbic acid) have the ability to reverse such aggravating epigenetic modifications and induce classical subtype characteristics that are associated with improved prognosis (Eyres *et al.*, 2020). Most importantly, as with other research involving bioinformatics, one of the greatest advantages is that the genomic and epigenetic data retrieved can be reused countless times to study a wide range of scientific questions regardless of time and location.

Hence, the .idat files containing raw BS and oxBS data from the Illumina Infinium HumanMethylationEPIC microarray were the starting materials for this project. Details regarding the workflow and methodologies, including DNA extraction and purification from FFPE tissue blocks, BS and oxBS, followed by the use of Illumina microarray (coverage: >850k CpG sites) can be found in Michael's paper (Eyres *et al.,* 2020). The relevant clinical data for the 144 treatment naïve (R0/R1) patients with no previous malignancy, no cancer metastatic spread and ascites; from the Oxford Kidani PDAC Cohort were also available. Note that Michael's project was approved by the Oxford Radcliffe Biobank, University of Oxford (Project code: OCHRe 14/A176).

2.2 Workflow

Below is the overall workflow of this project, all the way from the differential analyses of raw 5mC/5hmC data to in silico validations, then following functional enrichment analyses, "core genes" identification and finally, in vitro validation via RT qPCR in classical and squamous PDAC cell lines.



Chapter 3: Pairwise Differential Analyses

As described previously, in order to identify specific CpG sites or regions that were significantly differentially hydroxy/methylated in patients with distant recurrence within 12 months after surgery, 2 types of pairwise differential analyses were performed to compare the quantified hydroxy/methylation levels. Shortlisted CpGs and regions can then be used for scoring (chapter 4). Note that the RStudio package "ChAMP" (R version 4.1.2; RStudio version 2021.09.0+351) was used for the differential analyses (Morris *et al.*, 2014). The overall workflow and basic structures of different codes can be found within the ChAMP pipeline (Tian *et al.*, 2017), with a brief step-by-step guide available online: https://bioconductor.org/packages/release/bioc/vignettes/ChAMP/inst/doc/ChAMP.html#section-filtering-data

3.1 Pre-processing of Raw Data

3.1.1 Filtering Samples and Probes

As stated in chapter 2, the Oxford Kidani cohort is consisted of 144 PDAC patients. Nevertheless, due to various practical or technical issues (e.g., insufficient FFPE scrolls; the lack of high quality DNA samples or unsuccessful oxidation treatment), Michael only ran the Illumina EPIC microarray for 96 patients, resulting in 384 .idat files (4 per patient – red/green intensity readings for BS/oxBS samples); serving as the starting materials for this project. Note that before filtering according to phenotypes, 5 patients were removed due to conflicting/confusing clinical data. Then since the focus of this study is to identify patients with micrometastatic lesions (which are in this case, represented by patients with distant recurrence within 12 months), 12 patients with local

recurrence (i.e., supposedly no micrometastatic lesions) and 24 patients with distant recurrence after 12 months were also omitted (as patients being disease free for >12m are being defined as benefited from surgery for this study). Finally, as ChAMP was only capable of pairwise comparisons (no recurrence vs distant recurrence <12m), data collected from 3 healthy patients were also not included in the following analyses.

In fact, prior to the pre-processing of raw intensity data, a .csv document commonly known as the "pd file" or "Sample Sheet" which is available as part of the EPIC microarray output data, was edited to make sure it was in the correct format and contained the required data for the samples used in the study. These include the sample names, microarray positional information, batch number and variable groupings for each sample. It was crucial that data related to the samples' location was not altered as those are the information that the system uses to match the raw .idat intensity files to the corresponding samples. Although suggestions on how to construct the document is available online, since this study looks at both BS and oxBS data, extra attention was paid in the naming and order of the samples so that the filtered data would be easier to handle at later stages (see Fig. 3.1).

	A	B	C	D	E	F	G	H	I
1	[Header]								
2	Investigator Name	HM							
3	Project Name	P160246							
4	Experiment Name	P160246							
5	Date	03.01.18							
6	[Manifests]								
7	A	MethylationEPIC_v-1-0_B2.bpm							
8	[Data]								
9	Sample_Name	Sample_Well	Recurrence_Time	Recurrence_Site	BS_oxBS	Sample_Group	Sample_Plate	Sentrix_ID	Sentrix_Position
10	NR_01a	B10	NR	NR	oxBS	NR	P160246_P01	200527000000	R02C01
11	NR_02a	F10	NR	NR	oxBS	NR	P160246_P01	200527000000	R06C01
12	NR_03a	D05	NR	NR	oxBS	NR	P160246_P01	200527000000	R04C01
13	NR_04a	D07	NR	NR	oxBS	NR	P160246_P02	200527000000	R04C01
14	NR_05a	F07	NR	NR	oxBS	NR	P160246_P02	200527000000	R06C01

Fig. 3.1 pd file
This is a screenshot capturing the start of the revised pd.

Finally, the pd and raw .idat files saved within the same directory were then imported to R using the champ.load() function integrated from the "minfi" package, specifying [arraytype = "EPIC"] (Aryee *et al.*, 2014; Fortin, Triche and Hansen, 2016). It is noteworthy that files from both BS and oxBS were loaded at the same time to make sure data were pre-processed under the same conditions. 4 more patients were removed due

to high detection p.value [filterDetP=TRUE, SampleCutoff=0.1]; indicating that those sample contained a high number of probes with background signals, leaving a final sample size of 48: patients with no recurrence [NR] = 21 (by the end of the tracking programme: 30th of December 2017); whereas patients with distant recurrence & both local + distant recurrence (within 12 months) [DR_B (0-12m)] = 27.

The champ.load() function also filtered out probes according to 6 parameters: those with a high detection p-value [detPcut=0.01], bead-count lower than 3 in at least 5% of all samples [filterBeads=TRUE, beadCutoff=0.05], all non-CpG probes [filterNoCG = TRUE], those that were associated with SNPs [filterSNPs = TRUE] (based on Zhou, Laird and Shen, 2017 - default), all multi-hit i.e., non-specific probes [filterMultiHit = TRUE] (based on Nordlund et al., 2013 - default) and probes that detect CpGs located in chromosome X and Y [filterXY = TRUE]. As a result, there were 623660 probes left for following analyses and the final output of the function is a sheet of raw β-values = M/(M+U+α); in which M, U and α denotes the methylated intensities, unmethylated intensities and constant offset – 100 by default, respectively. In other words, β-values represents the fraction of methylation measured at specific CpG sites. The function also made minor adjustments to the data to avoid outliers: β-values must lie between 0 and 1. Hence, values that were smaller or equal to 0 were replaced by the smallest positive value; whereas values that were bigger or equal to 1 would be replaced by the largest value smaller than 1.

3.1.2 Normalisation

An extra normalisation step is required as Illumina EPIC array relies on two types of Infinium assays. Although including type 2 probes (Infinium II assays) greatly expands genome coverage, studies like the one conducted by Dedeurwaerder et al. (2011)

revealed a trade-off in measurement accuracy. This is due to a reduced sensitivity in type II probes for extremely low or high methylation values (Fig. 3.2). The decreased dynamic range observed for type II probes can be problematic as this may introduce "type I enrichment bias" in downstream statistical analyses that involve probe ranking. Dedeurwaerder et al.'s study also demonstrated that data measured with type II probes are less reproducible and have a greater variance (the standard deviation for replicates were 0.008 for type I probes but 0.029 for type II probes).

As such, while data retrieved from type 2 probes are still informative, it is recommended to correct them based on type I data. In ChAMP, this can be achieved by running the champ.norm() function. BMIQ (Beta Mixture Quantile) dilation is a common strategy for EPIC array normalisation in which type II probe β-values are transformed in a monotonic manner, making sure that while the adjusted density distribution is now comparable to that of type I probes, the relative ranking of the probes would not be altered (Teschendorff *et al.*, 2013) (Fig. 3.2). The fact that BMIQ normalisation in ChAMP can be run with multiple cores, which speeds up the process significantly, also makes it a popular choice (Tian *et al.*, 2017).

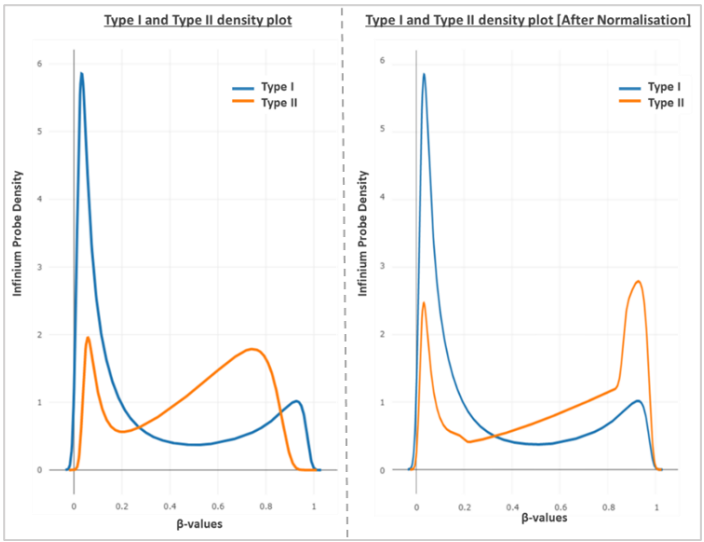


Fig. 3.2 Normalisation
These plots exported as part of the quality control output show how the normalisation with BMIQ corrects type II probe bias and affects the β-values density distribution.

3.1.3 Batch and Positional Effect Correction

Lastly, in order to examine the significance of different technical variables, including batch and positional effects (Fig. 3.3) that can possibly cause variation in the data, singular value decomposition (SVD) was applied to the normalised β-matrix via the champ.svd() function. Since all variables of interest were categorical, Kruskal Wallis Test would be used for the calculation (Teschendorff *et al.*, 2009). It can be seen from the output (Fig. 3.4) that the 3 technical variables: "Sentrix ID", "Sentrix Position" and "Sample Plate" were indeed playing a significant part in data variation. Therefore, there was a need to use the champ.runCombat() function (integrated from the "SVA" package) to correct for the positional and batch effects (Johnson, Li and Rabinovic, 2007; Leek *et al.*, 2012; Jiao *et al.*, 2018). This method is based on an empirical Bayesian framework, in which the effect size (coefficient) would be calculated and then removed for each batch/positional effect. It can be seen in the SVD heatmaps before and after correction that batch and positional effects were removed successfully, and "Sample Group" is now the variable accounting for the majority of variation observed in our data (see Fig. 3.4).

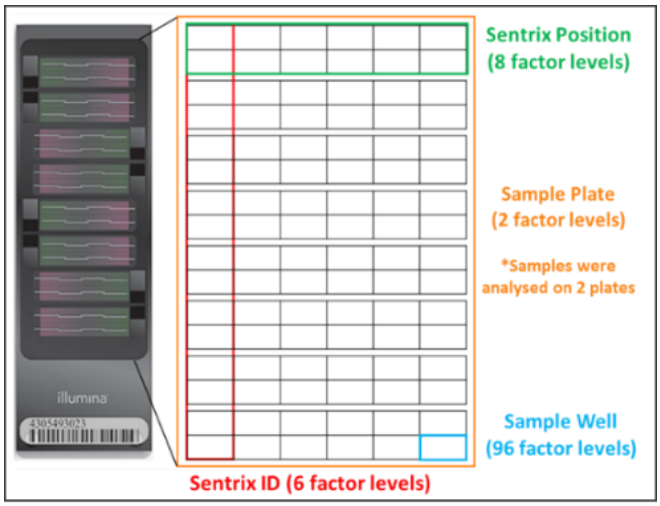


Fig. 3.3 Epic Array Positional Effects
This is an illustration of what different positional effects represent on a microarray.

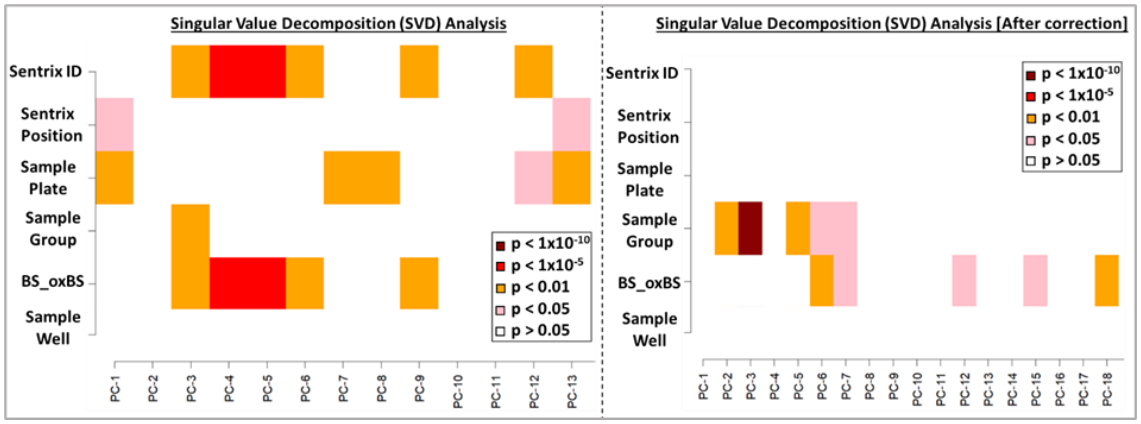


Fig. 3.4 Batch and Positional Effect Correction
These quality control SVD heatmaps generated by the SVA algorithm showcase how each variable contributes to the variance observed in the data. The before/after comparison shows how the batch/positional effects were removed successfully by the champ.runCombat() function.

3.1.4 Constructing Final Matrices

Once pre-processing was finished, the final matrices were constructed by extracting the appropriate data from the champ.runCombat() output datasheet which consists of normalised and corrected β-values from both BS and oxBS. As described earlier in the introduction section, this was straightforward for methylation data as it was just the matter of selecting the correct columns (those from oxBS) from the output sheet. However, for hydroxymethylation data, an extra step to subtract oxBS β-values from BS β-values would be needed for each patient and CpG. Note that for clarification, the methylation (oxBS) data would be referred to as 5mC data or β-values, whereas the hydroxymethylation (BS - oxBS) data would be referred to as 5hmC data or hβ-values from now on. Another important point to note is that it is assumed in this study (and most studies in the literature) that negative hβ-values were related to background signals. Hence, two other versions of the hβ-matrix: "NA_hβ-matrix" and "zero_ hβ-matrix" were generated by converting negative values to "NA" or "0", respectively, accounting for "bumphunter" compatibility issues at later stages.

3.2 Differential Hydroxymethylation Analysis

3.2.1 DhMP

To identify significantly differentially hydroxymethylated CpGs – DhMPs, the "limma"-based champ.DMP() function was then used (with the NA_hβ-matrix). Resembling the principles of well-established differential gene expression analyses, the statistical comparisons underlying DhMP analyses between the 2 data groups (categorical variables: NR or DR_B) are based upon the variable effect size/coefficient, or in simpler terms, the differences between their group average hydroxymethylation levels ($\Delta h\bar{\beta}$-values) (Smyth, 2004; Wettenhall and Smyth, 2004) (Fig. 3.5).

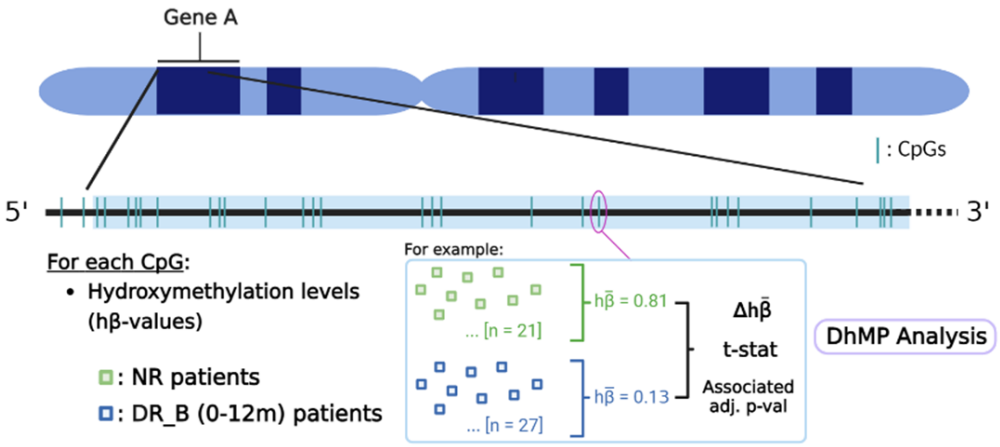
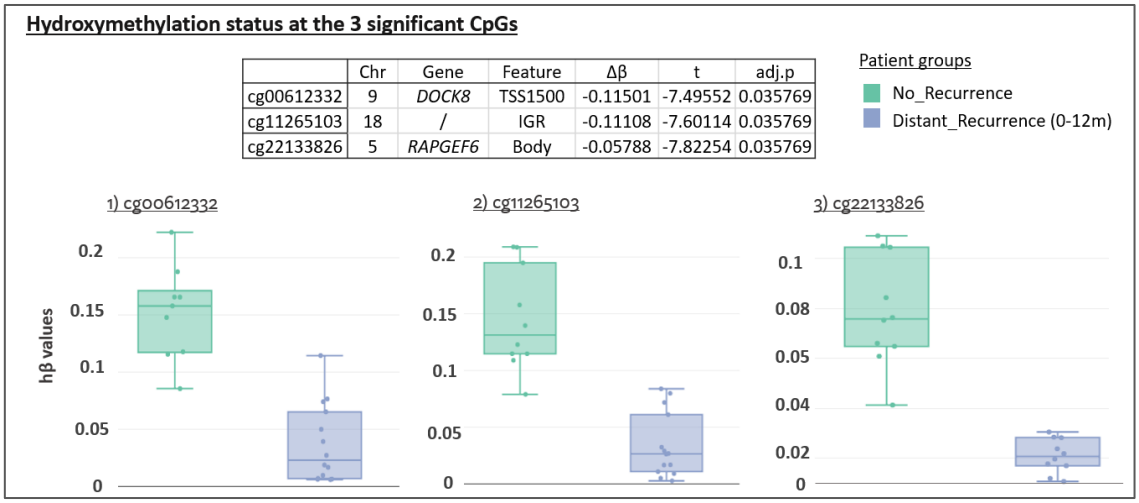


Fig. 3.5 DhMP Analysis
This is a schematic illustration of the statistical inference underlying DhMP analysis. It can be seen that our data would be considered as 2 groups: NR/DR_B (as defined in the pd file), in which t-test would be applied at each probe to see if the average hydroxymethylation level for the 2 groups are significantly different (multi-testing corrected automatically with the Benjamini-Hochberg Procedure)

The results for DhMP are presented in the format of a list of investigated CpGs with their $\Delta h\bar{\beta}$-values the t-test statistic values and corresponding adjusted p-values (multi-testing corrected automatically with the Benjamini-Hochberg Procedure). The complete output can be visualised easily with the DMP.GUI() function that brings up a graphical

user interface (GUI) which also includes different graphs that might be useful for reporting results. An example would be the boxplots showing the hydroxymethylation levels for the 2 patient groups at each significant CpGs.

In our case, the analyses identified 3 significant CpGs (adj. p < 0.05): cg00612332, cg11265103 & cg22133826 in which patients with distant recurrence were all hypo-hydroxymethylated (as displayed by the boxplots within Fig. 3.6). The output also showed the exact location of the CpGs and the genes that they are located within: *DOCK8* for cg00612332, and *RAPGEF6* for cg22133826. Note that the cg11265103 is not associated with any genes as it is located within an intergenic region. The $\Delta h\overline{\beta}$-values, t-statistic values and their associated p-values were also available (see Fig. 3.6).



	Chr	Gene	Feature	Δβ	t	adj.p
cg00612332	9	*DOCK8*	TSS1500	-0.11501	-7.49552	0.035769
cg11265103	18	/	IGR	-0.11108	-7.60114	0.035769
cg22133826	5	*RAPGEF6*	Body	-0.05788	-7.82254	0.035769

Fig. 3.6 DhMP Outputs

Here are the boxplots, corresponding statistical output and annotations for each significant CpG (DhMP) identified. These are all available as part of the DhMP outputs. It can be seen in the boxplots that all 3 CpGs are hypo-hydroxymethylated in DR_B patients.

3.2.2 DhMR

To identify significantly differentially hydroxymethylated regions – DhMRs, the champ.DMR() function was used (with the zero_hβ-matrix), in which the "bumphunter" package approach implemented in ChAMP was selected [method = "Bumphunter"] (Jaffe *et al.*, 2012). Although similar, the underpinning principle is slightly different from "limma", and extra steps were performed automatically as part of the algorithm prior to statistical evaluation. These include transforming hβ-values into hM-values: hM = log2(hβ/1-hβ), calculating the average hM-values for each CpG & each patient group (h$\bar{M}$-values), then calculating the differences between the averages (Δh$\bar{M}$-values) for each probe. The whole DhMR analysis is depicted again, in Fig. 3.7. Note that the logit transformation was performed so that the values used for following statistical analyses would not be bounded between 0 and 1 (Jaffe *et al.*, 2012).

On the other hand, the whole genome would be split into different "clusters" according to customised setting ("maxGap" and "minProbes" arguments within the champ.DMR function). In this case, genomic locations that are located within 300 base pairs would be considered as one cluster and only clusters that contained at least 7 EPIC microarray probes/CpGs would be included for the analyses. Once the clusters were defined, "smoothing" would be carried out (Jaffe *et al.*, 2012).

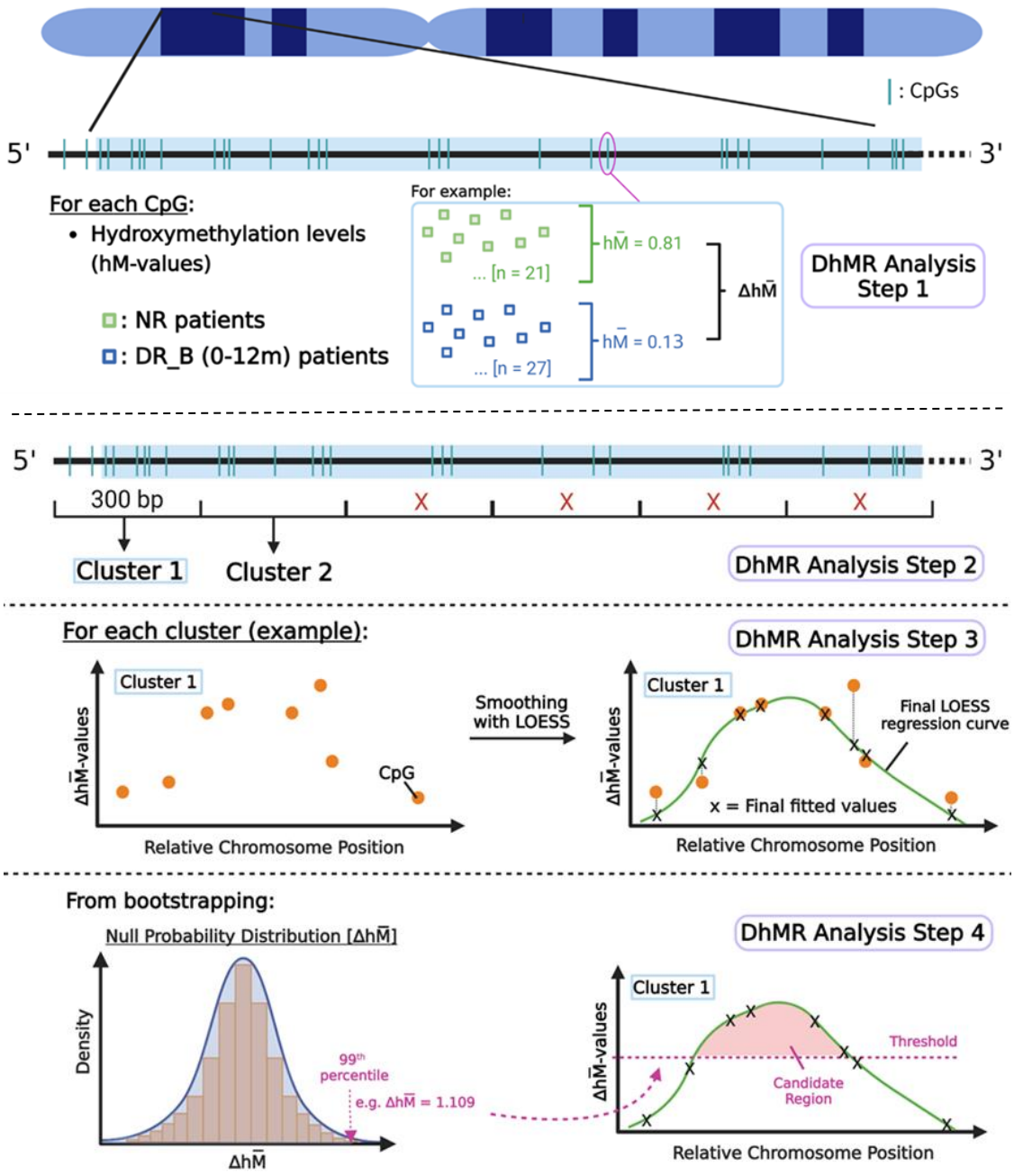
In bumphunter, the smoothing process takes the locally estimated scatterplot smoothing (LOESS) approach – a non-parametric strategy which fits a locally weighted polynomial regression model to the Δh$\bar{M}$-values within a cluster [smooth=TRUE, smoothFunction=loessByCluster] (Jaffe *et al.*, 2012). This smoothing approach is based on "weighted least squares", meaning that instead of fitting the regression curve for the entire cluster in one go, regression fitting would occur in sections, each time

considering a number of neighbouring CpGs (data points) only. Points further away therefore have less weight to the section of interest. Fitting would be repeated multiple times (in the 5' to 3' direction), every time taking the new fitted values as neighbours for the next CpG. Once this was done for the whole cluster, the whole regression fitting would be repeated a number of times to pick up points that consistently had large residuals i.e., outliers, which would be given less weight to consecutive fittings. Finally, all fitted values from the last fitting would be used to produce the final overall loess curve (Jaffe *et al.*, 2012).

Next, bootstrapping would be conducted to estimate the null distribution. Bootstrapping is achieved by resampling from a dataset that is derived from the original data (which had a $\Delta h\bar{M}$-value equal to 0, hence representing the situation when the null hypothesis is true). This is known as sampling with replacement (the same data point can be included more than once) and would be repeated until the final dataset has the sample size as the original dataset (Jaffe *et al.*, 2012; Aryee *et al.*, 2014). The whole bootstrapping process was repeated 250 times so that a null distribution can be estimated and now the system would be able to compute the estimated probability distribution for getting different $\Delta h\bar{M}$-values when null is true i.e., when there is no different in the hydroxymethylation levels between NR and DR_B. It is important to construct the null distribution so that a threshold can be established for the identification of candidate regions. In this case, the $\Delta h\bar{M}$-value at the 99^{th} percentile of the null distribution was selected as a cut-off to detect regions that may have real significance [pickCutoff = 0.99].

Once the candidate regions had been defined, those regions would be compared to the 250 null datasets in regard to 2 parameters: "AVG" which is based on the regional $\Delta h\bar{M}$ average and "AREA" which is based on the area under the regression curve. For AVG,

the significance (p-value) for each region can be estimated based on the same null probability distribution described previously, whereas a null probability distribution specifically for area would have to be constructed for estimating significance for the AREA parameter (using the same null datasets) (Jaffe *et al.*, 2012; Tian *et al.*, 2017).



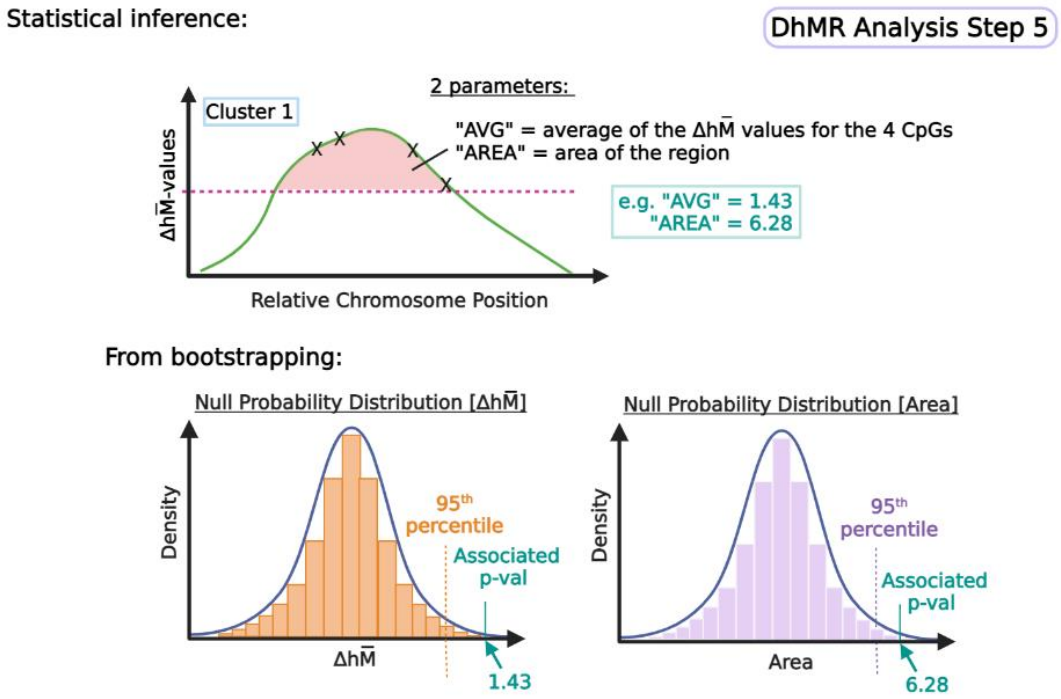


Fig. 3.7 DhMR Analysis

Step 1: Similarly, our data would be considered as 2 groups: NR/DR_B. Then for each probe, the average hydroxymethylation level for the 2 groups were calculated so that the difference between the 2 groups can be computed ($\Delta h\bar{M}$) and used in the following steps for statistical interference.

Step 2: While the $\Delta h\bar{M}$-values were computed for each probe, the system splits the whole genome into clusters (according to predefined settings e.g., > 300 bp).

Step 3: For each cluster, smoothing based on LOESS is applied to correct for outliers ($\Delta h\bar{M}$ values for probes that are not in the same range as its neighbours) in order to obtain the final fitted $\Delta h\bar{M}$-values for each probe.

Step 4: By performing bootstrapping (resampling according to null hypothesis), the null $\Delta h\bar{M}$ distribution can be generated to obtain the threshold, which allows the system to detect "bumps" or in other words, regions within the clusters that are very likely to be significantly hydroxymethylated.

Step 5: Finally, statistical interference based on bootstrapping would be performed to determine if the parameters associated with each region (AVG/AREA) are significantly different from the estimated null parameter values. The associated p-values are simply represented by (1-the percentile), which would then be adjusted automatically for multi-testing.

Similar to DhMP, the results for DhMR analysis are presented in the format of a list of investigated regions with their AVG and AREA values, alongside with their corresponding adjusted p-values. Again, users can access the complete output with the DMR.GUI() function and visualise via a GUI with different graphs. The exact location and length of the DhMRs would also be listed. In this case, 21 DhMRs were identified with their associated adj.p-values < 0.05 for both AVG and AREA. For example, Table 1 contains the top 10 DhMRs ranked according to the associated adj.p-values for AREA. Note that AREA is considered to be the better parameter here as it does not only consider the "height" of the region i.e., the effect size of the phenotype, but also the "length" of the region i.e., the number of CpGs within the region. Hence, the selection of DhMRs for later analyses were based on the AREA parameter.

	Chr	Gene	Region Start	Region End	Region Length	Cluster Length	AVG	adj.p [AVG]	AREA	adj.p [AREA]
DhMR_6	17	FBXO39	6679476	6679541	6 CpGs	10 CpGs	-1.26	0.00397	7.56	0.00674
DhMR_3	1	PDZK1	145727247	1.5E+08	5 CpGs	8 CpGs	-1.49	0.00255	7.44	0.00722
DhMR_2	5	EDIL3	83680253	8.4E+07	4 CpGs	13 CpGs	1.71	0.00186	6.85	0.0102
DhMR_1	7	PRKAR1B/ LOC101926963	641283	641413	3 CpGs	8 CpGs	2.15	0.00055	6.44	0.0127
DhMR_8	1	TWFRSF8	12123105	1.2E+07	4 CpGs	10 CpGs	1.51	0.00619	6.02	0.0165
DhMR_9	12	FBRSL1	133065724	1.3E+08	4 CpGs	14 CpGs	1.46	0.00814	5.84	0.0185
DhMR_10	22	ARFGAP3	43253521	4.3E+07	4 CpGs	10 CpGs	-1.44	0.00911	5.76	0.0194
DhMR_11	4	MSATD1/ C4orf44	3250690	3250745	4 CpGs	8 CpGs	-1.43	0.00976	5.71	0.0201
DhMR_12	20	LOC284798	25129520	2.5E+07	4 CpGs	10 CpGs	1.4	0.0115	5.59	0.0216
DhMR_5	5	GDNF	37835506	3.8E+07	3 CpGs	9 CpGs	1.83	0.00288	5.5	0.0227

Table 1 DhMR Output
This is a simplified table containing the output for the top 10 significantly differentially hydroxymethylated regions (ranked according to AREA adj.p-values).

3.3 Differential Methylation Analyses

3.3.1 DMP

Likewise, the champ.DMP() function was used to identify significantly differentially methylated CpGs/DMPs. Everything mentioned earlier for DhMP analysis still applies except that this time, the β-matrix can be used directly, with the subject matter being the differences between the group average methylation levels ($\Delta\bar{\beta}$-values). In addition, there is a striking difference in the magnitude of the outcome – the DMP analysis identified 35,459 significantly differentially methylated CpGs, in both directions (adj. $p < 0.05$). Shown in Table 2 is the top 10 DMPs ranked according to significance – those with the smallest adj. p-value of 9.7×10^{-6}, then in the order of the absolute effect size (considering both hypermethylation and hypomethylation). Again, note that cg06452234 and cg02566288 are located within an intergenic region, hence not associated with any genes.

	Chr	Gene	Feature	$\Delta\bar{\beta}$	t	adj.p
cg07685512	15	ISLR2	Body	0.189725505	-8.05496785	9.70E-06
cg18918831	3	MUC4	Body	-0.18465276	8.375399722	9.70E-06
cg04461228	5	FLJ42709	Body	0.174152798	-8.51900489	9.70E-06
cg06452234	12	/	IGR	-0.16998391	8.111393858	9.70E-06
cg04546999	1	SPRR1A	TSS200	-0.15170023	8.020353068	9.70E-06
cg02452944	5	PCDHGA4	Body	0.145685446	-8.43713815	9.70E-06
cg02566288	6	/	IGR	0.143030936	-8.14622541	9.70E-06
cg24129977	2	PXDN	TSS1500	0.131800956	-8.01831026	9.70E-06
cg24139737	10	C10orf72	TSS200	0.129507849	-8.18813427	9.70E-06
cg02186782	15	RYR3	TSS1500	0.118305491	-8.1048832	9.70E-06

Table 2 DMP Output
This is a simplified table containing the output for the top 10 significantly differentially methylated probes/CpGs (all with the smallest adj.p-value of 9.7×10^{-6} then ranked according to their AREA adj.p-values).

Note that besides the boxplots for each significant DMPs, the bar plot obtained from the GUI, showing the top 70 genes that are mostly enriched by the 35,459 significant DMPs, is also informative and would serve as a great tool for selecting genes at later stages (Fig. 3.8). Moreover, if necessary, the methylation status of all significant CpGs within a particular gene can be visualised, e.g., *PCDHGA4* consisted of 77 significant CpGs (76 hypermethylated; 1 hypomethylated) (Fig. 3.9).

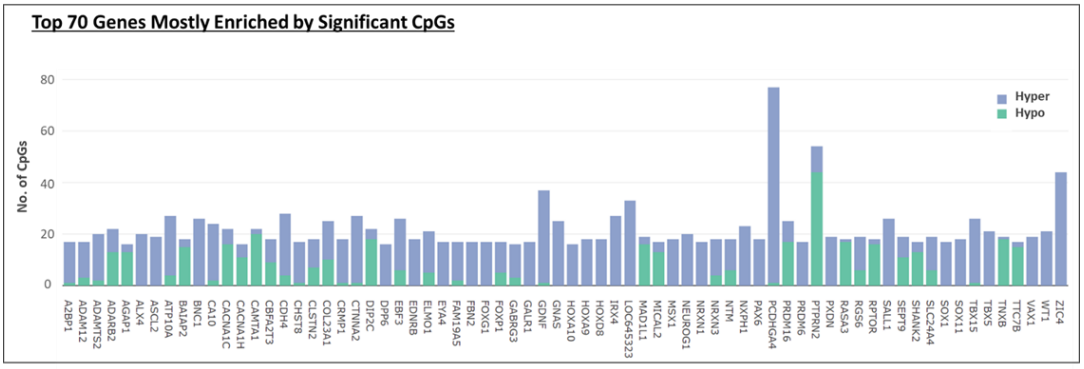


Fig. 3.8 Top 70 Genes Mostly Enriched by Significant CpGs
This bar plot is a useful output of DMP as it highlights the genes that are enriched with significant DMPs, which could be taken as candidates for further analyses like interaction or enrichment analyses. From the plot, the top enriched gene is *PCDHGA4* and the exact number of hypo/hyper-methylated CpGs are available within the interactive GUI: total = 77; hypo = 1; hyper = 76).

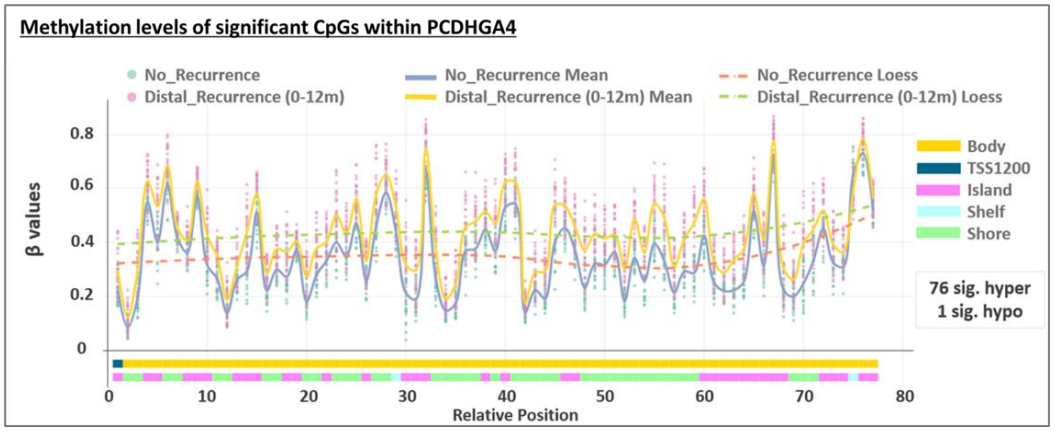


Fig. 3.9 *PCDHGA4* Methylation Status
This is another useful output for users who wish to visualise the methylation status of CpGs within the gene of interest. In this case, *PCDHGA4* is taken as an example and a consistent hypermethylation trend can be seen from the plot.

3.3.2 DMR

Similarly, the workflow for the DMR analysis is the same as that for DhMR analysis except that the M-matrix derived from the β-values would be used; with the subject matter being the differences between the regional group average methylation levels ($\Delta\bar{M}$ -values): AVG and the regional area: AREA. Again, a lot more significantly differentially methylated regions were identified (adj. p [AREA] < 0.05) and Table 3 is showing the top 10 out of the 551 significant DMRs. They were ranked according to their significance associated with the AREA parameter, then in the order of the regional area. It is also notable from the results that all of the top 10 significant DMRs were hypermethylated. The top gene enriched bar plot for DMR analysis is also available in the interactive GUI (Fig. A1). One thing that is key to bear in mind is that despite being part of a significant region, some CpGs might not necessarily be significantly differentially hydroxy/methylation themselves.

	Chr	Gene	Region Start	Region End	Region Length	Cluster Length	AVG	adj.p [AVG]	AREA	adj.p [AREA]
DMR_1	16	SALL1	51183988	51185772	28 CpGs	28 CpGs	0.786401	0	22.01924	0
DMR_2	13	EDNRB	78492216	78493958	35 CpGs	39 CpGs	0.567557	0	19.86451	0
DMR_3	12	/	115134201	115136308	42 CpGs	42 CpGs	0.460742	0	19.35115	0
DMR_4	15	ATP10A	26108391	26109215	18 CpGs	28 CpGs	0.898323	0	16.16982	4.78E-05
DMR_5	3	SPSB4	140769610	140771372	18 CpGs	18 CpGs	0.841993	0	15.15588	7.65E-05
DMR_80	6	EYA4	133561614	133562594	30 CpGs	30 CpGs	0.501182	5.74E-05	15.03547	8.61E-05
DMR_6	5	NEUROG1	134871462	134872315	18 CpGs	22 CpGs	0.756703	0	13.62065	0.000182
DMR_7	7	ELMO1	37488575	37488991	12 CpGs	12 CpGs	1.113253	0	13.35903	0.000191
DMR_8	7	IGFBP3	45960806	45961537	16 CpGs	27 CpGs	0.827402	0	13.23844	0.000191
DMR_9	11	ASCL2	2291652	2292787	19 CpGs	37 CpGs	0.696446	0	13.23247	0.000191

Table 3 DMR Output

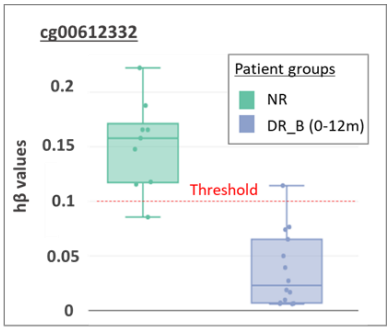
The table at the top of the panel contains the output for the top 10 significantly differentially methylated regions (ranked according to their adj.p-values, then according to their AREA value).

Chapter 4: Scoring

As stated previously, one of the goals of this study is to identify 5mC and 5hmC signatures – candidate CpGs and regions in which their hydroxy/methylation levels can be exploited to distinguish patients with no recurrence and those likely to develop distant recurrence within 12 months after surgery; so that the scoring of such signatures can be used as a non-invasive prognostic tool in the future. Now that we have acquired the list of significantly differentially hydroxy/methylated CpGs and regions, the next step would be to narrow down the list for the scoring of patients, which can then be used to determine the best signature combination via unsupervised hierarchical clustering (chapter 5). This again was performed in RStudio for efficient data handling.

4.1 DhMP

Things were straightforward for DhMP since there are only 3 significant CpGs. As preparation for scoring, a data frame was constructed listing the hβ-values at each specific CpG for each patient. Then thresholds for scoring were determined by sight according to the DhMP boxplots: 0.1 for both cg00612332 and cg11265103, and 0.04 for cg22133826 (see Fig. 4.1). Hence, patients with a hydroxymethylation level lower/higher than the threshold would be scored as 0/1, respectively.

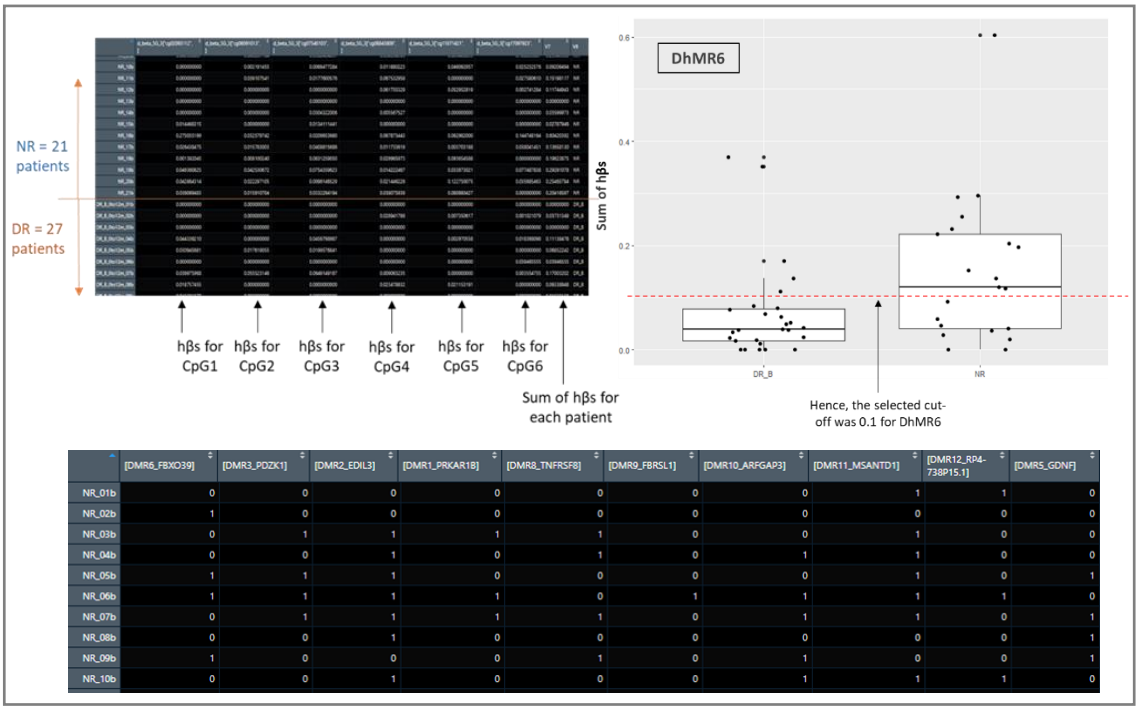


	[cg00612332_DOCK8]	[cg11265103]	[cg22133826_RAPGEF6]
NR_01b	0	1	1
NR_02b	0	0	0
NR_03b	1	1	1
NR_04b	0	0	1
NR_05b	1	1	1
NR_06b	0	0	0
NR_07b	0	0	0
NR_08b	0	1	0
NR_09b	1	0	1
NR_10b	1	1	0
NR_11b	1	0	1

Fig. 4.1 Scoring by DhMP Signature Thresholds
Taking cg00612332 as an example, the panel on the left shows how the threshold for scoring is determined. On the right is a screenshot capturing part of the score table for DhMP, where patients with a 5hmC level below/above the threshold were scored as 0/1, respectively.

4.2 DhMR

On the other hand, while there are 21 significant regions, only the top 10 DhMRs ranked according to AREA-associated significance would be kept for this study as 5hmC/DhMR signatures (those listed previously in Table 1) due to limited timeframe. Similarly, data frames were constructed for each DhMR, each listing the hβ-values for all CpGs within the specific region for each patient. For threshold establishment, the sum of the hβ-values for each patient would be calculated and used for generating the boxplot for the specific region. Then according to the selected threshold, each region would be scored as 0/1 for each patient. For instance, the scoring of DhMR 6 for each patient would depend on the sum of the 6 hβ-values measured for the 6 CpGs located within the region (see Fig. 4.2).



	[DMR6_FBXO39]	[DMR3_PDZK1]	[DMR2_EDIL3]	[DMR1_PRKAR1B]	[DMR8_TNFRSF8]	[DMR9_FBRSL1]	[DMR10_ARFGAP3]	[DMR11_MSANTD1]	[DMR12_RP4-738P15.1]	[DMR5_GDNF]
NR_01b	0	0	0	0	0	0	0	1	1	0
NR_02b	1	0	0	0	0	0	0	0	0	0
NR_03b	0	1	1	1	1	0	0	1	0	0
NR_04b	0	0	1	0	1	0	1	1	0	0
NR_05b	1	1	1	0	0	0	0	1	0	1
NR_06b	1	1	1	1	0	1	1	1	1	0
NR_07b	0	1	1	1	1	0	1	1	0	1
NR_08b	0	0	1	0	0	0	0	0	0	1
NR_09b	1	0	0	0	1	0	1	0	0	1
NR_10b	0	0	1	0	0	0	1	1	1	0

Fig. 4.2 Scoring by DhMR Signature Thresholds
Taking DhMR6 as an example, the 2 panels on the top show how the sum of hβs for each patient can be used as the comparison parameter and plotted for threshold determination. The panel at the bottom is a screenshot capturing part of the score table for all 10 DhMRs, where patients with a 5hmC level below/above the varying thresholds were scored as 0/1, respectively.

4.3 DMP

As mentioned in the chapter 3, much more significant CpGs were detected by the DMP analysis. In order to narrow down the list to a manageable size for further analyses, a number of criteria were set to include as much interesting CpGs as possible. These include: 1) 10 with the highest significance; 2) 10 with the greatest phenotype effect size or in other words, the greatest difference between the 2 patient groups (and with an adj. $p < 0.001$); and 3) CpGs that are associated with genes highlighted by DhMP or DhMR signatures (with adj. $p < 0.01$). Examples would be the 2 significant DMPs that are associated with *RAPGEF6*, the gene that was picked up by the DhMP analysis; and the 29 significant DMPs that are located within the gene *GDNF*, which is also the gene where DhMR 5 resides in. This allowed us to narrow down from 35459 to 64 CpGs. The procedure for scoring was the same as that for scoring DhMP identified significant CpGs – the appropriate data frame was constructed, and the thresholds were determined by studying corresponding boxplots for each CpG.

4.4 DMR

Similarly, the long list of significant DMRs were shortlisted according to 2 criteria: 1) 10 regions with the highest AREA associated significance; and 2) regions that were associated with genes highlighted by DhMP/DhMR signatures. An example would be DMR 279 that is associated with the gene *FBXO39*, which is also where DhMR 6 is located. Note that the DMR ranking was the same when ordered according to the AREA parameter. Threshold establishment and scoring were carried out the same way described for DhMR. In the end, 15/551 of the DMRs were included.

Chapter 5: Testing the Predictive Power of Selected Signatures

Now with the list of scored candidate CpGs and regions, unsupervised hierarchical clustering analysis (HCA) and multiple correspondence analysis (MCA) was performed to check if the framework indeed worked out up till this stage and most importantly, if the 5hmC/5mC signatures can really distinguish patients with distinct prognosis.

5.1 Methodology

5.1.1 Unsupervised HCA

Unsupervised HCA works under the principle that the system try to separate individuals (samples) into a certain number of clusters based on the provided data (CpG/region scores in this case), without knowing which phenotype group the samples are actually from. The algorithm simply groups them according to the dissimilarity in their scoring. There are multiple ways of doing this in RStudio and this time, the "pheatmap" package was used (Raivo, 2019). Note that "binary" and "ward.D2" were selected as the distance matrix generating algorithm and the agglomeration clustering method, respectively (with the number of cluster specified as 2). The results were presented as clustered heatmaps. To start, unsupervised HCAs was performed using all 92 signatures. Then the process was repeated 4 times for the signatures attained from the 4 differential analyses in order to examine the power of each type of analysis in patient stratification.

At the same time, a second package known as "pvclust" was used to visualise the clustering results in the form of dendrogram (Suzuki and Shimodaira, 2006). This also works as validation since the analyses were independent of the results obtained from "pheatmap". Similarly, the function works in 2 steps: computation of distance matrices ("binary" method selected) and agglomeration clustering via the "ward.D2" algorithm.

The strength of this package is that multi-scale bootstrapping can be applied to the input data (nboot=100,000 in this case), so that "Approximately Unbiased" (AU) p-values can be computed for each clustering. Dendrograms annotated with AU p-values can be then generated via the plot() function, followed by the pvrect() function with the "alpha="argument to highlight clusters identified. While it is true that having a high AU p-value means that the particular clustering is strongly supported by the data; dendrograms containing clusters with a relatively low AU p-value does not necessarily mean that the clustering analysis is invalid and that the patients belong to the same cluster. Instead, since a bottom-up clustering algorithm was employed, it actually suggests that the variance within the data can be explained better with more than 2 clusters, for example due to within-group heterogeneity (Suzuki and Shimodaira, 2006).

5.2 MCA

On the other hand, MCA was also used to examine the scoring data. MCA in fact is an adapted correspondence analysis which explores the variance between a group of individuals according to a list of categorical (binary) variables. The "Factoshiny" package which is implemented with key "FactoMineR" functions was used in this case; in which the data would be analysed automatically via the indicator matrix (or complete disjunctive table) approach (Lê, Josse and Husson, 2008; Kostov, Bécue-Bertaut and Husson, 2015). Any underlying patterns can be visualised on an "MCA factor map", where individuals are plotted in a 2 dimensional Euclidean space according to their similarities. The fact that the analysis results can be inspected on an interactive webpage makes this package appealing to users because it is relatively easy to fine-tune any parameters or graphical settings. For instance, it is possible to colour individuals according to specific variables (Fig. 5.1) or plot variables according to their

contributions to the positioning of individuals (Fig. 5.2). Note that the output also includes information for the different dimensions such as eigen-values and the percentage of variance explained by each dimension. The "contribution", "cos2" and "v-test" values which are indicators of the variable's/ individual's contribution to each dimension, how well the dimension represents the variable/individual, and the significance of the contributions (the critical value is usually set as -/+1.96), respectively are also available for further analysis.

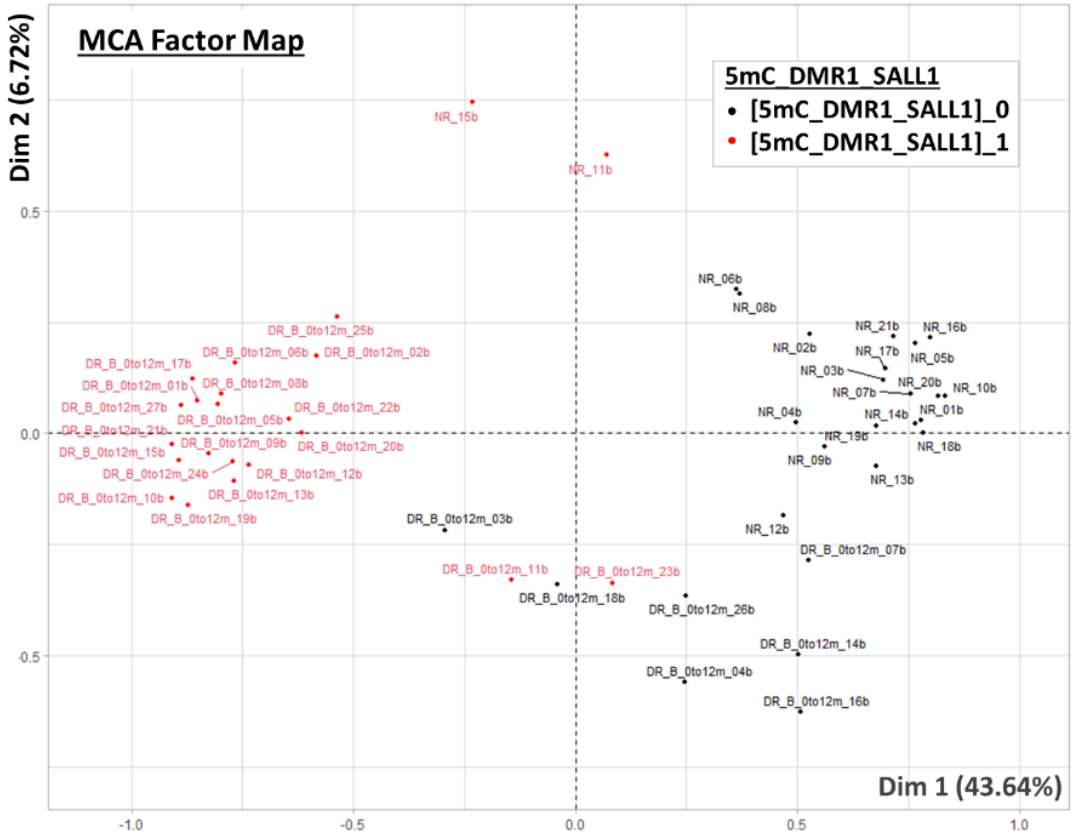


Fig. 5.1 Visualising MCA Results (Example 1)
Users can explore the results on the interactive webpage, e.g., this factor map shows the distribution of patients solely according to their score for DMR1. Looking at the individual labels, we can see that this variable on its own has got the power to distinguish patients – majority of DR_B in red (0) clustered to the middle left; NR in black (1) clustered to the middle right.

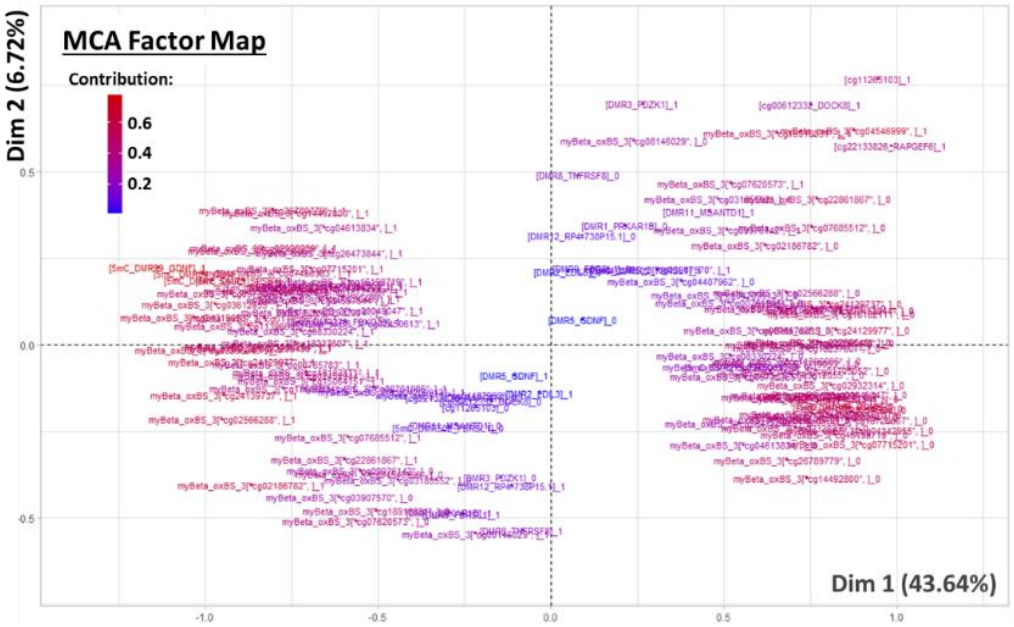


Fig. 5.2 Visualising MCA Results (Example 2)

This is another example of how users can explore the results on the interactive webpage: this factor map is for the visualisation of the strength and relationship between each variable i.e., the scoring for each signature e.g. [5mC_DMR1_SALL1]_0. For example, the entries on the right characterise DR_B patients (when comparing to the factor map for patients – as shown later in Fig. 5.4). They are also coloured according to their contribution towards the positioning of patients. Hence, being coloured in red and positioned away from the centre indicates that they have a greater influence statistically on the overall positioning of individuals.

5.1.3 Optimisation

In an attempt to search for a signature combination that poses an enhanced predictive ability, the signature list was reviewed in sections, for each analysis until 100% accuracy was achieved. It is however important to note that there might in fact be more than one suitable combination, due to each or combinations of variables posing varying degrees of predictive power. Ideally, when all possible combinations can be tested, this would not be a concern. Our limited computational power however restricted our ability to try out the impact of removing every possible combination of signatures as according to the combination formula, there are 92!/91!(92-91)! = 92 possible 91 variable-combinations; 92!/90!(92-90)! = 4186 possible 90 variable- combinations and so on. A

strategy based on strengthening the DMP signature was therefore developed accordingly to identify a combination that does not only involves a high number of variables, but also with an improved predictive ability. Besides these 2 principles, this approach was chosen so that signatures resulting from all 4 differential analyses were included and that the workflow can be handled relatively easily in R.

5.2 Results

5.2.1 All 92 Variables

As expected, including all 92 selected variables for clustering and MCA produced good results. The clustering results from pheatmap and pvclust are the same, with 40/48 patients being classified into the correct group, i.e., 83% accuracy (Fig 5.3/Fig. A2).

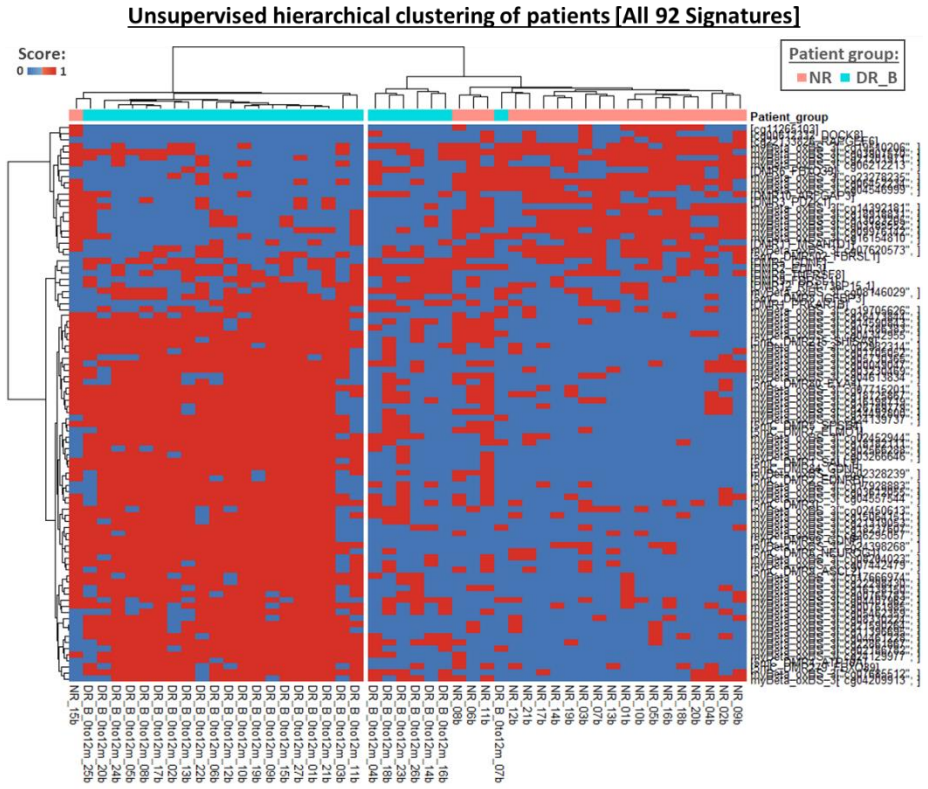


Fig. 5.3 Clustered Heatmap [92 Variables]
This clustered heatmap shows that the patient scorings of the selected 5hmC/5mC signatures altogether (92) had the predictive ability to separating patients into their corresponding prognostic group, with 83% accuracy.

The MCA factor map on the other hand, also showcased that the signatures indeed have the power to stratify patients into distinct groups: patients with no recurrence clustering to the middle left and patients with distant recurrence clustering to the middle right, respectively (Fig. 5.4). However, as suggested by the wrong groupings in the heatmap, and patients plotted in the middle of the MCA plot, there were still room for improvement considering the confidence and accuracy of the stratification.

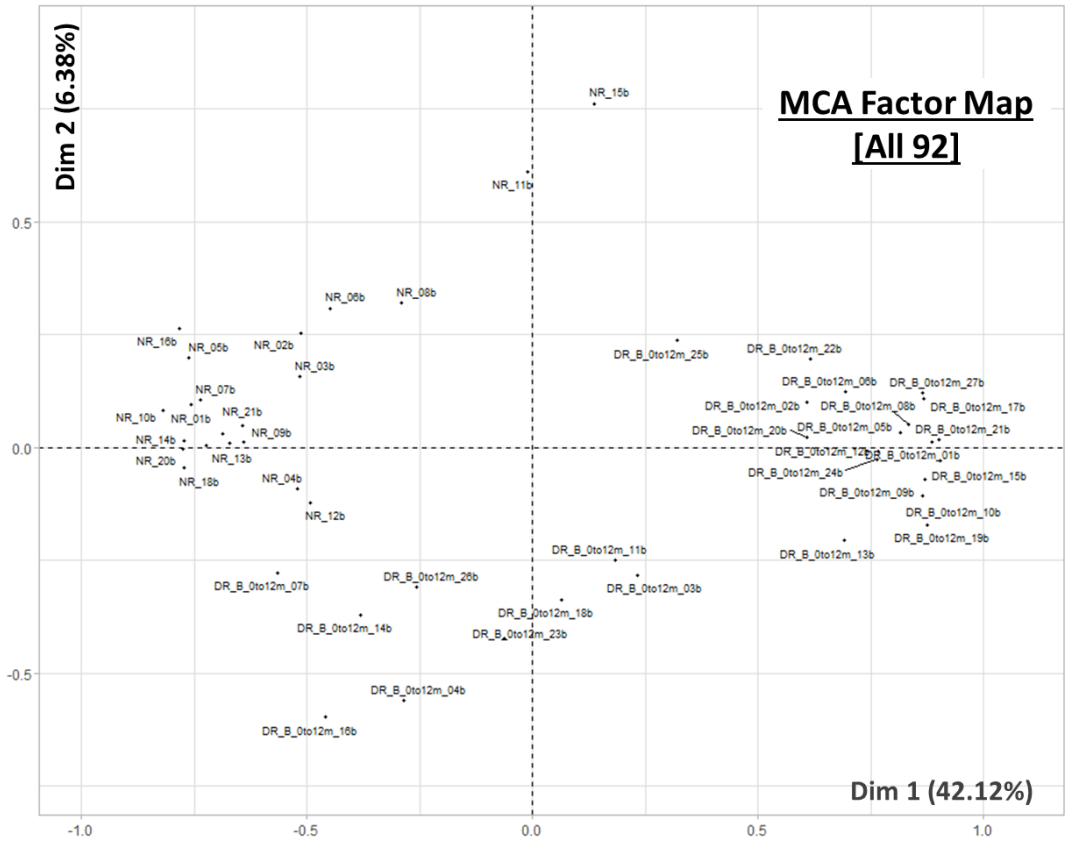


Fig. 5.4 MCA Factor Map [92 Variables]
This MCA factor map displays the overall positioning of individuals in relation to the scoring of all 92 variables of other patients. It can be concluded from the plot that these variables worked relatively well in splitting the patients, with the majority of NR individuals clustering to the middle left and DR_B individuals clustering at the middle right. The individuals locating at the bottom and middle of the plot however suggests that some variables may be contributing negatively towards the stratification of patients, amplifying the impact of outliers.

5.2.2 Individual Analyses for Signatures from the 4 Differential Analyses

On the other hand, the 4 clustered heatmaps (Fig. 5.5) shows that while DhMR and DMR signatures did not achieve high quality stratification - only with ~71% (34/48) and ~77% (37/48) accuracy, respectively; DhMP and DMP signatures can distinguish patients, with ~81% (39/48) and ~96% (46/48) accuracy, respectively. All results from pvclust (Fig. A3) are again the same except a minor difference in the patient order in the DhMP dendrogram since the corresponding distance matrix was directly used as the input data due to compatibility issues. The MCA factor maps also demonstrated matching results, with DhMR and DMR having the worst results as no particular clustering of individuals can be seen and it is impossible to accurately separate individuals into their corresponding groups on both maps (see Fig. A4).

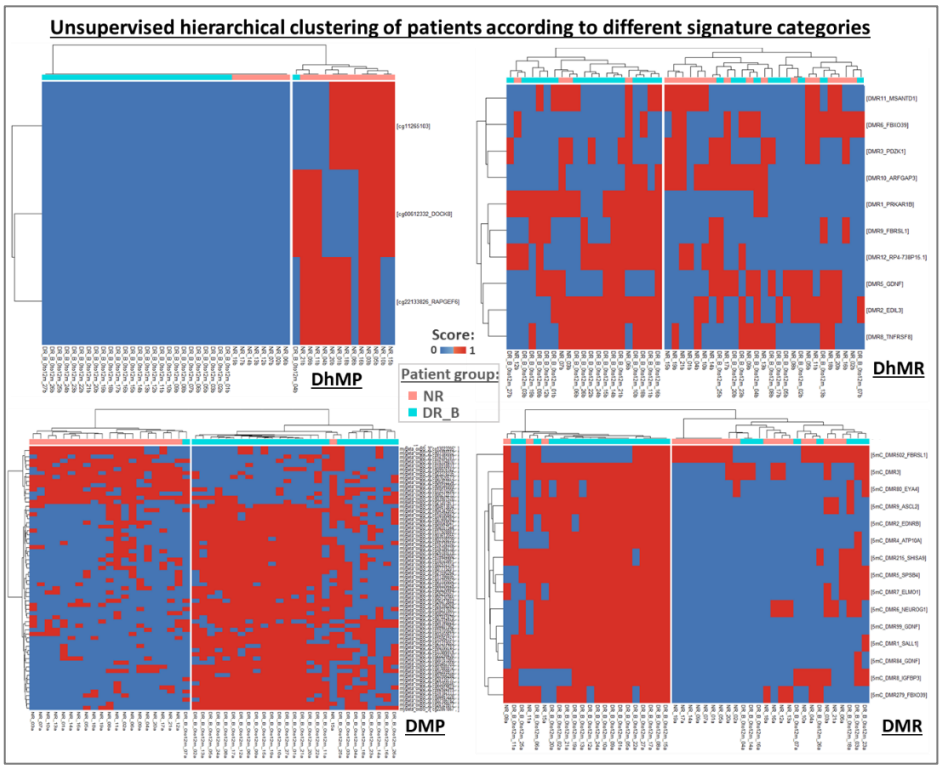


Fig. 5.5 Individual Clustered Heatmaps
These clustered heatmaps display the predictive power of each signature type: DhMP with ~81% (39/48), DhMR with ~71% (34/48), DMP with~96% (46/48) and DMR with ~77% (37/48) accuracy, respectively. It is obvious that DhMP & DMP signatures performed better.

5.2.3 Final Signature List

After numerous rounds of optimisation attempt, the final signature list which can be used to distinguish patients with 100% accuracy contains 68 variables. The ideal clustering results were obtained from unsupervised HCA via both pheatmap (Fig. 5.6) and pvclust (Fig. A5). Besides the fact that most of the individuals were distinguished into the 2 clusters correctly on the MCA factor map (Fig. 5.7), it is also apparent that the clusters are relatively tighter when compared to the map produced with 92 variables.

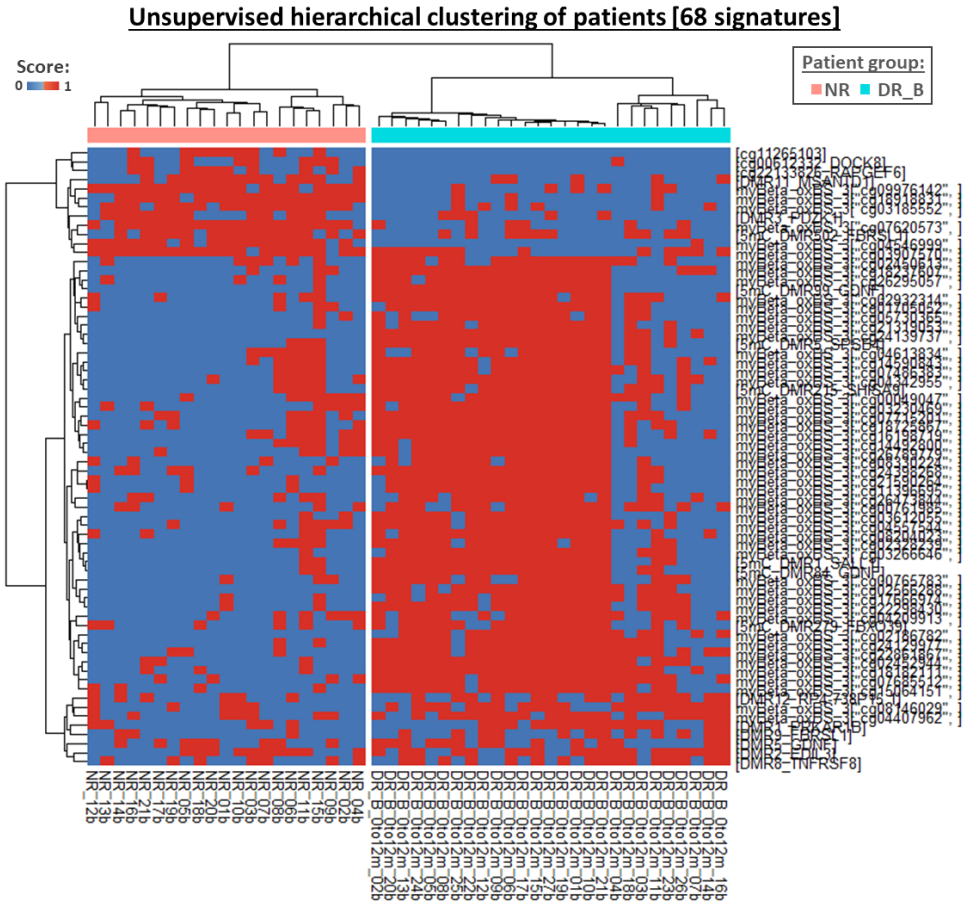


Fig. 5.6 Clustered Heatmap [68 Variables]
The clustered heatmap displayed that the scoring of this combination of 68 signatures was able to stratify patients into their correct groups, with 100% accuracy.

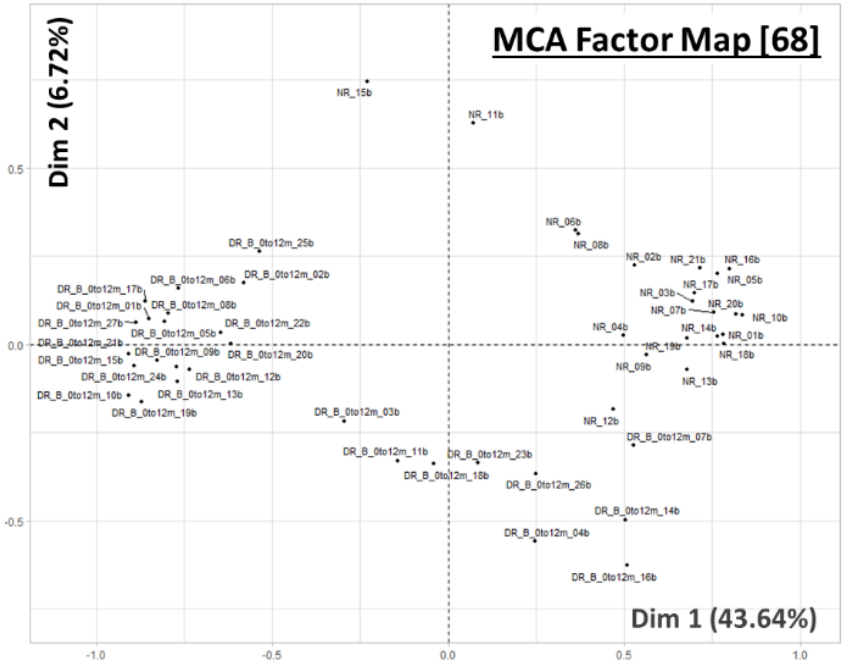


Fig. 5.7 MCA Factor Map [68 Variables]

Similarly, the factor map shows that this combination of signatures has a better predictive ability when compared to Fig. 5.4 as the clusters are tighter. The stratification, however, is still not perfect.

Nevertheless, as mentioned previously, the number of variables is probably not really a key factor in achieving 100% accuracy – in our attempts, signature lists of 11 or 73 variables both achieved perfect stratification. In fact, with sufficient computation power, it might be worth trying to obtain all signature combinations of different lengths and use all those together as the "ultimate test" for patient stratification e.g., a patient has to be classified into the same group for 95% of the combinations. It is also intriguing that HCA and MCA did not produce exactly matching results. This might be in fact due to the impact of heterogeneity in patients within the same group. While this might be resolved by increasing the sample size, it also raises a point – there is actually no need to put too much effort in searching for combinations that gives 100% accuracy at this stage, > 95% accuracy might be sufficient. Instead, cross-validation with multiple data sets should have been the priority once the 5mC/5hmC signatures were identified.

Chapter 6: Preparation for Validation

Now that we have checked the differential analysis output and confirmed that they accomplish their role as signatures for patient stratification, the next step was to test if the signatures retain their power in vitro. One way of doing this is performing RT qPCR to compare the expression of specific genes in different cell lines – here, classical vs squamous PDAC, in which the squamous type is known to be more aggressive and invasive. However, instead of directly using the analyses results, extra work was done to examine the outputs from another perspective, in an attempt to gain more insight into underlying mechanisms that may be contributing to the distinctive phenotypes.

6.1 Candidate Genes: Differential Analyses Output

To begin, the pool of candidate genes consisted of results from previous analyses (see Fig. 6.1): those associated with the final signature list (26 in total), genes that are enriched by all of the significant DMPs (top 10) and genes that are enriched by all of the CpGs within the significant DMRs (top 10). Note that although it was not possible to easily download the list of enriched genes directly as part of the ChAMP output, the enriched genes bar charts available in the DMP/DMR GUIs (shown previously in Fig. 3.8/Fig. A1) are in fact interactive. Hence, detailed information can be obtained by zooming in for genes of interest.

Final Signature list [68]

DMP_cg07685512_ISLR2	DMP_cg03907570_PRKAR1B	DMP_cg11396695_GDNF
DMP_cg18918831_MUC4	DMP_cg04342955_SHISA9	DMP_cg07486383_GDNF
DMP_cg04546999_SPRR1A	DMP_cg01705052_SHISA9	DMP_cg21319053_GDNF
DMP_cg02452944_PCDHGA4	DMP_cg08330224_SHISA9	DMP_cg03230469_GDNF
DMP_cg02566288	DMP_cg04613834_SHISA9	DMP_cg05730365_GDNF
DMP_cg24129977_PXDN	DMP_cg00765783_SHISA9	DMP_cg26473844_GDNF
DMP_cg24139737_C10orf72	DMP_cg04557544_SHISA9	DMP_cg14492800_GDNF
DMP_cg02186782_RYR3	DMP_cg03266646_GDNF	DMP_cg24398268_GDNF
DMP_cg09976142_CIZ1	DMP_cg18182111_GDNF	DMP_cg00761985_GDNF
DMP_cg03185552	DMP_cg03612055_GDNF	DMP_cg18725867_GDNF
DMP_cg07620573_FGF12	DMP_cg21590264_GDNF	DMP_cg26295057_GDNF
DMP_cg17666974_LOC728392	DMP_cg14590843_GDNF	DMP_cg16198719_GDNF
DMP_cg02932314_C17orf104	DMP_cg08204023_GDNF	DMP_cg02450613_GDNF
DMP_cg22861867_UNC79	DMP_cg18237607_GDNF	DMP_cg02328239_GDNF
DMP_cg22298430_LOC728392	DMP_cg04407962_GDNF	DMP_cg26789779_GDNF
DMP_cg08146029_RAPGEF6	DMP_cg00049047_GDNF	DMP_cg04209913_GDNF
DMP_cg15064151_RAPGEF6	DMP_cg07715201_GDNF	

DMR1_SALL1
DMR5_SPSB4
DMR279_FBXO39
DMR502_FBRSL1
DMR84_GDNF
DMR99_GDNF
DMR215_SHISA9

DhMP_cg00612332_DOCK8
DhMP_cg11265103
DhMP_cg22133826_RAPGEF6

DhMR1_PRKAR1B
DhMR2_EDIL3
DhMR3_PDZK1
DhMR5_GDNF
DhMR8_TNFRSF8
DhMR9_FBRSL1
DhMR11_MSANTD1
DhMR12_RP4-738P15.1

[TP] Genes enriched by sig. DMPs

Gene	Hyper	Hypo	Sum
PCDHGA4	76	1	77
PTPRN2	10	44	54
ZIC4	44	/	44
GDNF	36	1	37
LOC645323	33	/	33
CDH4	24	4	28
ATP10A	23	4	27
CTNNA2	26	1	27
IRX4	27	/	27
BNC1	26	/	26
EBF3	20	6	26
SALL1	26	/	26
TBX15	25	1	26
COL23A1	15	10	25
GNAS	25	/	25
PRDM16	8	17	25
CA10	22	2	24

[TR] Genes enriched by CpGs in sig. DMRs

Gene	Hyper	Hypo	Non-sig.	Sum
PCDHGA4	32	/	4	36
EDNRB	18	/	17	35
WT1	15	/	17	32
BNC1	25	/	6	31
EYA4	14	/	16	30
MSX1	17	/	11	28
SALL1	23	/	5	28
ALX4	10	/	16	26
NEU1	10	/	15	25
CA10	21	/	3	24
GALR1	14	/	10	24
FOXG1	16	/	6	22
HOXA5	5	/	17	22
RASGRF1	10	/	12	22

Fig. 6.1 Candidate Genes from Previous Analyses

This figure simply contains lists of candidate genes from previous analyses for validation.
List on top: Genes associated with the final signature combination (In the end, only highlighted 26 genes due to duplication and some DMPs/DhMPs being in IGRs)
List on bottom left: Genes enriched by all of the significant DMPs (those ranked as top 10)
List on bottom right: Genes enriched by all of the significant DMRs (those ranked as top 10)

6.2 Candidate Genes: PPI and Functional Enrichment Analysis

6.2.1 PPI Analysis with STRING

Next, in order to explore the 5mC and 5hmC signatures in the context of pathways that may play a great part in the characterisation of the PDAC subtypes, a protein-protein interaction (PPI) analysis based upon the "STRING" database (version 11.5) was conducted. The STRING webtool can be accessed freely and easily by copy and pasting the list of genes (proteins) into the "Multiple proteins" search box under the "Search" tab (Szklarczyk *et al.*, 2019, 2021). The initial query list consisted of 406 items which are genes associated with the DhMP (2), DhMR (25), DMP (20), DMR (13) results, as well as those enriched by DMP (70) or DMR (276) (using default settings). After the removal of duplicated items and those that were not yet incorporated into the database at the time of study, 306 were left for the analysis and the construction of the PPI network (see Fig. 6.2); with predicted or known physical and functional interactions between proteins being visualised as edges between different labelled nodes. As expected, the analysis report revealed a PPI enrichment p-value of $< 1x10^{-6}$, declaring that our network had a significantly higher number of interactions between our proteins of interest than that expected for a random protein list of the same size (expected no. of edges: 185; observed no. of edges: 520). This serves as another evidence that previous differential analyses produced meaningful results which can be exploited to deepen our understanding of the pathology behind the different subtypes.

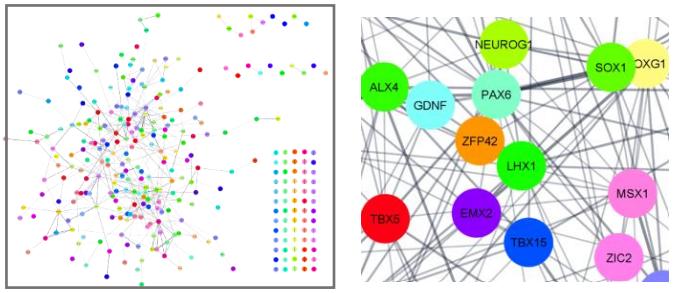
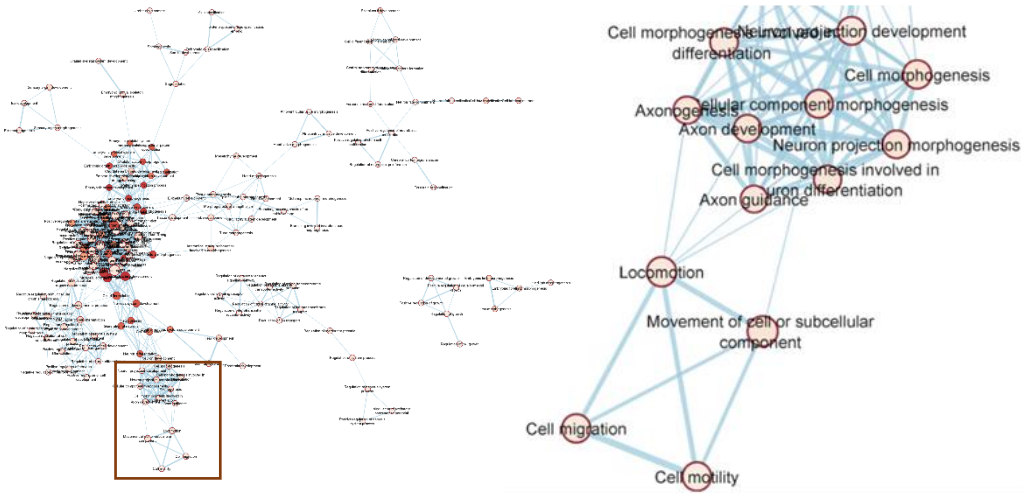


Fig. 6.2 PPI Network
On the left is just a capture of how a PPI network might look like, in which the name of the proteins (genes) can be seen by zooming in.

Actually, the strength of STRING lies within the fact that functional enrichment analyses were carried out automatically as part of the workflow. In this case, significant enrichments (adj. $p < 0.05$) were identified in 13 different database categories such as ones from the well-established Gene Ontology (GO) resource: "Biological Process", "Molecular Function" and "Cellular compartment". While multiple enrichments across different categories are in relation to Homeobox, development and cell fate specification, we decided to focus on the Biological Process enrichments (212 GO_BP terms) with the intention of capturing a broader view of the underpinning mechanisms.

6.2.2 Investigation of GO_BP Functional Enrichments with Cytoscape

In fact, another great feature of STRING is the option to transport the interactive PPI network along with all analysis results and functional enrichment data to the StringApp within Cytoscape (v3.9.1) for easy examination, while all 13 enrichment tables can also be downloaded from STRING/Cytoscape for further analysis (Shannon *et al.*, 2003). While it is worth looking at all terms e.g. via the functional enrichment map on Cytoscape and study possible clustering of pathways, an alternative approach is to classify each term into different themes (15 in this case), including "Cell signalling", "Proliferation", "Differentiation", "Motility/ Cell adhesion", "Development/ Morphogenesis", "Cell polarity/ Axis specification", "Ion transport ", "Molecular function", "Metabolic process", "Transcription", "Multicellular organism process", "Cellular process", "Behaviour", "Neural" and "Systemic biological process" (as shown in Fig. 6.3). Due to limited time, we are only focusing on "Motility/ Cell adhesion" (8 GO_BP terms), in which we believed to be one of the themes that would be specifically important for differentiating our PDAC phenotypes of interest (NR vs DR_B).



#Term ID	Term Description	Theme	Strength	False Discovery Rate
GO:0051414	Response to cortisol	Cell signalling	1.58	0.0142
GO:0008595	Anterior/posterior axis specification, embryo	Cell polarity/ Axis specification	1.2	0.016
GO:0001708	Cell fate specification	Differentiation	0.95	0.0000136
GO:0000578	Embryonic axis specification	Cell polarity/ Axis specification	0.91	0.0337
GO:2000648	Positive regulation of stem cell proliferation	Proliferation	0.87	0.0456
GO:0009952	Anterior/posterior pattern specification	Cell polarity/ Axis specification	0.86	2.52E-10
GO:0035282	Segmentation	Cell polarity/ Axis specification	0.86	0.00016
GO:0021515	Cell differentiation in spinal cord	Differentiation	0.83	0.0268
GO:1900449	Regulation of glutamate receptor signaling pathway	Cell signalling	0.78	0.0388
GO:0003002	Regionalization	Cell polarity/ Axis specification	0.75	2.7E-10
GO:0009953	Dorsal/ventral pattern formation	Cell polarity/ Axis specification	0.73	0.0298
GO:0009798	Axis specification	Cell polarity/ Axis specification	0.71	0.0388
GO:0045165	Cell fate commitment	Differentiation	0.62	0.00053

Fig. 6.3 Exploring Functional Enrichment Results

- Users can examine the functional enrichment results by visually selecting and focusing on particular clusters (consist of pathways positioned closely to one another due to their similar gene-set composition). For example, the panel on the top right is a zoomed in of the red rectangle on the top left panel (GO_BP functional enrichment network), in which it seems that pathways related to motility are positioned close to each other.
- An alternative approach is to look at the complete lists of GO_BP terms (212), then classifying the terms according to the 'biological theme' they are related to. The bottom panel is part of the resulting table.

Now going back to the PPI interaction network on Cytoscape, it is actually possible to remove all edges and colour each labelled nodes in terms of the number of "Motility/ Cell adhesion" GO terms it is associated with (the specific GO terms in the table can be located easily with the term IDs). The result showed that of the 306 genes, 56 of them are associated with 2 or more GO_BP terms i.e., being involved in 2 or more of the predefined gene lists which are specific for each of the terms (see Fig. 6.4). For example, the gene *CDH13* is associated with all 8 of the "Motility/ Cell adhesion"

terms, whereas *CTNNA2* and *ITGA4* are associated with 7. As such, these 56 genes were considered as "core genes" which may have a major role in the development of distinct PDAC phenotypes and were also incorporated into the pool of candidate genes.

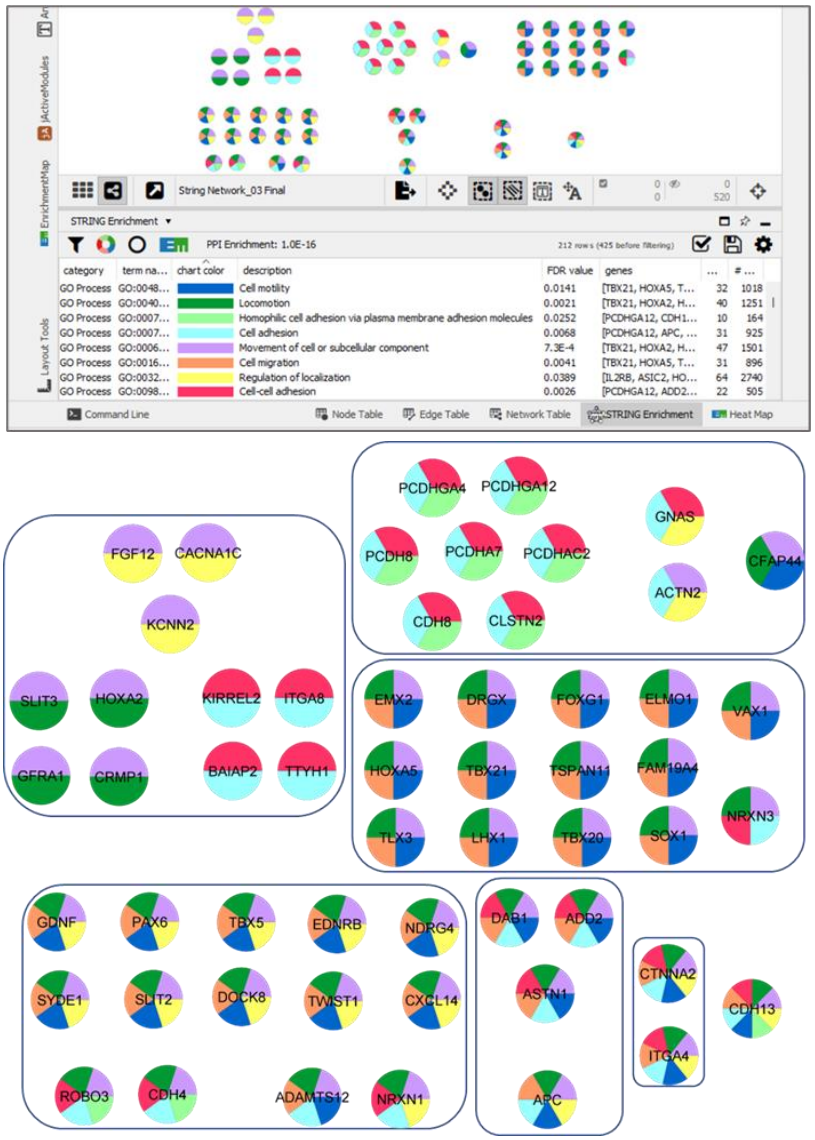


Fig. 6.4 Identification of "Core Genes"

- For the identification of "core genes", the nodes of the edge-less PPI network were coloured according to the number of pathways (GO_BP cell-motility themed terms) they are involved in. The top panel shows how this can be done in Cytoscape, with the operation panel displaying the 56 core genes.
- The bottom panel is a close up of the core genes. Taking *CDH13* as an example, since it is involved in 8 of the terms, the node contains 8 colours.
- As such, this approach allows an easy and focused evaluation of the functional enrichment results.

6.3 Selection of Genes

Finally, a number of factors were considered for the selection of genes for validation. These include: the number of times a gene appearing in analyses outcomes – for instance, *GDNF* was highlighted 4 times by DMP, top 10 genes enriched by significant DMPs (TP), top 10 genes enriched by CpGs of significant DMRs (TR) and functional enrichment core genes (FE); and *PCDHGA4* was highlighted 5 times by DhMR, DMP, DMR, TP and FE, respectively. Whereas some genes were selected since they were ones of the highest significance within their own categories e.g., *ISLR2*, *MUC4*, *SPRR1A*, *PXDN*, *C10orf72* and *RYR3* from *DMP*; *CDH13* and *ITGA4* from FE. On the other hand, primer availability and the depth of understanding for particular genes were also key factors. Lastly, some were picked after discussions with team members, for instance in relation to observations in previous projects or an interest in a particular protein family. In the end, a total of 27 genes were included for RT qPCR (see Fig. 6.5)

ISLR2 [DMP]	ATP10A [TP]	RAPGEF6 [DhMP, DMP]	CDH4 [TP, FE]
MUC4 [DMP]	COL23A1 [TP]	DOCK8 [DhMP, FE]	CTNNA2 [TP, FE]
SPRR1A [DMP]	EBF3 [TP]	SHISA9 [DMP, DMR]	EDNRB [TR, FE]
PXDN [DMP]	EYA4 [TR]	PRKAR1B [DhMR, DMP]	SALL1 [DMR, TP, TR]
C10orf72 [DMP]	NEU1 [TR]	FBRSL1 [DhMR, DMR]	PCDHGA4 [DMP, TP, TR, FE]
RYR3 [DMP]	CDH13 [FE]	BNC1 [TP, TR]	GDNF
FBXO39 [DMR]	ITGA4 [FE]	CA10 [TP, TR]	[DhMR, DMP, DMR, TP, FE]

Fig. 6.5 Final Gene List for RT qPCR
This is the list of genes (27) selected for validation. The terms within the square brackets are simply the types of analyses outcome that the particular gene previously appeared in. *PCDHGA4*, for instance, came up as the top results for 4 analyses: DMP, TP, TR, and FE.

Chapter 7: In vitro Validation – RT qPCR

As described previously, the last stage of the study explores the applicability of the established signatures in vitro and RT qPCR was selected as the approach to investigate if the methylation/hydroxymethylation patterns observed in the Oxford cohort translates to an alteration in transcription that fosters the development of distinctive phenotypes and patient prognosis. 5 cell lines were used, with BxPC-3 and Capan-1 representing the classical PDAC subtype; PANC-1, MIA PaCa-2 and PSN-1 representing the squamous i.e., more invasive and malignant PDAC subtype (according to lab group's previous work, Eyres et al., 2020). Note that Dr. Frances Willenbrock, my mentor made immense contributions to this stage of the project in which she shared her knowledge, experience, and supervised all the way from planning, the purchase and organisation of required materials, the execution of wet lab experiments, to the analysis of outputs.

7.1 Methodology

To begin, cells were harvested from their corresponding culture vessels (3 biological replicates per cell line except for Capan-1 – only 2 samples were available due to slow growth and limited timeframe). The Promega ReliaPrep^{TM} RNA Cell Miniprep System protocol was used for following RNA extraction and purification (Promega, 2012). The reverse transcription master mix was then prepared, and cDNA were synthesised as described in the LunaScript RT SuperMix Kit (E3010) protocol (New England BioLabs, 2018). Finally, the qPCR master mix was prepared for each of the primer pairs (27 genes of interest + 2 house-keeping genes: *GAPDH* and *ACTB*), each of the samples (14 cell line samples + 1 water/negative control), and qPCR was performed with diluted cDNA according to the Luna Universal qPCR Master Mix (M3003) protocol (New

England BioLabs, 2016). The quality of qPCR for each gene was monitored with the resulting amplification plots. Once the first set of results were analysed, extra rounds of qPCR were performed for selected samples, either due to having interesting trends, or no readings across multiple samples e.g., due to a relatively lower base expression level. Note that the list of primer sequences is available in the appendix (Table A1).

7.2 Output Interpretation

The qPCR results were presented as a spreadsheet of cycle thresholds (C_t) and *ACTB* was selected to be the normaliser since its expression was slightly more stable than *GAPDH* across all cell lines (standard deviation of 1.07 vs 1.22). Hence, the 27 gene of interest (GOI) C_t values for each cell line sample would be paired up with *ACTB* C_t values for the corresponding cell line sample (e.g., [Panc-1 sample 1 + *PCDHGA4*] paired with [Panc-1 sample 1 + *ACTB*], [Panc-1 sample 2 + *PCDHGA4*] paired with [Panc-1 sample 2 + *ACTB*] and so on). An adapted delta-delta Ct method would then be applied to calculate the relative fold change of GOI expressions in each cell line. The procedure starts with subtracting *ACTB* C_t values from their paired GOI C_t values to obtain the ΔC_t values. Next, after taking the antilog of the ΔC_t values [$2^{-\Delta Ct}$], the Panc-1 cell line (average of the 3 biological replicates) was used as the baseline (denominator) for the final fold-change calculation. These values were then used to generate boxplots (see to Table A2), illustrating how the relative fold-changes of the expression of GOIs varies in different cell lines. ANOVA and Tukey HSD test were also performed for selected genes to examine the significance in the difference in expression levels between cell line pairs.

Taking *DOCK8* as an example, the $2^{-\Delta Ct}$ values for the two Capan-1 samples (C1/C2) are 0.00209 and 0.00170, respectively; while the $2^{-\Delta Ct}$ average for the three Panc-1 samples is 3.46079×10^{-06}. Hence, there is a ($0.00209/3.46079 \times 10^{-06}$ =) 604- and ($0.00170/3.46079 \times 10^{-06}$ =) 492-fold change in the relative *DOCK8* expression levels in C1 and C2, respectively. Interestingly, similar/opposing results was observed in the rest of the cell lines and this trend is also visible in the *DOCK8* boxplot. It can therefore be concluded that the qPCR data suggests that *DOCK8* expression is substantially down-regulated in squamous PDAC cell lines.

7.3 Results

As such, calculations were done for each sample and the boxplot for each gene was examined. In the end, 6 out of the 27 genes demonstrated interesting trends, in both rounds of qPCR. These include: *PCDHGA4*, *ISLR2* & *RAPGEF6*, which were up-regulated in squamous cell lines; and *DOCK8*, *SPRR1A* & *ITGA4* which were down-regulated in squamous cell lines (as shown in Table A3, all with ANOVA F stat-associated p-values smaller than 0.05). It can be seen from the boxplots of the up-regulated genes (Fig. 7.1) that there is an explicit trend; whereas the trends for down-regulated genes is only obvious in *DOCK8* (Fig.7.2). In fact, the magnitude of change was especially large for *DOCK8* in all 4 Capan-1 samples (both rounds of qPCR) when compared to PANC-1 samples – the values for Capan-1 had to be rescaled (divided by 100) so that the trend within other cell lines can also be illustrated under the same scale. While the trends are not as obvious for *ITGA4* and *SPRR1A*, the fact that a large magnitude of expression difference can be seen in BxPC-3 samples (an average of around 8-fold and 65-fold increase, respectively), still makes them intriguing candidates for further analyses.

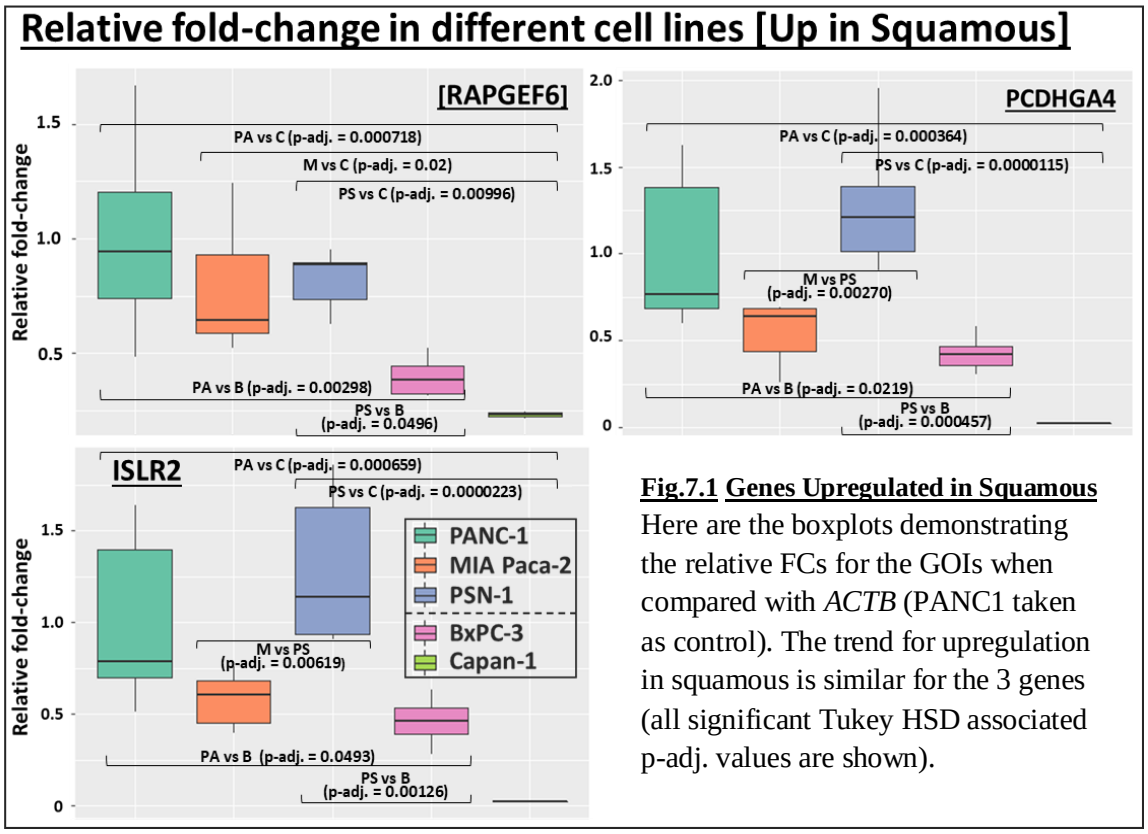


Fig.7.1 Genes Upregulated in Squamous
Here are the boxplots demonstrating the relative FCs for the GOIs when compared with *ACTB* (PANC1 taken as control). The trend for upregulation in squamous is similar for the 3 genes (all significant Tukey HSD associated p-adj. values are shown).

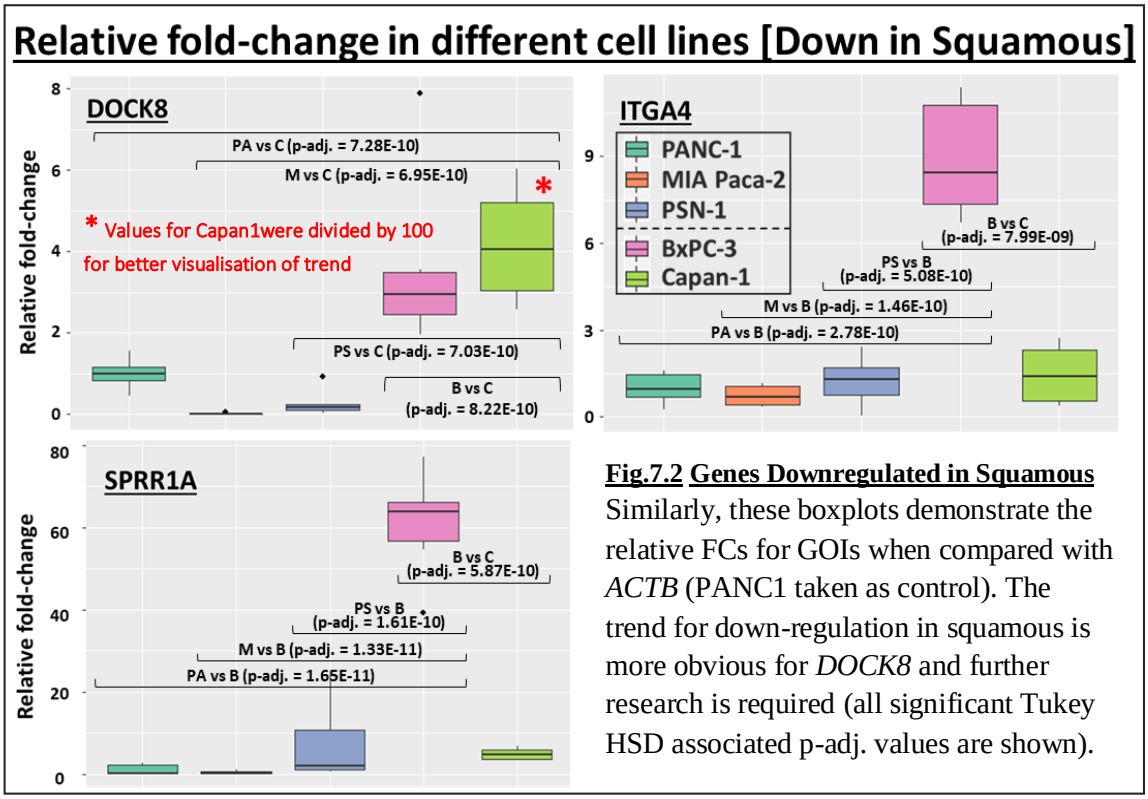


Fig.7.2 Genes Downregulated in Squamous
Similarly, these boxplots demonstrate the relative FCs for GOIs when compared with *ACTB* (PANC1 taken as control). The trend for down-regulation in squamous is more obvious for *DOCK8* and further research is required (all significant Tukey HSD associated p-adj. values are shown).

Chapter 8: Discussion

8.1 Project Review

This project started with the objective to determine if pateients with different prognostic outcomes (no recurrence vs distant recurrence within 12 months) have varying hydroxy/methylation profiles; and if the signatures pose the power to achieve high quality unsupervised patient stratification, hence carrying the potential to be candidates of biomarkers for improving prognostic accuracy in localised PDAC patients.

8.1.1 Challenges

Before focusing on individual analyses, it would be good to look into some of the limitations of the study (besides those mentioned in previous chapters). Despite efforts to follow widely accepted methodologies and algorihtms for data processing, and designing the workflow via the intergration of rational knowledge, there were still uncertainty regarding the feasibility of the study (especially with respect to the overall statistical signnificance). For example, one of the greatest challenge was to establish a framework that suits our data and addresses the objectives at the same time. With oncoepigenetic studies still being a rather new area of research, there were only a limited number of studies available for reference. While the procedures for studying differentially methylated CpGs and regions were relatively well-defined, the study of differential hydroxymethylation is still in its preliminary phase and it is widely agreed that the standard approach needs to be reviewed. As described in Slynko & Benner's paper (2019), one of the problems lies within the ambiguity in directly taking the differnces between BS and oxBS β-values (hβs) as the measurement for 5hmC levels. This approach is based on the equation:

$$h\beta = \beta BS - \beta oxBS = (M_{BS} / M_{BS} + U_{BS} + 100) - (M_{oxBS} / M_{oxBS} + U_{oxBS} + 100)$$

with the common interpreataion of $h\beta > 0$ indicating a notable level of 5hmC; $h\beta = 0$ indicating absense of 5hmC; and $h\beta < 0$ being considered as background noise. This interpretation however might be problematic as in some rare cases the methylation and unmethylation intensity measurements in BS (M_{BS}; U_{BS}/ M_{oxBS}; U_{oxBS}) may work out to give false negative or false positive results. In addition, considering the reduced dynamic range of 5hmC hβ-values, this calculaltion is also vulnerable to background signals and measuring errors. Although it is beyond the scope of this study to deal with this matter, it is a good idea to keep in mind that while this is now the most common approach for estimating hydroxymethylation levels, a technology that allows direct quantification (with high resolution) is preferred and should be used instead when one becomes available at a reasonable cost.

On the other hand, the novel interpretation of functional enrichments results for the identification of core genes again brings uncertainty to the project and further study with other data sets would be required to explore the validity and significance of the results produced via this approach. The fact that there are numerous argument setting options for the various bioinformatics algorithms throughout the entire workflow e.g., determining thresholds, selecting clustering methods, deciding how far to go for batch correction and so on; also increases the chances of introducing errors to the study.

Nevertheless, one of the most important conclusions from the project is that after all, the workflow indeed served its purpose. Whilst extra work, such as analysing the data with another tool/programme and cross validation with another cohort, should be conducted to address the uncertanties described above, the fact that the application of outputs from

the differential hydroxy/methylation analyses to PPI (the following functional enrichment) and RT qPCR managed to highlight significant mechanisms within the context of interset (cancer and metastasis-related) and yield patterns that make biological sense in vitro; is strong evidence that the experiment design worked fine practically. It is also arguably not as crucial at this point to focus too much on the details regarding application settings as long as the signatures function well as prognostic biomarkers. It would in fact be intersting to do follow-up studies to determine how different strategies might reach different conlcusions, how signifcant are those differences, and if it would later be possible to develop a system to accumulate and decipher useful information from multiple analyses in order to tackle scientific queations from various perspectives.

8.1.2 Result Summary

From the 2 sets of differential hydroxy/methylation analyses, it can be concluded that patients who did not suffer from a relapse indeed have a DNA hydroxy/methylation profile that are significantly different to that of patients who recurred within 12 months after surgery, i.e., the group of patients that might benefit more from other treatment plans. Both in silico (HCA & MCA) and in vitro validation (RT qPCR) then showcased the signatures' ability to distinguish samples of varying subtypes, with *PCDHGA4*, *ISLR2* & *RAPGEF6* being overexpressed and *DOCK8*, *SPRR1A* & *ITGA4* being underexpressed in vitro, respectively.

8.2 Limitations

While these are promising results that would foster the application of personalised interventions in PDAC, some concepts should be considered and reviewed eventually as

good scientific pratice. The first one is the fact that cross validation with another cohort was not yet performed due to the lack of freely accessible PDAC patient cohort data that does not only consist of its complete clinical data and DNA methylation data, but also its DNA hydroxymethylation data. As a compromise, it might be acceptable to randomly split the Oxford Kidani cohort into 2 sets of data (half of the patients from NR, and half from DR together as one set) so that one can be used for training (differential hydroxy/methylation analyses) and the second set can be used for validation (unsupervised HCA and MCA). The downside of this is that the sample size would be even smaller, which also means that the heterogenity of PDAC might have an amplified effect in the results.

The second complication is that these samples originated from biopsies collected during surgery. It is therefore irrational to simply assume that patients would be at the same state, with the same epigenetic profile during the earlier stages of their medical care, especially at the point when the doctors are constructing their treatment plan. For these signatures to be used as prognostic tools, research should to be done to compare the DNA hydroxy/methylation profiles of percutaneous biopsy samples collected at an earlier timepoint before the doctor made a decision on the type of cancer treatment to go for; with the profiles of samples collected after surgery. Further studies (e.g. with organoids or Ex vivo culture of pancreatic tumor slices) and clinical trials can be conducted to see if these predictions are accurate, with follow-up studies to see if adapting the treatment plan accordingly brought any significant improvements to the prognosis and the quality of life during or after treatment, which is difficult to measure but a crucial factor for patients to decide whether or not they should go for surgery.

8.3 Implications and Future work

8.3.1 Overlap with Classical/Squamous Subtype

The results of this project suggest that these 5hmC/5mC signatures have the potential to be used for the stratification of patients and may play a role in determing the course of treatments for localised PDAC patients. The fact that signatures arised from differentiating patients according to their prognostic outcome can be used to distinguish squamous PDAC cell lines from classical ones also hints that there might be an overlap of signatures, hence a matching stratification: the tumours of patients with no recurrence corresponding to the PDAC classical subtype; while the tumours of patients with distant recurrence within 12 months corresponding to the PDAC squamous subtype. Further studies would need to be done to clarify this but if this turned out to be the case, research performed for classical vs squamous studies would also be useful for the understanding of the drastic differences in prognostic outcomes observed.

8.3.2 RT qPCR Follow-up

On the other hand, while the first goal of this project is to capture 5hmC/5mC signatures for prognostic purposes, it would also be a good idea to spend some time on the 6 genes that were found to be up-regulated or down-regulated when comparing classical and squmous cell lines (regretfully, only 2 were investigated in detail due to limited timeframe). This would be valuable as they might be contributing to mechanisms that foster metastasis or the transformation of classical PDAC tumours into squamous forms, hence deepening our understanding of the pathology of PDAC.

Taking *DOCK8* (Dedicator Of Cytokinesis 8) as an example, although it is not known to be prognostic for PDAC, research has proved that *DOCK8* is a marker for renal and urothelial cancer, in which a high expression is generally associated with a better outcome. Numerous studies also demonstrated that DOCK8 deficiency e.g., due to loss-of-function mutation, causes immune defects – DOCK8 immunodeficiency syndrome, hence a susceptibility to conditions such as atopic dermatits, severe allergies and malignant tumours (Zhang *et al.*, 2022). Intriguingly, our data indicated that *DOCK8* was significantly hypo-hydroxymethylated in patients with distant recurrence and downregulated in squamous cell lines. This result does not only make biological sense since hypo-hydroxymethylation is associated with reduced transcription activation and as expected, a diminished expression; but most importantly, our hydroxymethylation/ transcriptional trend for *DOCK8* observed in patients with varying prognosis is indeed in concordence with those described in the literature.

Indeed, DOCK8 was known to play a pivitol role in the activity, migration or survival of various immune populations including T cells, B cells, NKT cells, macrophages and dendritic cells (Biggs, Keles and Chatila, 2017). As a member of the guanine nucleotide exchange factors (GEF) family, upon receptor-activation, DOCK8 employs parts of its fuctionality via the specific binding, hence activation of CDC42 (Rho GTPase) and downstream WASP/WIP-mediated signalling (Zhang *et al.*, 2022). Previous studies linked DOCK8 deficiency to a disruption in cytoskeleton reconstruction (actin polymerisation), cell adhesion/migration mechanisms, phagocytosis and immunological synaptic interactions (Blundell *et al.*, 2010; McGhee and Chatila, 2010; Randall *et al.*, 2011). Randall's research (2011) also showed that cytotoxic T cells of patients with DOCK8 deficiency have impaired effector functions with most of them in the state of

exhaustion, which intriguingly is one of the major features in advanced PDAC, as described in the introduction section. On top of that, while it might be challenging to investigate, it would be interesting to consider whether the relative *DOCK8* downregulation is a genetic disposition specific to certain patients groups or a pattern only happening at the tumour microenvironment, after or during a specific stage in PDAC progression. It would also be crucial to check if a coherent alteration is observed in DOCK8 protein expression and activities, while doing further research to study how such findings would fit into the bigger picture in PDAC tumorigenesis.

Similar evaluations can also be done for *RAPGEF6*, which is another gene containing 1 significantly hypo-hydroxymethylated and 2 hypermethylated CpGs. The situation for *RAPGEF6* however is not as straightforward since an upregulation was observed in the squamous cell lines. This might seem to be a contradicting result at first as hypo-hydroxyrmethylation and hypermethylation of a gene "should" both be associated with a less-activated gene, hence a decrease in expression. This result however reminded us of the fact that such assumption is rather naïve as differential hydroxymethylation or methylation of a single site within a gene might not necessarily bring a direct effect on the downstream transcription, which is also a reason why regional differential hydroxy/methylation analysis might sometimes be more relevant in the sense of direct correlation with transcriptional alterations. With that being said, it might also be too reckless to cross out all results obtained from single site analyses, especially with our results showing potential significance of *DOCK8*, as well as numerous previous studies proving the validity of such approach, particularly in differential methylome analyses.

Furthermore, the upregulated pattern in cancer cells with a more aggressive phenotype observed in vitro is in fact in line with previous studies which focuses on *RAPGEF6*'s

role in cell motility, where mutational and differential expression analysis defined *RAPGEF6*, also known as *PDZ-GEF2* and *RAGEF2*, as an oncogene for mesothelioma and triple-negative breast neoplasm (Jung *et al.*, 2017; Torricelli *et al.*, 2020). Most importantly, Huang and his group's work (2012) revealed RAPGEF6's potential significance in PDAC aggressiveness and progression towards metastasis. As the specific GEF for RAP1 and RAP2, RAPGEF6 facilitates RAP1/2-GTP loading e.g., upon EGFR stimulation and activates downstream oncogenic mechanisms such as proliferation, EMT, cytoskeleton rearrangement and invasion (Huang *et al.*, 2015; Yi-Lei *et al.*, 2017). Indeed, the knockdown of *RAPGEF6* in FG (metastatic) cell line blocked EGFR-induced migration (Huang *et al.*, 2012). The research by Dubé et al., (2008) also shed light on *RAPGEF6*'s necessity in the formation of cell adheren junctions and the control of E-cadherin levels in the A549 lung cancer cell line via RAP1 activation. Interestingly, work by Qian et al. (2022) on the other hand described how different KRAS mutations e.g., M67L might enhance RAPGEF6 activities (pre-print). Nevertheless, the way how RAPGEF6 exactly functions in cancer is still very poorly understood and a lot more research would be required to get through to actual practical applications.

In fact, the intriguing results obtained for *RAPGEF6* brought up an interesting question for further investigation: would it be possible that the alterations in hydroxy/methylation levels at particular sites (on its own or in combinations) be able to cause critical changes in the chromatin structure which can lead to the oncogenic activation/silencing of downstream target genes. This also proposes the significance in pairing epigenetic analyses with chromatin-related studies, as well as studies involving different types of epigenetic mechanisms (e.g., DNA hydroxy/methylation, histone modifications and

noncoding RNAs), since they might provide insight into the effects of specific types of epigenetic alterations have on the chromatin organisation, and other epigenetic markers.

8.4 Conclusion

In conclusion, this preliminary study achieved its objective to propose a user-freiendly framework (including the adaptation of well-established procedures) which successfully highlighted promising results – the existance of a significant differential hydroxy/methylaiton profiles for patients of varying prognostic outcomes (no recurrence vs distant recurrence after 12 months). After further validations, these signatures may serve directly as prognostic biomarkers via pateint scoring and the application of various clustering techniques (like HCA and MCA). These signatures may also serve as phenotype-specific candidates for targeted-drug discovery, for instance via the integration of multiple bioinformatics and machine learning tools. Most importantly, as shown in the later part of the project, it was exciting that results generated from the differential analyses can be exploited for protein-protein interaction analysis and gene set enrichment analyses, just like conventional transcriptomic analyses outputs, which facilitates an even deeper understanding of the biological processes underlying such opposing clinical outcomes and PDAC tumorigenesis in general. The fact that the themes of the extending research can be fine-tuned by selecting gene sets of the specific topic such as apoptosis, metabolic reprograming and cell motility in this case, also brings great flexibility to the study. Hence, this framework have the power to function as a robust tool with remarkable labour-efficiency and shed light on interesting and intertwining concepts underlying different phenotypes of interest, bringing more insight into the complex biology PDAC via diverse perspectives.

As such, with further research on the establishment of appropriate frameworks integrating multiple data types (mutational, epigenetic, transcriptional, proteomic and metabolomic), it might be beneficial to eventually include bioinformatics-based multi-omics studies as one of the gold-standard basic research prior to the conduction of clinical trials for PDAC and all types of cancer or conditions.

Word Count: 13,639

Bibliography

Aryee, M.J. *et al.* (2014) 'Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays', *Bioinformatics*, 30(10), pp. 1363–1369. Available at: https://doi.org/10.1093/bioinformatics/btu049.

Bailey, P. *et al*. (2016) 'Genomic analyses identify molecular subtypes of pancreatic cancer', *Nature*, 531(7592), pp. 47–52. Available at: https://doi.org/10.1038/nature16965.

Biggs, C.M., Keles, S. and Chatila, T.A. (2017) 'DOCK8 deficiency: Insights into pathophysiology, clinical features and management', *Clinical immunology (Orlando, Fla.)*. 2017/06/15, 181, pp. 75–82. Available at: https://doi.org/10.1016/j.clim.2017.06.003.

Blundell, M.P. *et al.* (2010) 'The Wiskott-Aldrich Syndrome: The Actin Cytoskeleton and Immune Cell Function', *Disease Markers*, 29, p. 781523. Available at: https://doi.org/10.3233/DMA-2010-0735.

Brabson, J.P. *et al.* (2021) 'Epigenetic Regulation of Genomic Stability by Vitamin C', *Frontiers in Genetics*, 12. Available at: https://doi.org/10.3389/fgene.2021.675780.

Bulle, A. and Lim, K.-H. (2020) 'Beyond just a tight fortress: contribution of stroma to epithelial-mesenchymal transition in pancreatic cancer', *Signal Transduction and Targeted Therapy*, 5(1), p. 249. Available at: https://doi.org/10.1038/s41392-020-00341-1.

Ciernikova, S. *et al*. (2020) 'Epigenetic Landscape in Pancreatic Ductal Adenocarcinoma: On the Way to Overcoming Drug Resistance?', *International journal of molecular sciences*, 21(11), p. 4091. Available at: https://doi.org/10.3390/ijms21114091.

Collisson, E.A. *et al*. (2011) 'Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy', *Nature Medicine*, 17(4), pp. 500–503. Available at: https://doi.org/10.1038/nm.2344.

Daemen, A. *et al.* (2015) 'Metabolite profiling stratifies pancreatic ductal adenocarcinomas into subtypes with distinct sensitivities to metabolic inhibitors', *Proceedings of the National Academy of Sciences*, 112(32), pp. E4410–E4417. Available at: https://doi.org/10.1073/pnas.1501605112.

Daniel, S.K. *et al*. (2019) 'Hypoxia as a barrier to immunotherapy in pancreatic adenocarcinoma', *Clinical and translational medicine*, 8(1), p. 10. Available at: https://doi.org/10.1186/s40169-019-0226-9.

Dedeurwaerder, S. *et al.* (2011) 'Evaluation of the Infinium Methylation 450K technology', *Epigenomics*, 3(6), pp. 771–784. Available at: https://doi.org/10.2217/epi.11.105.

Dell'Aquila, E. *et al.* (2020) 'Prognostic and predictive factors in pancreatic cancer', *Oncotarget*, 11(10), pp. 924–941. Available at: https://doi.org/10.18632/oncotarget.27518.

Deng, D. *et al.* (2022) 'Role of the Tumor Microenvironment in Regulating Pancreatic Cancer Therapy Resistance', *Cells*. MDPI. Available at: https://doi.org/10.3390/cells11192952.

Dubé, N. *et al.* (2008) 'The RapGEF PDZ-GEF2 is required for maturation of cell–cell junctions', *Cellular Signalling*, 20(9), pp. 1608–1615. Available at: https://doi.org/https://doi.org/10.1016/j.cellsig.2008.05.006.

Ene-Obong, A. *et al.* (2013) 'Activated pancreatic stellate cells sequester CD8+ T cells to reduce their infiltration of the juxtatumoral compartment of pancreatic ductal adenocarcinoma', *Gastroenterology*. 2013/07/25, 145(5), pp. 1121–1132. Available at: https://doi.org/10.1053/j.gastro.2013.07.025.

Eyres, M. *et al.* (2020) 'TET2 drives 5hmc marking of GATA6 and epigenetically defines pancreatic ductal adenocarcinoma transcriptional subtypes', *bioRxiv*, p. 2020.10.22.342436. Available at: https://doi.org/10.1101/2020.10.22.342436.

Fortin, J.-P., Triche, T.J. and Hansen, K.D. (2016) 'Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi', *Bioinformatics*, p. btw691. Available at: https://doi.org/10.1093/bioinformatics/btw691.

Gao, H.-L. *et al.* (2020) 'Molecular drivers and cells of origin in pancreatic ductal adenocarcinoma and pancreatic neuroendocrine carcinoma', *Experimental Hematology & Oncology*, 9(1), p. 28. Available at: https://doi.org/10.1186/s40164-020-00184-0.

Geng, X. *et al.* (2021) 'Cancer-Associated Fibroblast (CAF) Heterogeneity and Targeting Therapy of CAFs in Pancreatic Cancer', *Frontiers in Cell and Developmental Biology*, 9. Available at: https://www.frontiersin.org/articles/10.3389/fcell.2021.655152.

Groot, V.P. *et al.* (2019) 'Defining and Predicting Early Recurrence in 957 Patients With Resected Pancreatic Ductal Adenocarcinoma', *Annals of surgery*, 269(6), pp. 1154–1162. Available at: https://doi.org/10.1097/SLA.0000000000002734.

Guler, G.D. *et al.* (2020) 'Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA', *Nature*

Communications, 11(1), p. 5270. Available at: https://doi.org/10.1038/s41467-020-18965-w.

Hanahan, D. and Weinberg, R.A. (2011) 'Hallmarks of cancer: The next generation', *Cell*, pp. 646–674. Available at: https://doi.org/10.1016/j.cell.2011.02.013.

He, Q., Liu, Z. and Wang, J. (2022) 'Targeting KRAS in PDAC: A New Way to Cure It?', *Cancers*, 14(20), p. 4982. Available at: https://doi.org/10.3390/cancers14204982.

Huang, H. *et al.* (2015) 'Role of Rap2 and its Downstream Effectors in Tumorigenesis', *Anti-Cancer Agents in Medicinal Chemistry*, 15(10), pp. 1269–1276. Available at: https://doi.org/10.2174/1871520615666150518092840.

Huang, M. *et al.* (2012) 'EGFR-dependent pancreatic carcinoma cell metastasis through Rap1 activation', *Oncogene*, 31(22), pp. 2783–2793. Available at: https://doi.org/10.1038/onc.2011.450.

Huang and Wang (2019) 'Cell-Free DNA Methylation Profiling Analysis—Technologies and Bioinformatics', *Cancers*, 11(11), p. 1741. Available at: https://doi.org/10.3390/cancers11111741.

Huber, M. *et al.* (2020) 'The Immune Microenvironment in Pancreatic Cancer', *International journal of molecular sciences*, 21(19), p. 7307. Available at: https://doi.org/10.3390/ijms21197307.

Jaaks, P. *et al*. (2022) 'Effective drug combinations in breast, colon and pancreatic cancer cells', *Nature*, 603(7899), pp. 166–173. Available at: https://doi.org/10.1038/s41586-022-04437-2.

Jaffe, A.E. *et al.* (2012) 'Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies', *International Journal of Epidemiology*, 41(1), pp. 200–209. Available at: https://doi.org/10.1093/ije/dyr238.

Jiao, C. *et al.* (2018) 'Positional effects revealed in Illumina methylation array and the impact on analysis', *Epigenomics*. 2018/02/22, 10(5), pp. 643–659. Available at: https://doi.org/10.2217/epi-2017-0105.

Johnson, W.E., Li, C. and Rabinovic, A. (2007) 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics*, 8(1), pp. 118–127. Available at: https://doi.org/10.1093/biostatistics/kxj037.

Jones, S. *et al*. (2008) 'Core signaling pathways in human pancreatic cancers revealed by global genomic analyses', *Science (New York, N.Y.)*. 2008/09/04, 321(5897), pp. 1801–1806. Available at: https://doi.org/10.1126/science.1164368.

Jung, H.C. *et al*. (2017) 'Gene Regulatory Network Analysis for Triple-Negative Breast Neoplasms by Using Gene Expression Data', *Journal of Breast Cancer*, 20(3), p. 240. Available at: https://doi.org/10.4048/jbc.2017.20.3.240.

Koppensteiner, L. *et al.* (2022) 'Cancer Associated Fibroblasts - An Impediment to Effective Anti-Cancer T Cell Immunity', *Frontiers in Immunology*, 13. Available at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.887380.

Kostov, B., Bécue-Bertaut, M. and Husson, F. (2015) 'Correspondence Analysis on Generalised Aggregated Lexical Tables (CA-GALT) in the FactoMineR Package', *The R Journal*, 7(1), p. 109. Available at: https://doi.org/10.32614/RJ-2015-010.

Lê, S., Josse, J. and Husson, F. (2008) 'FactoMineR: An R Package for Multivariate Analysis', *Journal of Statistical Software*, 25(1), pp. 1–18. Available at: https://doi.org/10.18637/jss.v025.i01.

Leek, J.T. *et al.* (2012) 'The sva package for removing batch effects and other unwanted variation in high-throughput experiments', *Bioinformatics (Oxford, England)*. 2012/01/17, 28(6), pp. 882–883. Available at: https://doi.org/10.1093/bioinformatics/bts034.

Liu, M. *et al.* (2020) 'Regulation of metabolic reprogramming by tumor suppressor genes in pancreatic cancer', *Experimental Hematology & Oncology*, 9(1), p. 23. Available at: https://doi.org/10.1186/s40164-020-00179-x.

Lomberk, G. *et al.* (2018) 'Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes', *Nature communications*, 9(1), p. 1978. Available at: https://doi.org/10.1038/s41467-018-04383-6.

Mace, T.A. *et al.* (2013) 'Pancreatic cancer-associated stellate cells promote differentiation of myeloid-derived suppressor cells in a STAT3-dependent manner', *Cancer research*. 2013/03/20, 73(10), pp. 3007–3018. Available at: https://doi.org/10.1158/0008-5472.CAN-12-4601.

Malinova, A. *et al.* (2021) 'Cell Lineage Infidelity in PDAC Progression and Therapy Resistance', *Frontiers in Cell and Developmental Biology*, 9. Available at: https://www.frontiersin.org/articles/10.3389/fcell.2021.795251.

Maneshi, P. *et al.* (2021) 'Targeting Tumor-Stromal Interactions in Pancreatic Cancer: Impact of Collagens and Mechanical Traits', *Frontiers in Cell and Developmental Biology*, 9. Available at: https://www.frontiersin.org/articles/10.3389/fcell.2021.787485.

Martens, S. *et al.* (2019) 'Different shades of pancreatic ductal adenocarcinoma, different paths towards precision therapeutic applications', *Annals of Oncology*, 30(9), pp. 1428–1436. Available at: https://doi.org/10.1093/annonc/mdz181.

Masugi, Y. (2022) 'The Desmoplastic Stroma of Pancreatic Cancer: Multilayered Levels of Heterogeneity, Clinical Significance, and Therapeutic Opportunities', *Cancers*, 14(13). Available at: https://doi.org/10.3390/cancers14133293.

Mattiolo, P. *et al.* (2022) 'Refining targeted therapeutic approaches in pancreatic cancer: from histology and molecular pathology to the clinic', *Expert Opinion on Therapeutic Targets*, 26(1), pp. 1–4. Available at: https://doi.org/10.1080/14728222.2022.2021397.

McGhee, S.A. and Chatila, T.A. (2010) 'DOCK8 Immune Deficiency as a Model for Primary Cytoskeletal Dysfunction', *Disease Markers*, 29, p. 397291. Available at: https://doi.org/10.3233/DMA-2010-0740.

Moffitt, R.A. *et al.* (2015) 'Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma', *Nature Genetics*, 47(10), pp. 1168–1178. Available at: https://doi.org/10.1038/ng.3398.

Morris, T.J. *et al.* (2014) 'ChAMP: 450k Chip Analysis Methylation Pipeline', *Bioinformatics*, 30(3), pp. 428–430. Available at: https://doi.org/10.1093/bioinformatics/btt684.

Narayanan, S., Vicent, S. and Ponz-Sarvisé, M. (2021) 'PDAC as an Immune Evasive Disease: Can 3D Model Systems Aid to Tackle This Clinical Problem?', *Frontiers in Cell and Developmental Biology*, 9. Available at: https://www.frontiersin.org/articles/10.3389/fcell.2021.787249.

New England BioLabs (2016) *Luna® Universal qPCR Master Mix Protocol (M3003)*. Available at: https://international.neb.com/protocols/2016/11/08/luna-universal-qpcr-master-mix-protocol-m3003.

New England BioLabs (2018) 'Protocol for LunaScript® RT SuperMix Kit (E3010)'. Available at: https://international.neb.com/protocols/2018/01/31/protocol-for-lunascript-rt-supermix-kit-e3010.

Nordlund, J. *et al.* (2013) 'Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia', *Genome Biology*, 14(9), p. r105. Available at: https://doi.org/10.1186/gb-2013-14-9-r105.

O'Kane, G.M. *et al.* (2020) 'GATA6 Expression Distinguishes Classical and Basal-like Subtypes in Advanced Pancreatic Cancer', *Clinical Cancer Research*, 26(18), pp. 4901–4910. Available at: https://doi.org/10.1158/1078-0432.CCR-19-3724.

Orth, M. *et al.* (2019) 'Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches', *Radiation Oncology*, 14(1), p. 141. Available at: https://doi.org/10.1186/s13014-019-1345-6.

Principe, D.R. *et al.* (2021) 'The Current Treatment Paradigm for Pancreatic Ductal Adenocarcinoma and Barriers to Therapeutic Efficacy', *Frontiers in Oncology*, 11. Available at: https://www.frontiersin.org/articles/10.3389/fonc.2021.688377.

Promega (2012) *Promega ReliaPrepTM RNA Cell Miniprep System protocol*. Available at: https://www.promega.co.uk/-/media/files/resources/protcards/reliaprep-rna-cell-miniprep-system-quick-protocol.pdf?rev=749be353a9ca481bbe37f4b15b8d6d7b&sc_lang=en.

Qian, X. *et al.* (2022) 'A cancer-associated K-RAS mutant that preferentially binds GDP displays gain-of-function and atypical signaling'. Research Square. Available at: https://doi.org/10.21203/rs.3.rs-1521657/v1.

Qian, Y. *et al.* (2020) 'Molecular alterations and targeted therapy in pancreatic ductal adenocarcinoma', *Journal of Hematology & Oncology*, 13(1), p. 130. Available at: https://doi.org/10.1186/s13045-020-00958-3.

Rahib, L. *et al.* (2021) 'Estimated Projection of US Cancer Incidence and Death to 2040', *JAMA Network Open*, 4(4), pp. e214708–e214708. Available at: https://doi.org/10.1001/jamanetworkopen.2021.4708.

Raivo, K. (2019) 'Reference Manual for Package "pheatmap"'. Available at: https://cran.r-hub.io/web/packages/pheatmap/pheatmap.pdf (Accessed: 11 January 2023).

Randall, K.L. *et al.* (2011) 'DOCK8 deficiency impairs CD8 T cell survival and function in humans and mice', *The Journal of experimental medicine*. 2011/10/17, 208(11), pp. 2305–2320. Available at: https://doi.org/10.1084/jem.20110345.

Roalsø, M.T.T. *et al.* (2022) 'Emerging Role of Epigenetic Alterations as Biomarkers and Novel Targets for Treatments in Pancreatic Ductal Adenocarcinoma', *Cancers*, 14(3), p. 546. Available at: https://doi.org/10.3390/cancers14030546.

Shannon, P. *et al.* (2003) 'Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks', *Genome Research*, 13(11), pp. 2498–2504. Available at: https://doi.org/10.1101/gr.1239303.

Slynko, A. and Benner, A. (2019) 'Statistical methods for classification of 5hmC levels based on the Illumina Inifinium HumanMethylation450 (450k) array data, under the paired bisulfite (BS) and oxidative bisulfite (oxBS) treatment', *PloS one*, 14(6), pp. e0218103–e0218103. Available at: https://doi.org/10.1371/journal.pone.0218103.

Smittenaar, C.R. *et al.* (2016) 'Cancer incidence and mortality projections in the UK until 2035', *British Journal of Cancer*, 115(9), pp. 1147–1155. Available at: https://doi.org/10.1038/bjc.2016.304.

Smyth, G.K. (2004) 'Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments', *Statistical Applications in Genetics and Molecular Biology*, 3(1), pp. 1–25. Available at: https://doi.org/10.2202/1544-6115.1027.

Sophie John & John Broggio (2019) *Cancer survival by stage at diagnosis for England*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancersurvivalratescancersurvivalinenglandadultsdiagnosed (Accessed: 10 January 2023).

Sperb, N., Tsesmelis, M. and Wirth, T. (2020) 'Crosstalk between Tumor and Stromal Cells in Pancreatic Ductal Adenocarcinoma', *International journal of molecular sciences*, 21(15), p. 5486. Available at: https://doi.org/10.3390/ijms21155486.

Suzuki, R. and Shimodaira, H. (2006) 'Pvclust: an R package for assessing the uncertainty in hierarchical clustering', *Bioinformatics*, 22(12), pp. 1540–1542. Available at: https://doi.org/10.1093/bioinformatics/btl117.

Szklarczyk, D. *et al.* (2019) 'STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets', *Nucleic Acids Research*, 47(D1), pp. D607–D613. Available at: https://doi.org/10.1093/nar/gky1131.

Szklarczyk, D. *et al.* (2021) 'The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets', *Nucleic Acids Research*, 49(D1), pp. D605–D612. Available at: https://doi.org/10.1093/nar/gkaa1074.

Teschendorff, A.E. *et al.* (2009) 'An Epigenetic Signature in Peripheral Blood Predicts Active Ovarian Cancer', *PLOS ONE*, 4(12), pp. e8274-. Available at: https://doi.org/10.1371/journal.pone.0008274.

Teschendorff, A.E. *et al.* (2013) 'A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data', *Bioinformatics*, 29(2), pp. 189–196. Available at: https://doi.org/10.1093/bioinformatics/bts680.

Tian, Y. *et al.* (2017) 'ChAMP: updated methylation analysis pipeline for Illumina BeadChips', *Bioinformatics*, 33(24), pp. 3982–3984. Available at: https://doi.org/10.1093/bioinformatics/btx513.

Tiedemann, R.L. *et al.* (2021) 'Distinguishing Active Versus Passive DNA Demethylation Using Illumina MethylationEPIC BeadChip Microarrays', *Methods in molecular biology (Clifton, N.J.)*, 2272, pp. 97–140. Available at: https://doi.org/10.1007/978-1-0716-1294-1_7.

Torres, C. and Grippo, P.J. (2018) ‘Pancreatic cancer subtypes: a roadmap for precision medicine’, *Annals of medicine*. 2018/03/22, 50(4), pp. 277–287. Available at: https://doi.org/10.1080/07853890.2018.1453168.

Torricelli, F. *et al.* (2020) ‘Genomic analysis in short- and long-term patients with malignant pleura mesothelioma treated with palliative chemotherapy’, *European Journal of Cancer*, 132, pp. 104–111. Available at: https://doi.org/10.1016/j.ejca.2020.03.002.

Wettenhall, J.M. and Smyth, G.K. (2004) ‘limmaGUI: A graphical user interface for linear modeling of microarray data’, *Bioinformatics*, 20(18), pp. 3705–3706. Available at: https://doi.org/10.1093/bioinformatics/bth449.

Wood, L.D. *et al.* (2022) ‘Pancreatic Cancer: Pathogenesis, Screening, Diagnosis, and Treatment’, *Gastroenterology*. W.B. Saunders, pp. 386-402.e1. Available at: https://doi.org/10.1053/j.gastro.2022.03.056.

Xu, Z. *et al.* (2021) ‘Clinical Impact of Molecular Subtyping of Pancreatic Cancer’, *Frontiers in cell and developmental biology*, 9, p. 743908. Available at: https://doi.org/10.3389/fcell.2021.743908.

Yamasaki, A., Yanai, K. and Onishi, H. (2020) ‘Hypoxia and pancreatic ductal adenocarcinoma’, *Cancer Letters*, 484, pp. 9–15. Available at: https://doi.org/https://doi.org/10.1016/j.canlet.2020.04.018.

Yi-Lei, Z. *et al.* (2017) ‘Roles of Rap1 signaling in tumor cell migration and invasion’, *Cancer Biology & Medicine*, 14(1), pp. 90–99. Available at: https://doi.org/10.20892/j.issn.2095-3941.2016.0086.

Zhang, L. *et al.* (2022) ‘Deciphering the role of DOCK8 in tumorigenesis by regulating immunity and the application of nanotechnology in DOCK8 deficiency therapy’, *Frontiers in Pharmacology*, 13. Available at: https://www.frontiersin.org/articles/10.3389/fphar.2022.1065029.

Zhou, W., Laird, P.W. and Shen, H. (2017) ‘Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes’, *Nucleic acids research*, 45(4), pp. e22–e22. Available at: https://doi.org/10.1093/nar/gkw967.

Appendix

Note that all starting material or raw analyses data that are too large to be included in the appendix are available upon request.

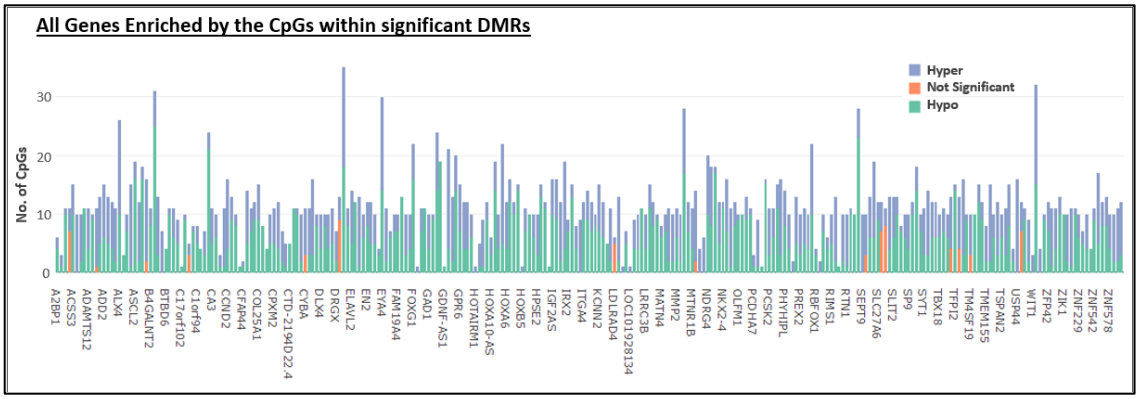


Fig. A1 All Genes Enriched by CpGs within significant DMRs

This bar plot is an output from the pairwise differential DMR analysis which highlights the genes that are enriched by all CpGs within the significantly differentially methylated regions. The list can be taken as candidates for further analyses like interaction or enrichment analyses – in this case, the top 10 ranked genes containing the greatest number of CpGs were considered as candidates for validation (14 in total as some genes were enriched with the same number of CpGs). From the plot, the top enriched gene is *PCDHGA4* and the exact number of hypo/hyper-methylated CpGs are available within the interactive GUI: total = 36; hypomethylated = 0; significantly hypermethylated = 32; non-significant = 4).

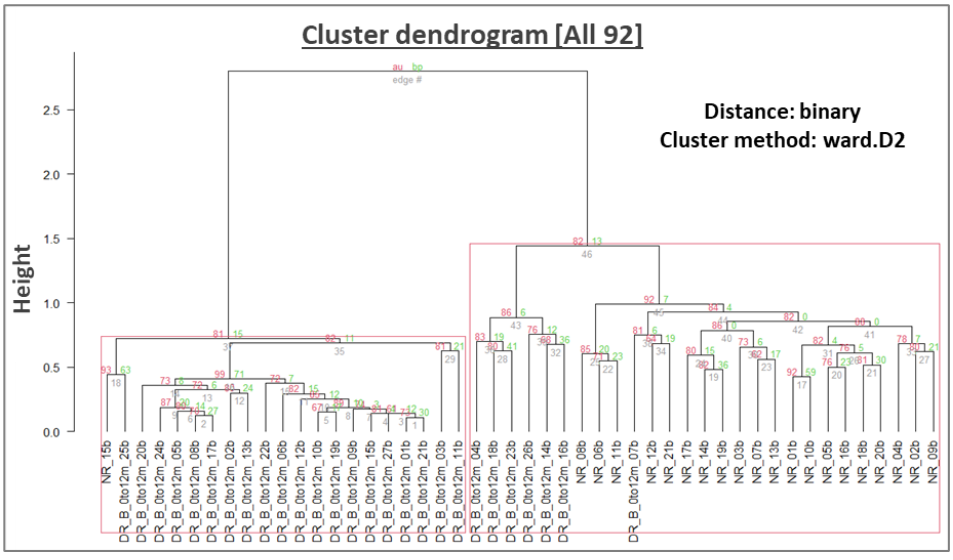


Fig. A2 pvclust Dendrogram [92 Variables]

As mentioned in the main body, pvclust computes distance matrices according to the selected method ("binary" in this case) and performs agglomeration clustering via the "ward.D2" algorithm. The result is exactly same as the one produced by pheatmap, with extra information – AU p-values presented for each clustering step (in red).

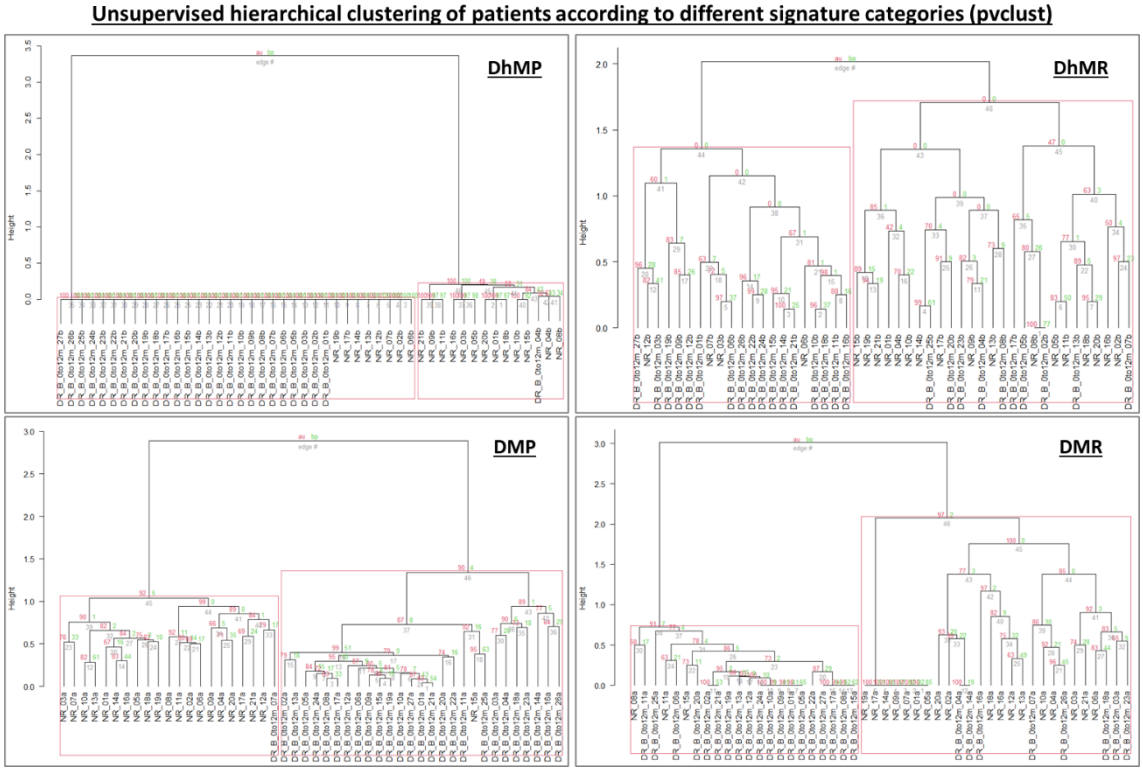


Fig. A3 pvclust Dendrograms for different signature categories

Similarly, pvclust was used for unsupervised hierarchical clustering of patients (methods = binary; ward.D2), based on the 4 signature categories – DhMP; DhMR; DMP and DMR. The results were again the same as the ones produced by pheatmap and hence, the same message – these signatures on its own are not powerful enough for 100% accurate clustering.

MCA Factor Maps for different signature categories

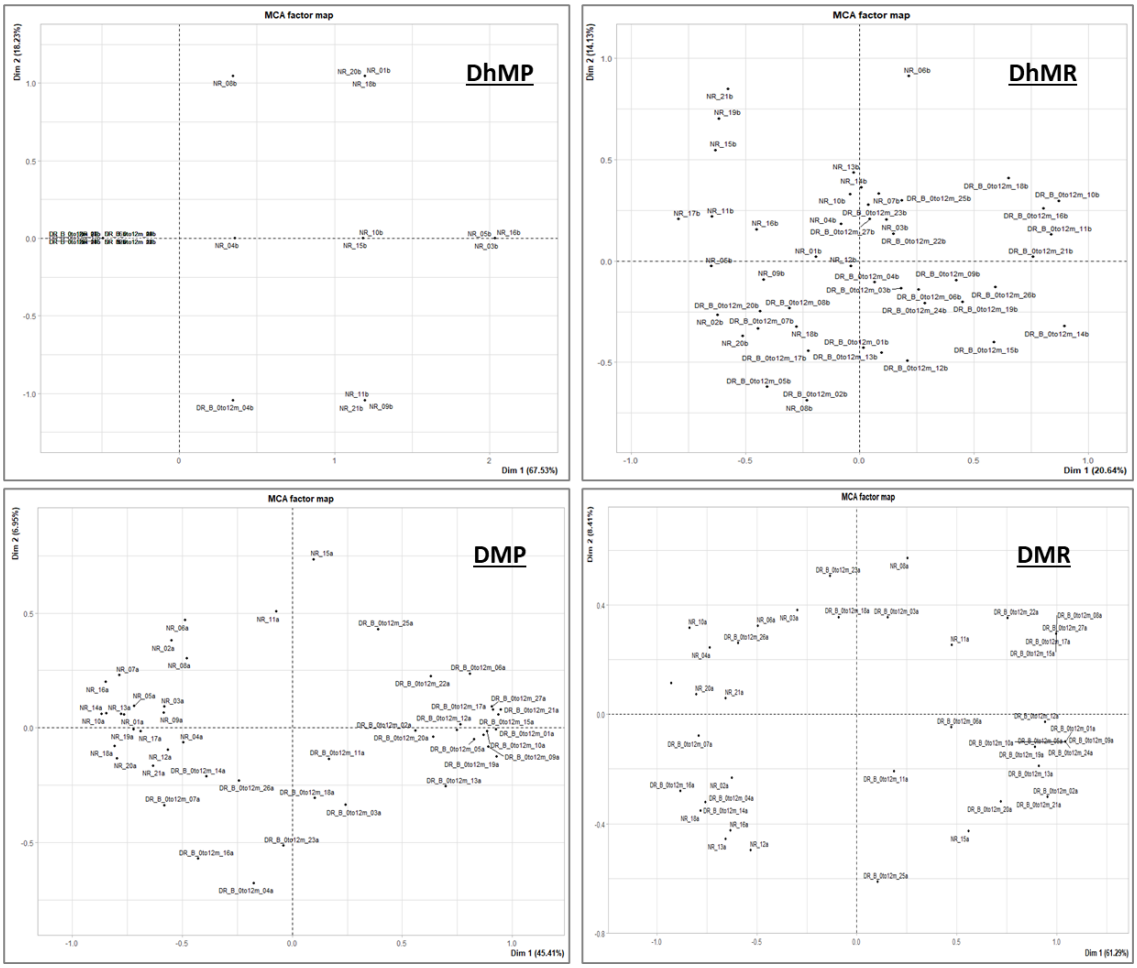


Fig. A4 MCA Factor Maps for different signature categories

These MCA factor map displays the overall positioning of individuals in relation to the scoring of selected variables of other patients (based on results from DhMP, DhMR, DMP and DMR analyses). Again, it can be seen in the plots that these signature combinations were not able to clearly stratify patients into their corresponding groups, especially DhMR and DMR; which matches the conclusion drawn from pheatmap/pvclust.

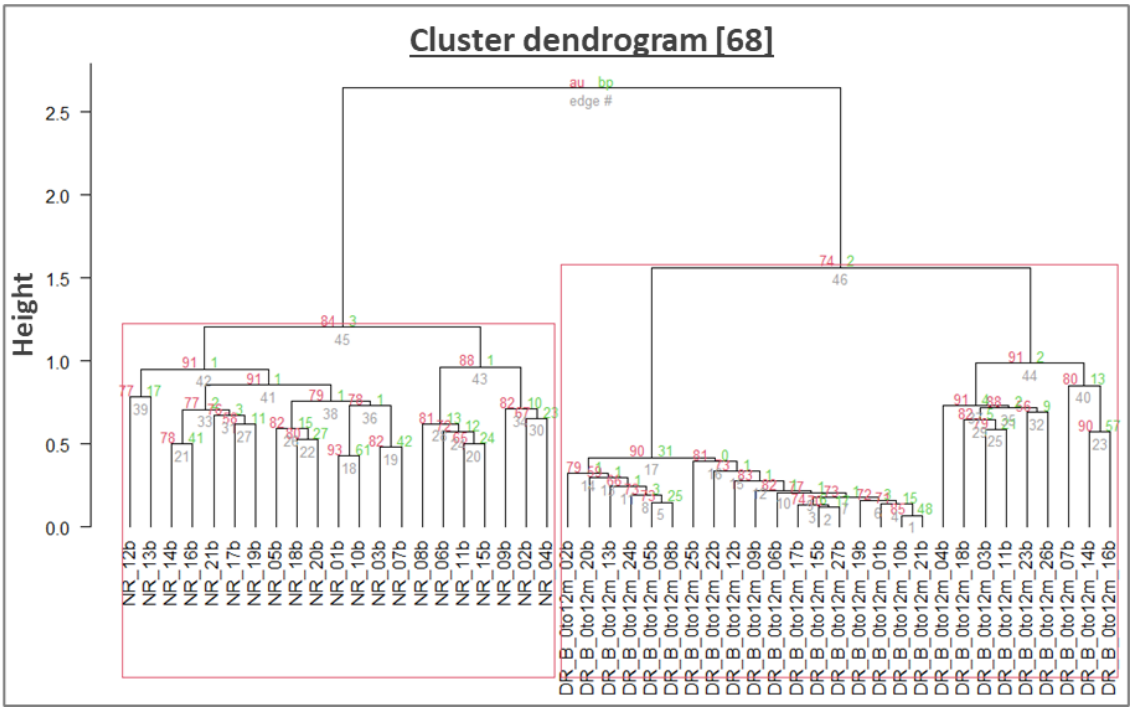


Fig. A5 pvclust Dendrogram [68 Variables]

As mentioned in the main body, unsupervised hierarchical clustering via pvclust produced the exact same result as the one produced by pheatmap, not only showing that the optimised signature list has the ability to correctly sort patients into their corresponding, but also gives insight into heterogeneity, highlighting possible subgroups within PDAC types, which may be useful in understanding the pathology or progression of PDAC.

1	RYR3	TCGGTCAGTGAGCCAGATATGG	F	
		AGGCAGAAGCTCTTAGGTCAGG	R	
2	C10orf72	TGTCACTAGCGTGACCAGCTTG	F	
		CAGCTTCGGTTTATGGAACGTGG	R	
3	PXDN	CGGACATTGCAGCTCATTCAGG	F	
		CCGACAGGTTTGCGATGAGGTT	R	
4	PCDHGA4	CAACACGGACTGGCGTTTCTCT	F	
		CATGGCTTGCAGCATCTCTGTG	R	
5	ISLR2	CAACACGGACTGGCGTTTCTCT	F	
		CATGGCTTGCAGCATCTCTGTG	R	
6	DOCK8	GCTCCTCTTCACAGAACTGGCT	F	
		TCACCTCTGGTTTGACACAGCG	R	
7	RAPGEF6	AGACAGATGAGGAGAAGTTCCAG	F	
		GACCTCATAGGCACTGGAGACA	R	
8	MUC4	AACACAGCCTGCTAGTCCAGCA	F	
		TGGAGAGGATGGCTTGGTAGGT	R	
9	SPRR1A	GTGAAACAACCTTGCCAGCCTC	F	
		TGGCAGGGCTCTGGAACCTTG	R	
10	GDNF	CGCCGAAGACCGCTCCCTCG	F	
		ATCCATGACATCATCGAACTGATC	R	
11	ATP10A	CTGCTCTTCAGTTGACGCCAGA	F	
		GCAACCAGCAGGCATACTCTTC	R	
12	BNC1	AGTGGAGGGCTGTAATGCTACC	F	
		CTCTCCAATGCTTCCTGGCTCA	R	
13	EDNRB	CAGAAAGCCTCCGTGGGAATCA	F	
		ACAGCCAGAACCACAGAGACCA	R	
14	EYA4	CAGCCTACTCACAGACAGGACA	F	Invitrogen
		GTAACTGAGGCAGCCACTCTGT	R	
15	SALL1	GCTTGCACTATTTGTGGAAGAGC	F	
		GAACTTGACGGGATTGCCTCCT	R	
16	CA10	TGAAGGCTGGTGGGCATACAAG	F	
		TGTTGACTGGCGACTGCCGTTT	R	
17	SHISA9	AGCAGCCACCAGGAAAAGAGCT	F	
		TGCCGTCGCTTGGTGTAGAAGT	R	
18	FBRSL1	CCGACATTCCAGCGTGAGTTTC	F	
		GACACCTTAGGCTGAAAAGCGC	R	
19	PRKAR1B	ACCTGGATGACAACGAGAGGAG	F	
		TGTTGGTCACCCACTCTCCGTT	R	
20	FBXO39	CCTGAAGAGTCAGAAAGAACGGC	F	
		CGATCAGCTTTCTGTACTTCCTG	R	
21	COL23A1	CTGGCGACACTGGGAAAGATGG	F	
		TGTCCCATCGTCGCCCTTCTTT	R	
22	CDH4	GTGCGGAAGAACTGGACCATCA	F	
		GGAGGGTTTCCAGAGTCTGTGA	R	
23	NEU1	TCCAAGGCTGAGAACGACTTCG	F	
		TCAGCAAAGGCGAGAAGAGTGC	R	
24	EBF3	CGGAACTATGTTGGTGTGGAGC	F	
		AGGAGCACCTTTGCAGAACTGC	R	
25	CTNNA2	GAGAAGCAGGTCCGAGTGTTGA	F	
		CCTCTTGGAGGGCTATCACACA	R	
26	CDH13	TGGCAGAACTCGTGATTGTCGG	F	
		GGGAGATACCACAATGGACCTC	R	
27	ITGA4	GCATACAGGTGTCCAGCAGAGA	F	
		AGGACCAAGGTGGTAAGCAGCT	R	
28	ACTB	CACCATTGGCAATGAGCGGTTC	F	Origene
		AGGTCTTTGCGGATGTCCACGT	R	
29	GAPDH	GTCTCCTCTGACTTCAACAGCG	F	
		ACCACCCTGTTGCTGTAGCCAA	R	

Table A1 Primer list

Here is the list of primer used for RT qPCR, with the biotech-company that they are ordered from.

	Panc1	MiaPaCa	PSN	BxPC	CAPAN		Panc1	MiaPaCa	PSN	BxPC	CAPAN
RAPGEF6	0.482012	0.570403	0.898142	0.5225	0.225441	DOCK8	1.033876	0.001118	0.238305	7.888102	6.036597
RAPGEF6	1.668529	0.52248	0.683969	0.320815	0.244174	DOCK8	0.769193	0.054625	0.091422	3.562253	4.923306
RAPGEF6	0.849459	1.023273	0.952577	0.328443	0.236648	DOCK8	1.19693	0.014092	0.921193	2.653861	2.573834
RAPGEF6	0.702575	0.658424	0.883334	0.443098	0.213666	DOCK8	0.461465	0.001164	0.127552	3.22788	3.192891
RAPGEF6	1.256699	0.632068	0.627936	0.315597	NA	DOCK8	1.568452	0.00484	0.028902	1.947232	NA
RAPGEF6	1.040726	1.240958	0.894134	0.445443	NA	DOCK8	0.970083	0.012833	0.212534	2.37518	NA
PCDHGA4	1.625838	0.259113	1.397983	0.471565	0.022597	ITGA4	0.745318	0.425411	0.574014	9.186586	0.606637
PCDHGA4	0.679496	0.592925	0.906927	0.305114	0.026849	ITGA4	1.606748	0.366736	0.039481	6.704216	0.396129
PCDHGA4	0.694666	0.692436	1.004374	0.450889	0.023058	ITGA4	0.647934	1.100312	1.325812	7.682587	2.176112
PCDHGA4	1.565139	0.384428	1.95386	0.579589	0.025218	ITGA4	1.149163	0.96724	2.396231	11.37654	2.724743
PCDHGA4	0.597059	0.681478	1.04743	0.349417	NA	ITGA4	0.256531	0.434482	1.242992	7.227682	NA
PCDHGA4	0.837803	0.691425	1.371152	0.386817	NA	ITGA4	1.594306	1.146419	1.841556	11.30823	NA
ISLR2	1.582389	0.398434	1.738443	0.633653	0.024516	SPRR1A	2.739496	0.159316	1.50484	39.22852	6.687721
ISLR2	0.687978	0.601279	0.922244	0.384699	0.017978	SPRR1A	0.18739	0.605324	0.670762	62.77407	3.335212
ISLR2	0.729633	0.616228	0.910886	0.416918	0.030624	SPRR1A	0.073113	0.094713	13.52033	64.83872	5.667048
ISLR2	1.642045	0.405633	1.860341	0.540559	0.026029	SPRR1A	2.695049	0.190958	2.479848	54.54041	3.54122
ISLR2	0.514635	0.763518	0.98439	0.283621	NA	SPRR1A	0.219372	0.994598	0.951672	66.67864	NA
ISLR2	0.84332	0.707013	1.299175	0.513781	NA	SPRR1A	0.085579	0.212533	24.90649	77.19863	NA

Table A2 Relative Fold Changes for genes showing interesting trends

As described in chapter 7.2, these were calculated from the qPCR results and used for the generation of boxplots. These were also inputs for following statistical tests – ANOVA and Tukey HSD test.

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RAPGEF6$`Cell line`	4	2.08457	0.52114	8.33125	0.0002625
Residuals	23	1.43871	0.06255	NA	NA
Tukey HSD		diff	lwr	upr	p adj
	BxPC-3-Capan-1	0.166	-0.31123	0.64323	0.8397463
	MIA PaCa-2-Capan-1	0.54462	0.06739	1.02185	0.0199621
	PSN-1-Capan-1	0.59337	0.11614	1.07059	0.0099567
	PANC-1-Capan-1	0.77002	0.29279	1.24725	0.0007183
	MIA PaCa-2-BxPC-3	0.37862	-0.04823	0.80546	0.0990807
	PSN-1-BxPC-3	0.42737	0.00052	0.85421	0.0496189
	PANC-1-BxPC-3	0.60402	0.17717	1.03086	0.0029795
	PSN-1-MIA PaCa-2	0.04875	-0.3781	0.47559	0.9969935
	PANC-1-MIA PaCa-2	0.2254	-0.20145	0.65224	0.5356785
	PANC-1-PSN-1	0.17665	-0.25019	0.6035	0.7382129

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DOCK8$`Cell line`	4	596074	149019	45.3598	1.37E-10
Residuals	23	75560.9	3285.26	NA	NA
Tukey HSD		diff	lwr	upr	p adj
	PSN-1-MIA PaCa-2	0.25521	-97.5658	98.0762	1
	PANC-1-MIA PaCa-2	0.98522	-96.8358	98.8062	0.9999998
	BxPC-3-MIA PaCa-2	3.59431	-94.2267	101.415	0.9999659
	Capan-1-MIA PaCa-2	418.151	308.784	527.518	6.95E-10
	PANC-1-PSN-1	0.73002	-97.091	98.551	0.9999999
	BxPC-3-PSN-1	3.3391	-94.4819	101.16	0.9999746
	Capan-1-PSN-1	417.896	308.529	527.263	7.03E-10
	BxPC-3-PANC-1	2.60908	-95.2119	100.43	0.9999905
	Capan-1-PANC-1	417.166	307.799	526.533	7.275E-10
	Capan-1-BxPC-3	414.557	305.189	523.924	8.224E-10

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PCDHGA4$`Cell line`	4	4.97781	1.24445	13.8879	6.70E-06
Residuals	23	2.06096	0.08961	NA	NA
Tukey HSD		diff	lwr	upr	p adj
	BxPC-3-Capan-1	0.39947	-0.17171	0.97065	0.2675487
	MIA PaCa-2-Capan-1	0.52587	-0.04531	1.09705	0.0812764
	PANC-1-Capan-1	0.97557	0.40439	1.54675	0.0003642
	PSN-1-Capan-1	1.25586	0.68468	1.82704	1.15E-05
	MIA PaCa-2-BxPC-3	0.1264	-0.38448	0.63728	0.9469011
	PANC-1-BxPC-3	0.5761	0.06522	1.08698	0.0218536
	PSN-1-BxPC-3	0.85639	0.34551	1.36727	0.0004573
	PANC-1-MIA PaCa-2	0.4497	-0.06118	0.96058	0.1030417
	PSN-1-MIA PaCa-2	0.72999	0.21911	1.24087	0.0027003
	PSN-1-PANC-1	0.28029	-0.23059	0.79117	0.4991877

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ITGA4$`Cell line`	4	290.837	72.7093	54.2262	2.22E-11
Residuals	23	30.8396	1.34085	NA	NA
Tukey HSD		diff	lwr	upr	p adj
	PANC-1-MIA PaCa-2	0.2599	-1.71633	2.23613	0.9948179
	PSN-1-MIA PaCa-2	0.49658	-1.47965	2.47281	0.9439948
	Capan-1-MIA PaCa-2	0.73581	-1.47369	2.9453	0.8595389
	BxPC-3-MIA PaCa-2	8.17421	6.19797	10.1504	1.46E-10
	PSN-1-PANC-1	0.23668	-1.73955	2.21291	0.9963862
	Capan-1-PANC-1	0.47591	-1.73359	2.6854	0.9673993
	BxPC-3-PANC-1	7.91431	5.93807	9.89054	2.78E-10
	Capan-1-PSN-1	0.23922	-1.97027	2.44872	0.9975562
	BxPC-3-PSN-1	7.67763	5.70139	9.65386	5.08E-10
	BxPC-3-Capan-1	7.4384	5.22891	9.6479	7.99E-09

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ISLR2$`Cell line`	4	4.83799	1.2095	12.2339	1.78E-05
Residuals	23	2.27387	0.09886	NA	NA
Tukey HSD		diff	lwr	upr	p adj
	BxPC-3-Capan-1	0.43742	-0.16254	1.03738	0.2319043
	MIA PaCa-2-Capan-1	0.55723	-0.04273	1.15719	0.0774299
	PANC-1-Capan-1	0.97521	0.37525	1.57517	0.0006592
	PSN-1-Capan-1	1.26113	0.66117	1.86109	2.23E-05
	MIA PaCa-2-BxPC-3	0.11981	-0.41681	0.65643	0.9629445
	PANC-1-BxPC-3	0.53779	0.00117	1.07441	0.0493172
	PSN-1-BxPC-3	0.82371	0.28709	1.36033	0.0012633
	PANC-1-MIA PaCa-2	0.41798	-0.11864	0.9546	0.1801032
	PSN-1-MIA PaCa-2	0.7039	0.16728	1.24052	0.0061862
	PSN-1-PANC-1	0.28591	-0.25071	0.82253	0.5272092

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SPRR1A$`Cell line`	4	15844.7	3961.16	68.223	2.03E-12
Residuals	23	1335.43	58.062	NA	NA
Tukey HSD		diff	lwr	upr	p adj
	PANC-1-MIA PaCa-2	0.62376	-12.3807	13.6283	0.9999016
	Capan-1-MIA PaCa-2	4.43156	-10.1079	18.971	0.893574
	PSN-1-MIA PaCa-2	6.96275	-6.04175	19.9672	0.5225689
	BxPC-3-MIA PaCa-2	60.5003	47.4958	73.5048	1.33E-11
	Capan-1-PANC-1	3.8078	-10.7317	18.3473	0.9354915
	PSN-1-PANC-1	6.33899	-6.66551	19.3435	0.608838
	BxPC-3-PANC-1	59.8765	46.872	72.881	1.646E-11
	PSN-1-Capan-1	2.53119	-12.0083	17.0707	0.9850232
	BxPC-3-Capan-1	56.0687	41.5292	70.6082	5.871E-10
	BxPC-3-PSN-1	53.5375	40.533	66.542	1.606E-10

Table A3 ANOVA/Tukey HSD test results

Here are the results for the ANOVA performed for *RAPGEF6*, *PCDHGA4*, *ISLR2*, *DOCK8*, *ITGA4* and *SPRR1A*; which all showed significance in explaining variance in GOI expressions with cell lines. Significance in the difference in expression levels between cell line pairs were then computed via Tukey HSD tests, with those that are <0.05 highlighted in red. Predominantly, the results support the hypothesis that the transcription trends for these genes are significantly different depending on the PDAC subtype (classical or squamous).