An Investigation of Mutational Signatures in the Evolution of Oesophageal adenocarcinoma

Sujath Abbas

Sidney Sussex College, University of Cambridge

Supervisor

Professor Rebecca Fitzgerald

Interim Director, MRC Cancer Unit

University of Cambridge

Secondary Supervisor

Dr.Maria Secrier

Division of Biosciences

University College London

This thesis is submitted for the degree of Doctor of Philosophy.

October 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed 60,000-word limit for the Clinical Medicine and Veterinary Medicine Degree Committee excluding figures, tables and bibliography

Sujath Abbas

October 2021

Summary

An Investigation of Mutational Signatures in the Evolution of Oesophageal adenocarcinoma

Sujath Abbas

Oesophageal adenocarcinoma (OAC) remains a public health challenge with dismal survival rates and increasing incidence. This PhD study aimed to investigate how mutational processes act across different stages of OAC development and in metastasis for better understanding of the influence of mutational forces during tumour formation. To identify these signatures in clinical samples this study also aimed to develop a cost-effective DNA sequencing method in clinical formalin fixed OAC samples. A large study cohort was assembled comprising of 161 Barrett's, 777 OAC primary tumours and 59 metastatic samples. Mutational signature analysis revealed 14 distinct single base substitution (SBS) mutational signatures in these genomes, SBS17b/a were most prevalent and presented early in Barrett's. Traces of BER (SBS30), MMR(SBS44) and colibactin associated signature (SBS41) were uncovered for the first time, as well as a platinum signature (SBS35). Mostly signatures increased in their proportions from Barrett's to invasive tumours and further in metastasis. SBS17 showed strong bias towards untranscribed and lagging strands. Nucleosome periodicity patterns were similar across the stages and SNVs were enriched in the inward facing minor groove suggesting a common mutational process throughout the disease evolution. Evaluation of evolutionary bottlenecks uncover a distinct SBS17b shift, with a decrease sub-clonally in Barrett's, OACs and metastasis and this was by far the most dominant signal during OAC evolution. Clinical risk factors including alcohol, smoking and NSAIDs were positively correlated with signature proportions. APOBEC and colibactin processes were informative for Barrett's and OAC classification, suggesting a role in transformation, and the BER signature (SBS30) was most prognostic in our cohort. Given that signatures have the potential to be clinically informative, a novel cost-effective DNA sequencing method to extract mutational signatures from archival FFPE tissues was developed successfully. Computational simulations on pan-cancer WGS and an experimental confirmation of the method showed very good concordance and mirrored the WGS-derived signatures (cosine similarity >0.9%). It is hoped that this work will pave the way for further studies to understand how mutations are laid down and determine their clinical application.

Contents

List of Figures

List of Tables

Collaborations

The Oesophageal Cancer Classification and Molecular Stratification Consortium (OCCAMS) close collaboration provided samples included in this study. Pathological quality control was done by a team of pathologists lead by Dr.Maria O' Donovan. Nucleic acid (DNA/RNA) extraction was performed by Jason Crawte and Alex Northrop. WGS libraries and sequencing was done at Illumina. RNA Seq library preparation and sequencing was undertaken by me.

WGS and RNA Seq data was managed by a team of Bioinformaticians including Ginny Devonshire, Dr Juliane Perner, Lawrence Bower, Dr SriGanesh Jammula and others. I used the output files for analysis as set out in this study.

Dr Maria Secrier from UCL supervised my bioinformatic analysis. Areas in which collaborations aided were listed in the attribution section of respective chapters in the thesis.

Acknowledgments

PhD at Cambridge was my dream from my school times when I first learned about DNA. Now, when I am expressing my gratitude to all who were part of this incredible journey, the excitement of dreams coming true, I feel is beyond words.

I am thankful to the almighty Allah (The GOD)!!

I am extremely grateful to **Prof.Rebecca Fitzgerald** for her guidance and support through out this journey! For introducing me to the real world of research here at Cambridge. Over this time, under her supervision I feel I have grown as a scientist and able to think and manage difficult scientific questions and strive towards contributing my bit in cancer research broadly. Thank you for all the personal support during my hard times.

I am so thankful to **Dr.Maria Secrier (UCL London)**, for training me on the informatics part of my thesis, I only learned this during my doctoral training as I did not had previous experience in this field.

All the past and present lab members, Dr Gianmarco Contino (now senior lecturer at University of Birmingham) who helped and guided me during my initial days in the lab. Dr Shona MacRae past lab manager for helping me settle in the lab and with my RNA Seq project. Then PostDocs Dr Juliane Perner and Dr Karol Nowicki-osuch for helping me with mutREAD project. Then Research Assistant Adrienn Blasko for her help with DNA extractions. Dr.SriGanesh Jammula for advise and discussions on my PhD project and help with informatics. I am thankful to our current lab manager Dr.Aisling Redmond for all the support. Thankful to Dr. Constanza Linossi for her help with mutREAD validation cohort. Calvin Cheah and Barbara Nutzinger for their help with clinical data. Pathologists Dr.Ahmad Miremadi and Shalini Malhotra for helping me with H&E slide review.

I am very lucky that I have got a lovely family who supported me by all means during this journey. Without their backing this would have not possible. I cannot thank enough to show my gratitude to my Queen, my mother **Zawwar Khadija Banu**, she always believed in me and encouraged me to pursue what I like. Love you Mumma!!

My lovely brothers **Vikhar Abbas Syed**, **Tawassul Abbas Syed,** they took my responsibilities to free me to pursue this career and my sister **Syeda Askari Banu,** brother in-law **Salman Raza** and my love to the new addition to the family **Shahr Bano** (4months) niece. Conclusion to this journey was also marked by an important turning point in my personal life, my wedding!!! I am getting married to my fiance **Dr.Muneeza Ali**.

Finally, I would like to fondly remember and dedicate this thesis to my late father

"Mir Shanawaz Hussain"

(1955-2014)

He was a dynamic leader, educationist, social reformer and served the needy in my native Holavanahalli, Karnataka India. He was instrumental in bringing a primary school to our village to promote literacy. He always wanted me to reach heights in my career, I believe he is happy in Heavens seeing my progress. I will strive to work towards his dreams for rest of my career.

Abbreviations

APOBEC- Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like

ATP- Adenosine triphosphate

BE- Barrett's Oesophagus

BER- Base excision repair

BMI- Body mass index

BRCA- Breast Cancer gene

CGC- Cancer gene census

CIN- Copy number instability

CNVs- Copy number variation

COSMIC- Catalogue of somatic mutations in cancer

CT- Computerized tomography

DBS- Double base substitution

DDR- DNA damage repair

DDRD- DNA damage repair deficiency

ECX- Epirubicin cisplatin and capecitabine

EDTA- Ethylenediaminetetraacetic acid

EOX- Epirubicin oxaliplatin and capecitabine

ESCC- Oesophageal squamous cell carcinoma

FFPE- Formalin fixed paraffin embedded

FLOT- 5FU,Folinic acid,Oxaliplatin,Docetaxel

GERD- Gastroesophageal reflux disease

GSVA- Gene set variation analysis

HR- Homologous recombination

ICGC- International cancer genome consortium

ID- Indel

MALT- Mucosa-associated lymphoid tissue

MMR- Mismatch repair

MMRD- Mismatch repair deficiency

MNU- N-methyl-N-nitrosourea

MP- Maximum power period

NHEJ- Non-homologous DNA end joining

NER- Nucleotide excision repair

NMF- Non-negative matrix factorisation

NSAIDs- Non-steroidal-anti-inflammatory drugs

OAC- Oesophageal adenocarcinoma

OCCAMS- Oesophageal Cancer Classification and Molecular Stratification Consortium

PCAWG- Pan Cancer Analysis of Whole Genomes Network

POLE- DNA polymerase epsilon catalytic subunit

PPI- Proton pump inhibitors

RAD-Seq- Restriction site associated DNA sequencing

ROS- Reactive oxygen species

RRS- Reduced representative sequencing

SBS- Single base substitution

SNPs- Single nucleotide polymorphism

SNR- Signal-to-Noise Ratio

SNVs- Single nucleotide variations

STR- Short tandem repeat

SigMA- Signature multivariate analysis

TCR- Transcription coupled repair

TE- Tris EDTA

TLS- Translesion synthesis

UMI- Unique Molecular Identifiers

UV- Ultraviolet

WES- Whole exome sequencing

WGS- Whole genome sequencing

dNTPs- deoxynucleoside triphosphate

mutREAD- Mutational signature detection by restriction enzyme-associated DNA sequencing

sWGS- Shallow Whole genome sequencing

5-FU- 5 Fluorouracil

1.Introduction

1.1 Cancer is a "Disease of Genome"

Cancer is a complex disease characterized by alterations in DNA. The transforming role of such alterations (mutations) has been under investigation since these were first reported in the HRAS gene (G>T in codon 12)[1]. With this discovery, the central focus of cancer research was to identify cancer associated genes with mutations that cause or "drive" the cancer. The COSMIC cancer gene census (CGC) has curated approximately 719 of 22,000 (3.2%) coding genes to be associated with cancer and mutations in these genes have also shown to be relevant in disease progression[2,3].

Advent of next generation sequencing technologies has enabled sequencing of whole cancer genomes[4,5]; including both non-coding (introns) as well as coding regions (exons). Sequencing has thus helped to catalogue and better understand the role of different types of mutations in genome, such as single nucleotide variants (SNVs), small insertions and deletions terms indels, large scale or structural rearrangements and DNA copy number changes[6,7,8,4] .

1.2 The concept of mutational signatures in human cancers

Non-inherited or somatic mutations are acquired throughout the life-time of an individual. It has therefore been important to distinguish between driver mutations that are involved in cancer progression, compared with inconsequential mutations associated with ageing termed “passenger mutations”. Mutations in general accumulate over time as a result of exposure to various endogenous and exogenous mutational processes such as defective DNA repair, replication errors, strand asymmetries and exogenous mutagens like UV, tobacco, alcohol and other chemical carcinogens[2,4,9,10]. However, sometimes it can be difficult to distinguish between passenger and driver mutations.

Exposure to different DNA damaging processes leaves a unique array of mutations called “mutational signatures”[11]. The number of mutations, termed the mutation burden, is generally linked to the length of exposure to a specific mutagen. Each mutational process, either from endogenous or exogenous processes, will have its own mutational pattern or foot print that is so specific that it is possible to decipher the events causing them[12] . The final mutational profile of a clinically discernible tumour is complex as it is derived from various mutational profiles accumulated during the life-history of the cancer. Thus the cancer genome can be thought of as an archaeological record of multiple mutational processes [13](Fig 1.1).

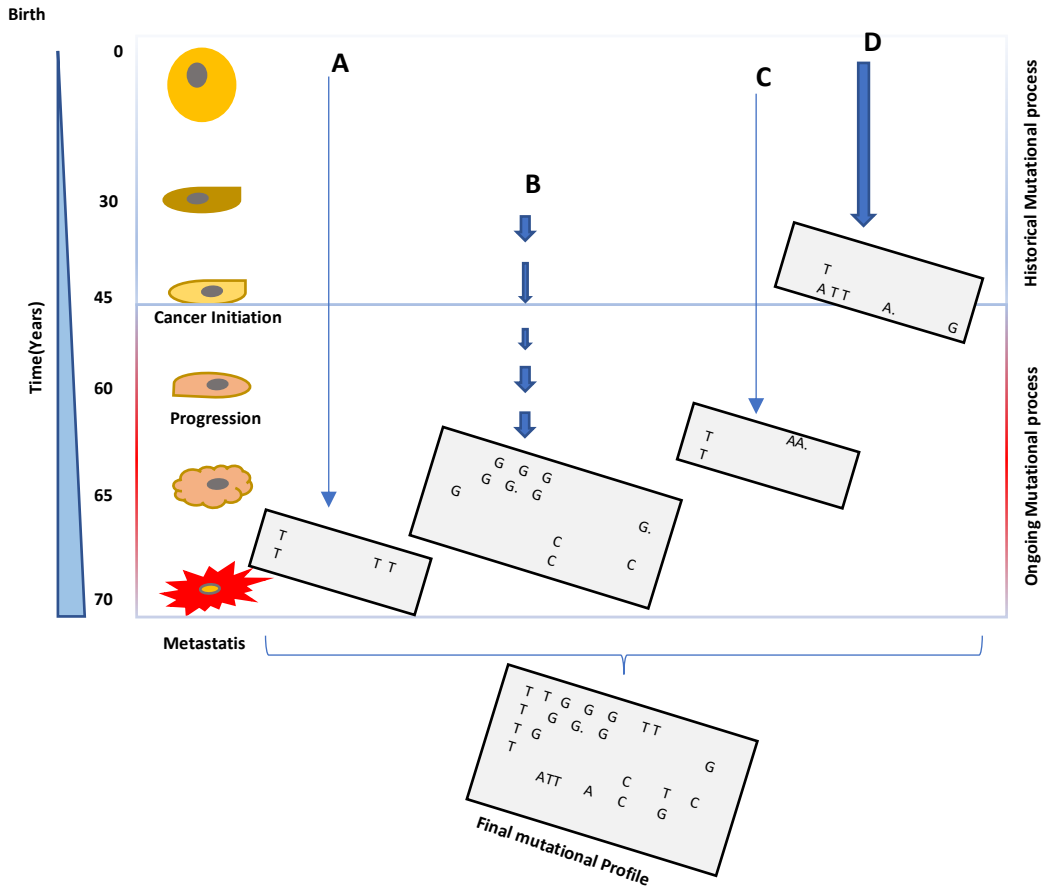


Figure 1.1: Mutational Processes during the phases of Cancer: Events **A**,**B**,**C** and **D** are the mutational processes. Event **A** can be due to deamination of methyl cytosines a life time event. **B** may be due to unknown mutational process(T>G at CTT), **C** can represent the defective mismatch repair like signatures. Where **D** can be due exposure to aristolochic acid. (Adapted from Serena Nik Zainal et al 2017)[14]

Single Nucleotide Variants(SNVs) can be categorized into six possible substitutions C>A, C>G, C>T, T>A, T>C, T>G. From a statistical perspective the mutational patterns can be best described when considering not only the individual base substitution but also the base on either side, the so called trinucleotide context[11,15]. Since, Crick pairing occurs there are therefore 96 possible combinations when a pyrimidine is mutated within its trinucleotide context (Figure 1.2). The proportion of substitutions within a given trinucleotide context can be quantified to determine the most prevalent mutational signatures within different cancer types[9,11,15].

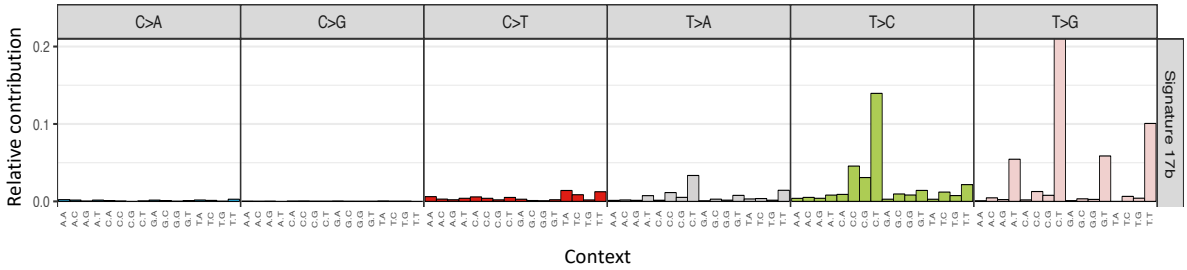


Figure 1.2: Representative SBS mutational portrait : Trinucleotide mutational portrait in 96 base substitution context(x-axis) (six pyrimidine base substitutions labelled on top of the figure and their flanking bases shown on x-axis) and their proportions (y-axis) observed in an oesophageal adenocarcinoma (OAC) patient. Plot was generated using Mutational patterns package in R[16].

1.3 A comprehensive catalogue of signatures of mutational processes in human cancers:

The concept of mutational signatures was proposed in 2012 and reported initially from 21 breast cancer cases[9]. Alexandrov *et al* have developed an algorithm based on non-negative matrix factorization (NMF) and identified 21 to 30 signatures[11]. Recently, the extensive efforts from the International Cancer Genome Consortium-Pan Cancer Analysis of Whole Genomes Network (PCAWG), released a substantial analysis of 4,645 whole genomes and 19,184 exomes across 23,829 cancer samples[15]. Mutational signatures were updated to 47 single base substitution (SBS) signatures in the COSMIC mutational signatures database[17]. The aetiology of these signatures is known for some, but the majority remain unknown. Some of these are associated with endogenous and exogenous exposures like, age, APOBEC enzyme activity, Homologous Recombination-associated DNA repair (BRCA), Ultraviolet light exposure, smoking and others (Table 1.1).

SBS Signature	Predominant feature	Associated mutational process	Cancer types
1	C>T at CpG	Deamination of 5 methyl-cytosine (age associated)	In most of cancers
2	C>T at TpCpN	APOBEC related	Common in cervical and Bladder cancers
3		HR deficient/BRAC1, BRCA2 mutation	Breast, Ovarian and Pancreatic cancers
4	C>A	Tobacco smoking	Head & neck,liver,lung and Oesophageal cancers
5	T>C	Uncertain (age associated)	In most of cancers
6	C>T (and C>A and T>C)	MMR deficient	Colorectal &Uterine cancers
7a	C>T at TCT	UV	Skin, Head & neck cancers
7b	C>T at CCC		
7c	T>A/T>C at TTT		
7d	T>C at GTT)		
8	C>A	amplified by HR deficiency?	Breast and medulloblastoma
9		Polymease η activity	Leukaemias & Lymphomas
10a	C>A	POLE ε mutation	Colorectal &Uterine cancers
10b	C>T		
11	C>T	Temozolomide treatment	Melanoma & Glioblastoma
12	T>C	Unknown	Liver cancer
13	C>G at TpCpN	APOBEC related	Cervical & bladder cancers
14	C>A and C>T	POLE mutataion and MMR deficient	Uterine cancers
15	C>T	MMR deficient	Stomach cancers
16	T>C at ATA	Unknown	Liver cancers
17a	T>G at CTT	Unknown	Oesophageal, stomach, breast, liver, lung & melanoma
17b	T>C at CTT		
18	C>A	ROS/loss of OGG1	Neuroblastoma &Stomach cancers
19	C>T	Unknown	Pilocytic astrocytoma
20	C>A (and C>T and T>C)	POLD1 mutation/ MMR deficient	Stomach & breast cancers
21	T>C at GTA	MMR deficient	Stomach cancer
22	T>A at CTG	Aristolochic acid exposure	Urothelial and liver cancers
23	C>T	Unknown	Liver cancer
24	C>A	Aflatoxin exposure	Liver cancer
25	C>A	Chemotherapy	Hodgkin lymphomas

26	T>C	MMR deficient	Breast, cervical, stomach & Uterine carcinoma
28	T>G at TTT	Unknown	Stomach cancers
29	C>A	Tobacco chewing	Gingivo-buccal oral squamous cell carcinoma
30	C>T	Defective base excision repair-NTHL1 mutation	Breast cancers
31	C>T at CCC	Platinum Chemotherapy	Liver and Pancreatic cancer
32	C>T	Azathioprine	Head SCC and Biliary Adenocarcinoma
33	T>C at TTG	Unknown	Cervix, prostate and Head SCC
34	T>A	Unknown	Oesophageal, Stomach and Breast cancers
35		Platinum Chemotherapy	Biliary Adenocarcinoma and Liver HCC
36	C>A	BER/ MUTYH mutation	Oesophageal SCC and Skin melanoma
37		Unknown	Colorectal and liver cancers
38	C>A	UV?	Skin melanoma
39	C>G	Unknown	Breast cancers
40		Age?	Ubiquitous
41	T>A/T>G	Unknown	Breast and Kidney cancers
42	C>A/C>T	Haloalkanes	Biliary adenocarcinoma
44		MMR deficient	Oesophageal, colorectal and Uterus adenocarcinoma

Table 1.1: Summary of COSMIC SBS signatures with prominent single base substitution type, their associated mutational process and the cancers in which these signatures were commonly reported[11,15,17].

Besides SBS signatures, other base substitution signatures such as 11 double base substitutions (DBS), and 17 small insertion and deletion signatures (ID) were reported in a recent update on pan-cancer data on behalf of International Cancer Genome Consortium (ICGC)[15]. Mutational signatures were also extended to large genomic rearrangements and copy number types, for example 6 rearrangement signatures were first reported in breast

cancers[18] and 7 copy number signatures in ovarian carcinomas[19]. SBS type mutational signatures have been more extensively studied to understand the underlying mutational events in cancer compared to other mutational signature types as their interpretation and validation in experimental models are less complex[20,10,18,21–24].

For Single Base Substitutions(SBS) signatures, analysis is generally performed according to six subtypes (C>A, C>G, C>T, T>A, T>C, T>G) in a trinucleotide context giving 96 classes as previously described[11,15]. In another classification, when two flanking bases at 5' and 3' of the mutated base are considered, it will lead to 1536 subclasses. Small indels are considered when a single base like a C or T is deleted or inserted in a mononucleotide repeat stretch. The length of the repeats in the vicinity of mutation are also considered and 83 subtypes are thus derived.

Non-negative factorization (NMF) based methods have been adapted for analysis. One such approach is SigProfiler[11,15], which was also used in the COSMIC signature analysis and the other was based on a Bayesian variant of NMF called Signature Analyzer[25,26,27,15]. Both of the methods perform consistently with little difference in the number of signatures extracted.

Fifty-two composite signatures have also been extracted taking together the mutation catalogues of SBS, DBS and IDs into 257 subclasses. For example, SBS4, DBS2(CC>AA) and ID3(delC at short runs of cytosine) have been found in lung cancers suggesting exposure to tobacco. In breast and ovarian cancers SBS3 and ID6 combined with ID8 have been identified. These are associated with defects in homologous recombination.

Rearrangement Signatures have been reported in a breast cancer cohort of 560 cases[18]. The deletions, inversions, tandem duplications and translocations were studied in 32 subclasses based on the length of these rearrangements. Usually the rearrangements were present

clustered in a region such as at the zone of gene amplification. When this information was also considered, and rearrangements were categorised as clustered or non-clustered, six such rearrangement signatures were reported[18](Table 1.2).

1.4 Genomic Features and associated mutational events

Mutational processes in the genome are influenced by cellular mechanisms such as DNA replication and transcription. Therefore, the distribution of mutations contributing to the signatures are varied. For example, replication timing plays a role i.e, some of these mutations aggregate in late or early replicating regions. Transcription strand bias is also observed, whereby the mutations are distributed on the transcribed or untranscribed strands preferentially[28,29,30,11]. In general, studies have shown that, substitutions are associated with late replicating regions with some transcription strand bias and rearrangements have generally been linked to early replicating regions in the cancer genomes (Table 1.2). These methods could help to shed some light on mechanistic links and may lead to discovery of the possible aetiology associated with the genomic distribution of mutation classes with replication and transcription changes[31,32,28] (Table 1.2).

Mutational signature	Predominant features of signature	Associated mutational process	Transcriptional strand	Replicative strand	Replication time	Chromatin organization
Single Base substitution mutational signatures						
1	C>T at CpG	Deamination of methyl-cytosine (age associated)		Some bias	Enriched late	
5	T>C	Uncertain (age associated)	Some bias	Some bias	Enriched late	Slight enrichment at linker
2	C>T at TpCpN	APOBEC related	Some bias	Strong lagging strand bias	Enriched late	
13	C>G at TpCpN	APOBEC related	Some bias	Strong lagging strand bias	Flat	
6	C>T (and C>A and T>C)	MMR deficient		Some bias	Flat	
20	C>A (and C>T and T>C)	MMR deficient		Some bias	Enriched late	
26	T>C	MMR deficient	Some bias	Strong bias	Enriched late	Enriched at linker
3		HR deficient	Some bias	Some bias	Enriched late	
8	C>A	amplified by HR deficiency?	Some bias		Enriched late	
18	C>A	ROS?	Some bias	Some bias	Enriched late	Enriched at nucleosomes and periodic
17	T>G	Uncertain		Some bias	Enriched late	Enriched at nucleosomes and periodic
30	C>T	Uncertain			Flat	
Rearrangement signatures						
RS1	Large tandem duplications (>100 kb)	Uncertain type of HR deficiency?	NA	NA	Enriched early	
RS2	Dispersed translocations		NA	NA	Enriched early	
RS3	Small tandem duplications (<10 kb)	HR deficiency (BRCA1)	NA	NA	Enriched early	
RS4	Clustered translocations		NA	NA	Enriched early	
RS5	Deletions	HR deficient	NA	NA	Enriched early	
RS6	Other clustered rearrangements		NA	NA	Enriched early	
Repeat-med	<3 bp indel at polynucleotide repeat tract	amplified when MMR deficient	NA	NA	Enriched late	Enriched at linker and periodic
Microhom	≥3 bp indel with microhomology at breakpoint junction	HR deficient	NA	NA	Enriched late	

Table 1.2: Summary of single base substitution signatures and rearrangement signature associated with genomic features in breast cancers (Modified from Nik-Zainal *et al* 2016[18,28]).

Distribution of SNVs across the regulatory regions of the genome provides hints of possible mutational processes linked to them. For example, a study of nucleosome periodicity in human cancers[33] shows how the somatic mutation rate exhibits a unique 10bp periodicity within nucleosomes, which follows the alternation of DNA minor groove facing toward and away from the histones. Mutational events govern the phase and strength of the mutation rate periodicity(Figure 1.3).

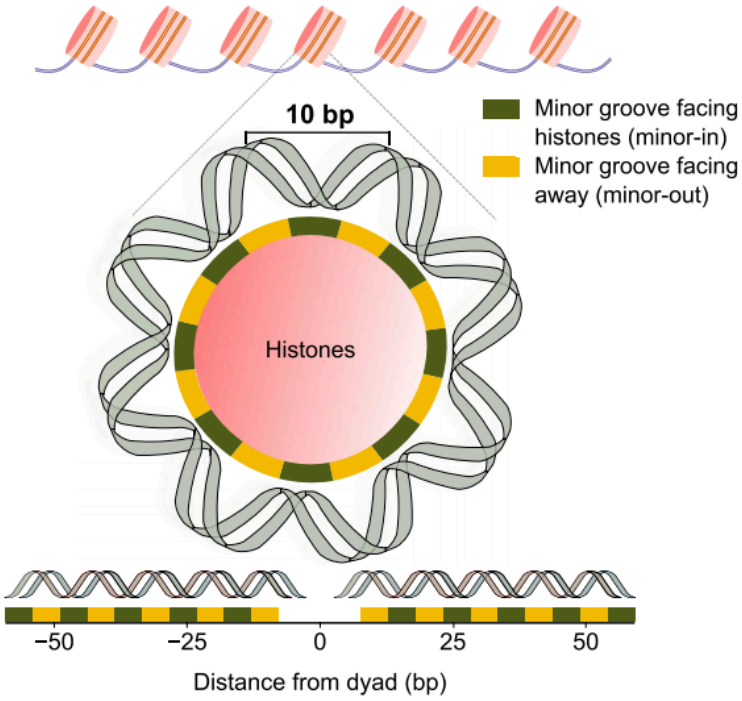


Figure 1.3: Nucleosome periodicity: Mutation Rate Periodicity between Minor-In and Minor-Out Nucleosome-Covered DNA Stretches (Adapted from Pich *et al* 2018[33])

A consistent 10bps periodicity patterns is followed between minor and major DNA grooves.

1.5 DNA damage as a result of exogenous exposures and impairment of endogenous processes

Mutational signatures are the consequences or result of specific DNA damaging exogenous exposure and impairment in associated repair pathways, which are critical in shaping the mutational signatures (Table 1.3).

DNA damage	Mutational Process (DNA repair pathway)	Mutational Signatures
C:G>T:A at methylated CpGs Correlates with age	Replicative polymerases(Deamination)	1A
APOBEC editing	Base excision repair with A-rule or with excess of deoxycytidyl transferease (REV1)?	2,13
UV radiation on pyridimines and dipyrimidines	Transcription coupled repair	7
Temozolomide induced O^6-methyl-guanine lesions	Direct repair using methylguanine DNA methyltransferease	11
Benzo[α]pyrene (B[α]P)adducts on guanine	Transcription coupled repair	4
Aflatoxin adducts on guanine	Transcription coupled repair	24
Aristolochic acid (AA) adducts on adenine	Transcription coupled repair	22
Natural errors	Mismatch repair pathway	6,20
Natural errors	Defective DNA polymerase ε	10

Table 1.3: Examples of some endogenous and exogenous mutagen induced DNA damage their affected repair pathway and associated mutational signatures.

In order to more precisely understand the interaction between exposures and repair processes a detailed study of 79 environmental agents exposed to iPSCs (Human induced Pluripotent Stem Cells), were reported recently. Forty one out of fifty three (41/53) agents generated stable SBS signatures with a good cosine similarity with the COSMIC mutational signatures[34] (Table 1.4), suggesting that the extent/dose of exposure to a mutagen, cell type

and their DNA repair efficiency may contribute towards generation of specific mutations. Also, mutational signatures may be an outcome of combined exposures, or the model system may not recapitulate the *in vivo* physiology.

Environmental Agent	**Dose**	**Category**	**Mutational Signature (Cosine similarity)**
Ellipticine	0.375uM	Drug therapy	8(0.83)
Temozolomide	200uM		12(0.83),21(0.83),26(0.89)
Mechlorethamine (nitrogen mustard)	0.3uM		30(0.8)
6-Nitrochrysene	12.5uM	Nitro-PAHs	16(0.8)
Aristolochic acid I	1.25uM	others	22(0.99)
3-Chloro-4-(di-chloromethyl)-5 hydroxy-2(5H)-furanone	7uM		24(0.8)
AZD7762(CHK inhibitor)	1.625uM	DNA Damage response inhibitor	25(0.82)
N-methyl-N-nitrosourea (MNU)	350uM	Alkylating agent	26(0.87)
Simulated Solar Radiation	1.25J	Radiation	7(0.94), 11(0.83)

Table 1.4: Representative environmental agents from different categories associated with stable SBS mutational signatures.

The study of mutational signatures in a disease helps to understand the history of mutational events that lead to precise DNA damage. These insights can help discover biomarkers for diagnostics, stratification, and potential targets for therapeutics. There are some cancer types such Oesophageal adenocarcinoma (OAC) with poor survival and limited therapy options. OAC is characterised by a high mutational burden[35,36] and is a good example for mutational signature based study approaches.

1.6 An ideal model for studying mutational signatures in cancer evolution is Barrett's and Oesophageal adenocarcinoma: Epidemiology and clinical pathology

Oesophageal cancer has two major subtypes, squamous cell carcinoma (ESCC) and adenocarcinoma (OAC). It is the sixth most common cause of death and eighth most frequent cancer type worldwide[37]. The ESCC subtype accounts for about ~90% cases globally and OAC of about ~10% with higher mortality rates than squamous cell carcinoma[38]. However, in the western world the incidence of OAC has increased 6-fold in the last 30 years and it is now the most common form of oesophageal cancer in white men and women in the UK, USA and Australia as well as in some parts of Europe.

Oesophageal cancer including OAC tends to present at a late stage when metastases have already occurred leading to poor survival, with a median overall survival of less than a year[39]. The disease tends to occur in older age, being most commonly diagnosed in individuals in their late 60s with a median age of 67 years. For OAC, race and sex are also risk factors such that white males are at higher risk than black women which is different for ESCC where Black and Asian men are more prone to the disease [40,41]. Common environmental risk factors are chronic gastro-oesophageal reflux disease (GERD), smoking and obesity[41,42]. These risk factors usually act over a long period of time during the genesis of the tumour and may have a substantial and unique influence on the development and progression of the disease for an individual patient. As these lifestyle exposures will leave a footprint on the genome these changes can be helpful to understand how the cancer developed and may even have clinical utility[11]. Although mainly linked with environmental factors, there is likely also a contribution

from germline predisposition in view of the clustering in some families and association of germline SNPs[43].

OAC is commonly preceded by Barrett's Oesophagus (BE) which is a metaplastic precancerous stage, with 0.12% to 0.5% progression per year to oesophageal adenocarcinoma[44]. Multistage progression from Barrett's to low grade dysplasia, high grade dysplasia and adenocarcinoma makes this disease an ideal study model to delineate the mechanisms causing the transformation from a precancerous lesion to OAC.

1.7 Somatic mutational signatures in Oesophageal adenocarcinoma (OAC)

Previously our laboratory has reported six commonly observed SBS signatures in an OAC cohort of 129 cases[35](Figure 1.4). Signature 17 (T>G, at CTT) was found in more than half of the cohort (53.3%). Signature 17 was reported as two subtypes. S17a with T>C substitutions in the CTT context and another S17b with T>G in the same context. The other signatures identified were: aging associated signature 1, reactive oxygen species linked signature 18, homologous recombination (HR) deficient associated signature 3 and APOBEC related signature 2. In this previous study the mutational signatures were used to classify the cohort into three molecular subtypes. The most prevalent of the three is the mutagenic subtype with T>G substitutions, signature 17. A homologous repair deficient, signature 3 (BRCA signature subtype) and age-related C>A/T molecular patterns-signature 1. The recent PCAWG mutational signature analysis with 97 OACs showed similar signatures with additional new signatures such as single base substitution signature SBS40, a double base substitution signature DBS8 and indel signatures ID1 and ID2 with unknown aetiologies[45].

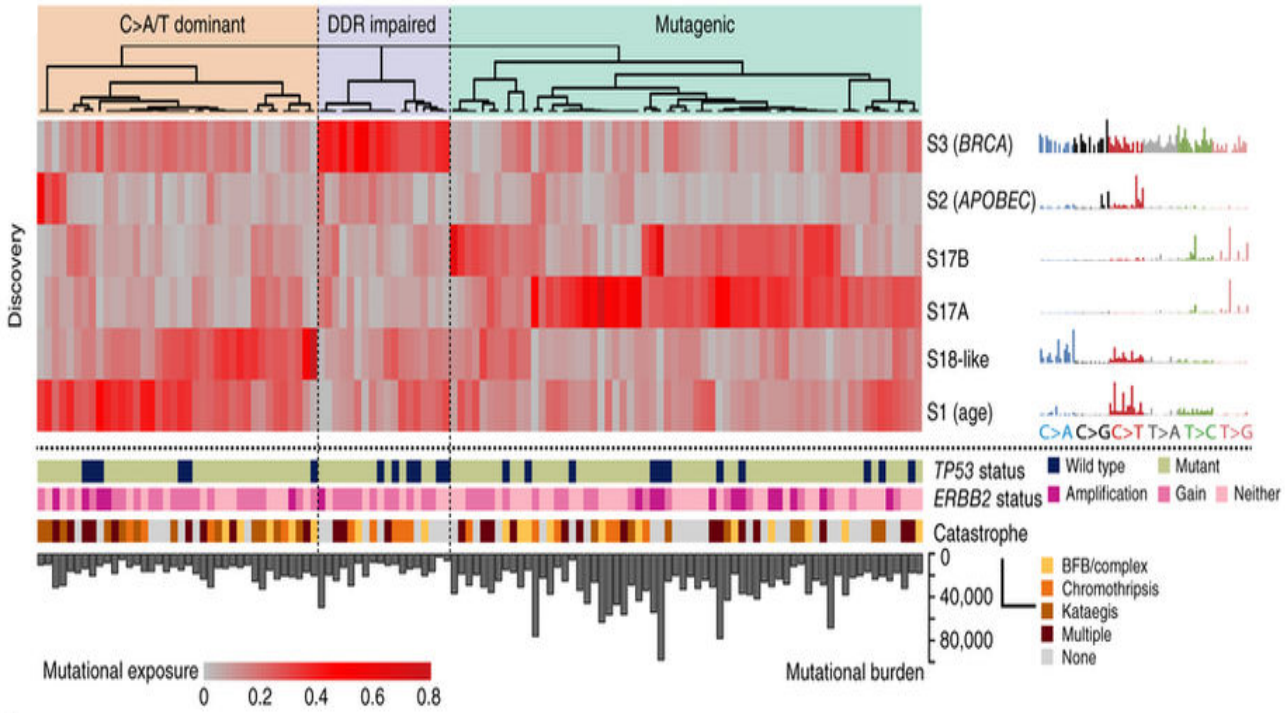


Figure 1.4: Predominant Mutational signatures in OAC: Six Single Base Substitution signatures were reported in 129 OACs (Adapted from Secrier *et al* 2016 [35])

1.8 Mutational signatures in the evolution of a tumour.

The evolution of a cancer can be traced by studying the somatic mutations acquired in normal and precancerous tissues in molecular time and space. A few of the somatic mutations are subject to selective pressures which initiate sustained clonal expansions (driver events). These clonal expansions further divide into sub-clonal expansions, when exposed to different mutagenic events or selection pressures and this cascade continues with accumulation of exposure specific somatic mutations, encrypting patterns of DNA damage in the genome. It is thought that this sequence of events leads to genome instability and tumour transformation[46,47,9,48].

Positive selection of a mutation will reduce the variability in the mutation profile which is called a selective sweep. Thus, the mutation is shared in all cells of the clone. This mutation

emerges as a common ancestor during evolution. So, the phylogenetic tree of these events can be built computationally by considering mutations before and after the common ancestor. Recently mathematical algorithms have been designed for investigating the clonal history of mutational signatures in a cancer genome[9,20,49,50,51]. This knowledge of clonal evolution of mutational signatures will help understand the early and late events in the tumour progression. These methods can be used to study the evolution of mutational signatures from Barrett's to OAC.

Previously our laboratory has reported whole genome sequencing analysis of 23 Barrett's and OAC pairs[48]. The single nucleotide variants (SNVs) were categorised based on the degree to which mutations overlapped or were shared into early (present in BE and OAC) and late SNVs (OAC only). A particular pattern driven by A:T>C:G at AAG context was shared between Barrett's and OAC. Presence of such unique and shared patterns suggests a common mutagenic-processes active in the tumourigenesis[48,52]. Also, the differences between these two states were driven by A>C at CAC; C>G at CCG; A>T at CAC and C>G at TCA trinucleotide contexts (Figure 1.5). However, a further detailed analysis of clonal dynamics of mutational signatures during development of OAC from Barrett's in much larger cohort will help better understand the disease progression.

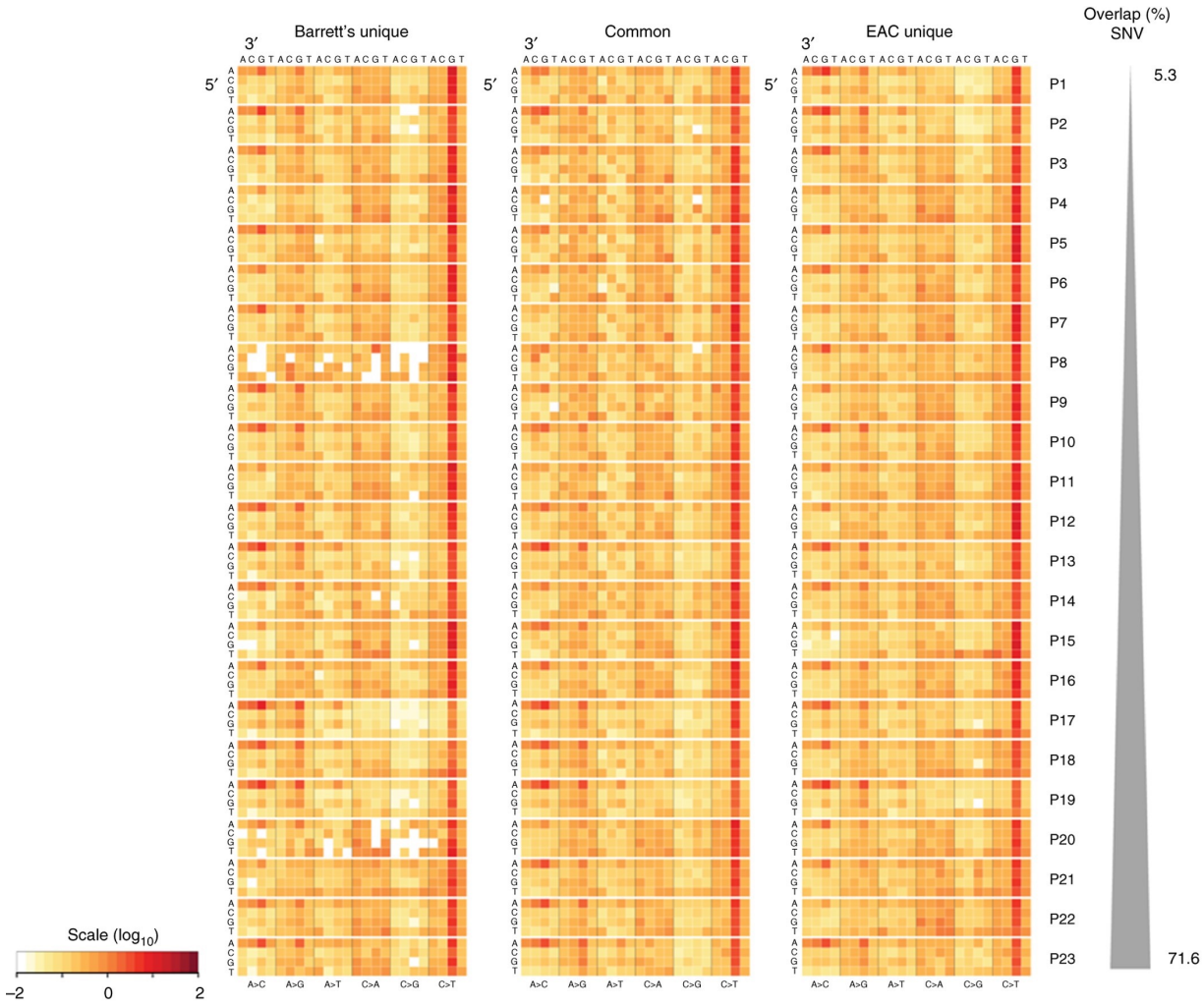


Figure 1.5: Shared and Unique SNVs between Barrett's with OACs : Summary of frequency of specific SNVs at all possible trinucleotide contexts (adapted from Ross-Innes C.S *et al* 2015[48]).

1.9 SBS 17 (T>G/C transversions at CTT trinucleotide context) in Oesophageal adenocarcinoma.

Signature 17 is characterised by T>G transversions at CTT tri nucleotide context. It predominates in almost half of the OAC cohort (53.3%), and is also observed early in the disease in Barrett's oesophagus[35,48]. The aetiology of this signature is poorly understood.

Understanding the possible causes of Signature 17 will likely help understanding how this cancer type arises and possibly inform sub-classification and options for cancer prevention strategies.

Signature 17 has been reported to be associated with some genomic features, including a bias towards the lagging strand and to be linked to late replicating DNA, where there are possible chances of replication-based errors which might have contributed to the mutations. There is also bias towards the untranscribed strand, suggesting impaired transcription coupled repair (TCR)[29].

Nucleosome periodicity patterns shows that Signature 17 occurs on the DNA groove facing towards the nucleosome suggesting a possible endogenous mutational process involvement in generation of these SNVs[53].

Chemotherapy with 5-Fluro Uracil (5-FU) treatment in tumours has also been associated with signature 17. These T>G substitutions were studied in metastatic tumours and in studies on chemotherapeutic exposure to Capecitabine (a 5FU derivative) in a Leshmania model[24]and 5FU exposure in intestinal organoids[54]. However, we have also reported that signature 17 is also observed in chemo-naïve tumours[35].

Since acid reflux is the most clearly defined risk factor for OAC, this signature has been dubbed the “acid signature” though this has not been causally proven. The role of gastro-oesophageal acid reflux is thought to be important in causing oxidative DNA damage. It has been suggested that reactive oxygen species (ROS) results in oxidation of the nucleotide pool and their mis-incorporation during replication[55,56]. Oxidation of Guanines is abundantly observed in dysplastic Barrett’s cells and on incubation of Barrett’s tissue with a cocktail mimicking bile acid [57,58,59]. The oxidation of Guanines in the double helical DNA strands will lead to 8-oxo-Guanine, and in the nucleotide pool will alter dGTP to 8-oxo-dGTP. For example, the

mutagenicity of 8-oxo-Guanine in DNA has been shown to cause C>A substitutions upon mispairing with Adenine during replication in E.coli[100]. Recently, chemical induction of ROS by Peroxynitrite and Potassium bromate in human iPSCs resulted in signature 18 like C>A substitutions[34]. Mis-incorporation of dGTP during replication by Translesion DNA polymerases will cause T>G transversions, as reported in HEK cells[60,61,62,63,64].

DNA base excision repair pathway enzymes such DNA glycosylases (OGG1 and MUTYH) prevent the mispairing. OGG1 prevent 8-oxoG to A mispairing and corrects to G to C pairing. The mispairing of 8-oxo-dGTP has been shown to reduce in the absence of MUTYH and cause C>A mutations[64]. Recently CRISPR-Cas9 based biallelic knockouts of DNA repair genes such as OGG1 and MUTYH resulted in G>T patterns at TGC>TTC, which is the mutational signature 18 associated with reactive oxygen species[22]. So, T>G at CTT substitutions might be caused by a combination of endogenous and exogenous factors.

1.10 Mutational Signatures detection in the Clinic.

Stratification of patients based on mutational subgroups may help to tailor therapies to improve survival[65]. As discussed, our laboratory has reported three subtypes in an OAC cohort of 129 patients[35]. However, application of high-depth WGS for signature-based patient classification for clinical use would be expensive. Furthermore, performing this in fresh frozen samples would present some logistical challenges due to the infrastructure required and since clinical practice currently relies on FFPE tissues. Therefore, a cost effective sequencing assay that could be readily applied to formalin fixed paraffin embedded tissues would be a major advance.

A few different computational algorithms are being developed to identify signatures. These include a method based on a lasso logistic regression model called HR- detect

that was reported to identify BRCA signatures and MMR detect for micro satellite instability in tumours[66,22]. Also, an algorithm based on multivariate analysis of signatures using the targeted panel data called SigMA is also developed[67]. These algorithms are dependent on WGS/targeted panel data and are not a cost-effective alternative solution for screening in larger cohorts.

The low coverage methods such as 10x sequencing are mostly used to detect copy number variations and single nucleotide polymorphisms (SNPs). Due to the low coverage this method is most likely to detect only mutations with high variant allele frequencies.

Exome sequencing is mostly used to identify SNVs that might disrupt gene function. It has been argued that the SNV frequencies in coding regions differ quite dramatically compared to the whole genome, possibly due to transcription-coupled repair. This might introduce a bias in the estimation of the exposure.

A method called reduced representative sequencing (RRS), also called genotyping by sequencing, uses restriction enzymes and a size selection step to target the sequencing power to random but reproducible set of regions in the genome. It is currently mostly used as part of a protocol for DNA methylation analysis and for SNP detection. This protocol offers the flexibility to choose which regions to sequence and the choice becomes a trade-off between sequencing as little of the genome as possible but covering as many SNVs as possible. It has the potential to circumvent the drawbacks of the other two methods in estimating the exposures (proportion of SNVs contributing for a signature) of the signatures. However, the resulting data might offer only a coarse picture of driver mutations in genes or CNVs. RRS is also referred to as restriction-site-associated DNA sequencing (RAD-Seq). There are several RAD Seq modified assays which were developed for SNP genotyping and phylogenetic analysis in plants and amphibians (Table 1.5)[68,69].

Protocols	No. of enzymes	Cut frequency	Shearing required	Size selection	Library prep time &required expertise	Subsequent library cost per sample
ezRAD (Any restriction enzyme based RAD seq)[70]	1 or more	Frequent	No	Yes	Low	Moderate
RAD tags (Restriction site associated DNA Sequencing)[71]	1	Rare	Yes	Yes	High	Low
GBS (Genotyping by Sequencing)[72]	1	Rare or frequent	No	No	Moderate	Moderate to very low
2-enzymeGBS (Genotyping by Sequencing involving two restriction enzymes)[73]	2	Rare + frequent	No	No	Moderate	Moderate to very low
ddRAD (double digest RAD seq)[74]	2	Frequent	No	Yes	Moderate	Very low
2b-RAD (Use of typeIIB restriction enzymes)[75]	1	Frequent	No	No	Moderate	Low

Table 1.5: Summary of Various RAD protocols for reduced representative sequencing.

Hypothesis:

For my PhD study I hypothesised that a more in-depth characterisation of the mutational signatures in a larger cohort of oesophageal adenocarcinoma cases, with addition of pre-malignant samples from Barrett's oesophagus and more advanced metastatic lesions will provide additional insights into the evolution and aetiology of this cancer with relevance to clinical management. Further, development of a cost-effective sequencing assay for ascertaining of mutational signatures from FFPE material will improve clinical applicability with potential to improve patient stratification strategies for prognosis and therapy in oesophageal adenocarcinoma.

Aims

Aim1: Comprehensive analysis of mutational processes in Barrett's, OAC and Metastasis.

This aim comprises several sub-aims:

- To characterize mutational signatures in the cohort of OACs, Barrett's and Metastasis;
- To determine the clonal evolution and timing of mutational signatures;
- To perform an analysis of genomic features including transcription, replication strand biases and nucleosome periodicity;
- To define the clinical characteristics and their possible associations with Barrett's and OAC mutational signatures.

Aim2: To develop a robust, cost-effective assay to identify mutational signatures from clinical samples (mutREAD) with application to formalin fixed and paraffin embedded samples.

Aim3 : To validate mutREAD in additional OACs (25 Cases with matched biopsy and resection tumours as available).

2.Methods

2.1 Study cohort

A cohort was assembled comprising 161 Barrett's, 777 OACs and 59 metastatic samples that had been collected through a multicentre UK wide study called OCCAMS (**O**esophageal **C**ancer **C**lassification **A**nd **M**olecular **S**tratification) and which have undergone whole genome sequencing (WGS) as part of the ICGC-International Cancer Genome Consortium. The study was approved by the Institutional ethics committee (REC 07/H0305/52 and 10/H0305/1) and included individual informed consent. I have also included 228/777 OACs from Mutograph project. Clinical data for tumours from Mutographs project is incomplete and is being collected (Figure 2.1). The samples from Barrett's and OAC's were procured from different patients as available, some of these are pairs (Barrett's-OACs) from the same patient. Metastasis samples were collected as available from OAC patients. A sample from the Barrett/tumour/metastatic sample was always matched with a germline reference, which was ideally matched blood or if not available normal squamous oesophagus as far away from the tumour as possible (at least 5cm) collected during surgical resection or at endoscopy. All samples were snap-frozen.

Prior to sequencing a systematic pathological review was performed by a Consultant Histopathologist to check the cellularity of the tumour samples using hematoxylin-and eosin-stained sections, and only samples with $\geq$70% cellularity were included. DNA was extracted from frozen tumours using the Allprep DNA/RNA mini kit (Qiagen, Hilden Germany) and DNA from blood was isolated using QIAmp DNA blood maxi kit (Qiagen, Hilden Germany).

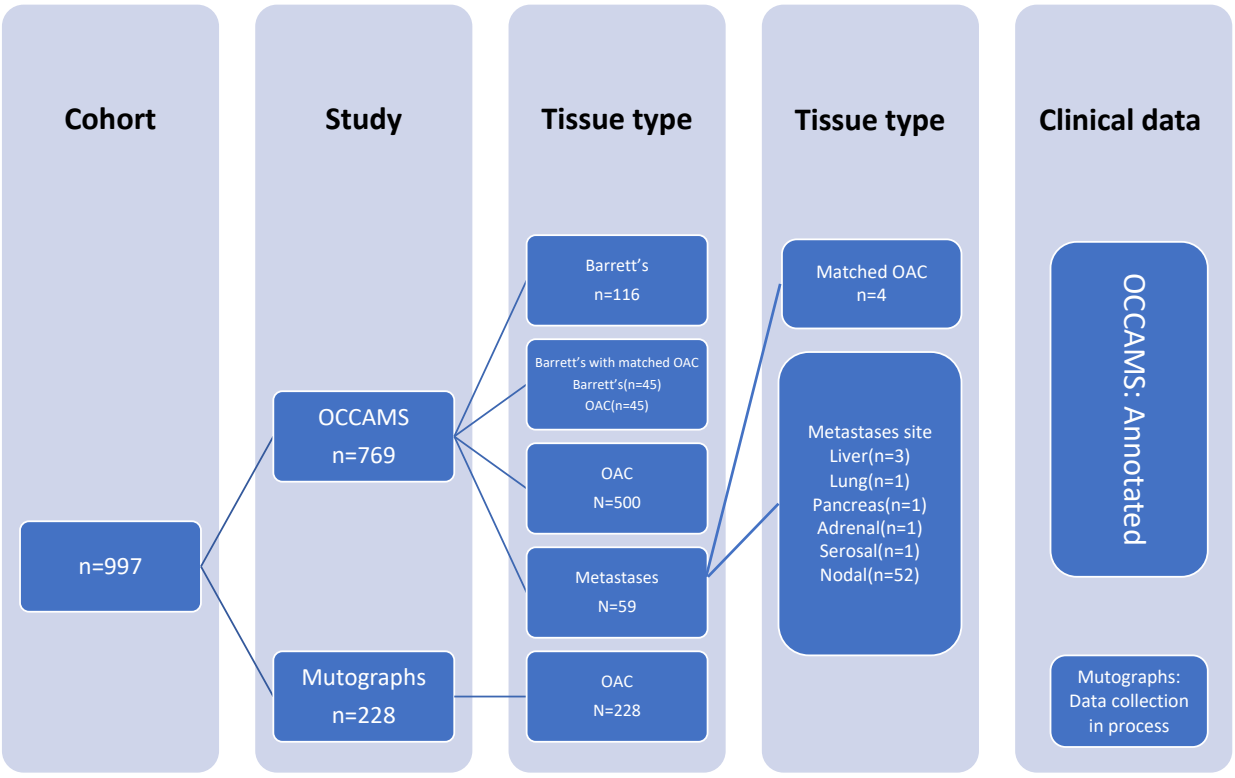


Figure 2.1: Consort diagram of the cohort: Overview of the study cohort describing details of number of samples, their respective studies, tissue type and clinical data availability.

2.2 Whole genome sequencing and mutation calling

100bp paired-end Whole Genome Sequencing(WGS) at 50X depth for tumours and 30X for matched normal (blood) was performed under contract by Illumina (San Diego,US) as part of the International Cancer Genome Consortium. Quality checks were performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and in-house tools.

For mutation calling, sequencing reads were aligned against the reference genome (hg19/Ensembl GRCh37) using the latest version of Burrows-Wheeler alignment algorithm, BWA-MEM. Aligned reads were then sorted into genome coordinate order and duplicate

reads were removed by using Picard (http://broadinstitute.github.io/picard). Strelka 1.0.13 software (Saunders 2012) was used for calling single nucleotide variants and Indels. Functional annotation of the resulting variants was performed using Variant Effect Predictor (VEP release 75).

2.3 Mutational signature discovery

Mutational signature discovery in the cohort was performed using SigProfilerExtractor [76]. The optimal signature configuration in the cohort was selected from a range of signature combinations from 5 to 17 based on the highest stability and lowest Frobenius reconstruction error for a signature combination. A total of 14 signatures were identified as the optimal configuration, and this was confirmed by independent analysis using the Bayesian methodology from Sigminer[77]. Once the main mutational processes in the cohort were defined, we used deconstructSigs [78] to infer the mutational contributions of these processes to each sample.

2.4 Transcription/Replication strand bias:

The MutationalPatterns package was used to map the SNVs to either the transcribed or untranscribed strand. Likewise for replication bias, SNVs were assigned to lagging or leading strands. Non negative matrix factorization (NMF) was performed on the matrix of annotated SNVs to respective strands and mutational signatures and their asymmetry was examined[16].

2.5 Nucleosome periodicity

This part of analysis was performed in collaboration with Nuria Lopez-Bigas laboratory at Institute for Research in Biomedicine(IRB) Barcelona, using methodology adapted from their previous study (Pich O *et al* 2018)[53].

In brief, the nucleosome positioning was obtained from the MNase Seq data (Gaffney.D J *et al* 2012)[79] and mapped to human reference genome. Phasing annotations for minor groove facing histones or away was obtained from (Cui.F and Zhurkin.U B *et al* 2010)[80]. Somatic mutations from the WGS data were mapped to the nucleosome positions and periodicity of change in mutation rate was calculated and plotted as described in (Pich.O *et al* 2018)[53].

2.6 Mutation clonality and timing analysis

To infer subclonality of mutations and mutational processes, first the likelihood for any sample containing subclonality was assessed on the distribution of purity-corrected variant allele frequencies, using the Hartigan's dip test[81]. Samples with no significant evidence of deviation from unimodal distribution were deemed as fully clonal. The remaining samples were assumed to contain subclones.

Next, MutationTimer[50] was used to infer the timing (early/late) of every mutation called in each genome as follows: for samples that were assumed to be fully clonal, MutationTimer was ran with default parameters (minimal read support = 3, 0 dispersion) and 100 bootstrap iterations; for samples with evidence of subclonality, MutationTimer was ran with modified input specifying the expected subclonal proportions (calculated from a Gaussian mixture

model with two components) and inferred both the clonality and timing of mutations. In both cases, the analysis was performed in a whole-genome doubling conscious manner.

We then used the MutationTimer results to split the mutations into clonal/subclonal and early/late and performed mutational signature inference using deconstructSigs[78] again on these separate populations. This allowed me to infer a time and clonality-depedent mutational prevalence of various signatures.

Finally, we corroborated the clonal composition results using TrackSig[51], which identifies cancer cell fraction bottlenecks where mutational signature proportions change. The cases where we observed at least one bottleneck were in agreement with cases where we observed subclonality using the approaches described above.

2.7 DDR signature discovery

To uncover signatures with DDR impairment in the cohort, we examined nonsynonymous mutations accumulated in >500 genes across 13 DDR pathways as described in Theo A *et al*. 2018 [82]. We employed NMF via the NMF package[83] in R to extract patterns of mutations accumulated in these pathways based on the following features: total mutation count per pathway per sample, total number of clonal/subclonal mutations and total number of early/late mutations. The optimal number of signatures in the cohort (5) was chosen based on the cophenetic coefficient statistic.

2.8 Positive selection

Groups were defined based on mutational signature dominance, as follows: samples where S17a+S17b contributed the majority of mutations in a sample were classed as “S17

dominant"; the rest of the samples were categorised as "Other dominance". The dndscv tool[84] was run separately on samples from the individual groups in order to infer genes that were under positive selection in the respective group. Finally, genes under positive selection were compared between the groups with/without dominance of a particular mutational signature, and common as well as specifically selected genes were extracted. Among these, cancer driver genes were identified by cross-referencing against the COSMIC Cancer Gene Census database[3]. For genes which had not previously been documented as cancer drivers, we used the GTeX database[85] to confirm their expression in oesophageal/gastric tissue. Olfactory receptors were discarded from the analysis as they are believed to be spurious hits.

2.9 Machine learning for OAC stage classification

We used a gradient boost classifier as implemented by the xgboost package(Version 1.4.1.1) in R to train two models to distinguish Barrett from primary tumours, and primaries from metastases, respectively, based on prevalence of all mutational signatures and including clonality and timing as covariates in the model. We split the cohort into 70% for discovery and 30% for validation, and used 5-fold cross-validation in 100 iterations to determine the optimal parameters for the training. The features ranked by importance were visualised using a Shapley plot. The modelling procedure was repeated in a similar manner but with prevalence of signatures detailed based on clonality and timing. The accuracies for testing were 87% and 94%, respectively. The analysis employed the code developed at the following github repository: https://github.com/pablo14/shap-values/blob/master/shap.R.

We also built a multinomial regression model which took as features mutational signature exposures, timing and clonality of signatures and trained a classifier on 70% of the data to predict the stage of the tumour (with the 3 stages of Barrett's, primary, metastases, predicted

simultaneously). We used the remaining 30% of the cohort to validate the model. This analysis was implemented using the glmnet package[86] in R.

2.10 RNA Seq

RNA was quantified using the Qubit High Sensitivity RNA kit (Thermo Fisher) and checked for quality (RNA integrity number; RIN) on the Agilent 2100 Bioanalyzer® (Agilent Technologies, USA) using the RNA 6000 Nano kit. Samples with insufficient material, or an incalculable RIN were excluded. There was no other lower limit for RIN inclusion.

Libraries were prepared with an input of 250ng RNA using the TruSeq Stranded Total RNA High Sensitivity protocol with ribosomal depletion. Samples with less than the specified input, but with >100ng total were included and this was noted for the analysis. Library quality and quantity were checked using the Agilent 2100 Bioanalyzer with the DNA 1000 kit and KAPA quantification (KAPA Biosystems, Roche, Switzerland), and were pooled according to the Illumina protocol. Samples were run on the HiSeq 4000 instrument to generate 75bp paired-end reads. A mixture of normal expression controls was run on each plate: squamous oesophagus, gastric cardia, duodenum. Barrett's oesophagus is a mosaic of gastric and intestinal cell types. Therefore, duodenum and gastric tissues are used as a control. Squamous oesophagus is a less useful comparison because it shares few features with the glandular epithelium of Barrett's, but it was included as an adjacent tissue.

RNA sequencing data was trimmed for poor quality bases using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and was then aligned using STAR using the ENSEMBL gene annotation. Reads per gene were quantified using the summariseOverlaps function from the GenomicRanges package[87], which was also later used for computing Transcripts per million (TPM).

2.11 Hallmarks of cancer and tumour microenvironment signatures

The cancer hallmark signatures were obtained from the CancerSEA database[88]. The tumour microenvironment signatures and composition were inferred using ConsensusTME [89].

2.12 Statistics

Group comparisons were performed using the Student's t test, Wilcoxon rank-sum test or ANOVA, as appropriate. Multiple testing correction using the Benjamini-Hochberg method was performed where appropriate.

Survival analysis was performed using univariate or multivariate Cox Proportional Hazards models as implemented in the ggforest R package. The optimal prognostic cut-offs for mutational signatures were determined using the maximally selected rank statistic, as implemented in the survminer package(Version 0.4.9) in R. Kaplan-Meier curves were plotted using the survminer package(Version 0.4.9).

2.13 Restriction enzyme selection criteria

The enzyme combination is an important parameter to optimize for the mutREAD (**mut**ational signature detection using **RE**striction enzyme **A**ssociated **D**NA sequencing) method. We focused on high-fidelity restriction enzymes provided by New England BioLabs Inc. (Ipswich, Massachusetts USA) to allow for fast DNA digestion and maximum target specificity under a broad range of experimental conditions. Since cancer samples frequently exhibit DNA hyper- or hypo-methylation, which could affect restriction enzyme sites, we required insensitivity to

CpG methylation status. To simplify the adapter design, only enzymes with a unique cut-site including only A, C, G and T were considered. Finally, cut sites were required to have a maximum length of six base pairs to increase the number of generated fragments. The tested list of enzymes is given in results chapter (mutREAD).

2.14 Simulations

We opted for a double-digest protocol to produce fragments that are reproducible between libraries. To simulate the performance of all possible enzyme combinations full-filling the above criteria, we use ddRADseqTools (v0.45) to perform *in silico* digestion of the human hg19 reference genome and size selection for fragments of expected length between 350-450bp. The expected fragment size range of 350-450 base pairs was chosen as the maximum fragment size such that the complete library fragments (insert, adapters and primers) could still be sequenced on a standard Illumina HiSeq system. WGS-based mutations were selected if they overlapped the resulting expected fragments and mutational signatures were calculated based on this selection. Similarly, Whole Exome Sequencing (WES) and expanded WES sequencing is simulated using the target regions provided by Nextera for the rapid capture exome/expanded exome kit (v1.2), where the exome kit comprises 45Mbps of coding regions and the expanded exome kit comprises 62Mbps of coding regions, untranslated regions and miRNAs. Further, the 21 simulated 10x shallow Whole Genome Sequencing (sWGS) libraries from a previous study were used. In short, the 10x sWGS were simulated by down-sampling the WGS libraries and re-running the mutational calling.

2.15 Cosine Similarity

We measure similarity between two mutational signature profiles P and Q using the cosine similarity. The cosine similarity between the non-zero vectors P and Q with n mutational signatures is defined as $cossim(P,Q) = \frac{\sum_{i=1}^{n} P_i Q_i}{\sqrt{\sum_{i=1}^{n} P_i^2}\sqrt{\sum_{i=1}^{n} Q_i^2}}$. Two mutational signature profiles that are independent have cosine similarity of 0. Conversely, identical mutational signature profiles obtain a cosine similarity of 1.

2.16 Computational simulations using Pan-Cancer Analysis of Whole Genomes data

Computational simulations on the WGS data from the PCAWG network was performed. The collection was downloaded from https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel. We have used the signature compendium from COSMIC (v3, downloaded from https://dcc.icgc.org/releases/PCAWG/mutational_signatures/Signatures/SP_Signatures/SigProfiler_reference_signatures) to capture all mutational signatures relevant to the different cancer types. Only cancer types with at least 10 samples present in the collection were analysed.

2.17 Cell lines/OAC patients DNA used for mutREAD

All optimization experiments were performed using 500 ng of genomic DNA from an OAC cell line (FLO-1), commercially available from culture collection of Public Health England. In-house STR analysis was done in the lab to confirm a >90% match prior to assay optimization. Experiments were then repeated with frozen tumour, matched blood and FFPE tumour DNA from OAC patients.

2.18 DNA extraction and Quantification

DNA was extracted from FLO-1 cell line and frozen tumours using the Allprep DNA/RNA mini kit (Qiagen, Hilden Germany) and DNA from blood was isolated using QIAmp DNA blood maxi kit (Qiagen, Hilden Germany). AllPrep DNA/RNA FFPE Kit (Qiagen, Hilden Germany) was used to extract DNA from FFPE tumours. DNA quantification was done using Qubit dsDNA Broad Range (BR) assay kit on Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham Massachusetts USA).

2.19 Somatic mutation calling for mutREAD

Mutation calling was performed using GATK Mutect2, taking into account for the SNV metrics: only reads with minimum mapping quality of 1, minimum base quality of 10 and excluding supplementary alignments, as well as discarding both reads in an overlapping read pair if they have different base calls at the locus of interest, or using just the read with highest base quality if they have the same base.

Additionally, Strelka (v 2.0.15) with disabled read depth filter was run on a subset of samples, taking into account for the SNV metrics only reads with minimum mapping quality of 1, minimum base quality of 10 and allowing a minimum alternate allele count of 2 and a minimum alternate allele frequency of 0.05 for a position to be considered in detecting SNV clusters.

For Mutect2- and Strelka-derived mutations, low-quality and spurious mutation calls were filtered by applying the following criteria: VariantAlleleCountControl > 1, VariantMapQualMedian < 40.0, MapQualDiffMedian < -5.0 || MapQualDiffMedian > 5.0, LowMapQual > 0.05, VariantBaseQualMedian < 30.0, VariantAlleleCount >= 7 && VariantStrandBias < 0.05 && ReferenceStrandBias >= 0.2. The parameter ReadCountControl was set to be < 20 for the three fresh-frozen and FFPE paired samples and <10 for the additional FFPE samples.

Additionally, based on the cosine similarity of WGS-derived mutational signatures and the mutational signatures derived for the initial three samples, we optimized the minimum number of reads supporting a SNV (fresh-frozen samples mutREAD = 5, WES = 7, 10x sWGS = 5, mutREAD FFPE = 10) and the minimal variant allele frequency of a SNV (fresh-frozen samples mutREAD = 0.03, WES = 0.01, 10x sWGS = 0.11, mutREAD FFPE = 0.13). The cut-offs were optimized separately for Strelka-derived mutations (fresh-frozen samples = 20 reads and 0.11 variant allele frequency, mutREAD FFPE = 11 and 0.03 variant allele frequency).

2.20 Mutational signature profile for mutREAD

The tri-nucleotide context for each SNV was determined using the SomaticSignatures[90] R package. Mutational signature profiles were derived for each sample using OAC-specific mutational signatures. Finally, non-negative least squares[91] in R was used to derive the contributions of each mutational signature to the overall mutational spectrum. The estimated coefficients were scaled to sum up to one.

For validation analysis I used the deconstructSigs[78] to obtain OAC specific signatures. Cosine similarities were calculated using 'cosine()' function in the lsa package (0.73.2) in R.

3.Results Chapter

Mutational processes unveil bottlenecks that shape evolution of oesophageal adenocarcinoma.

3.1 Attribution:

Prof. Rebecca C Fitzgerald (RCF) , Dr Maria Secrier(MS) from UCL and I designed the study. RCF and MS supervised the analysis. WGS and RNA Seq data was managed by Ginny Devonshire. Using VCF files, I performed signature extraction analysis, I curated the clinical data for cohort demographics and I performed the clinical correlation analysis with signatures, and I performed the survival analysis. I performed genomic feature analysis: transcription and replication strand bias analysis. Clonality/timing analysis, machine learning analysis and DNA repair signature analysis was performed by MS. Nucleosome periodicity analysis was performed in collaboration by Dr.Oriol Pich from Prof.Nuria Lopez Bigas lab IRB (Institute for Research in Biomedicine) Barcelona Spain.

3.2 Rationale:

The overall goal was to investigate how mutational processes shape the genome during OAC development from precancerous stages to advanced disease. We aimed to characterise mutational signatures in the precancer lesion (Barrett's), OAC and metastasis samples, so as to gain insights about their dynamics. Also, to study the influence of endogenous and external risk factors for better understanding of modulating forces of mutational signatures during the course of the disease. Then, to evaluate the status of DNA repair process and their impairment contributing to mutagenesis, we looked into mutational signatures in DNA repair pathways. In order to study the evolutionary bottle necks, we investigated clonality and timing of the mutational signatures at each stage of OAC development. Finally, to investigate prognostic value of these signature for guide patient outcome. (Figure 3.1)

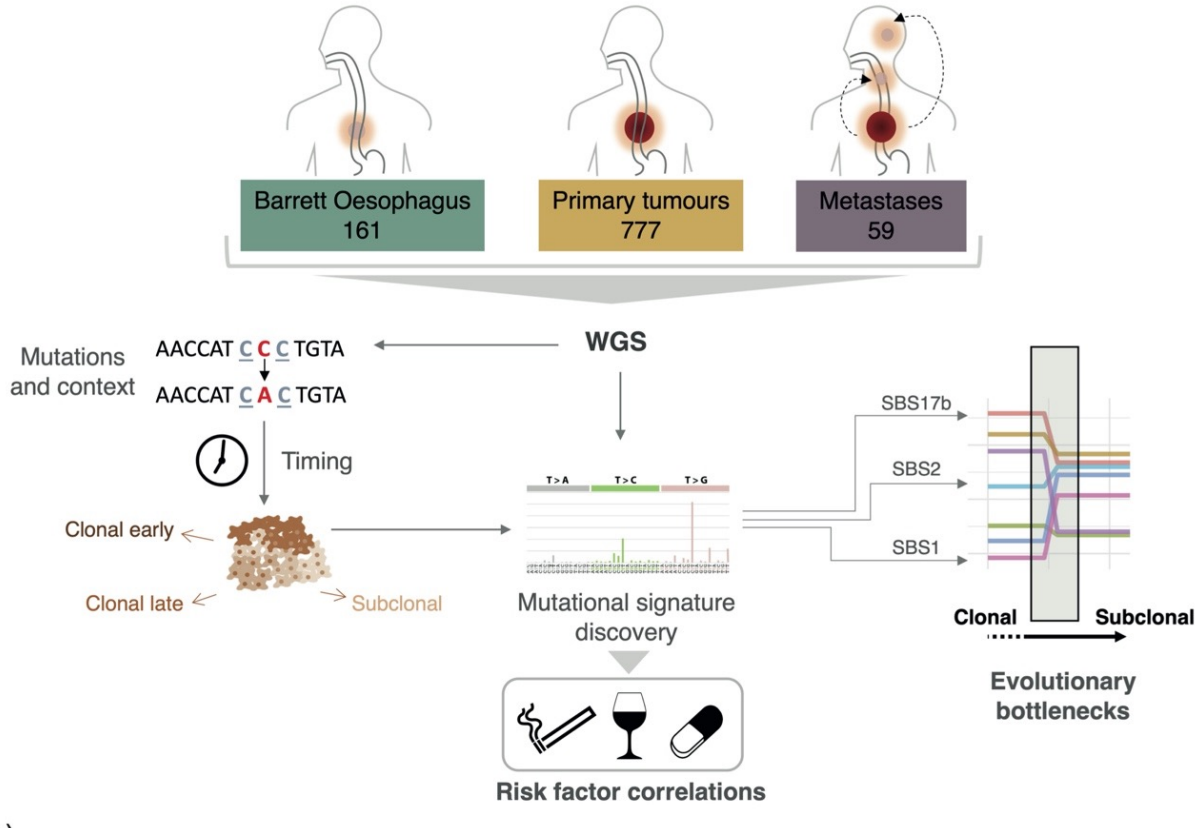


Figure 3.1: Overview of the study design

A study cohort was assembled, composed of 161 Barrett's oesophagus samples, 777 primary OAC tumour samples and 59 metastasis samples. 50X WGS was available for all these samples. Mutational signatures were extracted from WGS for all of the samples and these proportions were used for correlative analysis with risk/tumour factors. To another end, SNVs from the same 50X WGS data for all samples in the cohort were used to estimate the timing and then the clonality. Then evolutionary bottle necks were studied.

3.3 Mutational signature landscape from pre-malignant to advanced OAC and clinical associations

We employed whole-genome sequencing data from 161 Barrett Oesophagus samples, 777 OAC primary tumours and 59 metastatic samples to infer and compare the signatures of mutational processes that operate during the course of this disease. Clinical characteristics of the cohort, as available, are presented in Table 3.1. As expected, the majority of patients are males (OAC: 86.8%; Barrett's: 81.7%) and presented in older age (OAC: 67.1years; Barrett's: 68years).

We collected data on risk exposures/habits (smoking, alcohol, obesity/BMI) and relevant medications such as acid suppressants and anti-inflammatory drugs. Pre-treatment TNM staging was recorded, and most of the tumours (69%) were T3 (Invasion into adventitia), N1 (70.5% of nodes were positive) and since cases were recruited from a surgical pathway only 7.1% had evidence of distant metastases. Almost half (47%) of the Gastro-Oesophageal Junction type tumours (Siewert classification) were type 1 meaning that they were mainly in the oesophagus. Next major group (38.1%) was type 2, that is the tumours were located precisely at the oesophagogastric junction followed by small number (14.8%) of tumours representing type 3, in which the tumour extends into the gastric cardia and fundus. Overall, the median survival was 109(57-188) weeks.

Variable	Measure/Level	OAC(n=645)	Barrett's (n=148)
Age	Years (median,IQR)	67.1 (59.2-74.2)	**68 (62.0-75.9)**
Gender	Female	85 (13.2%)	**26 (17.5%)**
	Male	560 (86.8%)	**121 (81.7%)**
Exposures			
Smoking status		535 (82.9%)	**142 (95.9%)**
	Current	92 (14.2%)	**84 (56.7%)**
	Former	293 (45.4%)	**24 (16.2%)**
	Never	150 (23.2%)	**34 (22.9%)**
	Missing data	110 (17.0%)	**6 (4.0%)**
Alcohol (Units/week)	Mean(min-max)	6.2 (1-70)	**1.0 (1-3)**
Acid Suppressants		512 (79.4%)	**140 (94.5%)**
(PPI)	Current Use	255 (39.5%)	**126 (85.1%)**
	Past Use	57 (8.8%)	**1 (0.6%)**
	Never	200 (31.0%)	**13 (8.8%)**
	Missing data	133 (20.6%)	**8 (5.4%)**
Anti-inflammatory drugs		322 (49.9%)	**116 (78.3%)**
(NSAIDs)	Current Use	128 (19.8%)	**37 (25.0%)**
	Past Use	51 (7.9%)	**3 (2.0%)**
	Never	143(22.1%)	**76 (51.3%)**
	Missing data	323 (50.0%)	**32 (21.6%)**
BMI	Kg/m2(median, IQR)	27.3 (24.4-31.2)	**29 (13.5-43.9)**
Overall Survival	Weeks (median, IQR)	109 (57-188)	**Not Available**
Diagnosis			
Pre-treatment Tumour Stage		511 (79.2%)	**Not Applicable**
	T0	1 (0.1%)	
	T1	15 (2.3%)	
	T1a	3 (0.4%)	
	T1b	12 (1.8%)	
	T2	83 (12.8%)	
	T3	353 (54.7%)	
	T4	14 (2.1%)	
	T4a	8 (1.2%)	
	T4b	3 (0.4%)	
	Tx	19 (2.9%)	
	Missing data	134 (20.7%)	
Pre-treatment nodal involvement (CT)		571 (88.5%)	**Not Applicable**
	Positive	403 (62.5%)	
	Negative	168 (26.0%)	

	Missing data	74 (11.5%)	
Pre-treatment distant metastases (CT)		545 (84.5%)	
	Positive	39 (6.0%)	
	Negative	459 (71.2%)	
	Mx	47 (7.3%)	
	Missing data	100 (15.5%)	
Pre-treatment Siewert Classification		304 (47.13%)	
	Type I	143 (22.1%)	
	Type II	116 (18.0%)	
	Type III	45 (7.0%)	
	Missing data	341 (52.8%)	
Therapy			
NeoAdj.Chemotherapy			**Not Applicable**
5FU	Treated	57 (8.8%)	
	Not Treated	177 (27.4%)	
Capecitabine	Treated	276 (42.8%)	
	Not Treated	47 (7.3%)	
Cisplatin	Treated	272 (42.2%)	
	Not Treated	42 (6.5%)	
Epirubicin	Treated	271 (42.0%)	
	Not Treated	47 (7.3%)	
Oxaliplatin	Treated	49 (7.6%)	
	Not Treated	190 (29.4%)	
Lapatinib	Treated	5 (0.8%)	
	Not Treated	198 (30.7%)	
Surgery	Yes	452 (70.0%)	
	No	151 (23.4%)	

Table 3.1: Clinical Characteristics of the study cohort.

The first stage of the mutational signature analysis was to compare the different calling methods to check the consistency and to understand the technical differences in extracting signatures (Table 3.2). I used three packages to extract signatures, SigProfiler(Version 2.5.1.9), deconstructSigs (Version 1.8.0) and MutationalPatterns(Version 3.2.0) and webtool Mutalisk (http://mutalisk.org/). I was able obtain the OAC specific signatures (Signature1, 2,8 and 17) across all methods with additional traces of other signatures across different methods.

SigProfiler the gold standard method and deconstructSigs for exposure of all the SBS signatures and ease were chosen for mutational signature analysis.

COSMIC Signatures	SigProfiler (Secrier & Li et al)	SigProfiler	deconstructSigs	Mutational patterns	Mutalisk
Signature1 (Age)	✓ 32%(1 and 18)	✓ (20.4%)	✓ (12.5%)	✓ (17.1)	✓ (12.2%)
Signature2 (APOBEC)	✓	✓ (2.1%)	✓ (0.5%)	✓ (11.56)	
Signature3 (BRCA)	✓ 15%		✓ (1.2%)		
Signature5 (Unknown)		✓ (5%)	✓ (3.3%)	✓ (2.72)	✓ (18.5%)
Signature6 (MMR)		✓ (2.6%)	✓ (1.9%)	✓ (2.7)	
Signature8 (Unknown)		✓ (20.9%)	✓ (9.3%)	✓ (14.28)	✓ (12.8%)
Signature9 (Polymerase η)			✓ (1.4%)		✓ (16.6%)
Signature13 (APOBEC)					✓ (2.8%)
Signature17 (ROS?)	✓ 53.3%	✓ (48.8%)	✓ (66.6%)	✓ (51.4)	✓ (37%)
Signature18 (Unknown)	✓		✓ (1.4%)		

Table 3.2: Preliminary comparative mutational signature analysis: Mostly the main mutational processes were stable across methods.

We uncovered a total of 14 mutational signatures. Signatures SBS17a/b were the most prevalent, along with evidence for mutational processes linked with ageing (SBS1/5/40), oxidative stress (SBS18), APOBEC activity (SBS2) and DNA damage repair (DDR) impairment (SBS3/8). We also observed evidence of base excision repair mutagenesis (SBS30), mismatch repair deficiency (SBS44) and a colibactin-linked mutational process (SBS41), which have not been described extensively in this cancer (Figure 3.2, Figure 3.3).

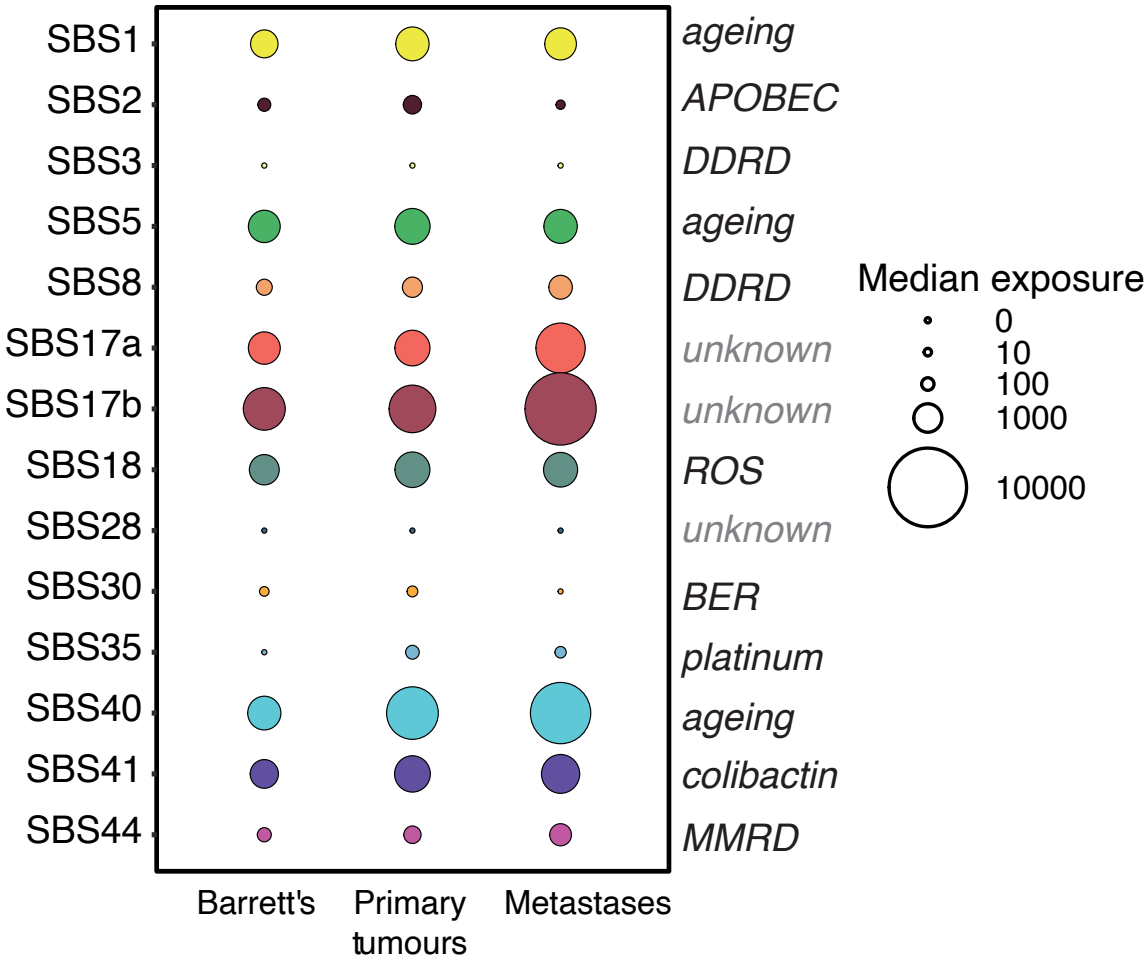


Figure 3.2: Landscape of Mutational signatures during course of OAC development.
Median prevalence of mutational signatures present identified in the three disease stages. The magnitude of the circles is proportional to the number of SNVs specific to that mutational signature in the samples.

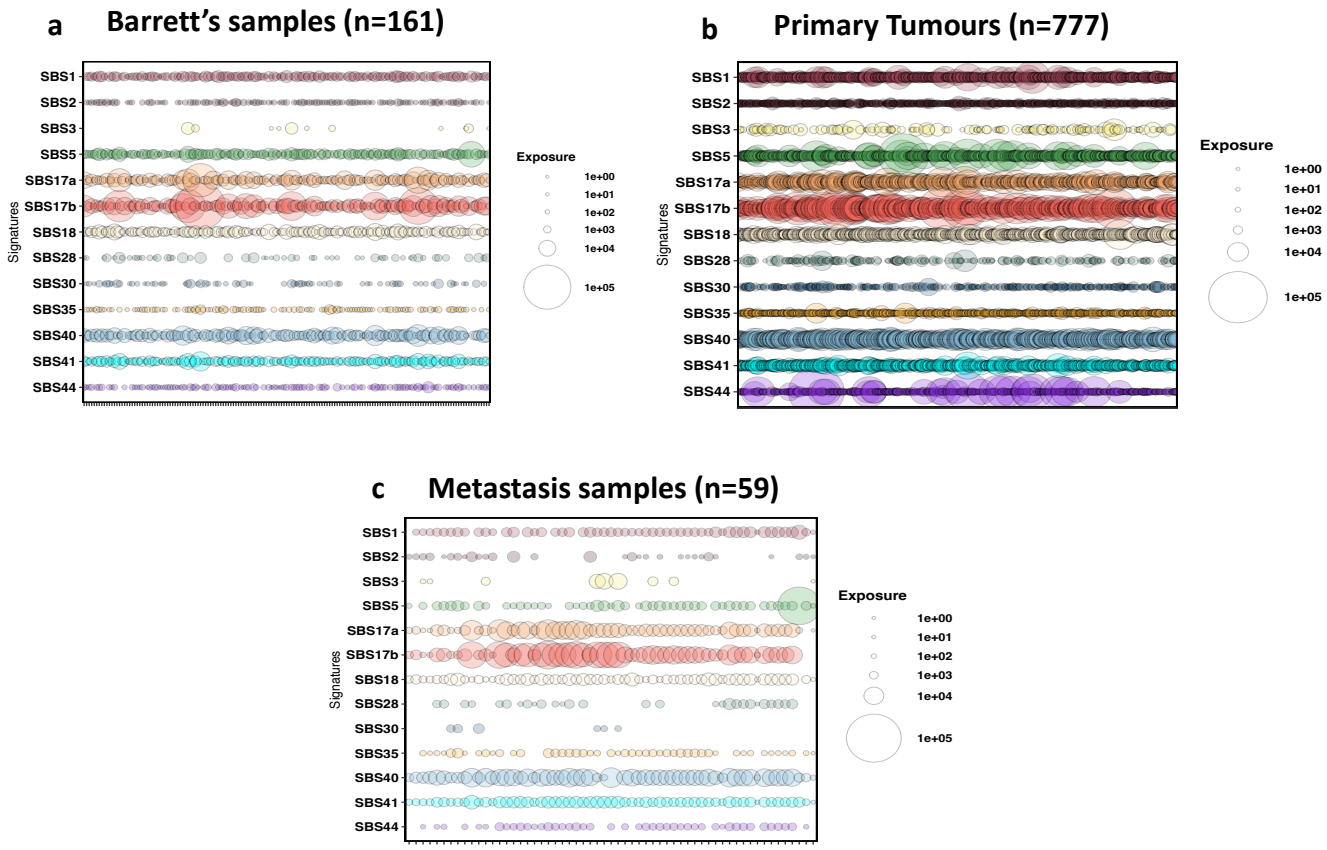


Figure 3.3: Mutational processes active across stages of oesophageal adenocarcinoma development. Relative contribution of mutational signatures, each bubble represents a patient and the size of the bubble is proportional to the number of mutations attributed to the respective mutational signature.

Additionally, we uncovered a signature of platinum treatment (SBS35) in primary tumours, which is expected, given that the majority of these tumours have been sequenced from the surgical resection specimen after treatment with chemotherapeutic agents and platinum is the backbone of treatment regimens. Indeed, this signature was increased specifically in chemotherapy treated samples (Figure 3.4a) and likely reflects the mutagenic effects of this therapy[24]. We also observed MMR linked SBS44 co-occurring in these treated samples. We looked at the proportions of SBS17B in Barrett's, chemo naive and treated primary tumours, SNVs associated with SBS17B were in abundance from early Barrett's and observed both in naive and treated primary tumours(Figure 3.4b).

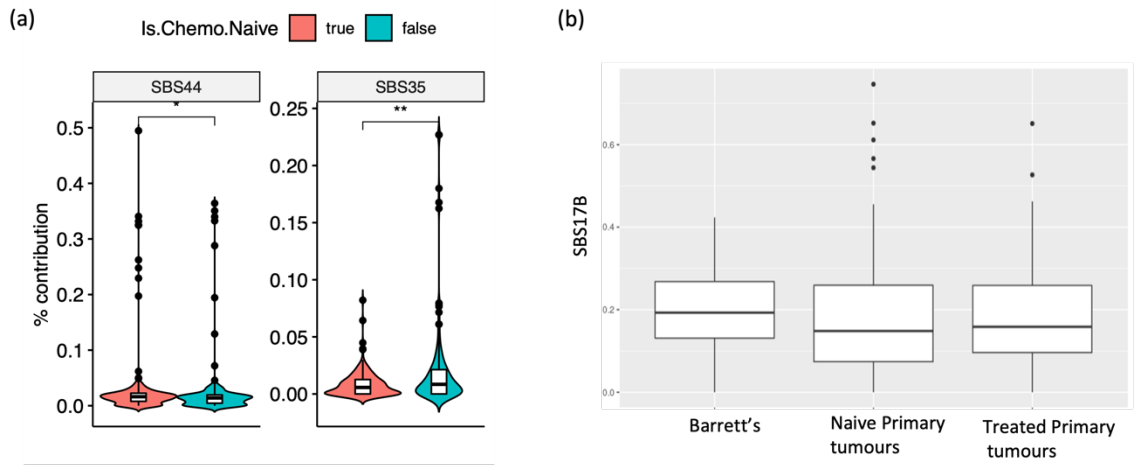


Figure 3.4: Influence of chemotherapy on mutational signatures:

(a)Proportions of mutational signatures significantly increased in chemo treated OACs for SBS35 (Platinum therapy associated signature) and SBS44 (Mismatch Repair signature)

(b) Proportions of SBS17B in Barrett's, chemo naive and treated primary tumours.

There was evidence for many of these mutational processes acting very early on in tumour evolution such that they were already present in Barrett Oesophagus, especially SBS17a/b and the ageing-linked signatures SBS1, 5 and 40. On average, the majority of signatures appeared increased in primary tumours, as expected, and tended to rise further in metastatic samples (particularly SBS17a/b, SBS40 and SBS41).

Mutation rates along the genome are highly variable and influenced by several chromatin features. We confirmed that SBS17-associated SNVs are enriched in the untranscribed and lagging strands. Other signatures did not show any significant strand or replication timing bias (Figures 3.5, 3.6, 3.7).

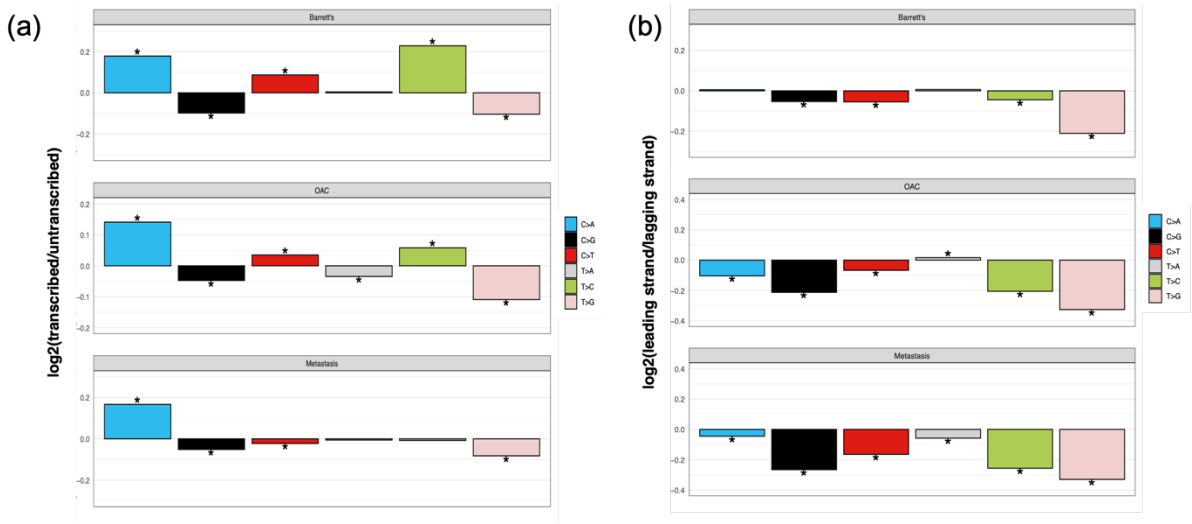
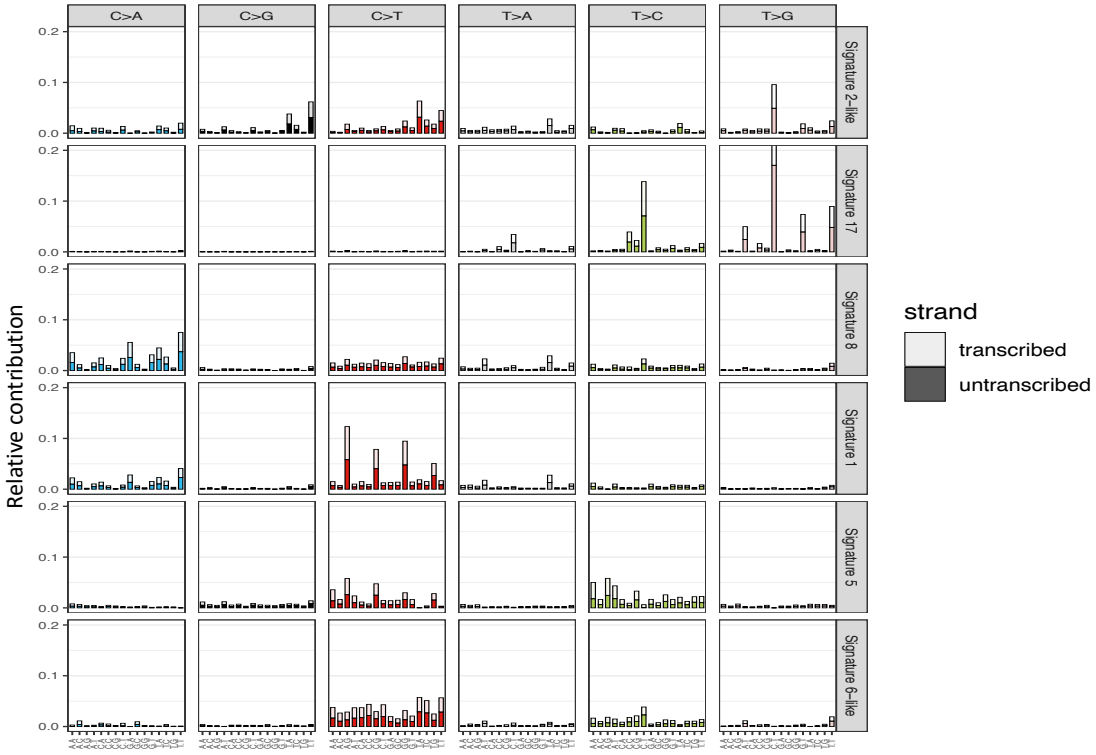
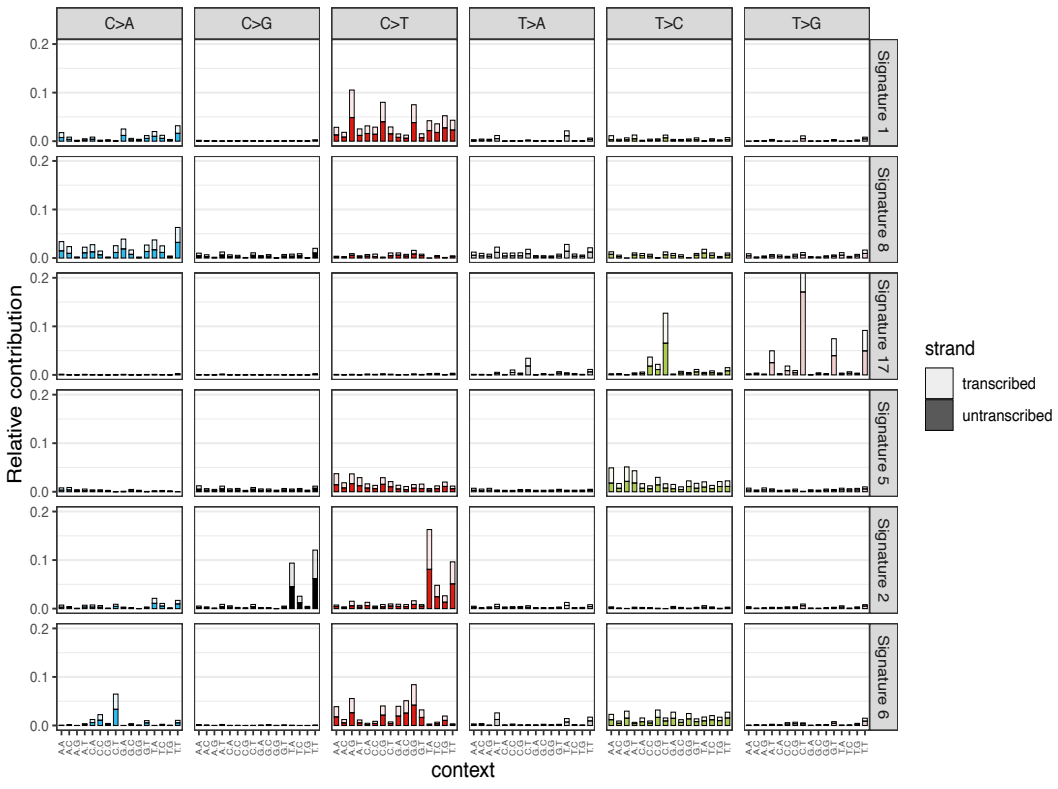


Figure 3.5: Transcription and Replication strand asymmetry: SNVs (C > X / T > X) were mapped to the transcribed/untranscribed strands and leading /lagging replicative strands. Significant strand asymmetries were marked by asterisks (p-value, Poisson test). (a)T>G mutations (Signature 17 associated) were mapped to untranscribed strands across Barrett's, OAC and Metastasis. (b)Lagging replicative strand is mostly mapped by all types of substitutions with T>G substitutions predominantly present on lagging strand across Barrett's, OAC and Metastasis.

Barrett's



OAC



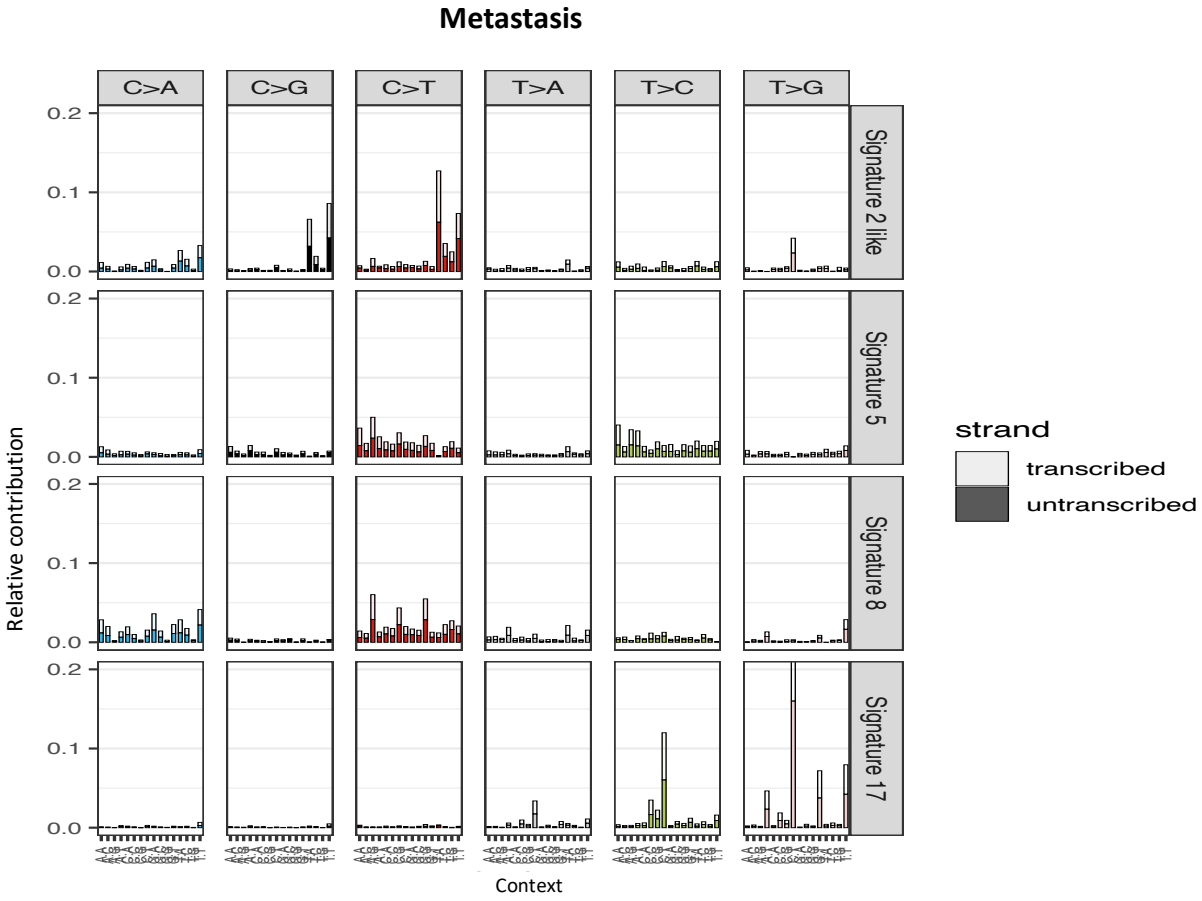
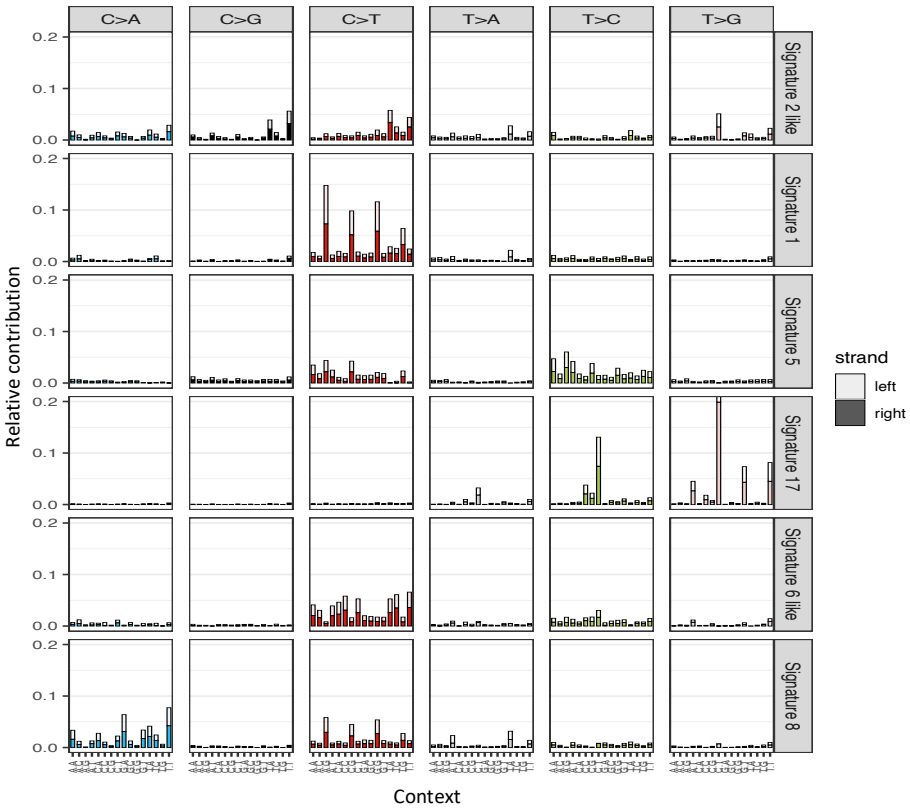
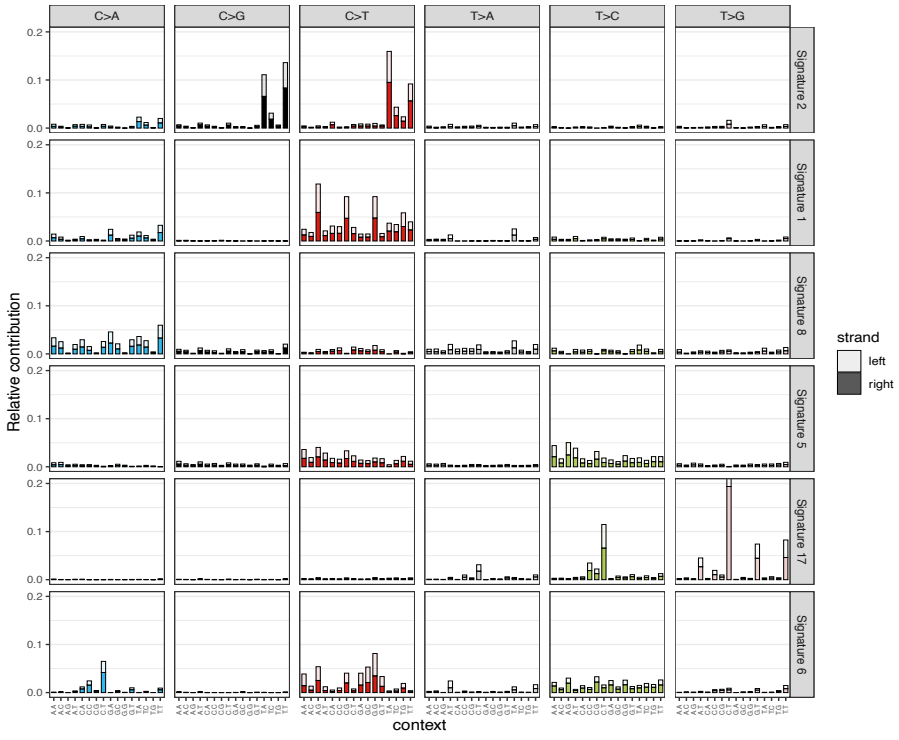


Figure 3.6: Mutational signatures and Transcription strand asymmetry: 96 base substitution profile. Mutational signatures were extracted using the strand annotated matrices. Across Barrett's, OAC and in Metastasis, signature17 showed transcription strand asymmetry.

Barrett's



OAC



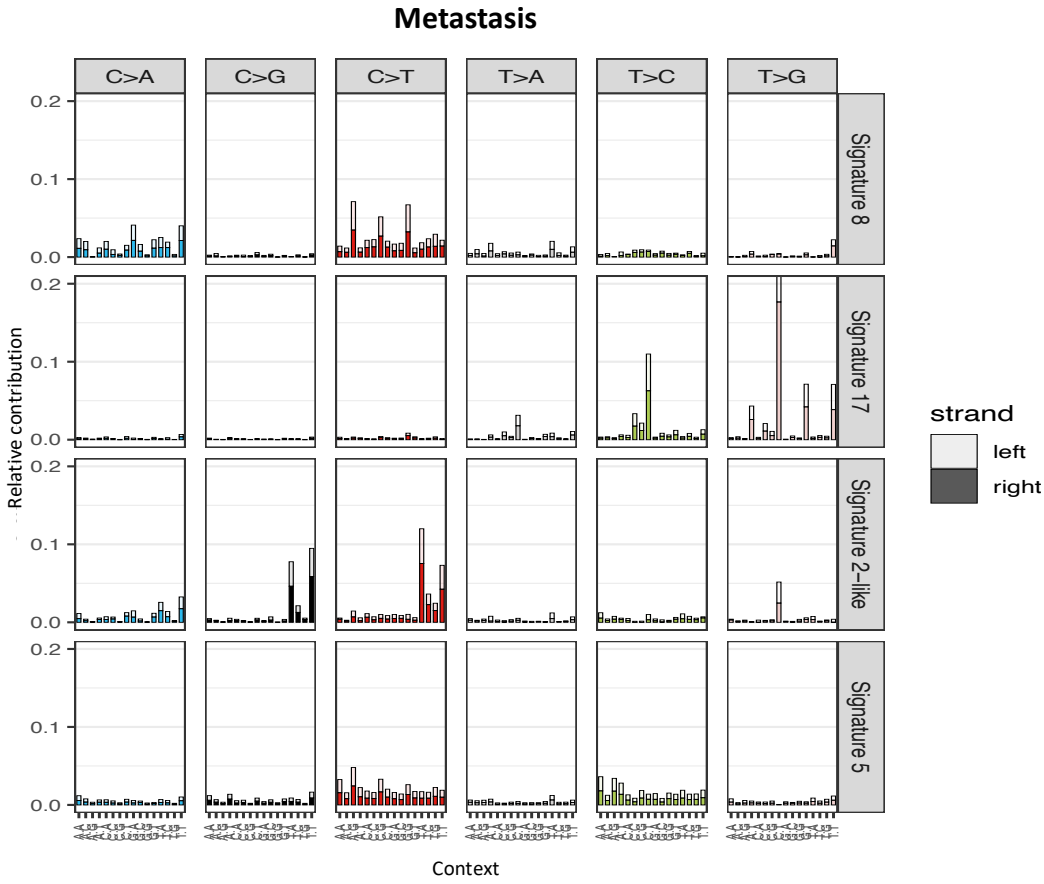


Figure 3.7: Replication strand asymmetry and Mutational signatures: 96 base substitution profile. Mutational signatures were extracted using the replication strand annotated matrices. Signature17 associated mutations were present on the lagging replicative strand across Barrett's, OAC and in Metastasis. Left: Leading strand, Right: Lagging strand.

The nucleosome periodicity patterns were similar across the stages, with maximum power period (MP) of 191.14bp between nucleosomes and linker DNA (Figure 3.8). The constant periodicity patterns between major and minor grooves with MP of ~10.3 , and with a significant SNR (Signal-to-Noise Ratio), The value of SNR explains the strength to the periodicity and provides a measurement of periodicity of a signal. The SNR values for Barrett's (SNR=433.5), Barrett's trios (SNR= 378) and OAC (SNR= 601.4) were estimated. These are significant SNR values for all three stages suggesting a strong periodicity of ~10.3bp across the stages, all were in phase 1 of the nucleosome orientation, enriched on minor-in (minor groove facing the nucleosomes) in keeping with what has been reported previously[53] (Figure 3.9).

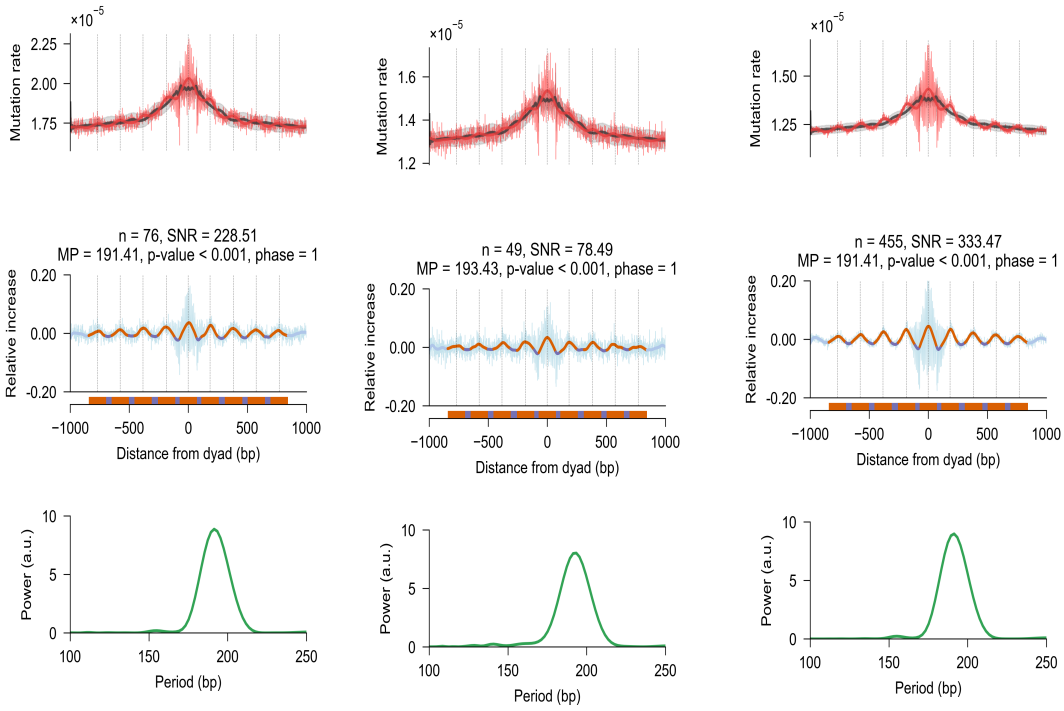
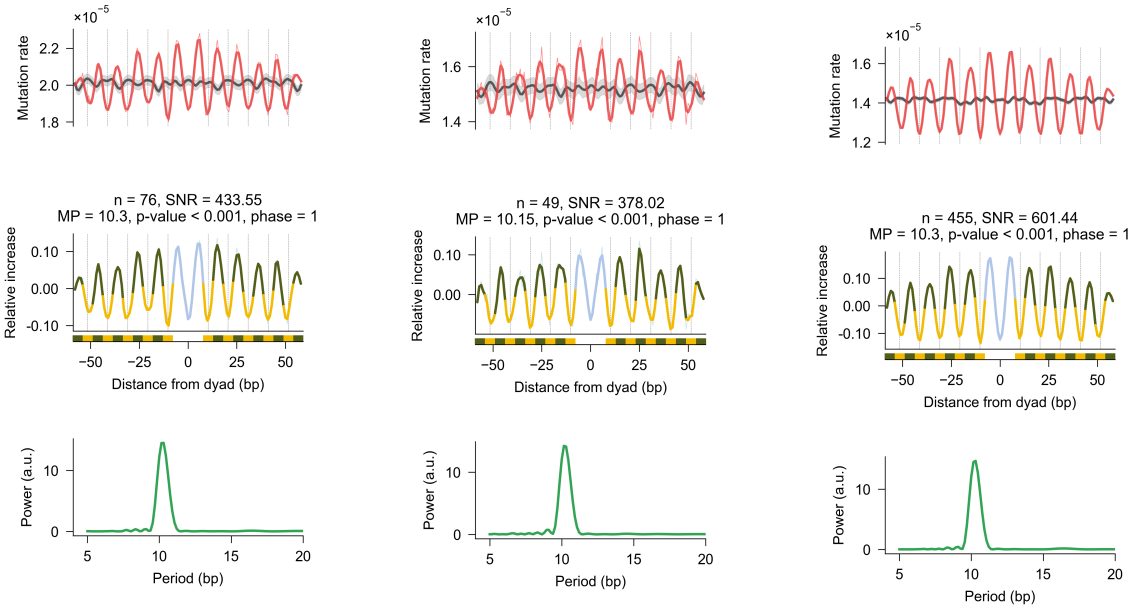


Figure 3.8: Nucleosome periodicity stable patterns across stages of OAC development (Zoom out): Mutation rate periodicity between nucleosome-covered and linker DNA



Barrett's (n=76)

Barrett's with adjacent tumour (n=49)

OACs (n=455)

Figure 3.9: Nucleosome periodicity stable patterns across stages of OAC development (Zoom in): Mutation rate periodicity between minor-in and minor-out nucleosome-covered DNA stretches.

Next, we correlated mutational signatures with reported exposures in the cohort. Alcohol consumption, which is a modest risk factor for this cancer type[92] nevertheless was correlated with SBS17a and SBS44 (MMR) prevalence already in Barrett Oesophagus and this association increased in primary tumours. Smoking was strongly associated with SBS17a in primary tumours only. NSAID usage was linked to increased mutagenesis from SBS2-APOBEC, SBS3-DDRDand SBS30-BER in Barrett Oesophagus and SBS35 and SBS40 in primary tumours. No link was found between any signature and PPI/acid suppressant usage but this might be affected by the fact that the majority of patients with this disease take these medications and the information on dose and duration of use is limited. DDRD-SBS3 was strongly associated with positive nodes(Figures 3.10, 3.11, 3.12).

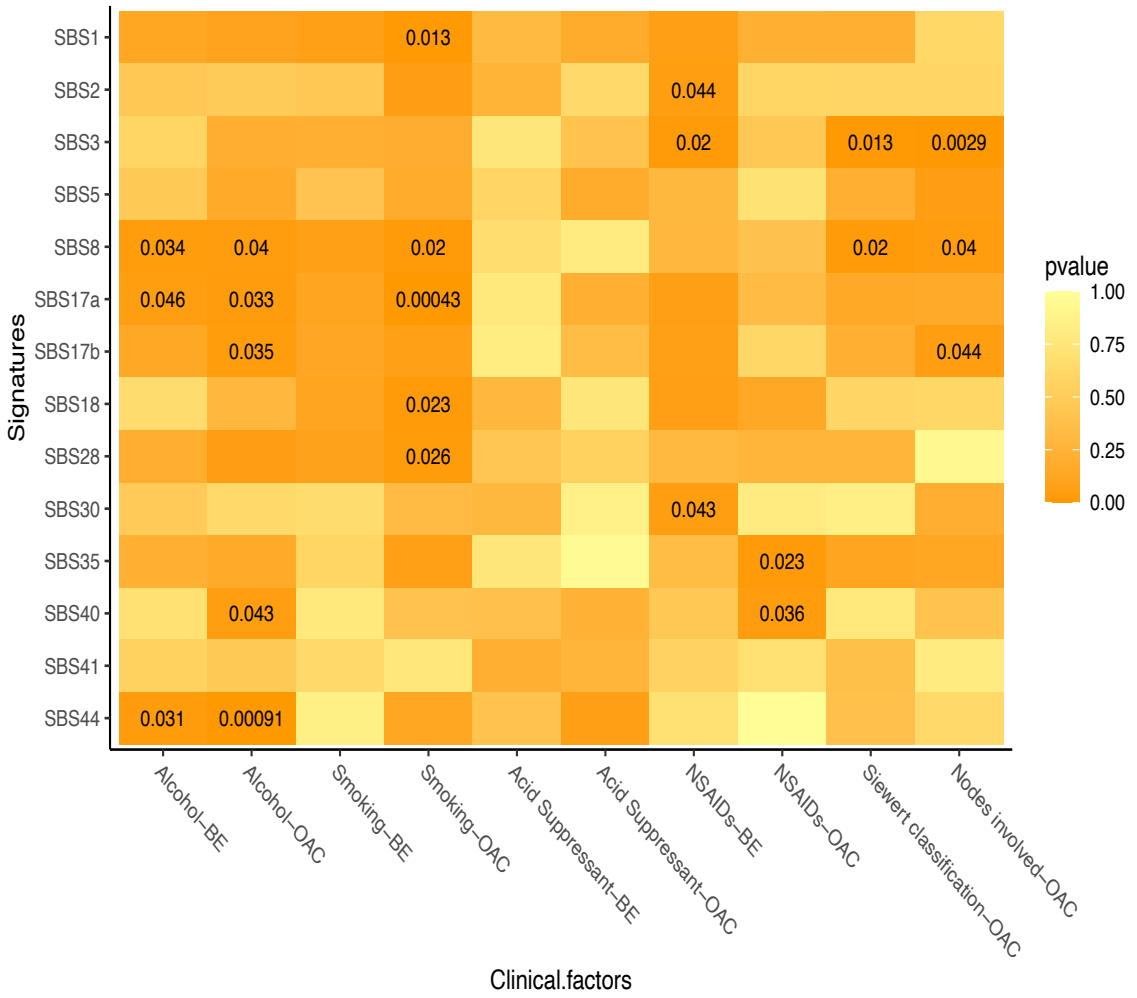


Figure 3.10: Comparative correlations between Barrett's and OAC Clinical Factors with the proportion of mutational signatures. Significant positive associations are denoted(p-value).

(a) **Alcohol**

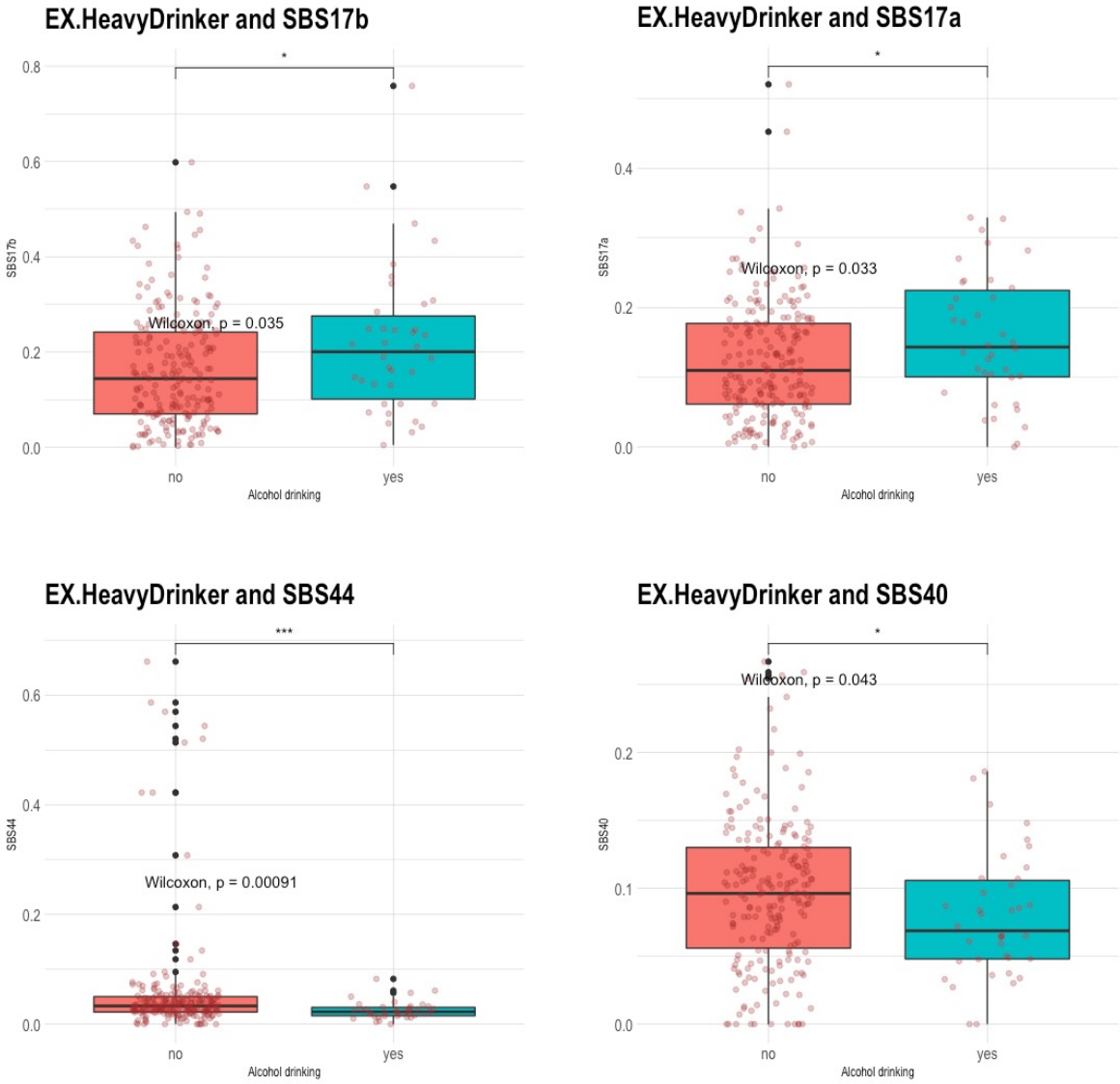


Figure 3.11: Representative Positively Correlated risk factors (Exposures): (a) Alcohol
Statistical correlations between different levels of a variable and the proportions of mutational signatures in OAC.

(b) **Smoking**

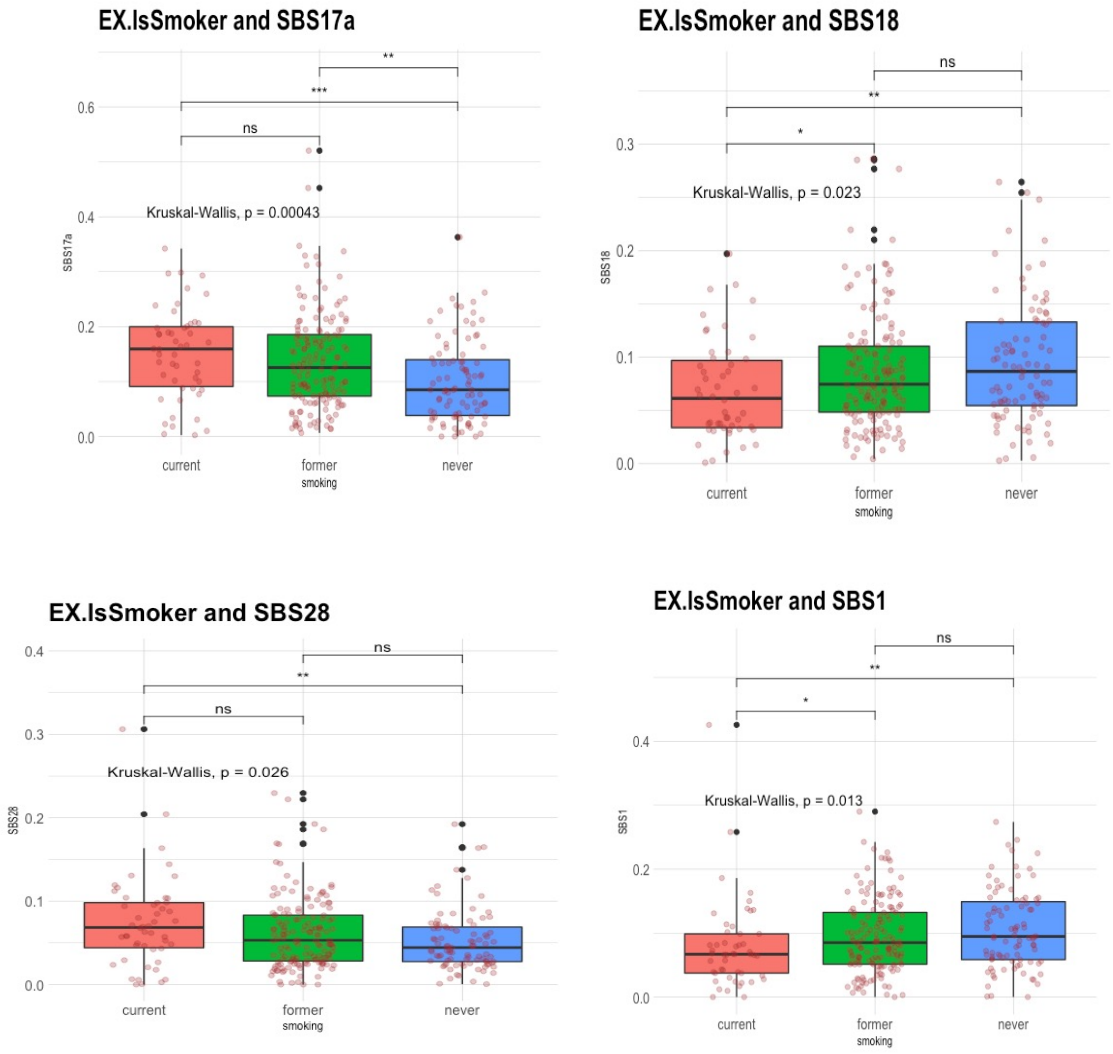


Figure 3.11: Representative Positively Correlated risk factors (Exposures): (b) Smoking
Statistical correlations between different levels of a variable and the proportions of mutational signatures in OAC.

(c) **NSAIDs**

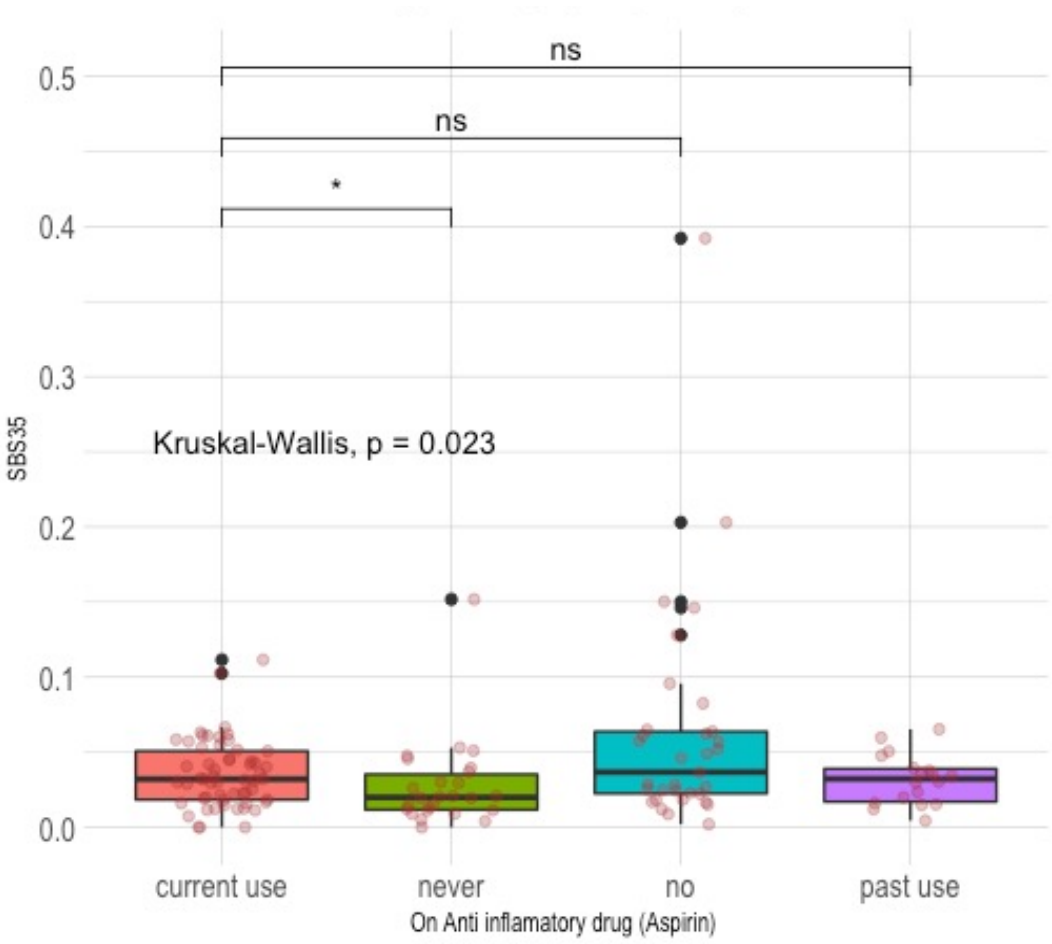
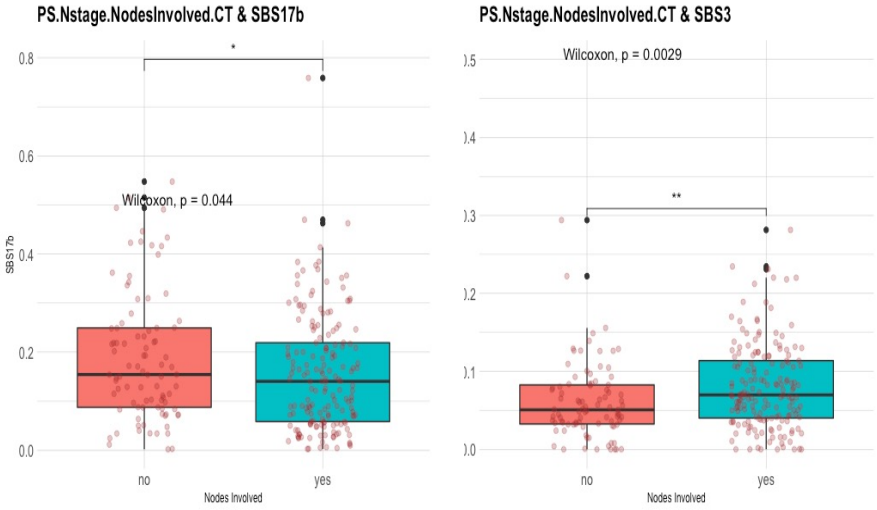


Figure 3.11: Representative Positively Correlated risk factors (Exposures): (c) NSAIDs
Statistical correlations between different levels of a variable and the proportions of mutational signatures in OAC.

Nodes Involved



Siewert Classification

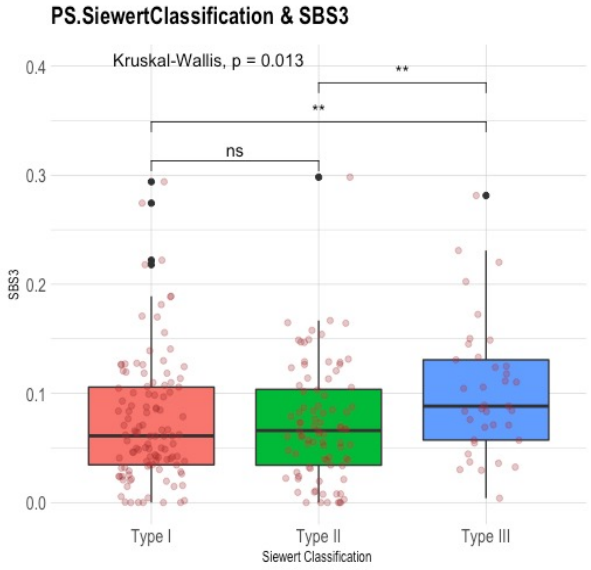


Figure 3.12: Representative Positively Correlated tumour factors of OAC with mutational processes. Statistical correlations between different levels of a variable and the proportions of mutational signatures in OAC.

3.4 Dynamics of mutational processes from pre-malignant to advanced disease

Next, we examined how the impact of mutational processes active in OAC varies across stages of the disease. In general, we observed an increase in the contribution from APOBEC-linked mutagenesis (SBS2), colibactin (SBS41) and platinum treatment (SBS35) most prominently in primary tumours, while the SBS17 processes and MMR were most increased in metastases (Figure 3.13a). Ageing-associated mutational events (SBS1 and 5) mostly appeared as a continuous background contribution that decreases in importance with increased cancer stage. However, SBS40, also thought to be linked with ageing, increased from pre-malignant to advanced disease – suggesting that the derivation of this mutational process is different and may have a higher impact in this disease than previously appreciated.

Comparing matched samples of Barrett Oesophagus and primary tumours from the same individuals further corroborated our previous findings: APOBEC mutagenesis, DDR impairment, the SBS40 process and the platinum signatures increase with disease progression, while the ageing signatures 1 and 5 and SBS18, linked to oxidative stress, seem to decrease (Figure 3.13b).

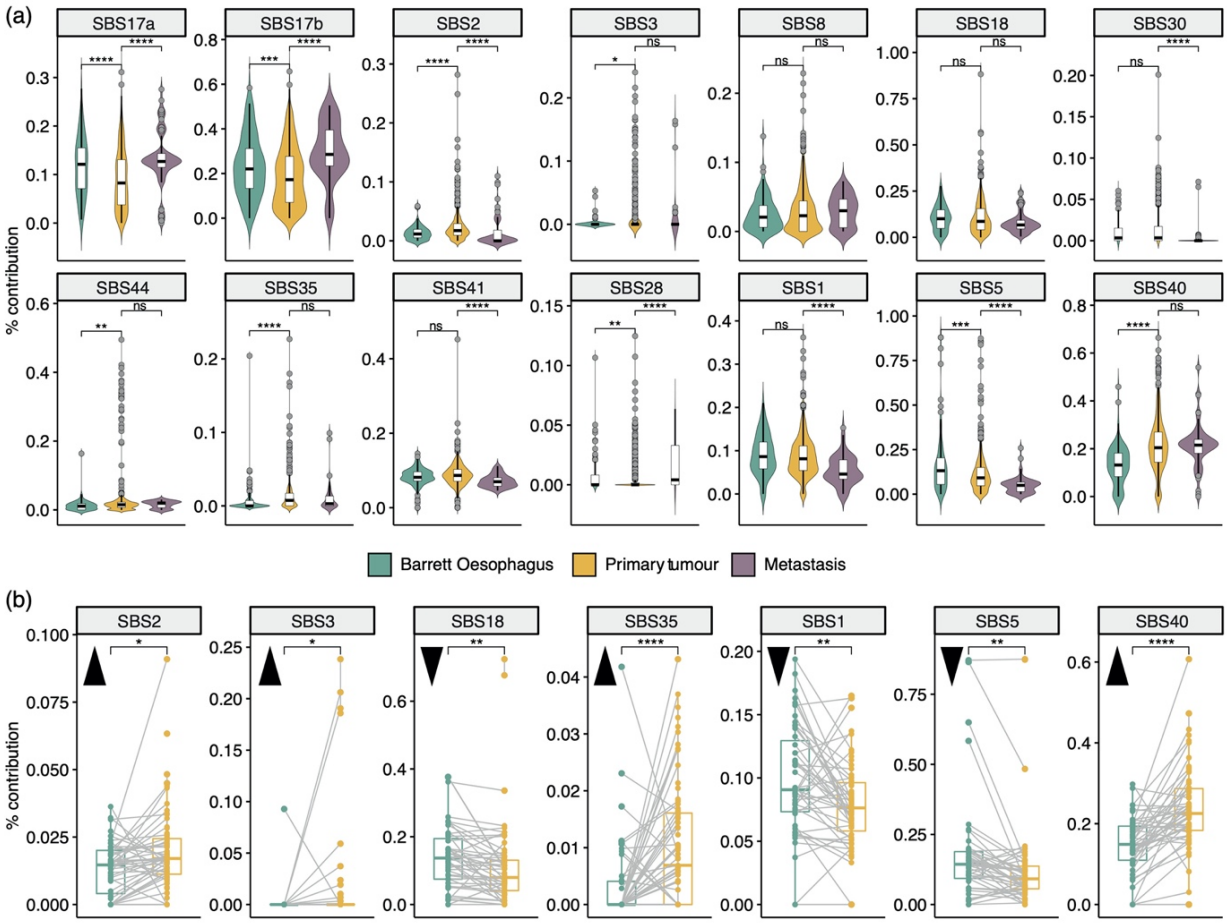


Figure 3.13: Mutational process dynamics from Barrett Oesophagus to primary tumours and metastases. (a) Mutational signature contributions compared across the three disease conditions in non-matched samples. (b) Changes in mutational signature prevalence between matched Barrett Oesophagus and primary tumour samples. Upward direction of triangles denote an increase in signature contribution in primary tumours; downward direction of triangles denote a decrease. Only signatures with a significant change are shown.

Within an individual disease stage, we observed various combinations of mutagenic processes acting in the genomes (Figure 3.14a), some of which were common between stages, such as the joint presence of SBS17a/b and SBS40, and some of which were unique, e.g. SBS41 and all ageing-linked signatures were only observed to co-occur in primary tumours. To make sense of this complexity, we asked whether we could prioritise signatures to hep distinguish between Barrett Oesophagus, primary tumours and metastases. To this end, we employed a gradient boost classifier approach to distinguish between cancer stages based on the mutational footprint alone (see Methods section 2.9). When considering the overall signature contributions in each cancer stage, the model distinguishing Barrett Oesophagus from primary tumour genomes had a performance of 87% AUC (Figure 3.14b). The APOBEC mutagenesis signature was ranked as the most predictive of primary tumour development, followed by the ageing-linked SBS40 and the colibactin signature SBS41, suggesting they may be more important in driving the malignant transformation of pre-neoplastic lesions. The ageing signature S1 appeared most specific to Barrett cases, which is not surprising given that it is the primary source of mutations in healthy tissues.

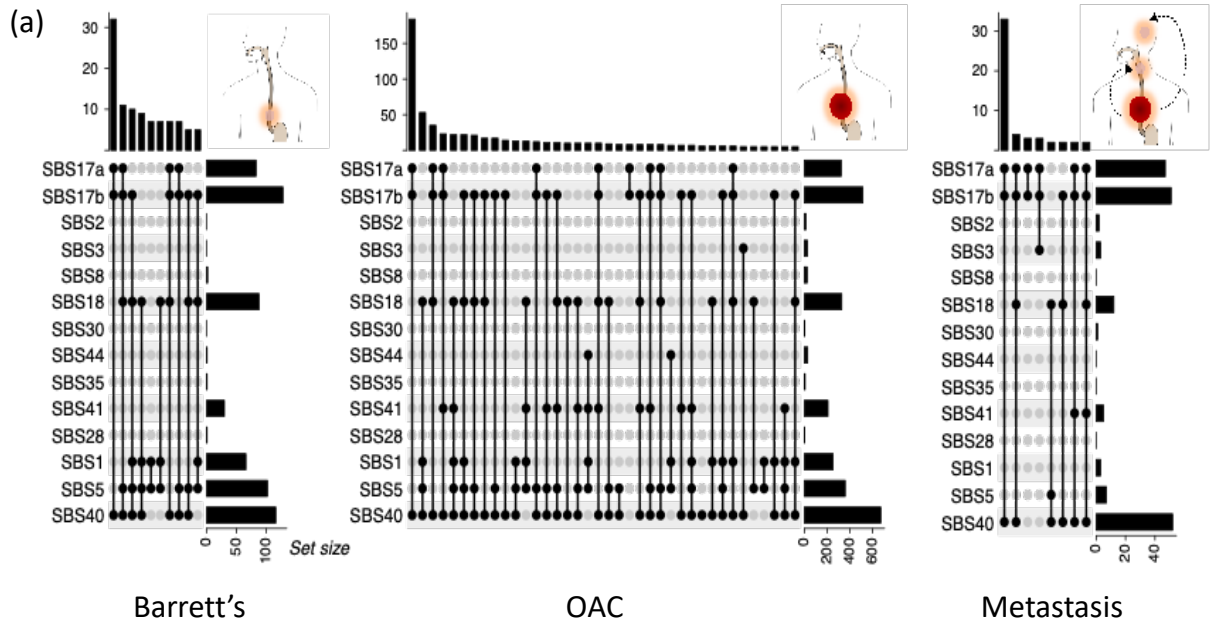


Figure 3.14: Mutational process dynamics from Barrett Oesophagus to primary tumours and metastases. (a) Prevalence of combination of mutational processes co-occurring in patients across stages of tumour development and spread.

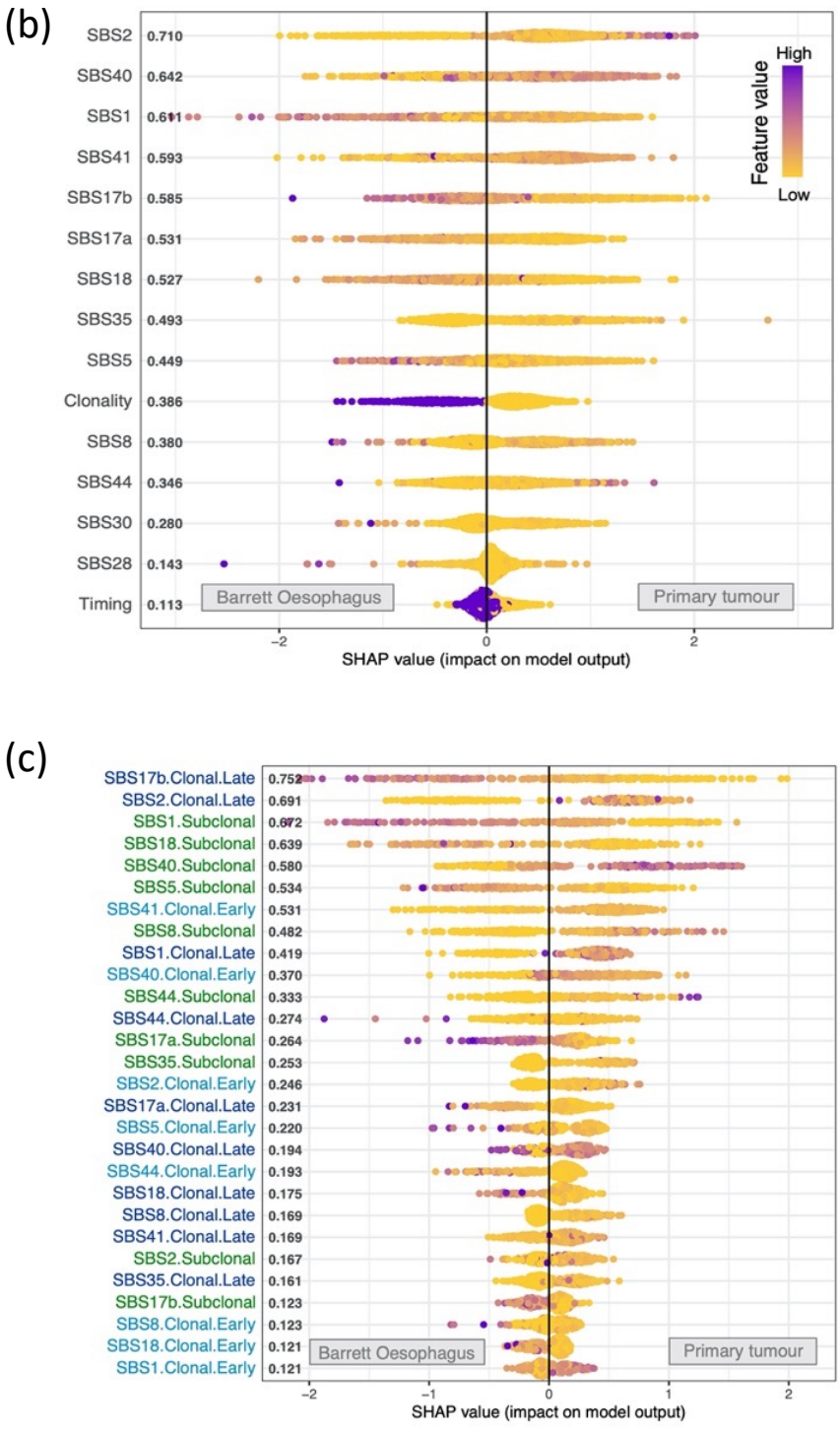


Figure 3.14: Mutational process dynamics from Barrett Oesophagus to primary tumours and metastases.(b) Output of xgboost model distinguishing Barrett Oesophagus from primary tumours based on overall signature prevalence, while accounting for clonality and timing.

Features are ordered according to their ranking in the model (top ranking features first). Every dot is a sample and the colour corresponds to the signature contribution in that sample.
(c) Output of xgboost model distinguishing Barrett Oesophagus from primary tumours based on detailed signature contributions split by clonality and timing. ns-not significant; *p<0.05; **p<0.01; ***p<0.001; ****p<0.0001

The signature associations with the primary tumour stage were further corroborated by a multinomial regression analysis (Figure 3.15).

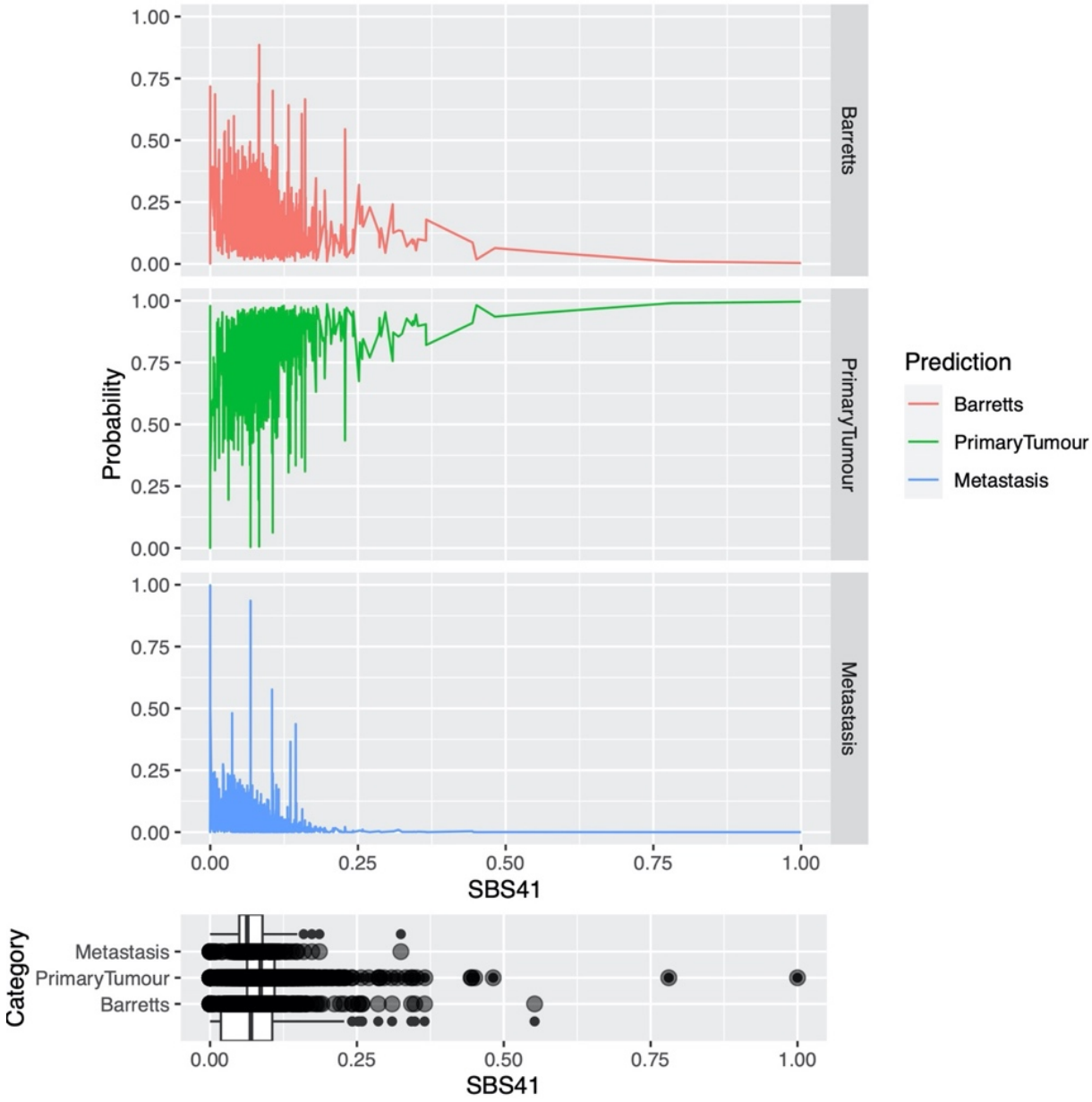


Figure 3.15: Multinomial regression classifier results distinguishing Barrett Oesophagus, primary tumours and metastases based on signature prevalence. The predictive power of SBS 41 in distinguishing primary tumours is exemplified.

Interestingly, it emerged from the model that the clonality of the mutations had a strong contribution to distinguishing between cancer stages (Figure 3.14b). As a result, we built a second gradient boost classifier that would enable us to highlight processes that act subclonally or later in evolution in a stage-specific manner, which had an accuracy of 86% (Figure 3.14c). This model confirmed the key signals from the previous analysis, but shed further light on the fact that the APOBEC and colibactin mutations that appear as a distinct signature in primary tumours are accumulated clonally later (APOBEC) and earlier (colibactin) in evolution, respectively. Furthermore, SBS17b clonal mutations that accumulated later in evolution emerged as the most specific for Barrett genomes.

Our power to detect signature differences when comparing primary tumours to metastases was reduced due to the smaller size of the metastatic cohort (despite an accuracy of 94%), but we could observe a prominent contribution from a subclonal signature SBS17b in metastases and a clonal SBS30 signature distinguishing primary tumours from the advanced stage (Figure 3.16).

While we are not proposing these classifiers for clinical application, this analysis does suggest that there are distinct contributions of mutational processes over a lifetime of a tumour which are prevalent enough to be somewhat predictable.

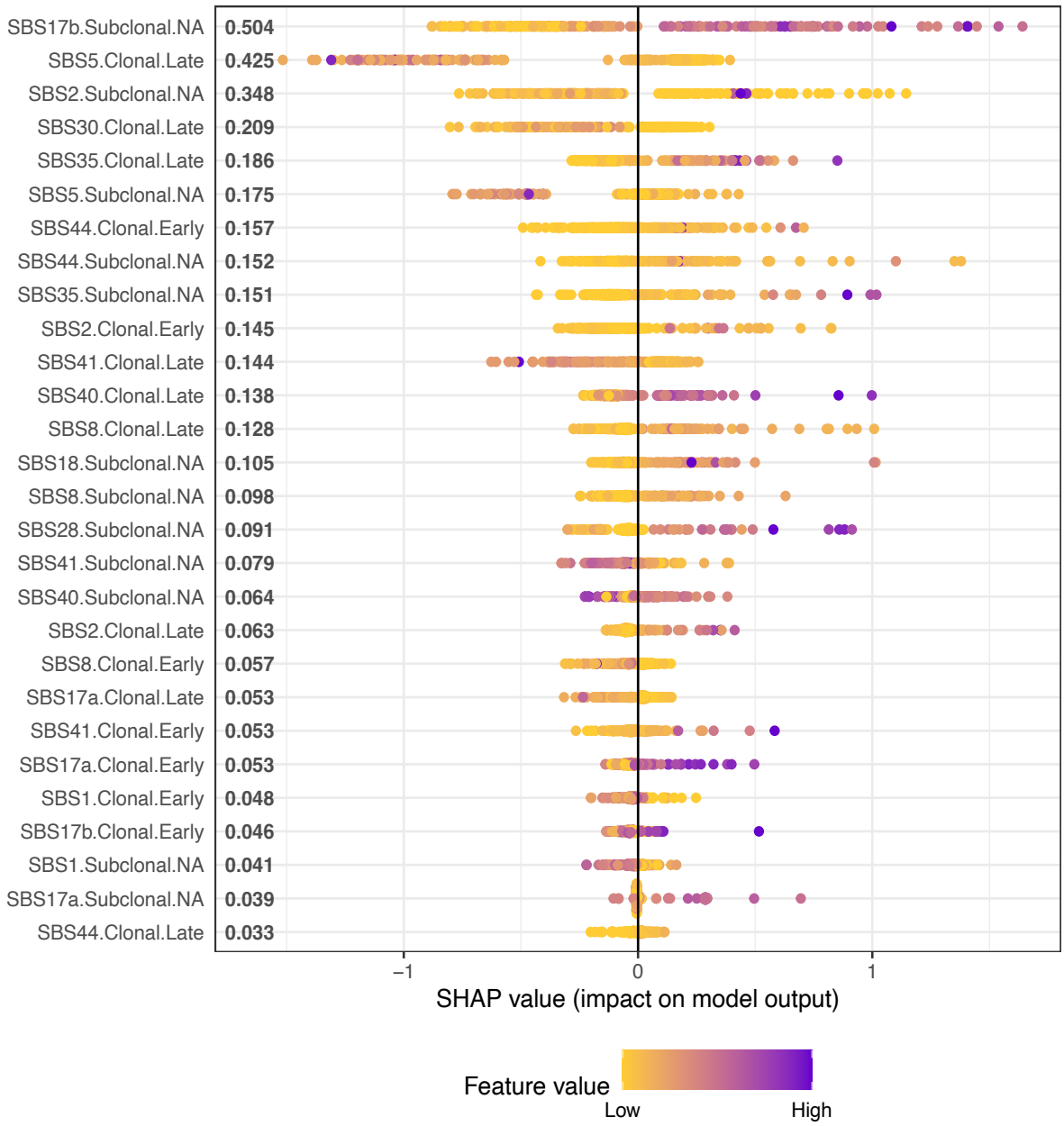


Figure 3.16: Gradient boost classifier results distinguishing metastases from primary tumours based on mutational signature prevalence, clonality and timing. Features are ordered according to their ranking in the model (top ranking features first). Every dot is a sample and the colour corresponds to the signature contribution in that sample. Features linked with metastasis have a positive Shapley score, those linked with primary tumours a negative score.

3.5 DNA repair pathway dysregulation modulates mutational events in OAC development

We next investigated how DNA damage repair (DDR) regulation might contribute to shaping the mutational landscape of this disease. First, we asked whether tumours that accumulate higher loads of mutations due to homologous recombination (HR) deficiency, as evidenced by presence of mutational signatures SBS3 and SBS8, may also harbour a genomic context that positively selects for specific cancer drivers. Indeed, by investigating the ratios of non-synonymous to synonymous mutational burden (dNdS) of cancer drivers in samples with a dominant DDR impairment phenotype versus the rest (see Method section 2.8), we identified the genes KMT2D, CNOT3 and ABI1 as positively selected specifically in the context of DDR impairment (Figure 3.17). KMT2D is a histone methyltransferase frequently altered in oesophageal squamous cell carcinoma[93,94,95] but less well characterised in adenocarcinoma. Mutations in this gene have been linked with transcriptional stress and genomic instability[96]. CNOT3, part of the CCR4-NOT complex, is involved in mRNA processing and degradation and was shown to contribute to DNA damage and replication stress responses in yeast[97,98].

ABI1 is part of the c-Abl system facilitating signal transduction of tyrosine kinases, has been shown to regulate DNA damage-induced apoptosis[99] and is downregulated in gastrointestinal cancers[100]. While none of these genes have been highlighted as major regulators in OAC, this analysis indicates they may play very specific roles in the context of high DDR deficiency.

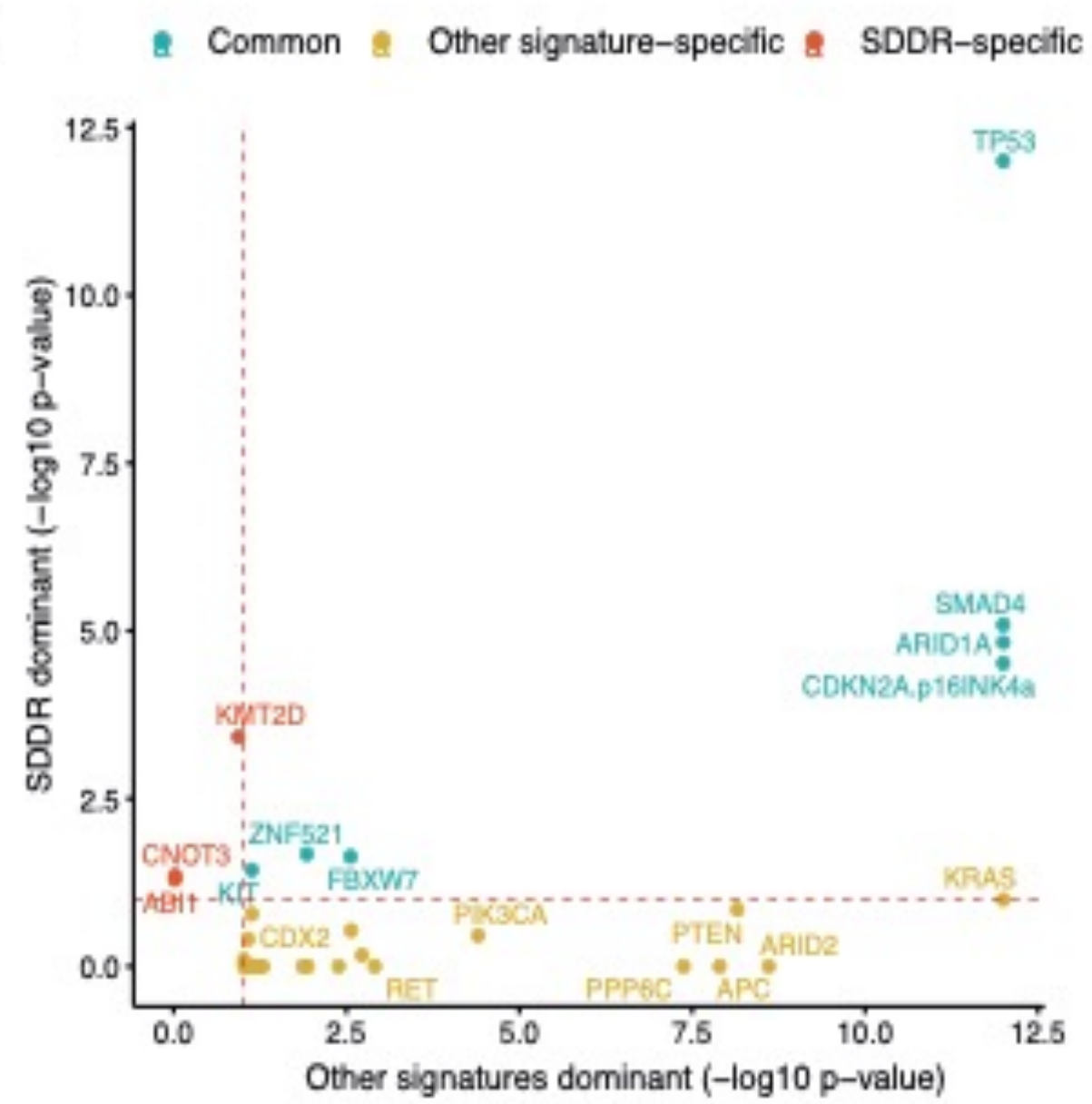
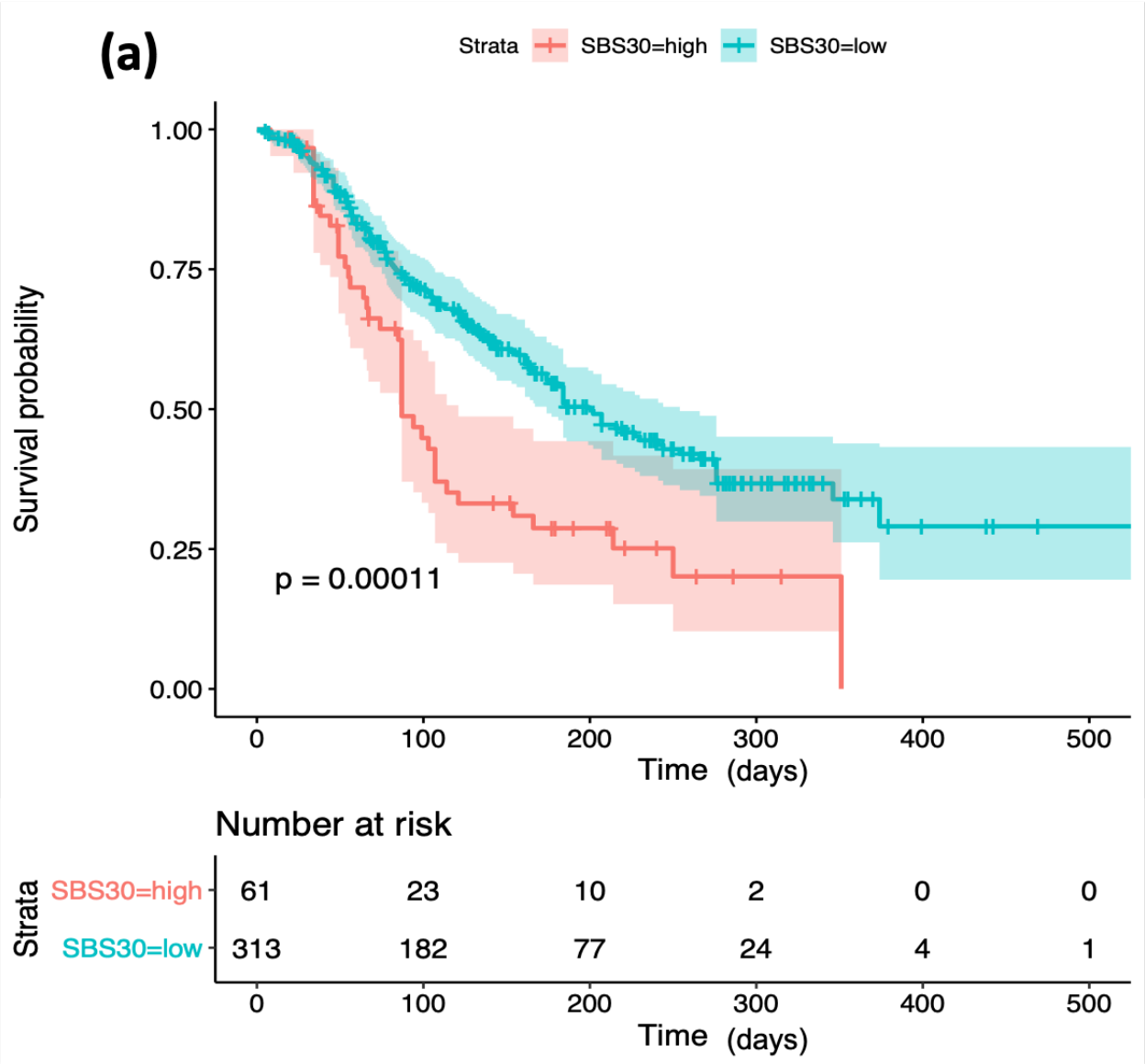


Figure 3.17: DNA damage repair signatures and associated driver genes. Positively selected genes in primary tumours with dominant DNA damage repair impairment signatures versus the ones positively selected in tumours with other dominant signatures. Genes commonly positively selected in both categories are highlighted in blue. Genes positively selected only in the SDDR dominant group are highlighted in red.


(a)
Strata
SBS30=high
SBS30=low
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.00011
0
100
200
300
400
500
Time (days)
Number at risk
Strata
SBS30=high
SBS30=low
61
23
10
2
0
0
313
182
77
24
4
1
0
100
200
300
400
500
Time (days)

(b)

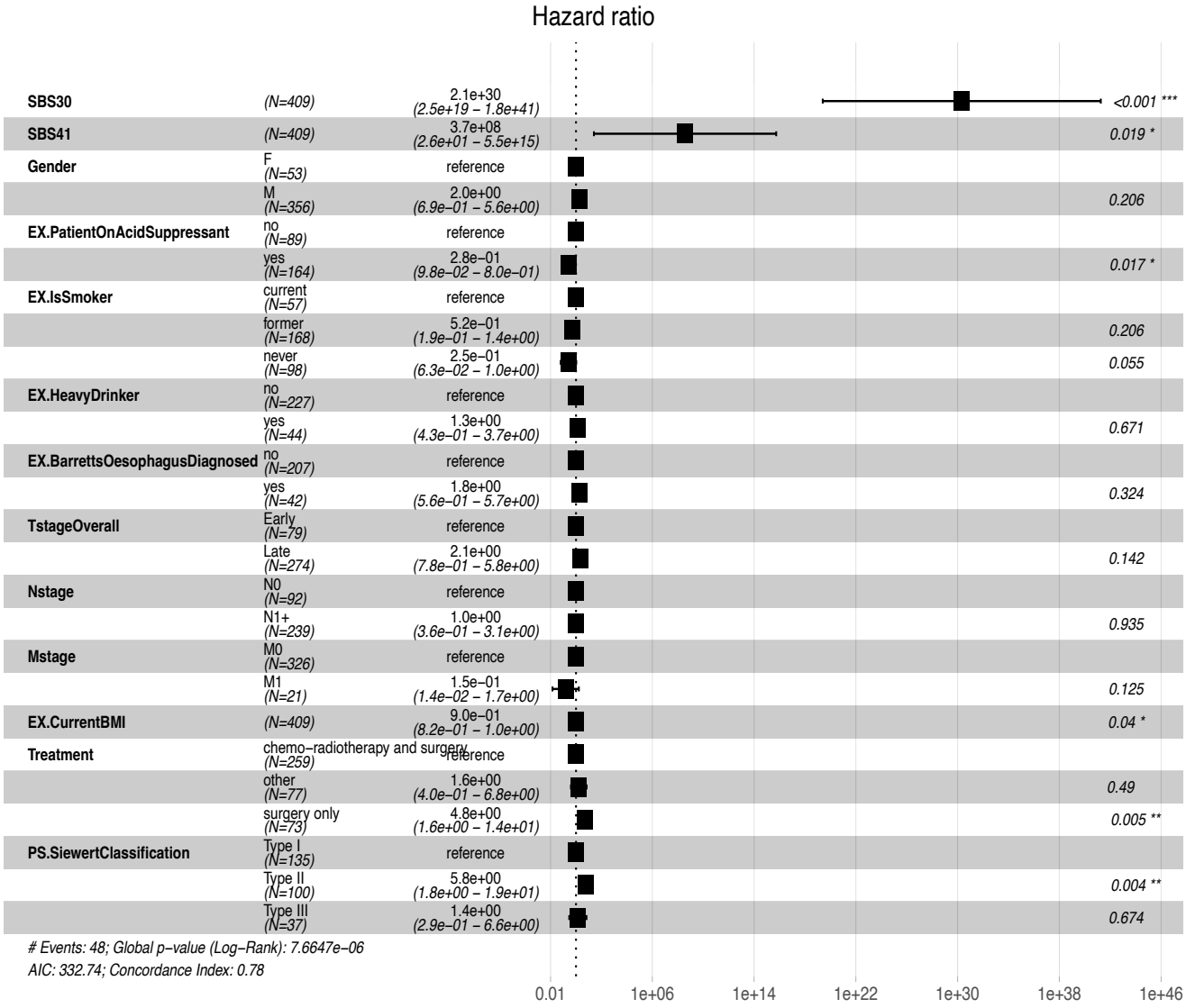


Figure 3.18: Clinical relevance of Base Excision Repair associated signature (SBS30).

(a) Patients with a BER signature prevalence >3% have a significantly worse overall survival outcome.

(b) Multivariate Cox plot of SBS30 with other clinical factors.

Beyond HR impairment, a newly discovered mutational process in the cohort was that linked with base excision repair impairment (SBS30). Remarkably, this signature was also the most prognostic in our cohort, even after accounting for confounding factors such as age, gender,

stage etc (Figure 3.18). Patients showing any evidence for BER deficiency in their tumours (>3%) had a worse overall survival (Figure 3.18a), suggesting a potential prognostic utility for this signature in the clinic.

Finally, we reasoned that DDR pathways might accumulate deficiencies earlier or later during evolution, thus enabling the fixation of mutations generated by various neoplastic processes. To investigate this further, we used non-negative matrix factorisation across >400 genes acting in 13 DDR-related pathways to describe the temporal distribution of putative driver mutations across these pathways. When surveying the 300 primary tumours, 27 Barrett's cases and 17 metastases which harboured nonsynonymous mutations in DDR-related genes, we could observe that the mutational insults favouring the evolution of OAC appear concentrated on certain key DDR processes, including nucleotide excision repair (NER), mismatch repair (MMR), homologous recombination (HR), base excision repair (BER), translesion synthesis (TLS) and the G1-S cell cycle checkpoint (Figure 3.19a). Across the entire cohort, we identified four time-dependent signatures of DDR deficiency-linked mutational fixation in genomes: a generic signature of coupled NER/MMR deficiency, an early clonal signature of G1-S checkpoint damage, a late clonal signature of deficiency in HR and Fanconi Anemia (FA) pathways and a subclonal signature of joint mutations across the BER/TLS pathways (Figure 3.19a). Surprisingly, the signature of early G1-S damage appeared relatively increased in metastatic samples, perhaps highlighting the importance of G1-S repair control in early clones seeding metastases from the primary tumour (Figure 3.19b). In contrast, subclonal TLS/BER deficiencies were more prevalent in Barrett Oesophagus and NER/MMR defects dominated primary tumours (Figure 3.19b).

The four DDR-deficient signatures appeared largely nonconcurrent in the cohort, as shown in the heat map of Figure 3.19c. We confirmed that the mutated pathways were inactive in the

respective groups, and further observed concomitant upregulation/downregulation of other DDR pathways (Figure 3.19c, top). There were striking differences in the patterns of activation of other DDR pathways between patients with different DDR signatures. In particular, a multitude of single strand and double strand break repair pathways were increased in activity in the NER/MMR subgroup, presumably to partly compensate for the lack of repair via these mechanisms. The TLS/BER deficient subgroup presented a wide downregulation of most DDR pathways with the exception of damage repair in S phase, which appeared upregulated.

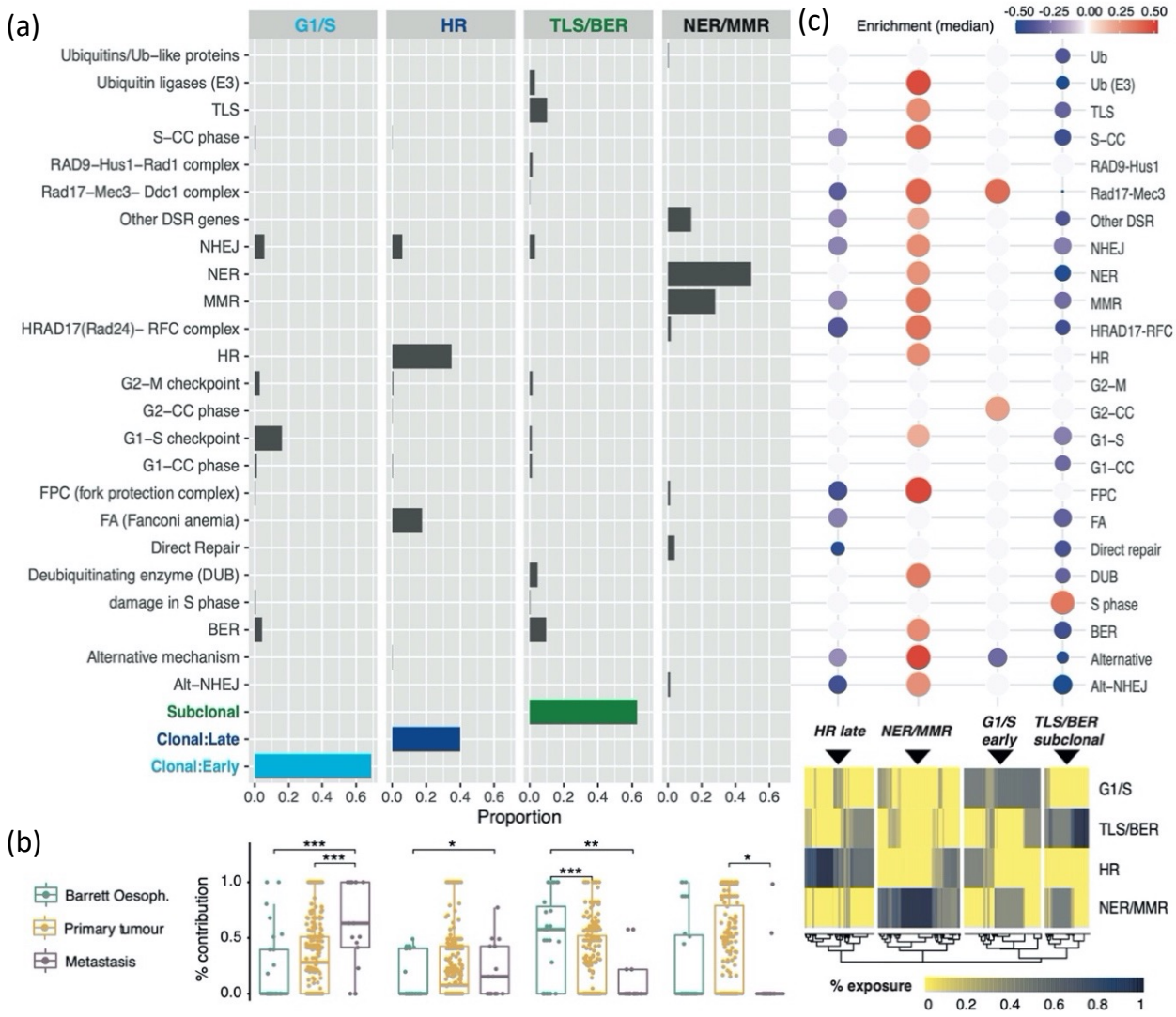


Figure 3.19: Mutational signatures of DNA damage repair, their timing and prevalence

(a) Signatures of DNA damage repair impairment in oesophageal adenocarcinoma. The frequency of nonsynonymous mutations in distinct DDR pathways is shown along with their timing during evolution and subclonality. (b) Prevalence of DDR signatures across the course of the disease. The mutational contributions are compared between Barrett Oesophagus, primary tumours and metastases for the four signatures identified previously keeping the same order as in (a). (c) Prevalence of signatures across the entire cohort (heat map) and corresponding median activity in every DDR-related pathway (balloon plot), as measured from expression of genes implicated in the pathway using GSVA. Blue circles indicate downregulation, red circles upregulation. *p<0.05; **p<0.01; ***p<0.001.

3.6 Evolutionary bottlenecks uncover a phenotypically distinct SBS17 mutagenic shift

To further understand how mutational processes shape evolutionary trajectories in OAC, we investigated the timing of mutation accumulation due to the different neoplastic processes identified in the cohort. We identified frequent evolutionary bottlenecks (~51% of samples) where mutational pressures change (Figure 3.20). Most of these changes were consistent across tumour stages, with the exception of SBS18 and SBS5, which increased only at Barrett's related bottlenecks, and decreased in primary tumour and metastasis subclones. The most notable change was a subclonal decrease in SBS17a/b mutations, corroborating the findings from the PCAWG consortium study in primary tumours[50].

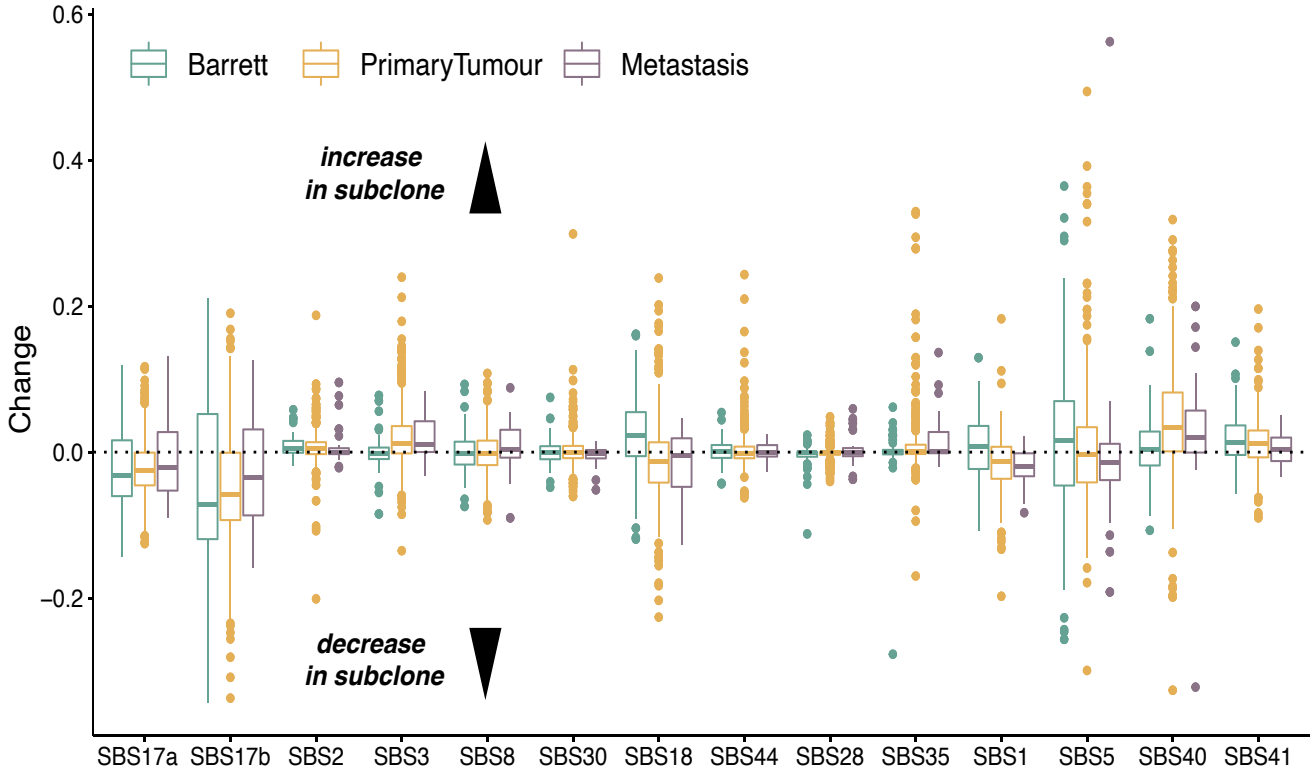


Figure 3.20: Evolutionary bottlenecks reveal widespread SBS17 shifts

Changes in signature exposure at evolutionary bottlenecks. Values below 0 indicate a decrease in signature exposure in the sublones, values above 0 an increase. Signatures SBS17a and b are the only ones showing a sublconal decrease across Barrett, primary and metastatic stages.

This decrease is observed across Barrett's, OAC and metastasis for SBS17a/b. These were by far the most dominant signals of dynamic shift observed during OAC evolution.

No association between SBS17a/b and p53 mutational status was found, but a higher prevalence of these signatures was more often found in tumours lacking mutations in several recurrent OAC drivers, including KRAS, PIK3CA, PTEN, ARID1A and APC ($p<0.01$). Instead, multiple cancer drivers involved in chromatin remodelling and transcriptional control, including SMARCA4, KMT2D and ARID2, were positively selected only in samples with abundant SBS17 signals (Figure 3.21).

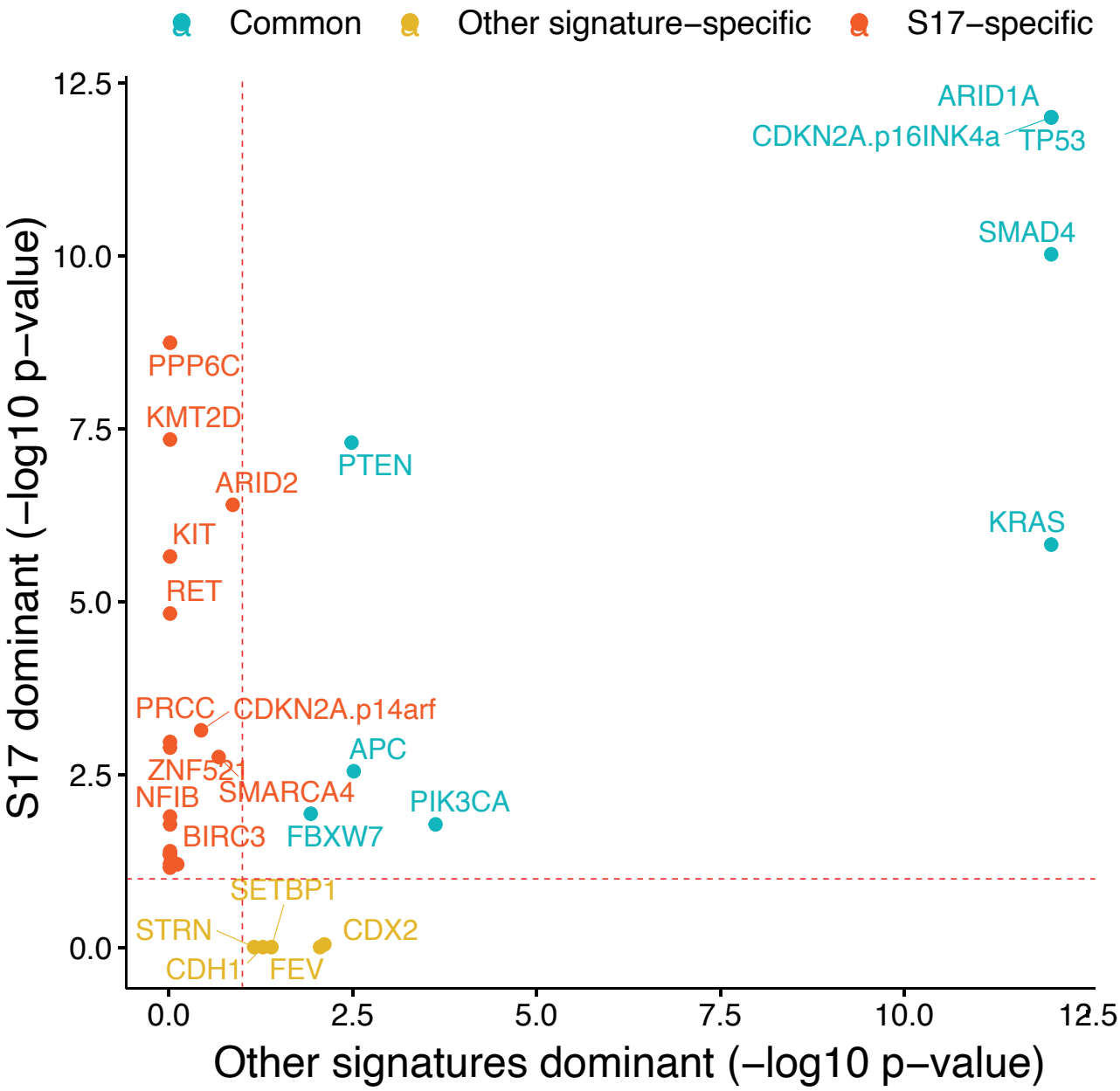
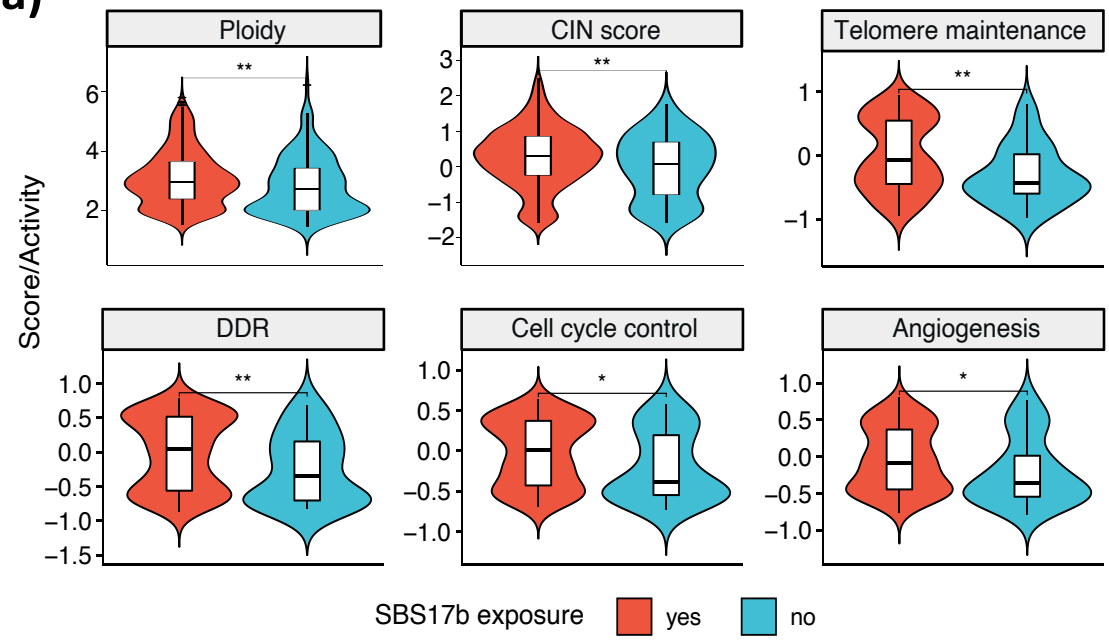


Figure 3.21: Positively selected genes in primary tumours with a dominant SBS17 signature versus the ones positively selected in tumours with other dominant signatures. Genes commonly positively selected in both categories are highlighted in blue. Genes positively selected only in the SBS17 dominant group are highlighted in red.

Tumours with SBS17b exposure displayed increased ploidy and chromosomal instability, as well as higher DDR activity, telomere maintenance, cell cycle control and angiogenesis (Figure 3.22a). Furthermore, samples where the SBS17 process increased in intensity at the bottleneck were more often p53 wild type (Fisher's exact test p = 0.0004, 1.9-fold enrichment) and showed a decreased CD8+/CD4+ T cell, regulatory T cell and monocyte infiltration (Figure 3.22b). All of these aspects seem to suggest a role for SBS17 in promoting tumour progression.

(a)


Ploidy
CIN score
Telomere maintenance
DDR
Cell cycle control
Angiogenesis
Score/Activity
SBS17b exposure
yes
no

(b)

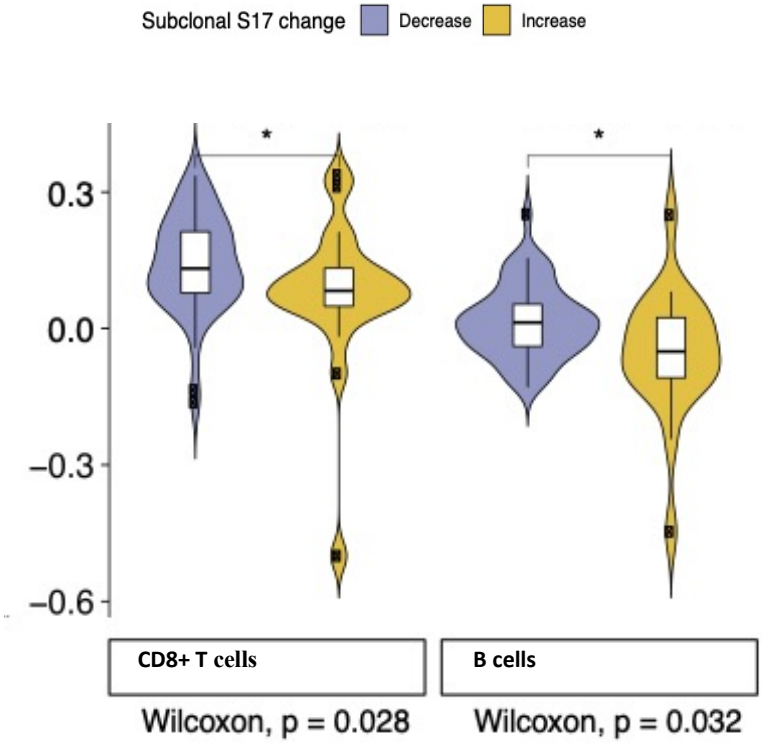

Subclonal S17 change
Decrease
Increase
CD8+ T cells
B cells
Wilcoxon, p = 0.028
Wilcoxon, p = 0.032

Figure 3.22: Signature 17 associated processes influence modulation of cellular and microenvironmental phenotypes. (a) The presence of SBS17b is associated with an increase in ploidy and chromosomal instability, as well as higher activity of telomere maintenance, DNA damage repair, cell cycle control and angiogenesis pathways. (b) Effect of signature 17 exposure on immune cell types.

Focusing on SBS17 impact, we found that SBS17a and SBS17b were individually prognostic in the cohort, with higher exposure correlating with better patient outcome (Figure 3.23).

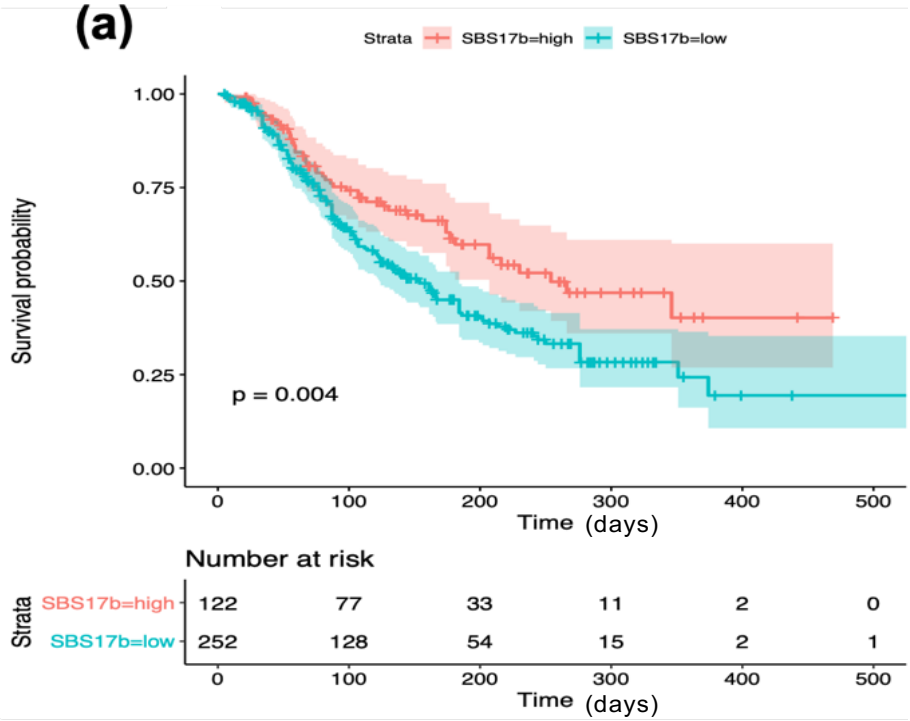
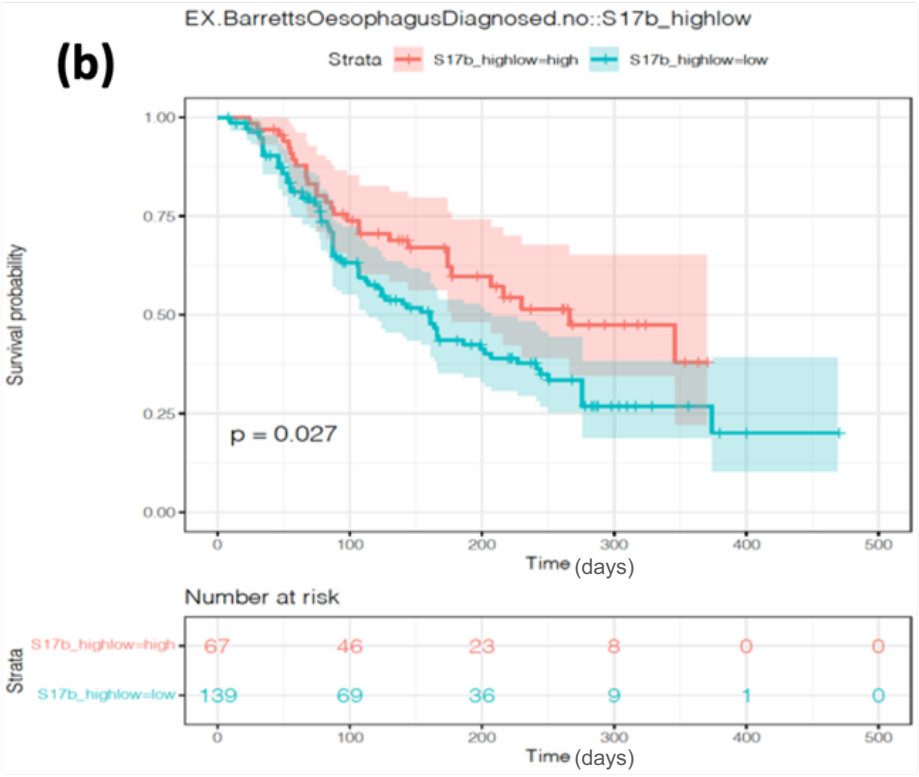


Figure 3.23: Prognostic relevance of SBS17a/b

(a)SBS17b proportions were linked to better out of the patients in OAC cohort



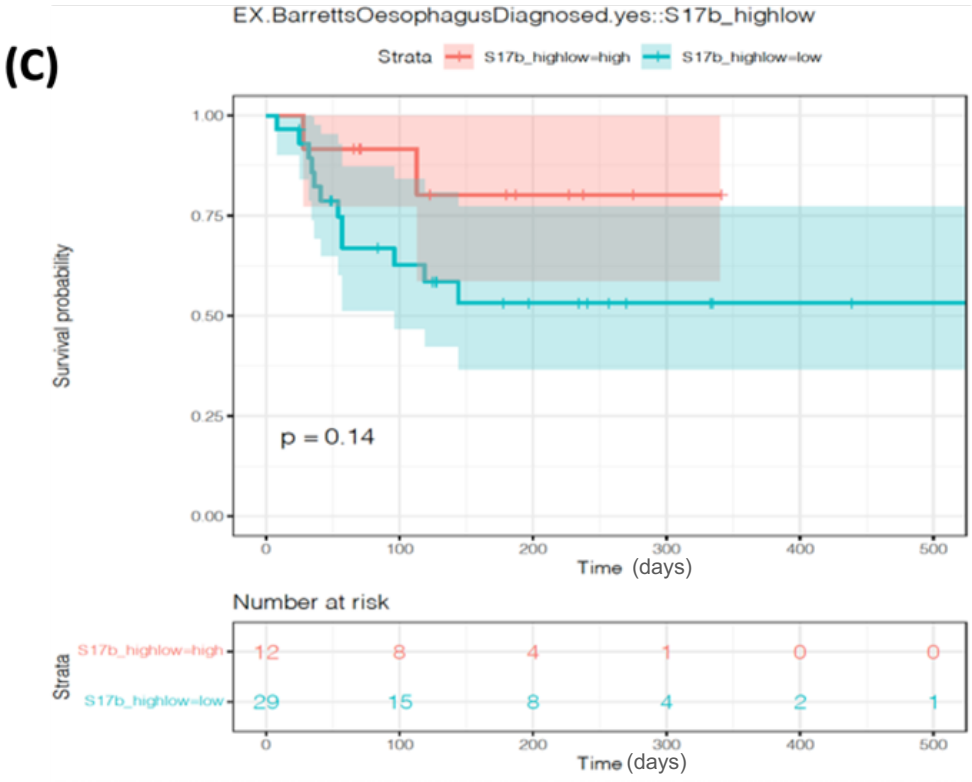


Figure 3.23: Prognostic relevance of SBS17a/b.(b)Higher proportions of SBS17b in OAC Patients with adjacent Barret's were associated with improved survival. (c) Similar trend is observed in patients without Barrett's, but the association is statistically not significant.

3.7 Summary

In this chapter, I have presented a comprehensive analysis of mutational signatures acting across different stages of OAC development. A graphical summary is presented in Figure 3.24.

In order to understand what mutational forces drive disease progression from pre-cancerous stages to advanced malignancy in OAC, we surveyed a cohort of 997 patients across different stages of oesophageal adenocarcinoma progression, from pre-malignant to advanced disease. Based on the pattern of single base substitutions observed from whole-genome sequencing data, I inferred the mutational processes that are likely to have acted during the evolution of this cancer. In addition, I have characterised the prevalence of mutational signatures across cancer stages and determined their association with a range of parameters including lifestyle factors and prognosis. I identify consistent evidence for specific DNA damage repair deficiencies and pinpoint evolutionary bottlenecks that play a key role in shaping the progression of this disease.

I observed that SBS17b starts early in the precancerous stage and tends to increase in the late stage. Most of these tumours are chemotherapy treated and therefore the increase in late stage suggests that these may be treatment effects. Endogenous processes such as APOBEC and BER are low in magnitude but further decrease in late stages. Colibactin signatures were identified, hinting at the possibility of *E.coli* colonisation during OAC development. SBS35-platinum therapy signature was enriched in treated samples along with MMR linked SBS44 co-occurring in these samples.

After characterising mutational signatures, I investigated how the endogenous and external risk factors influence and shape the mutational signatures in OAC development. I observed a

strand asymmetry during transcription and replication, which was explained by distribution of T>G SNVs to untranscribed and lagging strands of DNA. A common 10bp nucleosome periodicity pattern was observed across stages of OAC development, suggesting a common mutational process implicated in this disease. Further, I performed clinical correlation analysis with the risk factors (data from Barrett's and OAC patients) and the relative proportions of mutational signatures. I found positive correlations with alcohol consumption (SBS8, SBS17a and SBS44) smoking habits (SBS1,SBS8,SBS17a, and SBS18) and use of anti-inflammatory drugs (NSAIDs) (SBS2,SBS3,SBS30,SBS35 and SBS40).Tumour factors such as positive nodes(SBS3,SBS8 andSBS17b) and Siewert classification(SBS3 and SBS8) were also positively correlated with OAC signatures.

We also investigated mutational signatures co-occurring across different stages, these were informative to distinguish stages of cancer development. Further, we looked for SNVs and their timing in DNA repair pathways and identified four different signatures. These were clustered into four subgroups(HR-Late; NER/MMR, G1/S-Early and TLS/BER-Subclonal). Also identified, driver genes (KMT2D, CNOT3 and ABI1) associated with these signatures.

We then investigated the evolutionary bottle necks for signatures during OAC development. We identified that SBS17b decreases sub-clonally across all the stages of OAC development and other signatures remain unchanged, suggesting the most dominant signals of dynamic shift observed during OAC development. We also investigated cellular and microenvironmental phenotypes being associated with SBS17b. Chromosomal instability, and ploidy was increased along with higher DDR activity, telomere maintenance, cell cycle control and angiogenesis.

We then investigated the clinical relevance/prognostic value of signatures, BER associated SBS30 was found to be strongly prognostic in our cohort. Whereas SBS17b proportions correlated with better patient outcome.

Overall this study provides the evidence on how mutations shape the development of OAC and can be exploited to guide therapy and patient stratification.

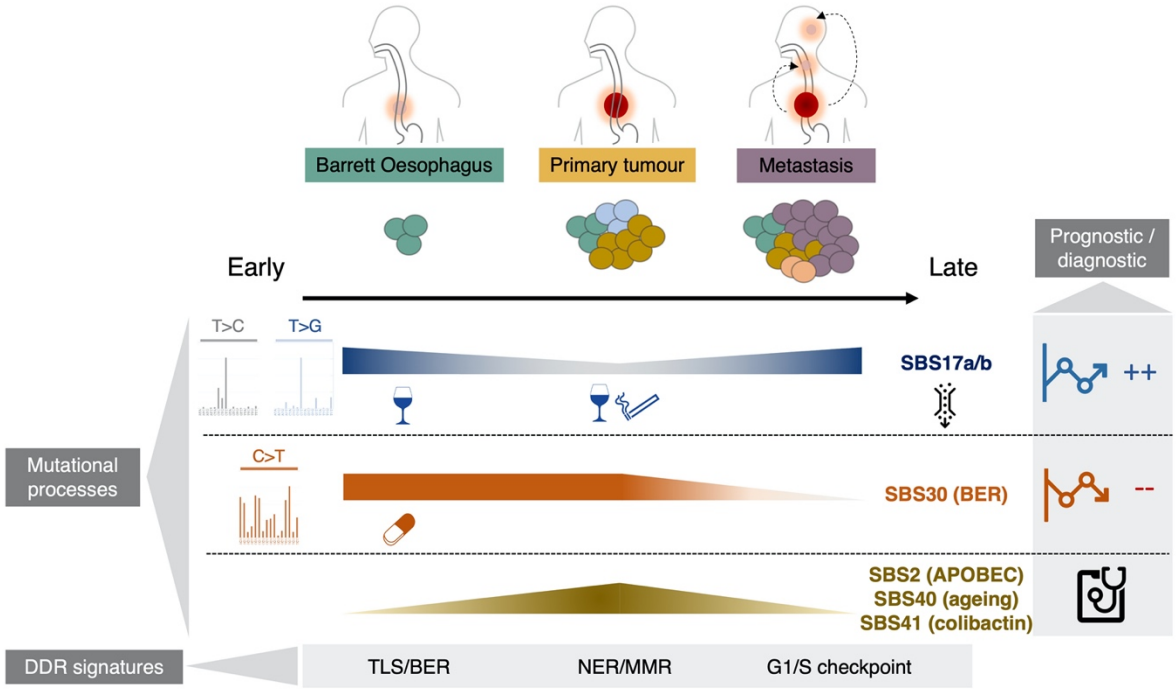


Figure 3.24: Key genomic signatures underlying distinct exposures, expansion and outcomes during OAC evolution from pre-cancerous to advanced disease. SBS17a/b processes show a relative decrease in the primary tumours compared to Barrett and metastasis cases, and are linked with alcohol and smoking exposures and better survival outcomes. Frequent bottlenecks appear for this signature, indicated by the symbol. SBS30 appears elevated in Barrett and primary tumours and is linked with NSAID usage in pre-cancerous stages and worse survival. Several signatures including SBS2/SBS40/SBS30 are most highly represented in primary tumours and can be used to distinguish this stage. DDR signatures specifically elevated at distinct points during OAC evolution are also highlighted.

4.Results Chapter

A novel DNA sequencing method for quantifying mutational signatures in clinical cancer samples

4.1 Attribution:

This chapter is adapted from a manuscript which was published in Nature Communications in January 2020:

Perner,J.[#], **Abbas, S.[#],** Nowicki-Osuch,K.[#], Devonshire,G., Eldridge,M,D. Tavaré,S., Fitzgerald, R,C.[*] The mutREAD method detects mutational signatures from low quantities of cancer DNA. *Nat Commun* 11,3166(2020).

https://doi.org/10.1038/s41467-020-16974-3

[#] Equally contributing authors; *Corresponding author

4.2 Author Contributions:

Juliane Perner and my supervisor Prof. Rebecca C. Fitzgerald conceived the project concept. I designed and performed the wet laboratory experiments to test the computational prediction. Under the supervision of postdoc Karol Nowicki-Osuch I developed the lab protocol to push the boundaries of making the assay as clinically relevant as possible. All the bioinformatic analyses were performed by Juliane Perner. Ginny Devonshire and Mathew Eldridge helped with sequencing data management and assisted Juliane in bioinformatic data analysis. Prof. Rebecca C. Fitzgerald and Prof. Simon Tavare supervised and obtained funding

for the study. I wrote the first draft of the paper then Juliane Perner, Karol Nowicki-Osuch and Prof. Rebecca C Fitzgerald helped to improve the manuscript.

The contributions as noted in the paper are adapted in the below paragraph:

J.P., S.A. and K.N. designed experiments, interpreted the data, and wrote the manuscript. S.A. and K.N. performed experiments and J.P. performed computational analysis. G.D. conducted sequencing data management. M.E. contributed expertise for sequencing data analysis for this study. R.C.F. and S.T. supervised the work and helped to write the manuscript. R.F. obtained funding for the study. All authors approved the final version of the manuscript.

4.3 Acknowledgements

The OCCAMS consortium for sample collection and sequencing was funded by a Programme Grant from Cancer Research UK. The laboratory of R.C.F. is funded by a Core Programme Grant from the Medical Research Council (RG84369). We thank the Human Research Tissue Bank, which is supported by the UK National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre, from Addenbrooke's Hospital. Additional infrastructure support was provided from the Cancer Research UK–funded Experimental Cancer Medicine Centre. We thank Chris Laumer (EBI UK), the Genomics core at CRUK CI and the bioinformatics core at CRUK CI for valuable discussion and support. We would like to thank Dr Maria O'Donovan, Dr Shalini Malhotra and Dr Ahmad Miremadi for histopathological assessment of samples.

4.4 Data availability

All mutREAD data generated for the article is available from European Genome-phenome Archive(accession number EGAD00001006170, https://ega-archive.org/datasets).

WGS data for the matched patient samples is available from the ICGC data portal (https://dcc.icgc.org/).

4.5 Code availability

All analysis code is freely available from https://github.com/jperner/mutREAD.

4.6 Rationale

Mutational processes acting on cancer genomes can be traced by investigating mutational signatures. Because high sequencing costs limit current studies to small numbers of good-quality samples, we developed a robust, cost- and time-effective method, called mutREAD, to detect mutational signatures from small quantities of DNA, including degraded samples. We also show that mutREAD recapitulates mutational signatures identified by whole genome sequencing and this will ultimately allow the study of mutational signatures in larger cohorts and, by compatibility with formalin-fixed paraffin-embedded samples, in clinical settings.

The method is based on the premise that obtaining a random subset of all mutations is sufficient to determine the presence of mutational signatures. To test this assumption, we first performed computational simulations (methods section 2.14) using available data from whole-genome sequencing of 129 esophageal adenocarcinoma (OAC) samples and the six mutational signatures derived from them[35].

4.7 Mutational Signatures analysis on computationally simulated data

The stability of the mutational signature profile was evaluated as a function of the number of randomly selected mutations detected in the WGS samples (Figure 4.1). The cosine similarity relative to the original mutational signature profile increases with the number of mutations available for estimation. A plateau is reached at 500 mutations, suggesting that fewer than the number of mutations derived from WGS (on average 26k mutations per OAC sample) are sufficient to obtain the mutational signature profile. The second assumption is that the mutation subset generated by Reduced Representative sequencing (RR-Seq) is an unbiased representation of the mutational spectrum. We simulated subsets of mutations for RR-seq using different enzyme combinations, as well as for 10x shallow Whole Genome Sequencing (sWGS) and Whole Exome Sequencing (WES)(methods section 2.14) In this simulation, RR-seq with at least 161 out of 169 enzyme combinations outperforms (expanded) WES and 10x sWGS in terms of average cosine similarity between the WGS-derived and simulated signature profile in OAC (Figure 4.2). This difference can in part be attributed to the number of mutations recovered by the different methods (Table 4.1). Notably, RR-seq derived mutations originate from a much lower proportion of the genome than (expanded) WES-based mutations (Table 4.2).

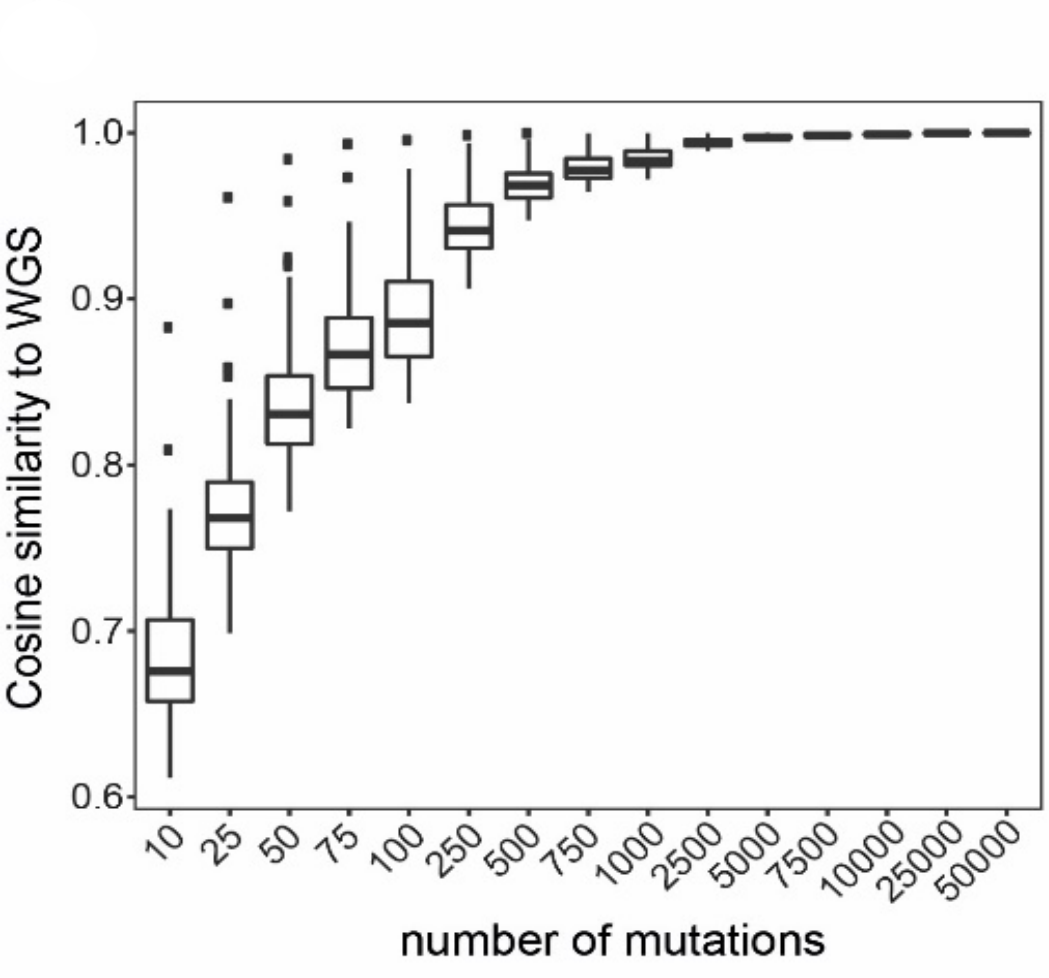


Figure 4.1 –Computationally simulated Mutational signatures: WGS v/s RR Seq

Cosine similarity (y-axis) of whole genome sequencing (WGS)-derived mutational signatures for 129 OAC samples and signatures derived from random subsets of mutations with increasing size (x-axis). Boxes show the 25% and 75% quartile with the median indicated by the bold line. Whiskers extend to 1.5 times the interquartile range and samples outside this range are indicated as points. Only samples having sufficient number of mutations (at least the number indicated on the x-axis) contribute to the boxes.

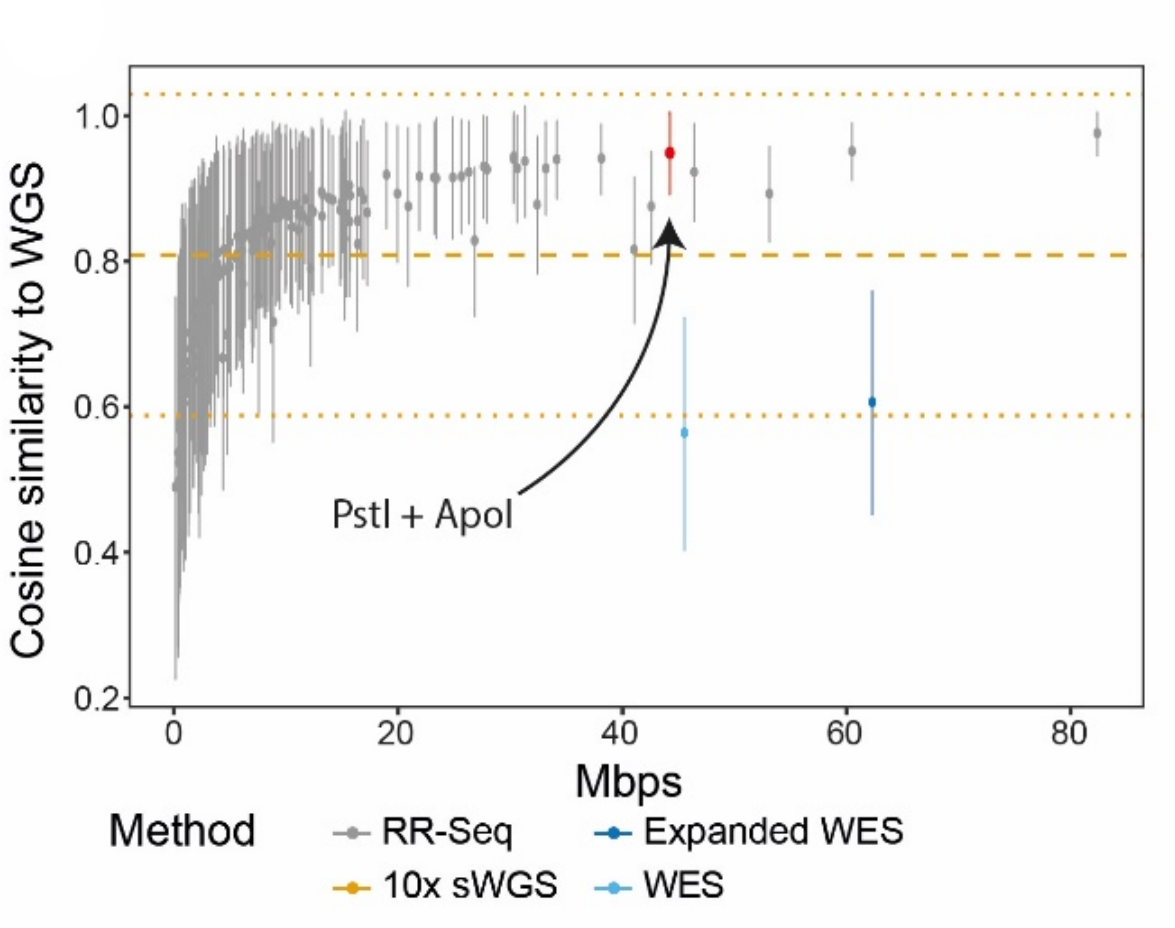


Figure 4.2 – Computationally simulated Mutational signatures across different methods. Cosine similarity (y-axis) of WGS-derived mutational signatures for 129 OAC samples and signatures derived from subsets of mutations simulating different sequencing approaches (x-axis). Points show the average cosine similarity and whiskers indicate the standard deviation across all 129 OAC samples. Different enzyme combinations were simulated for RR-seq, each shown as a different point. For 10x sWGS, the average across the 21 simulated samples is given as dashed horizontal line and the standard deviation given as dotted line. RR-Seq – reduced representation sequencing, 10x sWGS – 10x shallow whole genome sequencing, WES – whole exome sequencing, expanded WES – whole exome sequencing expanded to untranslated regions and miRNAs.

DNA Sequencing Method	Number of mutations recovered
Whole Exome Sequencing (WES)	211
Expanded WES	282
10X Shallow Whole Genome Sequencing (sWGS)	462
Reduced Representative Sequencing	381

Table 4.1: Comparative summary of mutations recovered by different DNA sequencing methods

DNA Sequencing Method	Proportion of genome covered Mean Mbps (percentage)
Whole Exome Sequencing (WES)	46(1.39%)
Expanded WES	62(1.88%)
Reduced Representative Sequencing	10 (0.3%)

Table 4.2: Comparative summary of portion of genome covered by different DNA sequencing methods

4.8 Computational simulations based Mutational Signature estimation using Pan-cancer WGS data

We further investigated the applicability of RR-seq for estimating mutational signatures in different cancer types using the WGS data collected by the Pan-Cancer Analysis of Whole Genomes (PCAWG) network[45]. RR-seq accurately estimated the mutational signature profiles across the majority of the 20 cancer types, including cancers with highly diverse mutational signature content, e.g. liver hepatocellular carcinoma (Liver HCC), and a non-solid tumour, i.e. B-cell non-Hodgkin lymphoma (Lymph-BNHL, Figure 4.3). As expected from our simulations above and in keeping with other signature algorithms, the performance of the method was correlated with the mutational load across cancer types (Figure 4.4). Finally, RR-seq outperformed (expanded) WES in all cancer types (Figure 4.5).

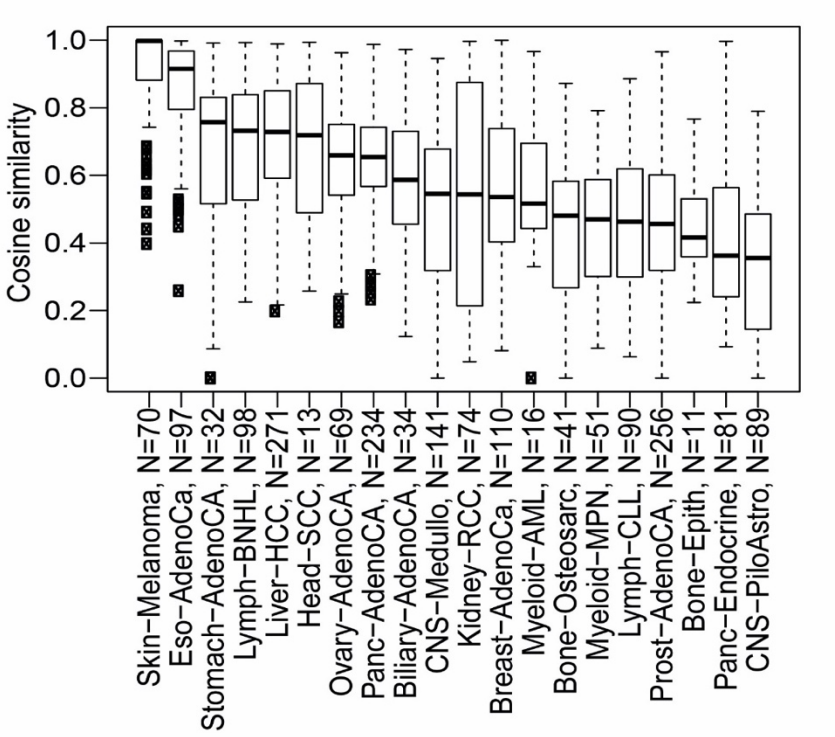


Figure 4.3: The efficiency of RR-seq-based mutational calling across the PCAWG tumour types. The distribution of cosine similarities between the RR-seq computational simulation-derived (best combination of enzyme per tumour type) and WGS-based mutational signatures.

Boxes show the 25% and 75% quartile with the median indicated by the bold line. Whiskers extend to 1.5 times the interquartile range and samples outside this range are indicated as points. For each cancer type the number of samples per group (N) is indicated within the x-axis labels.

Abbreviations: Eso-AdenoCa – Esophageal Adenocarcinoma; AdenoCA – Adenocarcinoma; Lymph-BNHL – B-cell Non-Hodgkin Lymphoma; HCC – Hepatocellular Carcinoma; Head-SCC – Head and Neck Squamous Cell Carcinoma; Panc-AdenoCA – Pancreatic Adenocarcinoma; CNS-Medullo – Medulloblastoma and variants; RCC – Renal Clear Cell adenocarcinoma, papillary type; Myeloid-AML – Acute Myeloid Leukaemia; Bone-Osteosarc – Osteosarcoma; Myeloid-MPN – Myeloproliferative neoplasm; Lymph-CLL – Chronic Lymphocytic Leukaemia; Prost-AdenoCa – Prostate Adenocarcinoma; Bone-Epith – Adamantinoma, Chordoma; Panc-Endocrine – Neuroendocrine carcinoma; CNS-PiloAstro – Pilocytic astrocytoma.

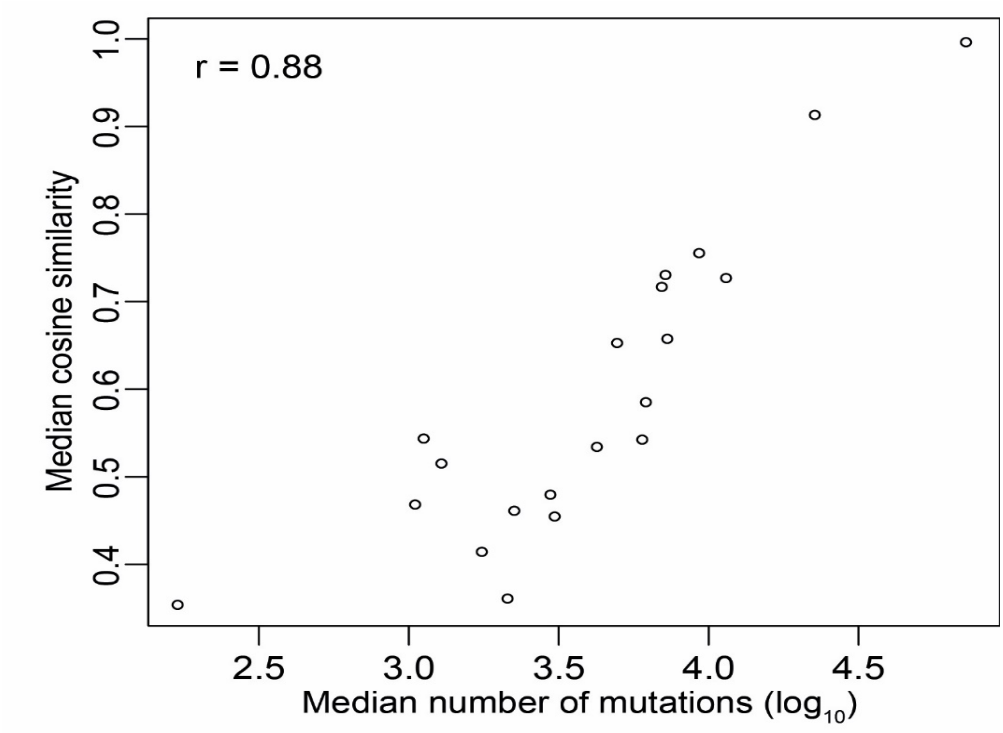
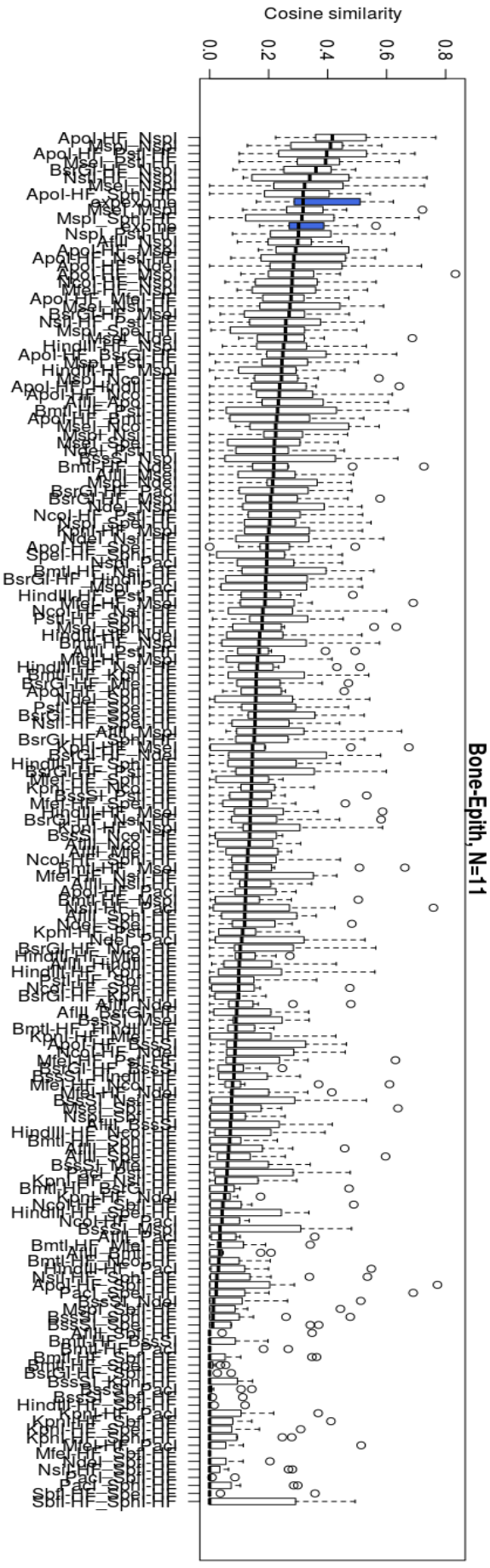
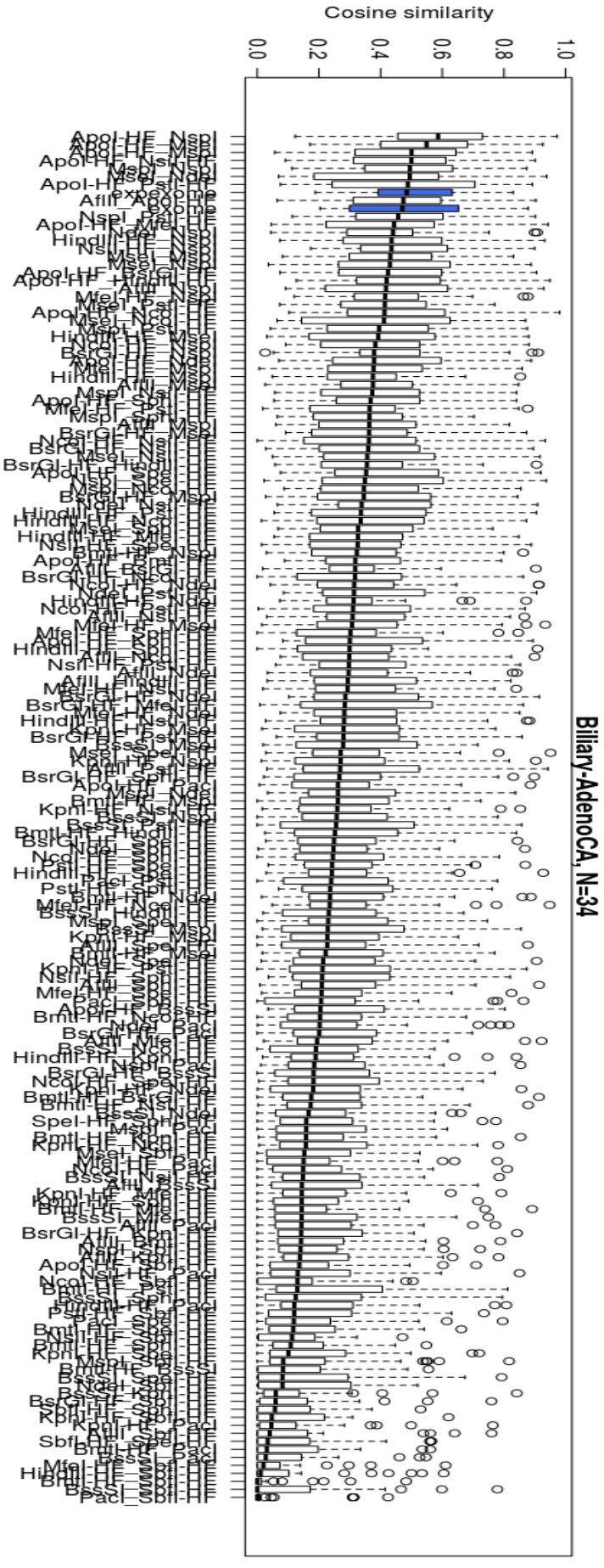


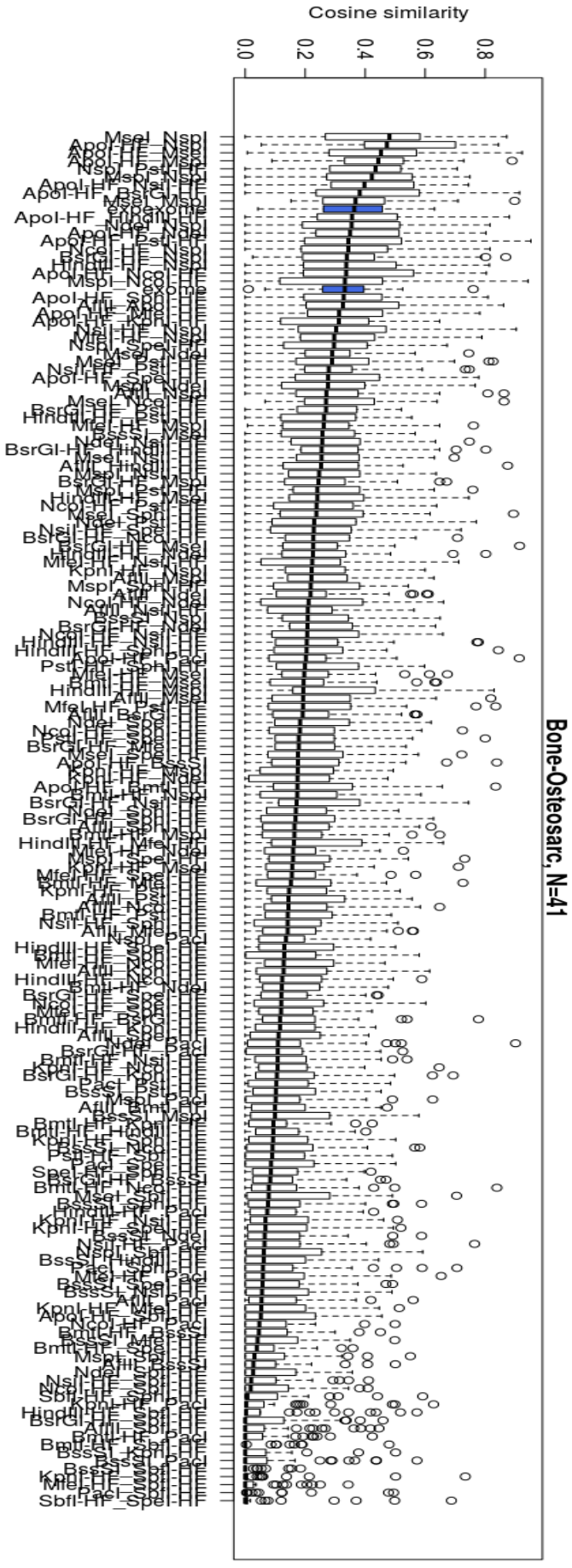
Figure 4.4: The efficiency of RR-seq-based mutational calling across the PCAWG tumour types(correlation). Scatterplot of the log10-scaled median number of mutations (x-axis) and the median performance of the RR-seq computational simulation-based mutational signatures measured by cosine similarity to the WGS-based mutational signatures (y-axis) per PCAWG cancer type. Each point represents one cancer type.

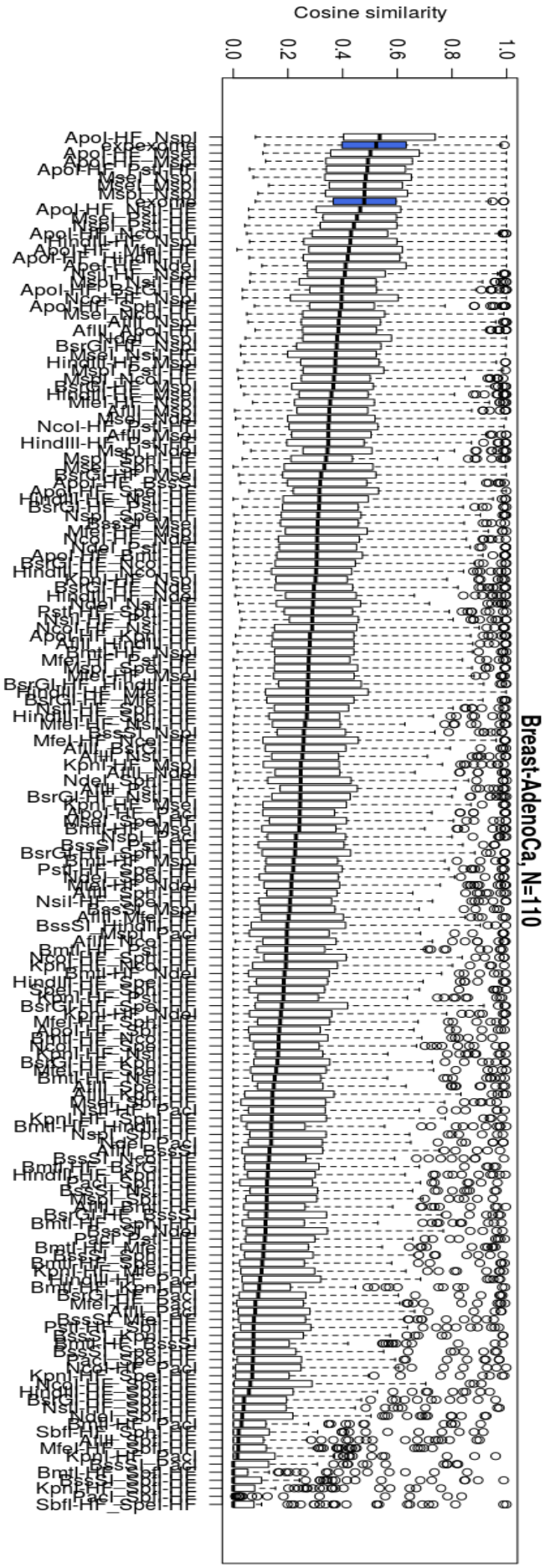
Figure 4.5: Mutational signatures computationally simulated across the PCAWG cohort.

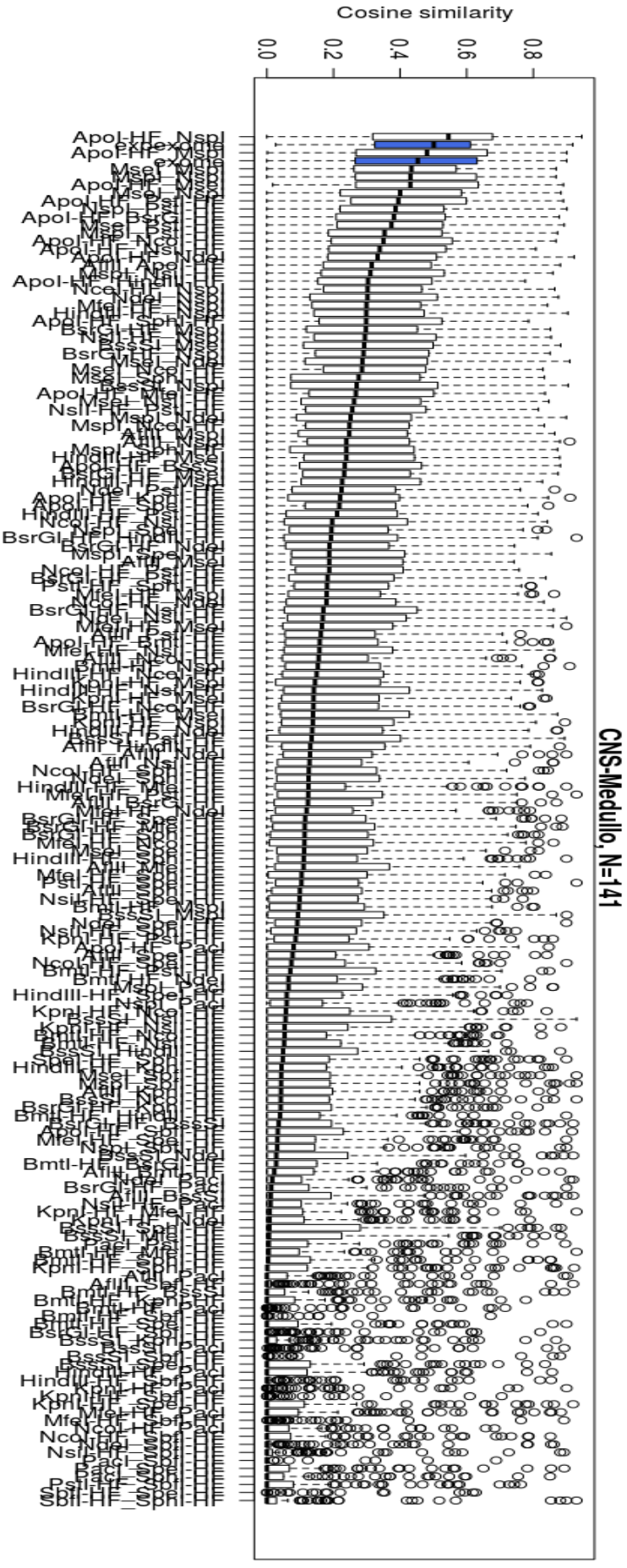
Summary of the cosine similarities (y-axis) of WGS-derived mutational signatures and mutational signatures derived from subsets of mutations simulating different sequencing approaches (x-axis) for each of the of individual tumour types from the PCAWG cohort. Boxes show the 25% and 75% quartile with the median across the samples indicated by the bold line. Whiskers extend to 1.5 times the interquartile range and samples outside this range are indicated as points. Different enzyme combinations were simulated for RR-seq, each shown as a different box. RR-Seq – reduced representation sequencing, WES – whole exome sequencing, expanded WES – whole exome sequencing expanded to untranslated regions and miRNAs. Title of each page contains abbreviated tumour name (explained in supplementary figure 1) and the number of samples used for the analysis.

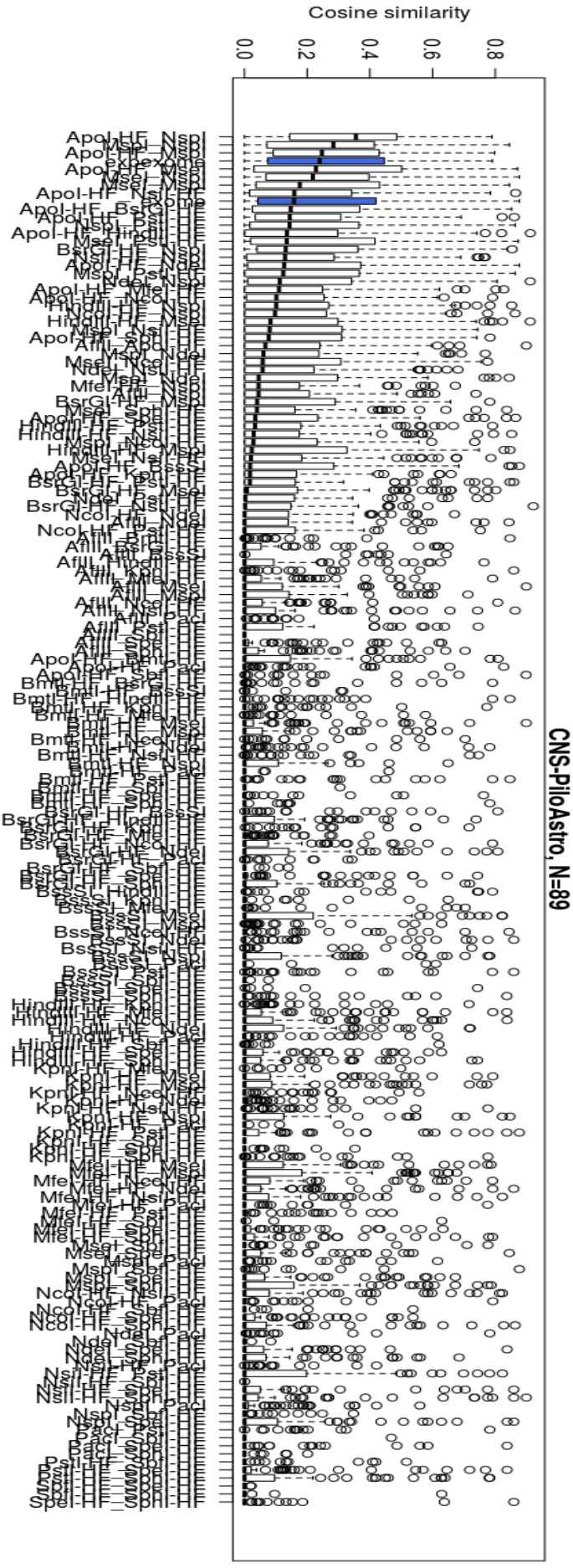

Cosine similarity
0.0
0.2
0.4
0.6
0.8
Bone-Epith, N=11

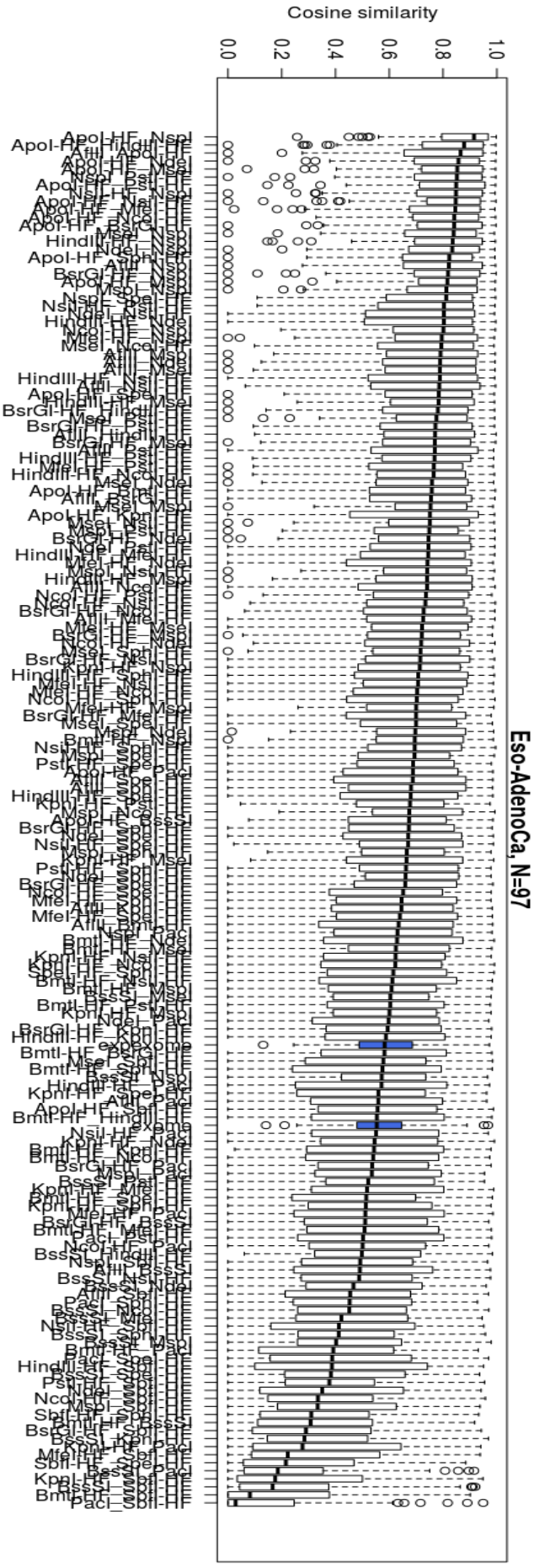

Cosine similarity
0.0
0.2
0.4
0.6
0.8
1.0
Biliary-AdenoCA, N=34

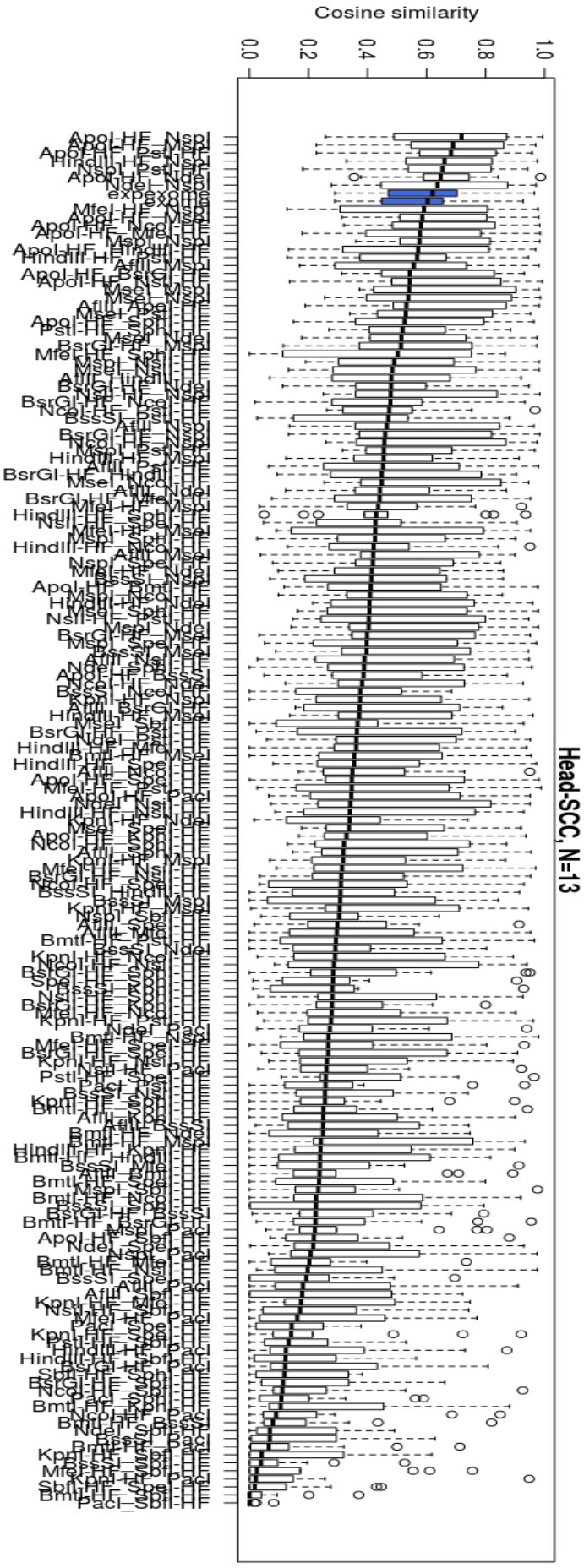

Cosine similarity
0.0
0.2
0.4
0.6
0.8
Bone-Osteosarc, N=41


Cosine similarity
0.0
0.2
0.4
0.6
0.8
1.0
Breast-AdenoCa, N=110


Cosine similarity
0.0
0.2
0.4
0.6
0.8
CNS-Medullo, N=141


Cosine similarity
0.0
0.2
0.4
0.6
0.8
CNS-PiloAstro, N=89


Cosine similarity
0.0
0.2
0.4
0.6
0.8
1.0
Eso-AdenoCa, N=97


Cosine similarity
0.0
0.2
0.4
0.6
0.8
1.0
Head-SCC, N=13

Having established superiority of RR-seq over other methods in the simulation, we implemented our approach, which we called mutREAD (**Mut**ational Signature Detection by **Re**striction **E**nzyme-**A**ssociated **D**NA Sequencing), by adapting and improving on the principles of the quaddRAD protocol[68]. Key features of the protocol include incorporation of Unique Molecular Identifiers (UMI) and inline barcodes, which allow for computational identification of PCR duplicates and larger multiplexing capabilities, respectively (Figure 4.6).

We further streamlined the protocol by simultaneous enzymatic digestion and adapter ligation and removal of unnecessary purification steps. Here, we optimized the protocol towards application to OAC, for the six mutational signatures that were previously identified from WGS on fresh-frozen samples(Secrier et al 2016)[35].

In particular, we chose the optimal pair of enzymes based on the simulation described above. The enzyme combination PstI and ApoI showed one of the highest cosine similarities to WGS results in OAC (Figure 4.2), as well as broad genome coverage and even distribution of target loci throughout the genome (Figure 4.7).

Hence, we designed adapter sequences that terminated with PstI and ApoI restriction enzyme compatible sites and that are devoid of PstI or ApoI restriction enzyme sites to avoid digestion of the adapters (Table 4.3).

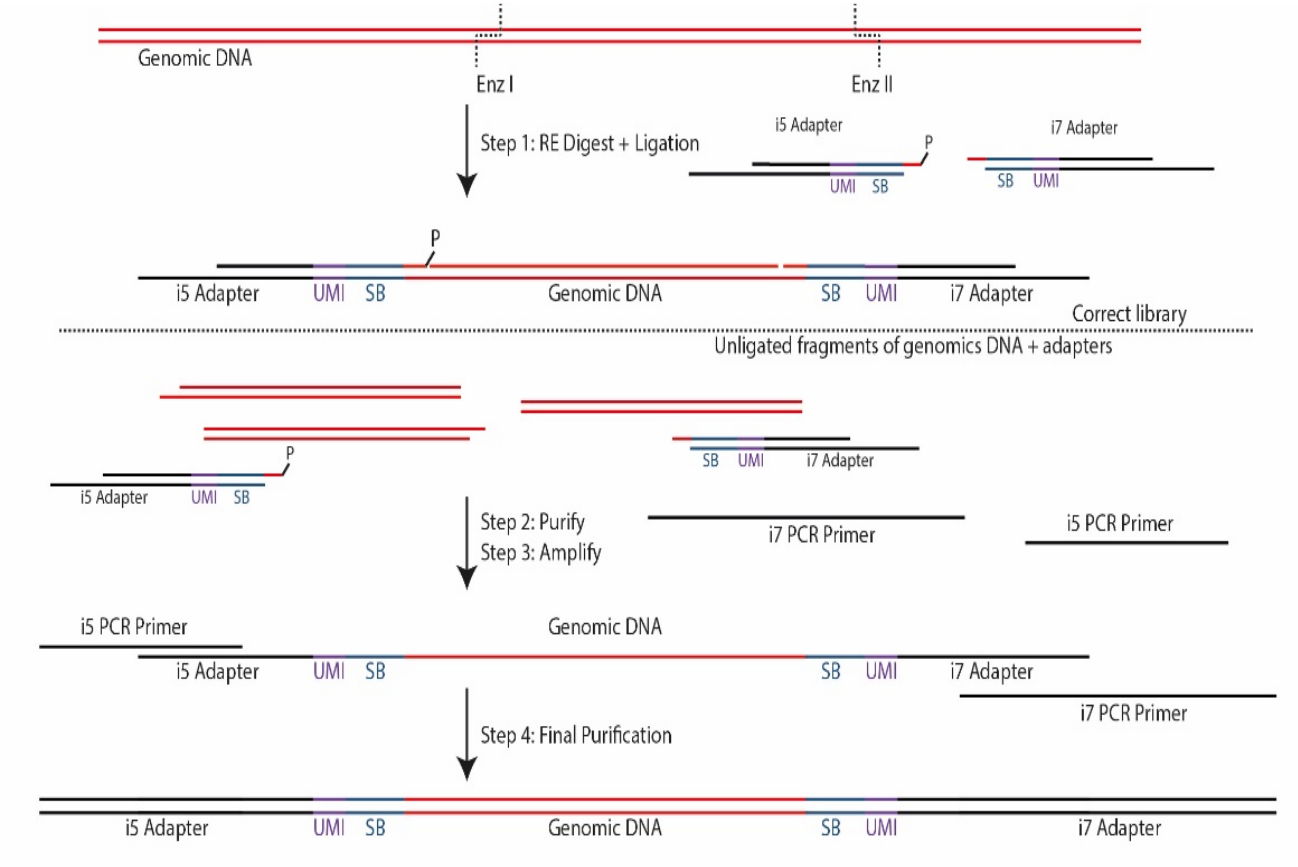


Figure 4.6: Method overview.

Schematic overview of the individual steps in mutREAD. Details for each step are given in the results section 4.9 below . SB – sample barcode, UMI – unique molecular identifier, RE – restriction enzyme.

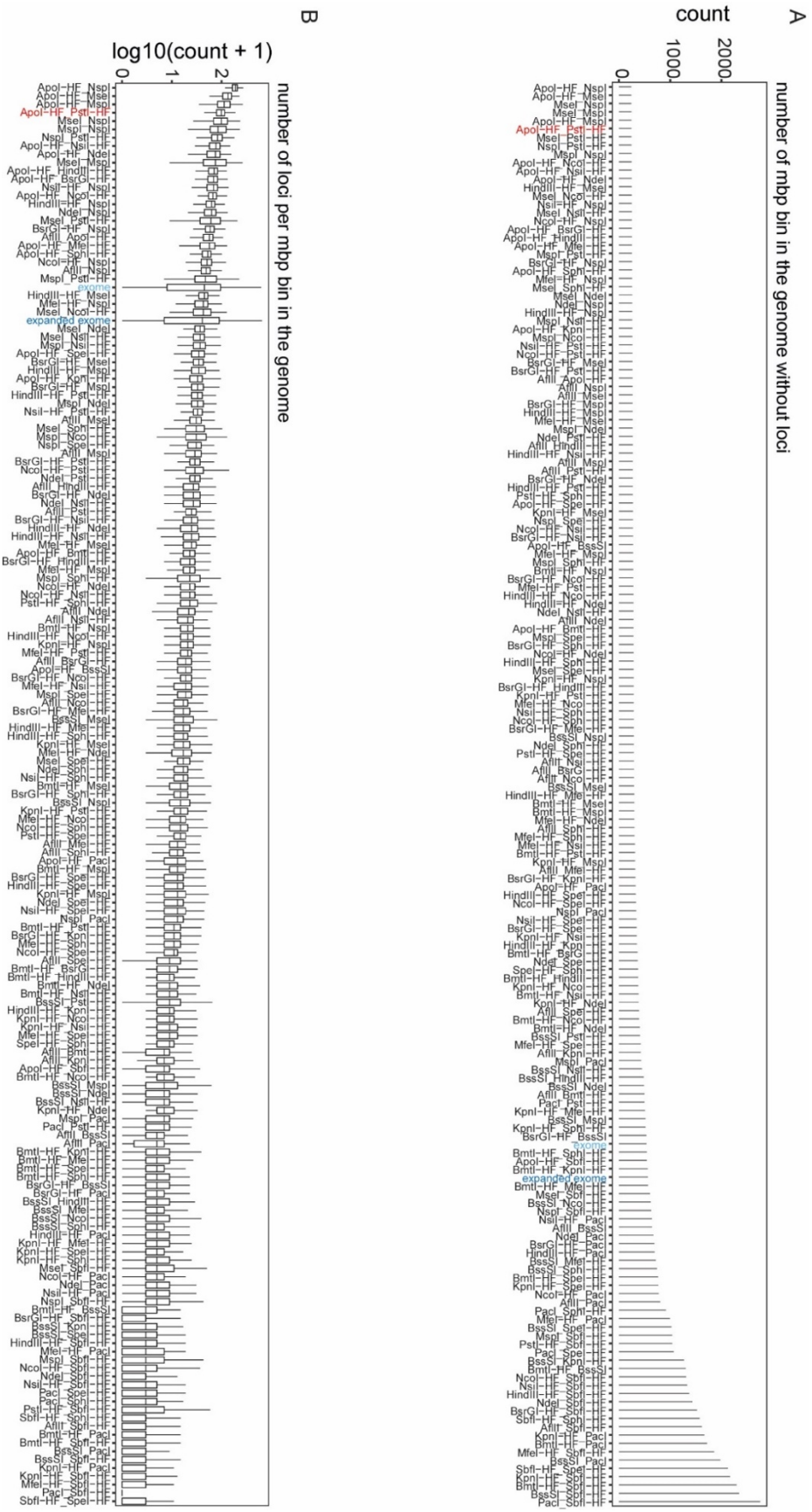

A
count
0
1000
2000
number of mbp bin in the genome without loci
B
log10(count + 1)
0
1
2
number of loci per mbp bin in the genome

Figure 4.7: Summaries of the genome-wide distribution of loci resulting from the different sequencing approaches

A) Bar plot of the number of genome-wide consecutive 1Mbps bins that are not covered by at least one expected loci in the computational simulation for each RR-seq with different enzyme combinations and (expanded) WES (x-axis).

B) Summary of the number of expected loci per 1Mbps bin on logarithmic scale (y-axis) for each RR-seq with different enzyme combinations and (expanded) WES (x-axis). Each box shows the 25% and 75% quartile with the median across all genome-wide consecutive 1Mpbs bins indicated by the bold line. Whiskers extend to 1.5 times the interquartile range and samples outside this range are indicated as points.

Here we present details of the laboratory assay development,

4.9 Assay optimization

All optimization experiments were performed using 500 ng of genomic DNA from an OAC cell line (FLO-1) that is commercially available from culture collection of Public Health England. In-house STR analysis was done in the lab to confirm a >90% match prior to assay optimization. Experiments were then repeated with frozen tumour, matched blood and FFPE tumour DNA from OAC patients.

4.9.1 Restriction digestion optimization for ApoI HF-PstI HF double digest

High-Fidelity (HF) ApoI and PstI restriction enzymes were obtained from New England BioLabs Inc. (Ipswich, Massachusetts USA). The optimization of restriction enzyme digestion (Figure 4.8) was performed on 500 ng of FLO1 cell line genomic DNA and included optimization of enzyme concentration, library purification procedure, PCR cycle optimization and removal of FFPE artefacts.

4.9.2 Adapter design and primers

Adapters (i5 and i7, Table 4.3) were designed to target DNA fragments with restriction overhangs for the selected restriction enzymes (PstI and ApoI) and achieve specific and uniform sampling of the genome by modifying Illumina adapter sequences[101] following the general principles of the quaddRAD protocol[68]. The random 4bp degenerate barcode included in both, i5 and i7, was designed to avoid creating new restriction sites. The 6bp unique inner barcode sequences were balanced for A/C and G/T content to increase the sequence diversity at each position across the inner barcodes. Additionally, PhiX control was spiked in to 20% to improve the overall sequencing quality. The i5 upper adapter was phosphorylated to abolish

the ligation at the 3’ end and the lower i5 adapter was phosphorylated for its ligation with the DNA insert. To avoid non-specific amplification during the PCR stage the i7 adapters were designed in a Y-shape conformation to amplify only those DNA fragments with specific adapters ligated to them. Illumina universal PCR primers (i5nn and i7nn) were used for amplification (Table 4.3). A phosphorothioate bond at the 3’ end of the outer barcodes/primers (i5nn/i7nn) was added to protect from nonspecific or proofreading nuclease degradation.

A) mutREAD adapter sequences

Adapter	Adapter name	Sequence
mutREAD-i5	mutREAD-i5-upper_1_ATGAGCGA mutREAD-i5-lower_1_TCGCTCAT	5'-C GCT CTT CCG ATC T HNNNATGAGCGATGCA-phos-3' 5'-phos-TCGCTCATNNNDA GAT CGG AAG AGC GTC GTG TAG GGA AAG AGT GT-3'
	mutREAD-i5-upper_2_GCCTAGCG mutREAD-i5-lower_2_CGCTAGGC	5'-CGCTCTTCCGATCTHNNNGCCTAGCGTGCA-phos-3' 5'-phos-CGCTAGGCNNNDAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3'
mutREAD-i7	mutREAD-i7-upper_1_CGTGTACC mutREAD-i7-lower_1_GGTACACG	5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTHNNNCGTGTACC-3' 5'-AATTGGTACACGNNNDAGATCGGAAGAGCA-3'
	mutREAD-i7-upper_2_GCACATGT mutREAD-i7-lower_2_ACATGTGC	5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTHNNNGCACATGT-3' 5'-AATTACATGTGCNNNDAGATCGGAAGAGCA-3'

B) mutREAD primer sequences

Primer	Primer Name	Sequence
mutREAD-i5nn	mutREAD-i501_TATAGCCT	5'-AATGATACGGCGACCACCGAGATCTACAC**TATAGCCT**ACACTCTTTCCCTACACGAC*G-3'
	mutREAD-i502_ATAGAGGC	5'-AATGATACGGCGACCACCGAGATCTACAC**ATAGAGGC**ACACTCTTTCCCTACACGAC*G-3'
mutREAD-i7nn	mutREAD-i701_ATTACTCG	5'-CAAGCAGAAGACGGCATACGAGAT**CGAGTAAT**GTGACTGGAGTTCAGACGTGTGC*T-3'
	mutREAD-i702_TCCGGAGA	5'-CAAGCAGAAGACGGCATACGAGAT**TCTCCGGA**GTGACTGGAGTTCAGACGTGTGC*T-3'

Legend: NNNNN = Unique Molecular Identifier

NNNNN = Inner sample barcode

NNNNN =Outer sample barcode

Table 4.3: mutREAD adapters and primers

A) Summary of the sequences of mutREAD adapters used for ligation to DNA fragments. The colour of the nucleotides indicates specific elements of the adapters (unique molecular identifiers, inner samples barcodes). Ambiguous base codes H and D translate to bases A/C/T and A/G/T, respectively. Adapter names include the arm of the adapter (i5 contains PstI compatible end and i7 – ApoI compatible end) and the sequence of the samples barcode.

B) Summary of the Illumina compatible primers used for amplification of the ligated libraries. Colour indicates the sequences and location of outer sample barcode. Adapter names include the arm of the primer (i5XX is compatible with i5 adapter and i7XX with i7 adapter) and the sequence of the samples barcode.

Abbreviations: phos – phosphorylation of the indicated nucleotide, * - phosphorothioate bond between the indicated bonds.

Restriction enzyme	Restriction site sequence
AflII	C*TTAAG
ApoI	R*AATTY
BmtI	GCTAG*C
BsrGI	T*GTACA
BssSI	C*ACGAG
HindIII	A*AGCTT
KpnI	GGTAC*C
MfeI	C*AATTG
MseI	T*TAA
MspI	C*CGG
NcoI	C*CATGG
NdeI	CA*TATG
NsiI	ATGCA*T
NspI	RCATG*Y
PacI	TTAAT*TAA
PstI	CTGCA*G
SbfI	CCTGCA*GG
SpeI	A*CTAGT
SphI	GCATG*C

Table 4.4: List of restriction enzymes tested in the computational simulation and their restriction site sequences

The table lists the enzymes selected as described in the Methods section2.13 and their restriction sites (5'→3'), with the cutting position indicated by *, highlighting the different possible overhangs. Ambiguous codes R and Y translate to A/G or C/T, respectively, and indicate that either base at this position is accepted by the enzyme.

4.9.3 Adapter preparation

Lyophilized adapters obtained from Integrated DNA Technologies (IDT, Leuven Belgium) were reconstituted in Tris-EDTA (TE pH:8) buffer to get 100 μM stock. Complementary upper and lower single strands of i5 and i7 were annealed at 10 μM each using annealing buffer (500 mM NaCl,100mM Tris-HCl, pH 7.5-8) on a thermal cycler with the following conditions: Denature at 97.5°C for 2.5 min and then bring down to 4°C at a rate of 3°C/min. Hold at 4°C. Adapters were stored in -20°C. This 10μM working dilution of adapters stock was used in ligation reaction.

4.9.4 Library preparation and sequencing

Double Restriction digestion and ligation reaction: Both restriction digestion and ligation reaction were performed simultaneously. 500ng of genomic DNA was digested with 50 U of PstI-HF and ApoI-HF in presence of 0.187 mM mutREAD i5 and i7 adapters, 400 U of T4 ligase and 1 mM ATP in 1X CutSmart buffer. The reaction was incubated on a thermal cycler at 30°C for 3 hours. Ligation reaction was stopped by addition of 10 μl of 50mM EDTA.

Size selection: Two step size selection for 400-500bp inserts (DNA fragments, excluding adapters) was performed using Agencourt AMPure XP beads (BECKMAN COULTER, Brea California US). Unwanted larger fragments were removed with 0.6x ratio of AMPure beads to ligation product and the short fragments were removed by 0.15x size selection.

PCR Amplification of Library:

The size selected DNA fragments ligated with adapters (20μl) were amplified using PCR primers (i5nn/i7nn) compatible with Illumina sequencing platform. The reaction was performed in total volume of 100 μl with 0.8 U of Phusion high-fidelity polymerase, in the

presence of 0.2 mM dNTPs and 1X Phusion High Fidelity buffer. PCR was performed in the following conditions: 98°C/2min denaturation, 12 cycles of amplification at 98°C/10sec, 65°C/30sec, 72°C/30sec and final extension at 72°C for 5min. Libraries were purified using 0.8X AMPure beads (80 μl beads+100 μl library), this step was repeated one more time to remove all unwanted leftover reactants during PCR. Libraries were eluted in 20μl TE buffer (Tris-EDTA buffer 10mM TrisHCl and 0.1mM EDTA, pH8) and stored at -20°C. Quality control was performed on Agilent 2100 Bioanalyzer using Agilent High Sensitivity DNA kit (Santa Clara, California, US) or High Sensitivity D1000 TapeStation kit (Agilent). Quantification of the libraries was performed using KAPA Library Quantification kit (KK4953-07960573001 for Illumina platforms, Kapa Biosysytems Roche Holding AG Basel Switzerland) on the Light cycler 480 (Roche Life Sciences, Basel Switzerland). Libraries with unique adapters were pooled and sequenced on the HiSeq4000 using paired end, 150 bps.

After developing the lab protocol using cell line DNA, we further optimized the protocol to suit either fresh-frozen or FFPE samples, the latter being the standard sample preservation strategy in clinical practice. Restriction enzyme double digestion, adapter ligation conditions and size selection were optimized for optimal digestion, adapter annealing and size selection using an OAC cell line (FLO-1). The protocol was further adjusted for FFPE derived DNA from the same OAC cell line (Figure 4.8).

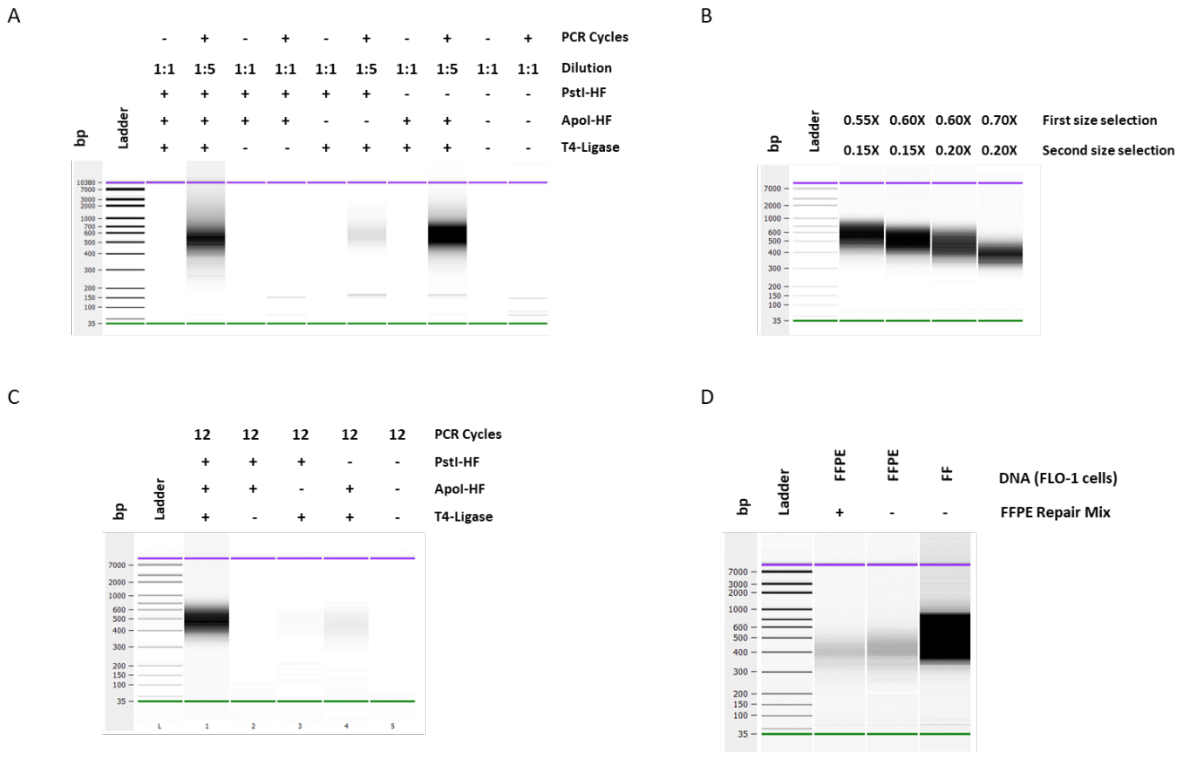


Figure 4.8: Optimization of mutREAD library preparation using FLO1 cell line

A) Bioanalyser traces for the optimization of the single step double digestion and ligation. 500 ng of FLO1 genomic DNA was used for ligation of mutREAD adapters in the presence of indicated enzymes and underwent PCR amplification (20 cycles) using Illumina compatible primers. Samples before (-) and after (+) PCR are shown for each enzyme combination. Dilution indicates dilution of samples for bioanalyzer analysis (for samples that exceeded recommended detection range).

B) Bioanalyser traces for different titration of ratios of AMPure beads and ligated DNA solution (50ul) to optimize the double size selection of the fragments in the library.

C) Bioanalyser traces prepared under optimised PCR cycles conditions. Note significant decrease in the level of ApoI only fragments when compared to 20 PCR cycles (A).

D) Bioanalyser traces showing improved bands for FFPE samples after treatment with FFPE repair mix and library preparation with optimized protocol.

All samples were run using DNA High Sensitivity Bioanalyzer kit with standard DNA ladder. Green and purple bands indicate lower and upper markers respectively.

We then applied mutREAD to fresh-frozen tumour , matched FFPE and blood samples from biopsies of three different OAC patients and evaluated the quality of the library under several criteria (Figure 4.9, Figure 4.10, Table 4.5).

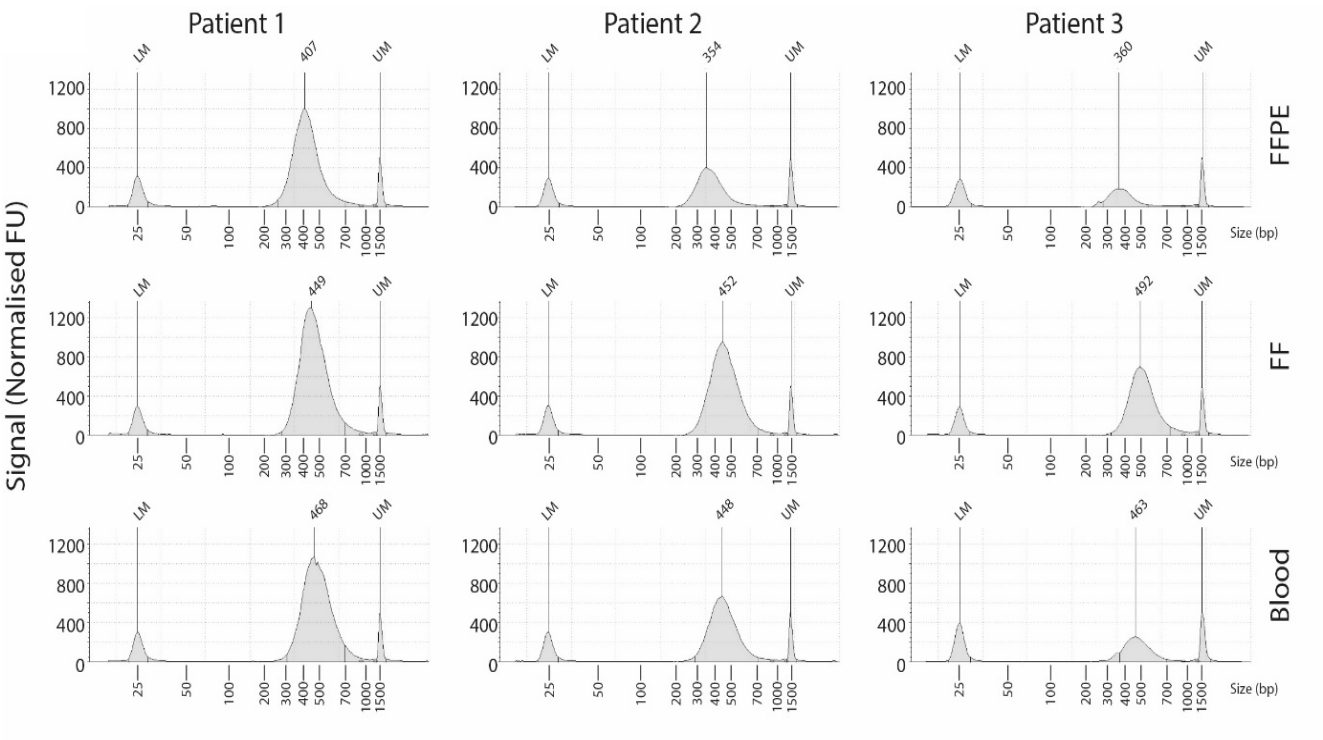


Figure 4.9: Fragment size distribution of mutREAD libraries

Fragment size (x-axis) distribution of sequencing libraries measured on the Tape-station. Electropherograms of DNA fragments from three samples derived from FFPE (neat), Fresh Frozen (FF, 1:4 dilution) and matching blood samples (1:4 dilution) with the average size of libraries highlighted above the plot. LM – lower marker, UM – upper marker, FU – fluorescent units.

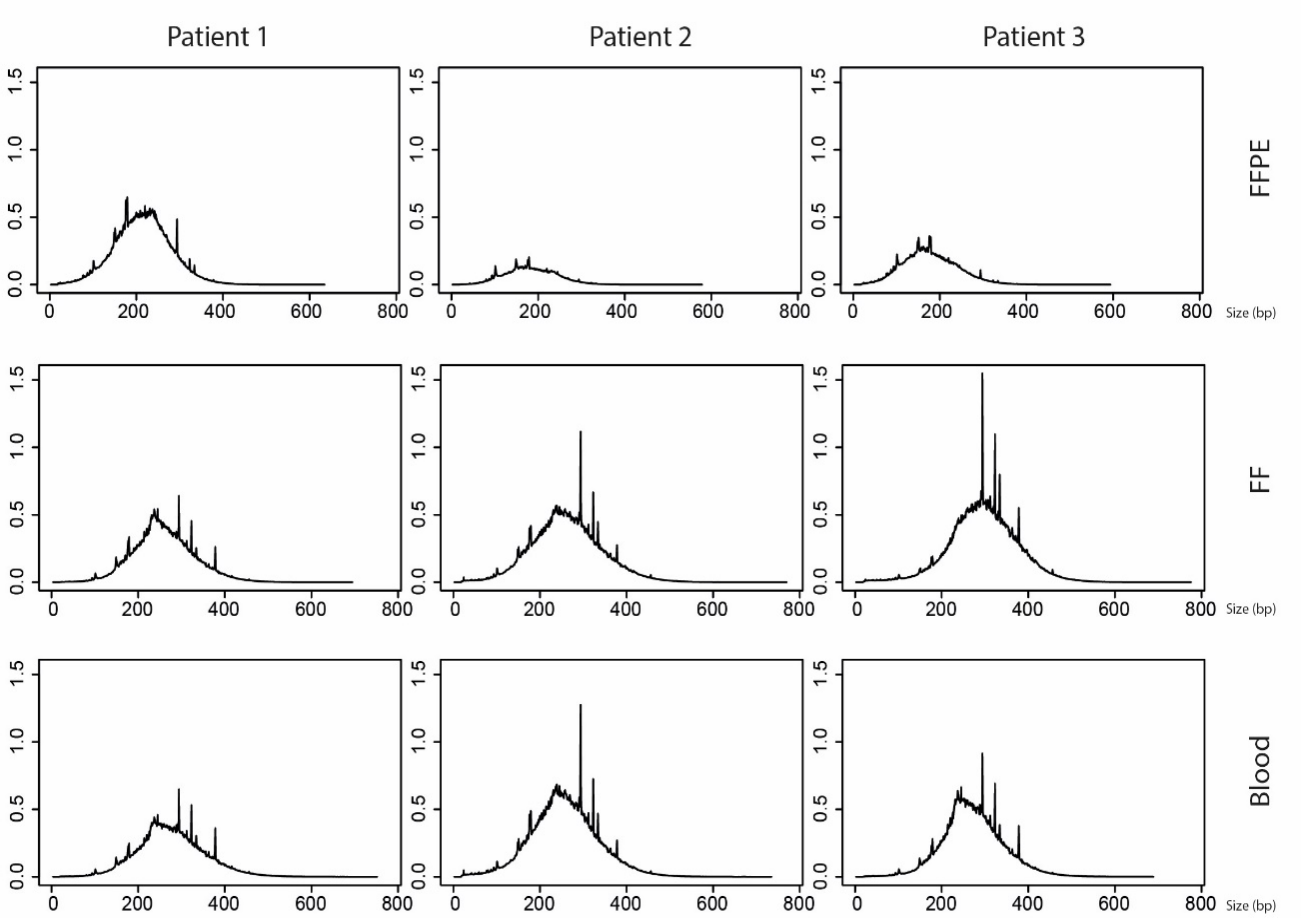


Figure 4.10: Fragment size distribution on sequencing.

Fragment size distribution derived from read-pairs mapped to the human genome. Each plot shows the number of fragments (y-axis) for each length in base pairs (x-axis). The fragment length was calculated as the number of base pairs between the 5' ends of the read mates (including restriction site parts but not adapters or barcode sequences) and summarized to a histogram using Picard's CollectInsertSizeMetrics function.

Patient ID (1)	Sample type (2)	Number of reads after outer barcode demultiplexing (3)	Percent retained after QC (4)	Percent lost due to unidentifiable barcode (5)	Percent lost due to low quality (6)	Percent of reads lost due to inner barcode mixup (7)	Percent lost due to ambiguous RAD-tag (8)	Estimated average fragment size (bp) (9)
Tumour1	FF	158,178,068	94.54	0.09	0.37	3.83	1.17	260
	FFPE	184,526,290	93.41	0.33	0.36	4.46	1.43	215
	Blood	155,543,840	94.14	0.07	0.42	3.52	1.85	276
Tumour2	FF	206,790,834	92.68	0.23	0.62	5.55	0.91	263
	FFPE	43,949,748	94.92	0.23	0.47	3.41	0.98	183
	Blood	230,847,264	96.21	0.13	0.48	2.37	0.80	257
Tumour3	FF	231,185,296	95.00	0.07	0.46	3.56	0.92	297
	FFPE	86,259,612	94.26	0.67	0.59	3.33	1.15	178
	Blood	194,066,264	96.39	0.08	0.47	2.02	1.05	273

Patient ID (1)	Sample type (2)	Base pairs covered with at least 10x (10)	Percent of retained reads contributing to 10x loci (11)	Base pairs in 10x loci shared in tumour/blood pair (12)	Base pairs covered with at least 50x (13)	Base pairs covered with at least 100x (14)	mutREAD - Number of mutations (15)	WGS - Number of mutations (16)
Tumour1	FF	175,049,803	96.54	166,936,824	98,935,164	60,297,417	1,050	28,732
	FFPE	170,810,606	96.91	143,331,473	122,274,170	86,782,166	383	
	Blood	186,266,055	96.26	-	103,044,362	60,323,363	-	-
Tumour2	FF	195,958,931	96.63	187,858,494	147,765,532	105,375,328	1,471	27,764
	FFPE	95,105,098	93.98	88,953,041	32,115,195	10,201,299	47	
	Blood	193,634,665	96.61	-	147,906,131	111,689,924	-	-
Tumour3	FF	198,984,001	96.57	170,614,310	146,079,968	106,880,654	530	11,068
	FFPE	131,586,722	95.11	113,854,654	77,474,870	36,663,830	90	
	Blood	190,613,393	96.49	-	114,092,331	73,822,016	-	-

Table 4.5: Quality metrics for mutREAD libraries derived from tumour, FFPE and blood samples of three patients.

The table summarizes quality metrics for each sample, including fresh-frozen (FF) and formalin-fixed paraffin-embedded (FFPE) tumour, as well as blood samples (column 2) from three patients (column 1). Sample groups of three were sequenced on one lane, where each sample had a unique outer barcode. Number of reads derived from the libraries de-multiplexed by outer barcode are listed in column 3. Percentages (with respect to column 3) of reads that are retained for further analysis (column 4) or filtered due to an unidentifiable inner barcode (column 5), low read quality (column 6), wrong/unexpected inner barcode (column 7), missing restriction site overhang (column 8) are listed in the respective columns. The average fragment size derived from read pair mates after mapping is given in column 9 (related to Supplementary Figure 4). The number of base pairs covered with at least 10x, 50x and 100x is listed in column 10, 13 and 14, respectively. The percentage of retained reads (column 4) contributing to loci defined in column 10 is given in column 11. Finally, the overlap between tumour and blood samples in loci defined in column 10 is shown in column 12. The number of mutations used for deriving the mutational signatures is given in column 15 and 16.

4.10 Comparative Mutational Signatures analysis across different methods on three OAC patients samples.

The mutational signatures, derived from 530-1471 mutations detected using GATK Mutect2[102], showed cosine similarities of 0.95-0.96 when compared with the WGS-derived mutational signature profiles (Figure 4.11). We observed similar cosine similarity between mutREAD and WGS when mutations were derived using an alternative mutation caller, Strelka[103] (Table 4.6). In summary, the mutREAD protocol results in reproducible, good quality, target-specific libraries from which mutational signatures can be successfully derived.

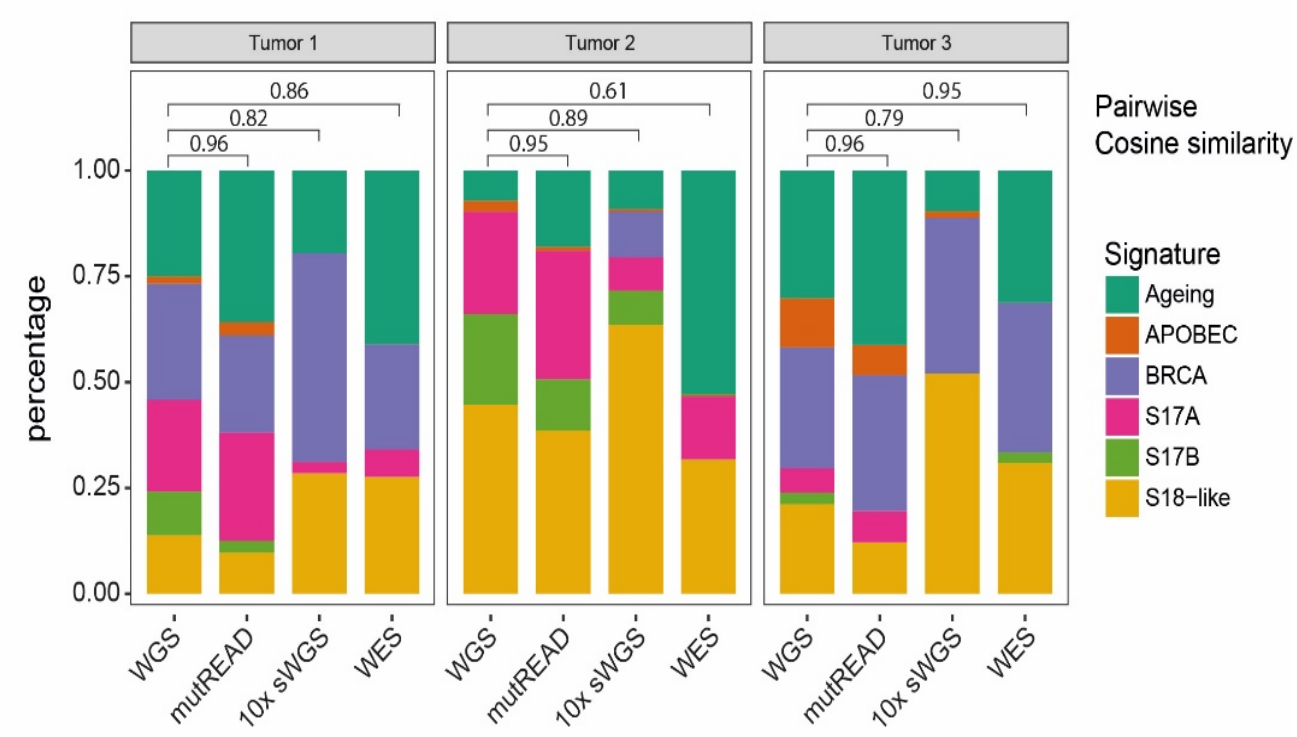


Figure 4.11: Mutational signatures derived with different sequencing methods.

Comparison of the mutational signature profiles for three OAC samples across different sequencing methods (x-axis). Each bar indicates the contribution of the mutational signature (y-axis) to the overall mutational spectrum. Pairwise cosine similarities to WGS for mutREAD, WES and 10x sWGS are indicated above the bars.

A) Number mutations (fresh-frozen)

	mutREAD			WGS		
	Mutect2	**Strelka**	**Consensus**	**Mutect2**	**Strelka**	**Overlap**
Tumour1	1050	520	440	28732	27370	26048
Tumour2	1471	839	714	27764	26540	25284
Tumour3	530	339	217	11068	10398	9950

B) Cosine similarity with WGS (fresh-frozen)

	Mutect2	**Strelka**	**Consensus**
Tumour1	0.96	0.94	0.92
Tumour2	0.95	1.00	0.99
Tumour3	0.96	0.84	0.83

C) Number mutations (FFPE)

	Mutect2	**Strelka**	**Consensus**
Tumour1	383	811	104
Tumour2	47	420	27
Tumour3	90	838	45

D) Cosine similarity with WGS (FFPE)

	Mutect2	**Strelka**	**Consensus**
Tumour1	0.89	0.83	0.76
Tumour2	0.93	0.81	0.89
Tumour3	0.96	0.81	0.88

Table 4.6:Comparison of Mutect2 and Strelka mutation calling pipelines

The tables A and C summarize the number of mutations detected by Mutect2 and Strelka, as well as the overlap/consensus between the two mutation callers, for the three fresh-frozen (mutREAD and WGS) and FFPE tumour samples, respectively. The cosine similarity of mutREAD-derived and WGS-derived mutational signatures is summarized in tables B and D for the fresh-frozen and FFPE samples, respectively. For each mutation caller and the consensus set, the mutational signatures were calculated from respective mutREAD-derived and WGS-derived mutation set and compared against each other using cosine similarity.

Next, we compared mutREAD with WES and 10x sWGS libraries of the same samples sequenced to similar depth. Quality measures for the resulting libraries of the different methods are summarized in Table 4.7. WES resulted in 46-325 mutations per sample and 10x sWGS identified 21-83 mutations per sample. mutREAD consistently achieved high cosine similarity to the corresponding WGS-derived signatures. Conversely, WES and 10x sWGS had lower cosine similarities and much higher variability between patients (Figure 4.11).

A)
10x
sWGS

Patient ID	Sample type	Number of reads after outer barcode demultiplexing	Properly paired reads	Base pairs covered with at least 10x	Percent of retained reads contributing to 10x loci	Base pairs in 10x loci shared in tumour/blood pair	Number of mutations
Tumour1	**FF**	215,680,416	206,317,606	2,432,340,493	90.21	99,329,518	42
	Blood	180,158,855	175,855,924	2,248,030,642	88.49	-	-
Tumour2	**FF**	329,742,782	321,605,192	2,646,459,519	88.55	77,553,550	21
	Blood	217,304,771	210,403,566	2,444,354,485	88.51	-	-
Tumour3	**FF**	139,225,587	134,967,944	1,914,933,524	89.19	94,152,562	83
	Blood	233,793,837	228,289,614	2,478,664,018	88.27	-	-

B)
WES

Patient ID	Sample type	Number of reads after outer barcode demultiplexing & PCR clone removal	Properly paired reads	Base pairs covered with at least 10x	Percent of retained reads contributing to 10x loci	Base pairs in 10x loci shared in tumour/blood pair	Number of mutations
Tumour1	**FF**	149,007,546	146,883,752	228,030,295	92.41	396,961,129	325
	Blood	72,706,935	69,490,692	165,989,267	89.08	-	-
Tumour2	**FF**	44,956,340	44,224,012	119,724,582	94.68	902,341,754	142
	Blood	64,528,333	63,067,844	145,330,607	92.69	-	-
Tumour3	**FF**	74,142,650	72,864,976	156,263,110	82.16	251,229,586	46
	Blood	84,750,728	79,661,310	187,169,707	92.76	-	-

C)
mutREAD

Patient ID	Sample type	Number of reads after outer barcode demultiplexing, PCR clone removal & mutREAD filtering	Properly paired reads	Base pairs covered with at least 10x	Percent of retained reads contributing to 10x loci	Base pairs in 10x loci shared in tumour/blood pair	Number of mutations
Tumour1	**FF**	150,115,473	142,432,706	175,049,803	96.54	166,936,824	1,050
	Blood	145,505,593	137,261,472	186,266,055	96.26	-	-
Tumour2	**FF**	196,879,399	195,040,496	195,958,931	96.63	187,858,494	1,471
	Blood	219,749,023	217,691,576	193,634,665	96.61	-	-
Tumour3	**FF**	216,676,031	214,418,984	198,984,001	96.57	170,614,310	530
	Blood	185,303,214	175,401,460	190,613,393	96.49	-	-

Table 4.7: Quality metrics for 10x sWGS, WES and mutREAD libraries derived from tumour and blood samples of three patients.

The table summarizes quality metrics of the libraries generated by (A) 10x sWGS, (B) WES, and (C) mutREAD for tumour and blood samples (column 2) of the same three patients (column 1). Column 3 and 4 gives the number of reads and properly paired read pairs used for mutation calling, respectively. The number of base pairs covered with at least 10x, the percentage of reads contributing to these loci and the overlap in these loci between tumour and normal samples is listed in column 5-7.

4.11 Comparative Mutational Signatures analysis between sample type: Frozen v/s FFPE samples

Finally, we investigated if mutREAD can be used to study historical samples by sequencing FFPE specimens matching the previously analysed frozen samples. Fresh frozen and FFPE-derived samples generated similar signature patterns (Figure 4.12), despite the lower sequencing depth and smaller fragment distribution of final FFPE-derived libraries (Figure 4.9, Figure 4.10, Table 4.5). Cosine similarities to WGS-derived mutational signatures were between 0.89-0.96 based on 47-383 detected mutations.

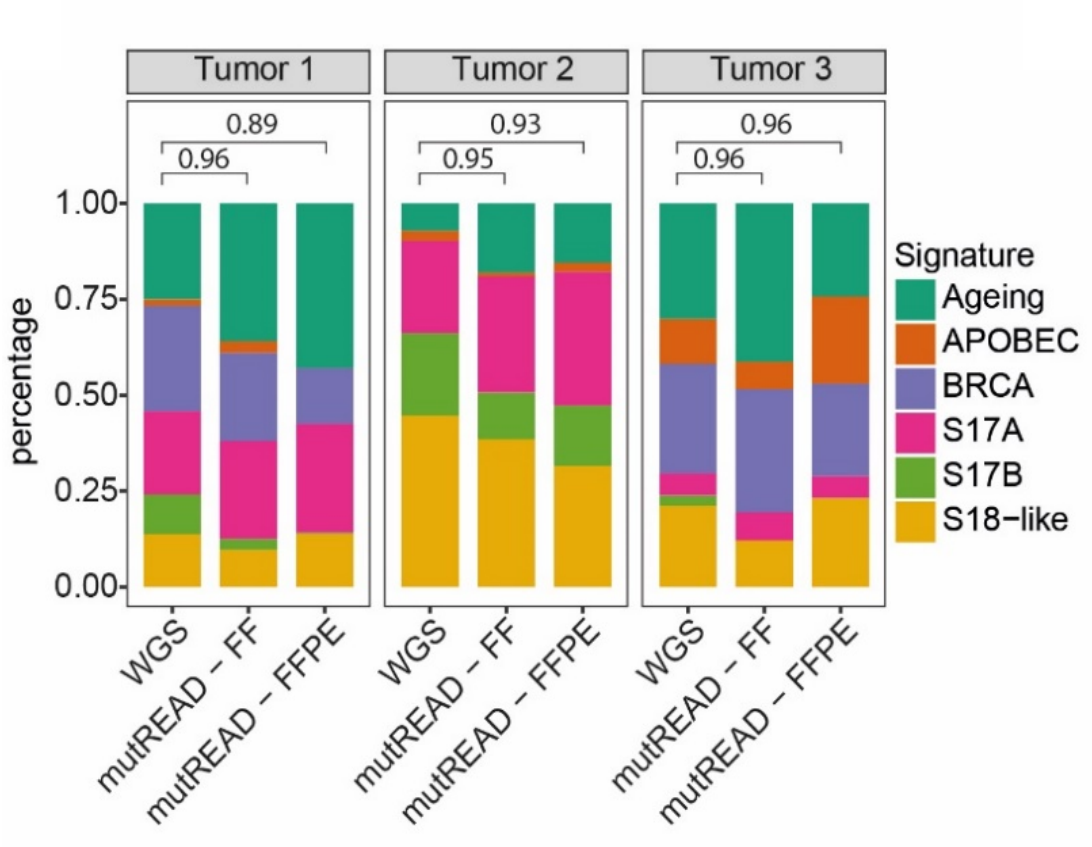


Figure 4.12: Comparative Mutational signature analysis between Frozen and FFPE tumour samples.

Comparison of the mutational signature profiles between WGS, fresh-frozen (FF) and FFPE samples for the same three OAC samples as in Figure 1. Each bar indicates the contribution of the mutational signature (y-axis) to the overall mutational spectrum. Pairwise cosine similarities to WGS for the two mutREAD libraries are indicated above the bars.

We replicated the good cosine similarity to WGS-derived mutational signatures in an additional nine FFPE samples (Table 4.8). Of note, samples were derived from tumour resections and pathology estimates for these samples show low tumour content (10-70%, Table 4.9), explaining the lower number of mutations and higher variability across samples compared to the previously tested biopsy samples.

	Number of SNVs	Cosine similarity to WGS
Patient 1	7	0.31
Patient 2	37	0.83
Patient 3	254	0.94
Patient 4	52	0.93
Patient 5	60	0.94
Patient 6	58	0.83
Patient 7	21	0.73
Patient 8	25	0.84
Patient 9	13	0.88

Table 4.8: mutREAD recapitulates mutational signatures in FFPE samples as per WGS
Cosine similarity between mutational signatures derived from nine additional FFPE and WGS sample pairs and the number of detected mutations in the FFPE samples used to derive the mutational signatures.

A) Tumour biopsies

Patient	Pathologist 1	Pathologist 2	Pathologist 3
Tumour1	70%	60%	45%
Tumour2	50%	55%	30%
Tumour3	30%	35%	15-20%

B) Tumour resections

Patient	Pathologist 1	Pathologist 2	Pathologist 3
Patient1	60%	60%	15%
Patient2	60%	70%	20%
Patient3	30%	60%	15%
Patient4	N/A	N/A	N/A
Patient5	30%	20%	10-15%
Patient6	70%	40%	20%
Patient7	20%	45%	20-25%
Patient8	25%	45%	50%
Patient9	50%	50%	20-25%

Table 4.9: Tumour cellularity of FFPE samples estimated by pathology

A) Estimated percent of tumour content for the three biopsy samples estimated by pathologist review of diagnostic slides.

B) Estimated percent of tumour content for the nine tumour resection samples estimated by pathologist review of diagnostic slide

Basic demographics of the patients from whom tumour samples taken are listed in Table 4.10

Patient ID	Anonymized Name (ICGC identifier)	Age at diagnosis	Sex	Tumour Stage	Node Stage
Tumour1	a504c27a1fd7af3a53e0b5108cc052cab5ff8d1a353800e85ea8eec766707bde	75.4	male	T3	N0
Tumour2	fa37be85256c1efbac0501eb13ae1daf9f0fcaff4a7cbb3d2f4420f70f75d334	77.1	male	T3	N1
Tumour3	da859c8e95cc5acefde4e70aaed8fc89449c2aa5d9a6cef6202041987840e0a3	75.3	male	T2	N1
Patient1	9a498e8b17034fd8bb534f0e65e12c83a73aa65908fd41f4a13f86cf35b3e0cd	80.8	female	T1a	Nx
Patient2	493bf7322b8c18365466c43bf7a9e119bd4d7782147f9bc368bf4909539c43de	68.7	male	T1b	N0
Patient3	078773cc36a9ab58ece5c92e50368462dea52ded70f1b0da8c66e65066a3ce53	72.4	male	T3	N1
Patient4	3180f8e34845d13e27dddd90486dba083cd87a340d919f75d29859766f7faee4	69.9	male	T3	N1
Patient5	a7376de8be895a08d2abb22b1e3ee2483fb47abc534d7c6a066b06b2f0d4459a	51.9	female	T1	N0
Patient6	75ce1bd6dbaf2d4ca50f51e6c1f2a09f3275b080539a07ac7ed962ae72c5179f	58.4	female	T3	N0
Patient7	1395dc4ab7c754e0a84c8daa3996f16e5caf11169f8a9be6c800f6da00474321	65.6	male	T2	N1
Patient8	934fe84809fc20a81f124747d5ed57817eb0f9120ea63a69a1548d514710978f	59.2	male	T4a	N1
Patient9	900fdae05c90e27aba521996cc05d0a83e32dec27c271a8f14e59fa84439ed34	73.5	male	T3	N0

Table 4.10: Patients Clinical Characteristics

The table lists the information about individual patients used in the study. Patient ID follows the convention established in figures 1 and 2. Anonymized Name provides ICGC patient ID that can be used to obtain the Whole Genome Sequencing data used in the study.

4.12 Reproducibility of mutREAD in identification of Mutational signatures from FFPE samples

Given the high degradation expected in FFPE samples which can result in variability, we also tested the reproducibility of FFPE-derived mutREAD libraries. Technical replicates of the nine FFPE samples showed high concordance in sequenced regions and fragment size distribution (Figure 4.13, Figure 4.14). Hence, while it is expected that the performance on FFPE is lower compared to fresh-frozen samples, our results suggest that mutREAD can also be applied to FFPE-derived DNA samples with low tumour content and leads to reproducible results.

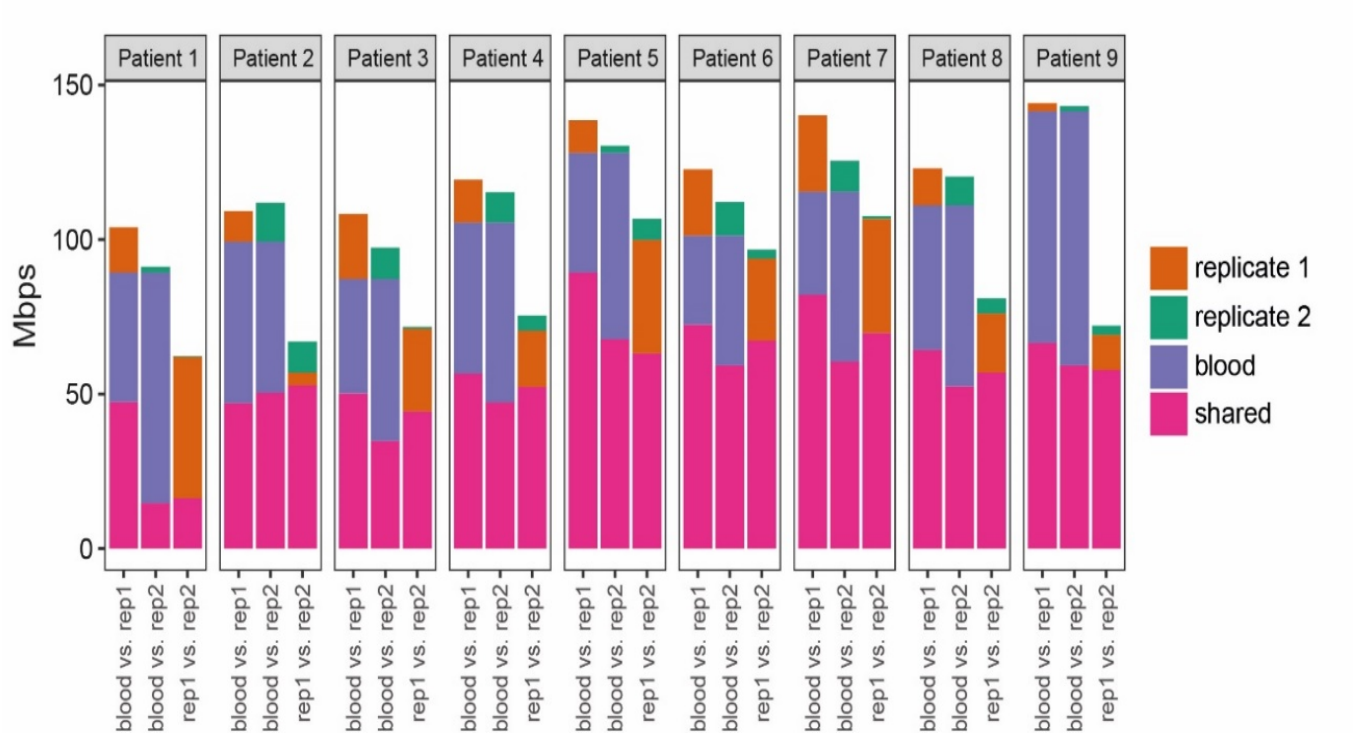


Figure 4.13: mutREAD reproducibly detects mutational signatures in FFPE samples

Reproducibility of the sequenced regions between the first FFPE-derived technical replicate and the blood sample, the second FFPE-derived technical replicate and the blood sample, and between the two technical replicates (x-axis). The bars indicate the size of the overlapping

regions in Mpbs (y-axis) for each comparison. Only regions covered at least 10x contribute to the comparison. The second technical replicate was sequenced to lower coverage and we down-sampled the first technical replicate by 50% to approximately match the sequencing coverage for comparison.

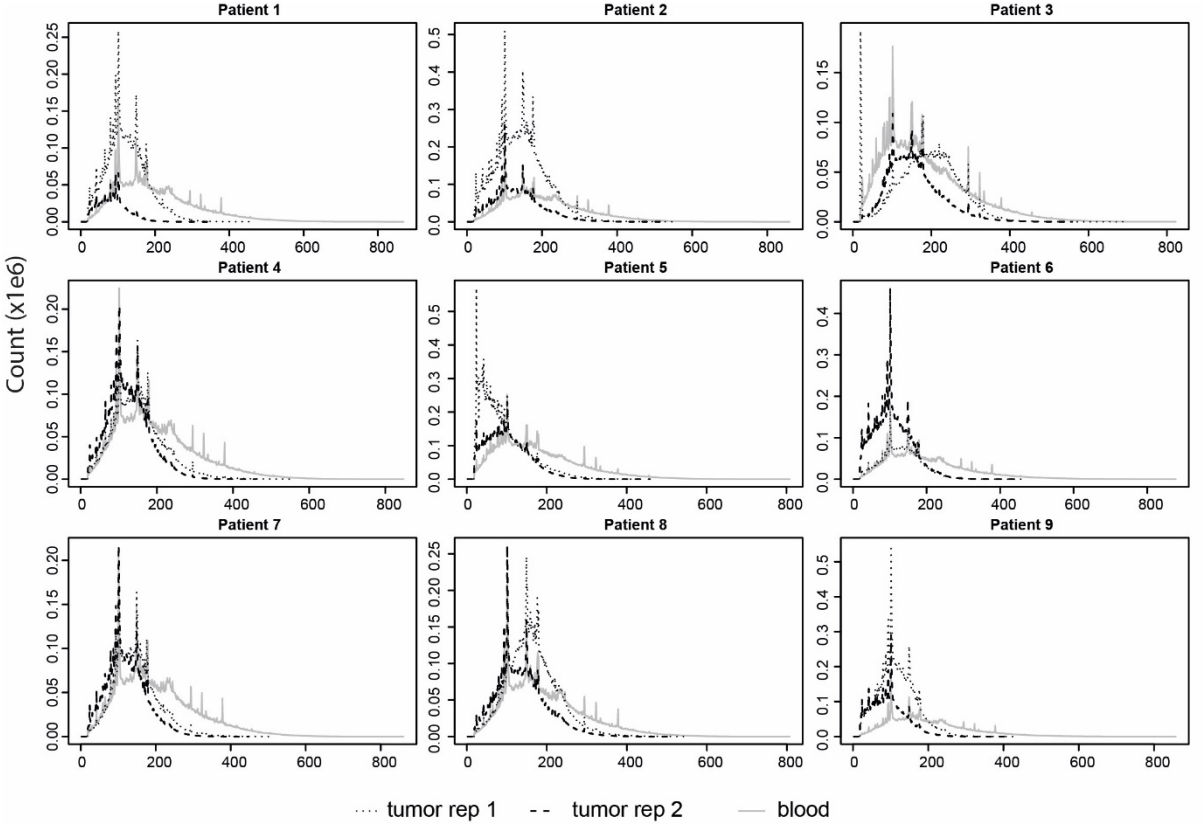


Figure 4.14: Comparison of the fragment size distributions for technical replicates of FFPE samples and blood Fragment size distribution derived from read-pairs mapped to the human genome. Each plot shows the number of fragments (y-axis) for each length in base pairs (x-axis) for the two technical replicates of FFPE tumour samples and the corresponding blood sample per patient. The fragment length was calculated as the number of base pairs between the 5' ends of the read mates (including restriction site parts but not adapters or barcode sequences) and summarized to a histogram using Picard's CollectInsertSizeMetrics function.

We then evaluated the estimate of cost of mutREAD for library preparation protocol and compared with other methods. These are tabulated below (Table 4.11)

A) mutREAD					
No.	**Reagents**	**Size**	**Cost (£)**	**Use per sample**	**Per sample (£)**
1	Adapter and Primers	400ul	--	9.5ul	2
2	T4 ligase	20,000Units	62.58	400units	1.25
3	ApoI-HF	1000 Units	49.6	50units	2.48
4	PstI-HF	10,000Units	41.6	50units	0.2
5	10mM ATP	1000ul	24	4ul	0.1
6	Ampure XP beads	5000ul	195.26	37.5ul	1.5
7	10mM dNTPs	800ul	49.6	2ul	0.1
8	Phusion High fidelity polymerase	100Units	61.6	1unit	0.6
				Total:	**£8.23**
B) WES					
Sl no	**Reagents**	**Size**	**Cost (£)**	**Per sample (£)**	
1	DNA library preparation and enrichment kit	16	3,589	199	
			Total:	**£199**	
C) 10x sWGS					
Sl no	**Reagents**	**Size**	**Cost (£)**	**Per sample (£)**	
1	Thruplex library preparation and enrichment kit	96	3,818.59	39.7	
2	Sonication	96	352	3.7	
			Total:	**£43**	

Table 4.11: Comparative cost evaluation for library preparation per sample

A) Estimated cost of individual elements used for library preparation using the mutREAD protocol. The cost is estimated using reagents provided by New England Biolab, Ipswich, MA 01938 USA

B) Cost estimate of enrichment-based whole exome sequencing provided by Agilent, Santa Clara, CA 95051 USA. The cost does not include AMPure XP and Streptavidin beads required for the selection of target sequences.

C) Cost estimate of the whole genome library preparation method provided by Takara Bio, Kusatsu, Shiga 525-0058, Japan. The cost does not include AMPure XP and Streptavidin beads required for the selection of target sequences.Cost of sequencing: Assuming 200x coverage for mutREAD, the per-sample cost is around £150. The costs for WES would be similar as both methods sample a comparable proportion of the genome. 10x sWGS would cost £300-£700 depending on the chosen sequencing platform.

4.13 Summary

In this chapter I have presented the development of a novel DNA sequencing method, mutREAD, a simple yet robust protocol for detection of mutational signatures from low quality and quantities of DNA. We hypothesised that sampling a random subset of mutations from a genome will suffice the detection of mutational signatures. The RR-Seq based mutREAD method performed well on in-silico mutational signature analysis and we showed that with a small random subset of mutations we were able to recapitulate the mutational signatures present in WGS data. We then performed a comparative in-silico analysis across different methods such as WGS, WES and 10x sWGS, and RR-seq based mutREAD with different combinations of restriction enzymes. mutREAD out performed the other methods and we selected the top hit restrictions enzymes (PstI+ApoI) for development of a robust lab protocol. We also evaluated mutREAD across 20 different cancer types using the WGS data by PCAWG and showed that mutREAD accurately identified the mutational signatures specific to each type, thus expanding its application to other cancer types.

We then developed the lab protocol to prepare DNA libraries. Optimization experiments were performed on cell line DNA and evaluated on fresh frozen tumour DNA and later DNA from FFPE samples. We designed enzyme specific adapters and streamlined the protocol to accommodation, restriction digestion, adapter ligation in one step. Unwanted purification steps were removed, and reproducible fragment size selection step was optimised. Minimum PCR cycles were used to avoid duplicates in sequencing data. To increase the diversity of reads during sequencing PhiX DNA was spiked.

Two different mutation callers were employed to call mutations and we observed similar cosine similarity values. In the first set of experiments, we extracted mutational signatures

from three fresh frozen tumours and compared these with WGS, 10XsWGS and WES. mutREAD derived signatures showed 0.96 similarity to WGS derived signatures. Next we evaluated mutREAD on FFPE samples from the same three tumours, and we were able to get similar results (cosine value: 0.89-0.96). We then analysed data from additional 9 FFPE samples and were able to recapitulate the WGS based signatures with some variations in cosine similarities values, which were explained by the low tumour cellularity. Overall we showed that using mutREAD we reliably recapitulated mutational signatures specific to cancer type.

The key features of mutREAD are its simple library preparation work flow and the reduced time required for preparing DNA for sequencing. Its flexibility and easy-to-use protocol for customised usage are advantageous. The lab reagents used can be easily procured which helps for its wider application without any need for advanced instruments, only a thermal cycler can be sufficient for library preparation. This reduces cost and mutREAD is thus a cost effective scalable and can be tailored for different applications. We have developed mutREAD for OAC specific mutational signature detection. For future directions this method can be further improved to enable detection of other signature types such as indels, structural variants and copy number alteration

5.Results Chapter

Validation of mutREAD using archival samples from oesophageal adenocarcinoma patients.

5.1 Rationale:

With the development of a cost-effective method for large scale screening of mutational signatures, we extended the method on a new cohort of FFPE samples to provide confirmation that it was robust. I performed mutational signature analysis on FFPE samples from pre(diagnostic) biopsies and matched resection tumours and compared the data to WGS from fresh frozen tumour from the same patient. In addition, I ascertained the influence of therapy on the proportions of mutational signatures obtained from mutREAD data.

5.2 Study Cohort:

We assembled a cohort of 25 OACs patients, for whom we had WGS data. We procured pre (diagnostic) and post chemotherapy treatment FFPE blocks as available for the 25 patients. As expected, this validation cohort was male prediminant (92%) with a mean age of 68.3 years. Basic exposure data such as alcohol, smoking and BMI as available is presented. The majority of these tumours are T3(56%), more than half the patients had positive nodes (56%) and only a few have loco-regional metastasis since they were on a surgical pathway (16%). GEJ type II tumours were almost half of the cohort (56%). Overall survival was 29 weeks(21-51). (Table 5.1)

Variable	Measure/level	OAC (n=25)
Age	Years(median,IQR)	68.3 (57.1- 76.4)
Gender	Female	1 (4%)
	Male	23 (92%)
	Missing	1 (4%)
Smoking status	Current	2 (8%)
	Former	11 (44%)
	Never	4 (16%)
	Missing	8 (32%)
Alcohol(Units/week)	Mean(min-max)	3.6 (1-21)
BMI	Kg/m2(median, IQR)	26.2 (22.6-28.2)
Overall Survival	Weeks (median, IQR)	29 (22-68)
Pre-treatment Tumour Stage	T2	4 (16%)
	T3	14 (56%)
	T4a	1 (4%)
	Missing	6 (24%)
Pre-treatment nodal involvement	Positive	14 (56%)
	Negative	7 (28%)
	Missing	4 (16%)
Pre-treatment loco-regional metastasis	Positive	4 (16%)
	Negative	18 (72%)
	Missing	3 (12%)
Chemo treated	Yes	21 (84%)
	No	0
	Missing	4 (16%)
Siewert Classification	Type I	3 (12%)
	Type II	13 (52%)

	Missing	9 (36%)

Table 5.1: Cohort Demographics

5.3 mutREAD Library quality control measures

mutREAD libraries were prepared as per the protocol I developed detailed in the results chapter 2(section 4.9), with some minor changes to the protocol. Library preparation and sequencing for 23 pre(diagnostic) biopsies and 25 resection tumours were processed in two batches. First batch was composed of 3 pre(diagnostic) biopsies and 5 resection tumours and in second batch all the remaining samples: 20 pre(diagnostic) biopsies and 20 resection tumours were processed.Here I used only 200-300ng of FFPE DNA as input due to the scarcity of material. Quality checks were performed to confirm fragment sizes of the libraries within the range of 250-450bps and to ensure that I had clean fragments with minimal contamination from un-ligated adapters during PCR amplification. Representative individual library traces for the first batch were shown in Figure 5.1.

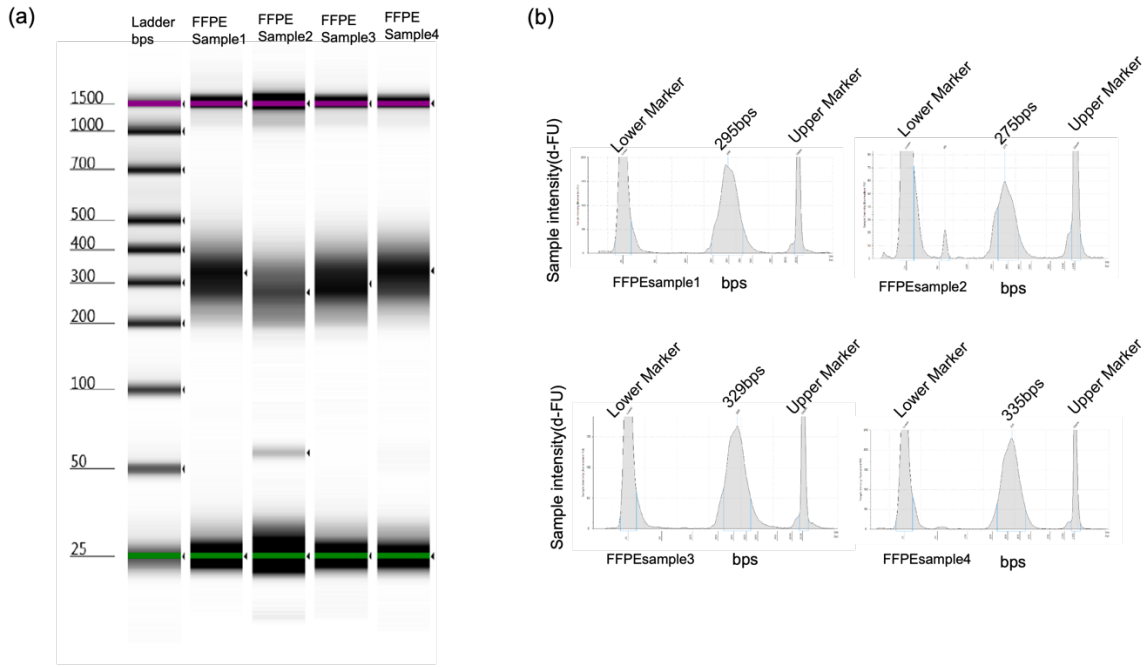


Figure 5.1: Representative mutREAD libraries from the first batch: Fragment size (x-axis) distribution of sequencing libraries measured on the Tape-station.

(a)Gel image of the libraries (b) Electropherograms of DNA fragments from four samples derived from FFPE (undiluted) with the average size of libraries highlighted above the plot.

DNA fragments in the libraries from the second batch of samples were checked individually on the bioanalyzer and then they were pooled for sequencing. In figure 5.2, I show the traces after pooling. Number of samples in a pool was selected to get 100Million reads/sample after sequencing. All 20 pre(diagnostic) were pooled together and 20 chemo treated were pooled. Libraries from matched blood for these cases were pooled as one. Due to varied degree of fragmentation of FFPE samples, different fragments length were observed. Pre(diagnostic) samples after pooling showed an average fragment length of 317bps. FFPE samples from resection tumours from treated patients showed relatively more degradation with average fragment size of 285bps. As expected, DNA from matched blood samples showed an average fragment length of 447bps.

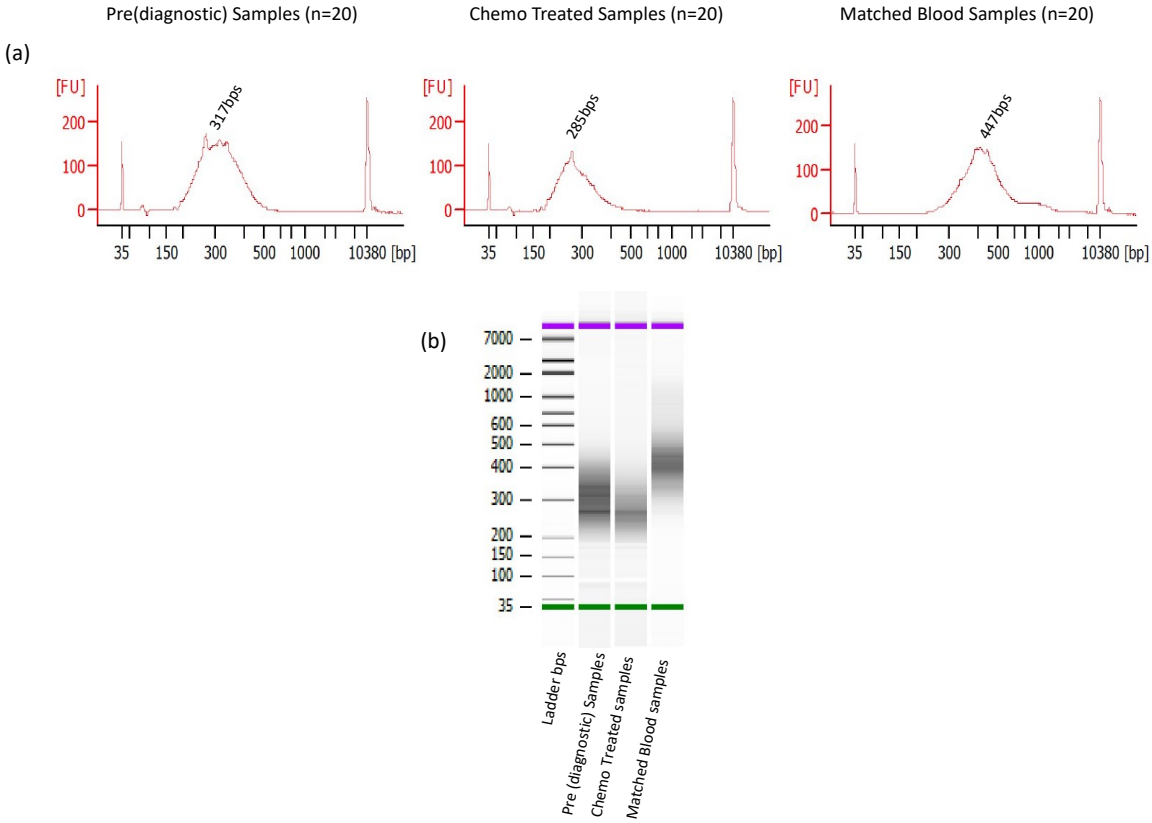


Figure 5.2: Fragment size distribution of mutREAD libraries. Bioanalyser traces for pooled libraries: (a)Electropherograms of DNA fragments from second batch of 20 patients taken from pre(diagnostic) biopsies, resection tumours after treatment and matched blood samples derived from FFPE (undiluted) with the average size of libraries highlighted above the plot. (b) Gel image of the libraries

5.4 Comparative analysis of mutREAD derived OAC specific mutational signatures with their WGS mutational profiles.

We were able to get pre(diagnostic) biopsies from 23 of 25 patients and matched resection tumours for all 25 patients. Mutational signatures specific to OAC (S1- aging; S2-APOBEC; S3-BRCA; S17A;S17B and S18-ROS) were extracted using the mutREAD- variant calling files (VCF) generated as described in the methods chapter section 2.20.

Using mutREAD sequencing, SNVs in the range of 53 to 591 (median = 191) were obtained. Using these SNVs mutational signatures were extracted. Mutational signatures profiles for individual samples from 23 pre(diagnostic) samples were compared with the WGS data using cosine similarity function and recorded (Table 5.2).

The number of samples in each signature-based comparison varies as per the prevalence of that signature among the pre(diagnostic) samples. In Figure 5.3, 5.4, 5.5 and 5.6 the WGS signature proportions are shown in comparison with the mutREAD proportions for S17, S18-like, S1-Aging and S2-APOBEC/S3-BRCA respectively. Also, the Pearson correlation(*R*) was computed to measure the linear correlation for proportions obtained from WGS and mutREAD for that signature along with the *p* value for statistical significance. Only the cases with non-zero proportions were considered for the correlation analysis, hence the denominator for individual signature comparison is less than 23.

S17A and B were merged into one as they are linked to one mutational process in view of their similarity and to ensure that there was signal for all cases. mutREAD data for 81.81% (9/11) of samples with non-zero signature proportions for S17 recapitulated WGS proportions, one outlier with four times the proportion of S17 in mutREAD data is recorded,

this is most probably technical artefact associated with filtering of SNVs during raw data processing. A strong positive linear correlation between mutREAD and WGS signature proportion estimates was observed with Pearson correlation *R* = 0.83 and significant *p* value of 0.0015(Figure 5.3).

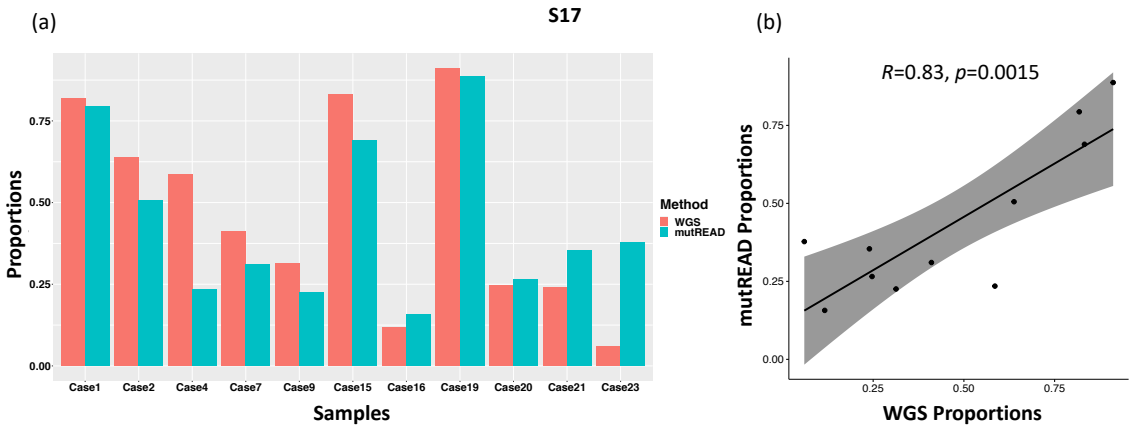


Figure 5.3: Mutational signature wise comparative analysis of mutREAD signature with matched WGS data: S17 comparison.

(a)Grouped bar charts for individual pre(diagnostic) samples, first bar is for the signature from WGS data and is followed by the bar for signature obtained from mutREAD. X-axis is pre(diagnostic) samples with non-zero signature proportions. Y-axis is S17 proportions. S17A and B were merged into one as they are linked to one mutational process in view of their similarity and to ensure that there was signal for all cases.

(b) Scatter plot showing strong positive linear correlation between WGS and mutREAD S17 proportions. Pearson correlation (*R*)=0.83 with significant *p*=0.0015.

S18 mutational signature was also successfully recapitulated with 10/11 (90.90%) samples recovering S18 proportions in mutREAD data. It also had a strong linear correlation with Pearson correlation *R* = 0.83 and significant *p* value of 0.0015(Figure 5.4). Aging associated

signature S1 was recovered in all non-zero samples (12/12). In half (6/12) of these samples, proportions in mutREAD data was double the proportions in WGS, hinting towards FFPE associated C>T contribution. Pearson correlation showed a moderate (*R*= 0.57) linear relation with a significant *p* value of 0.053(Figure 5.5)

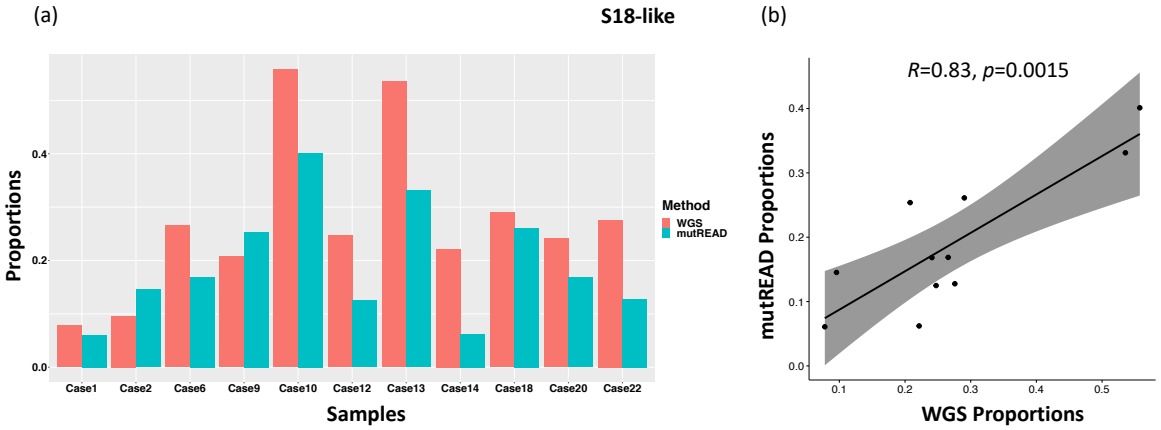


Figure 5.4: ROS linked S18 comparative analysis with correlation between WGS and mutREAD proportions.

(a)Individual pre(diagnostic) sample-based comparison for S18 proportions depicted in the bar chart, first bar is for WGS followed by mutREAD. X-axis for pre(diagnostic) samples and Y-axis for the S18 proportions.

(b) Strong positive Correlation between WGS and mutREAD for S18 proportions shown in the scatter plot. Pearson correlation (*R*)=0.83, *p*= 0.0015

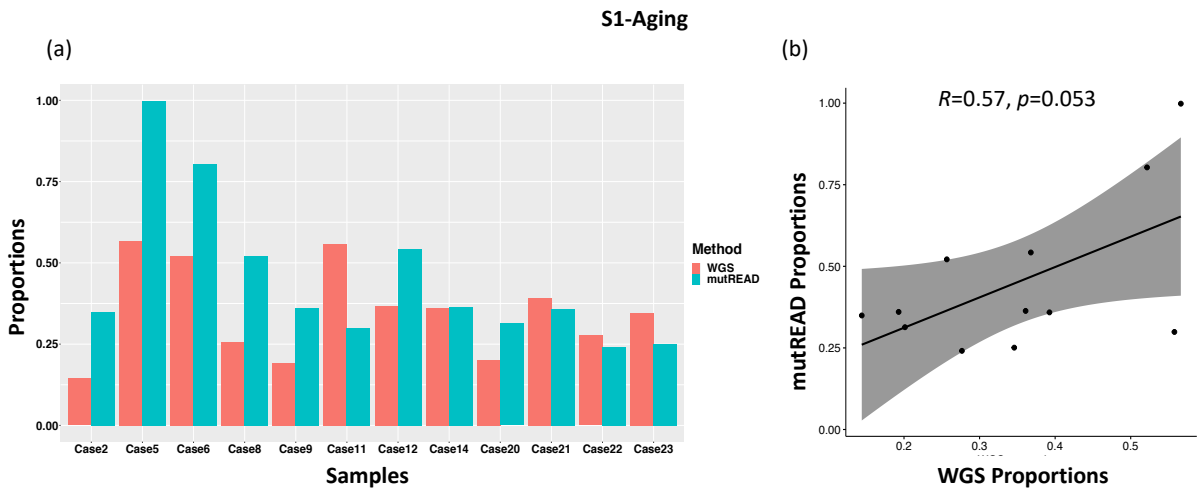


Figure 5.5: Aging associated S1 based comparative analysis of mutREAD with matched WGS data.

(a)X-axis is pre(diagnostic) samples with non-zero signature proportions. Y-axis is S1 proportions. First bar in the grouped bar chart is WGS and second bar chart for mutREAD S1 proportions.

(b) A moderate positive correlation is shown between WGS and mutREAD for S1 proportions. Pearson correlation (*R*)=0.57, *p*= 0.053

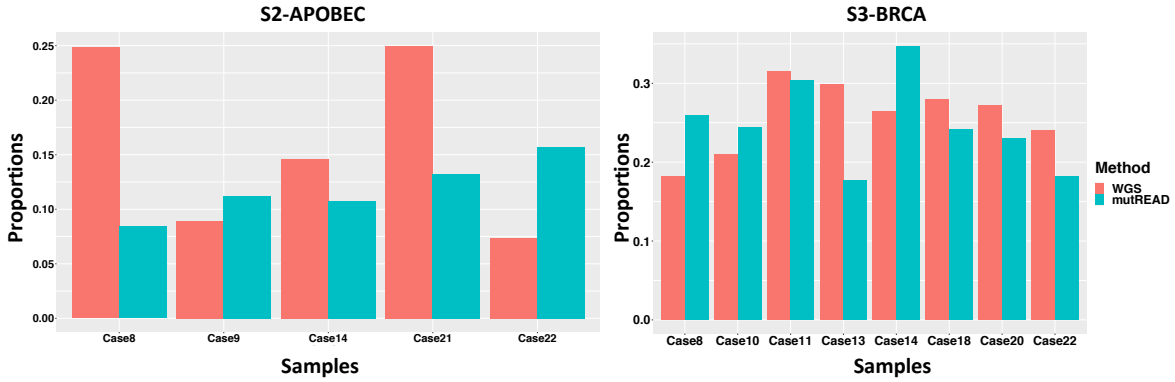


Figure 5.6: Sample based comparative analysis between WGS and mutREAD proportions for APOBEC and BRCA signatures.

Grouped bar chart depicting S2-APOBEC and S3-BRCA mutational signatures, samples are on X-axis and signature proportions are on Y-axis. First bar denotes proportions from WGS and second bar is for mutREAD.

APOBEC associated mutational signature S2 prevalence was low with only 5/23(21.7%) patients had a non-zero prevalence. mutREAD data was not able to recapitulate this signature (*R* = -0.49, *p*=0.4) as compared to the other dominant signatures like S17, S18 and S1-aging. Low prevalence and random sampling of SNVs in mutREAD further dilutes this signature and might account for this (Figure5.6).

S3-BRCA mutational signature was prevalent in 8/23(34.7%) cases, mutREAD data for this signature was recovered better than APOBEC but did not correlate with the WGS (*R*=0.054; *p*=0.9) (Figure5.6).

In Figure 5.7, an overview of the six mutational signatures predominant in OAC are shown across the pre(diagnostic) samples in the cohort in a comparative analysis with matched WGS and the cosine similarity is indicated

Overall good(above 0.5) cosine similarity values (Median=0.8) were obtained for all the pre(diagnostic) samples, 82.6% (19/23) of FFPE samples recapitulated the OAC signatures with a cosine similarity value ranging from 0.51 to 1.0. Only 17.4% (4/23) poorly matched to WGS with low cosine values ranging from 0 to 0.50 (Table 5.2; Figure 5.7). These are outlier samples with more than 50% of their proportions composed of S1-aging associated SNVs(C>T), may be contributed by FFPE artefacts.

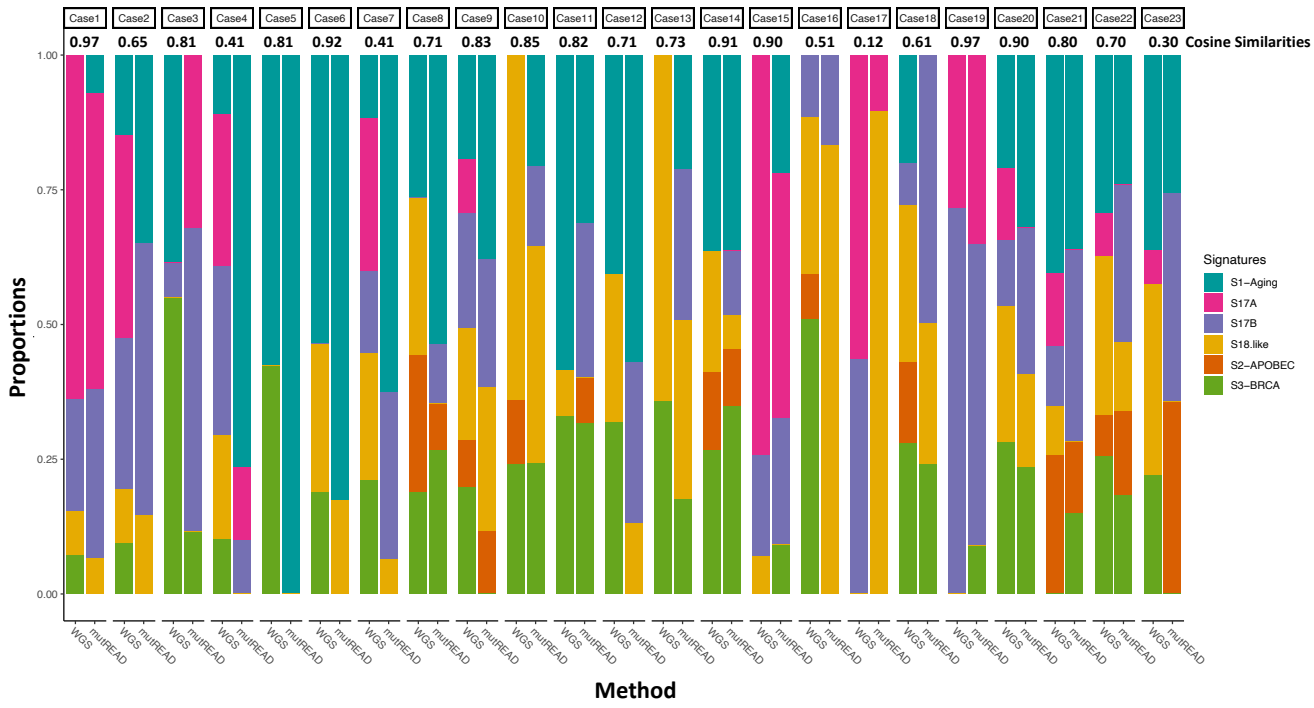


Figure 5.7: Comparitive landscape of mutational signatures obtained from mutREAD to WGS:

Comparative mutational signature analysis for six OAC mutational signatures from 23 pre(diagnostic) samples with their matched WGS data. Respective cosine similarity value for mutREAD vs WGS is shown on top of each stacked bar in the plot. First bar in the grouped bar is for WGS and second bar is for mutREAD. X-axis, samples were labelled on top of each pair of grouped bars and the method from which the proportions were obtained are labelled on x-axis. Y-axis is for signature proportions.

Patient ID	**Cosine Similarity with WGS** (Median =0.80)	**Total number of SNVs (mutREAD)** (Median =191)
Case1_Pre(diagnostic)	0.97	474
Case2_Pre(diagnostic)	0.65	66
Case3_Pre(diagnostic)	0.81	591
Case4_Pre(diagnostic)	0.41	343
Case5_Pre(diagnostic)	0.81	334
Case6_Pre(diagnostic)	0.92	298
Case7_Pre(diagnostic)	0.41	65
Case8_Pre(diagnostic)	0.71	258
Case9_Pre(diagnostic)	0.83	268
Case10_Pre(diagnostic)	0.85	158
Case11_Pre(diagnostic)	0.82	242
Case12_Pre(diagnostic)	0.71	102
Case13_Pre(diagnostic)	0.73	139
Case14_Pre(diagnostic)	0.91	152
Case15_Pre(diagnostic)	0.90	291
Case16_Pre(diagnostic)	0.51	191
Case17_Pre(diagnostic)	0.12	78
Case18_Pre(diagnostic)	0.61	97
Case19_Pre(diagnostic)	0.97	341
Case20_Pre(diagnostic)	0.90	388
Case21_Pre(diagnostic)	0.80	169
Case22_Pre(diagnostic)	0.70	94
Case23_Pre(diagnostic)	0.30	53

Table 5.2: Cosine similarities of mutational signatures obtained by mutREAD to WGS. Measure of similarity between all six mutational signatures extracted from mutREAD and WGS are recorded. SNVs obtained from mutREAD, which were used to extract mutational signatures are listed.

The cosine similarity trend in pre(diagnostic) samples was plotted in Figure 5.8, it shows the density is skewed towards the higher end of the range 0 to 1 (median= 0.8). Further, confirming that we were able to recapitulate the WGS mutational signatures by mutREAD. Some outliers as mentioned earlier contributed to low scores, mostly the SNVs associated with FFPE artefacts/aging (C>T).

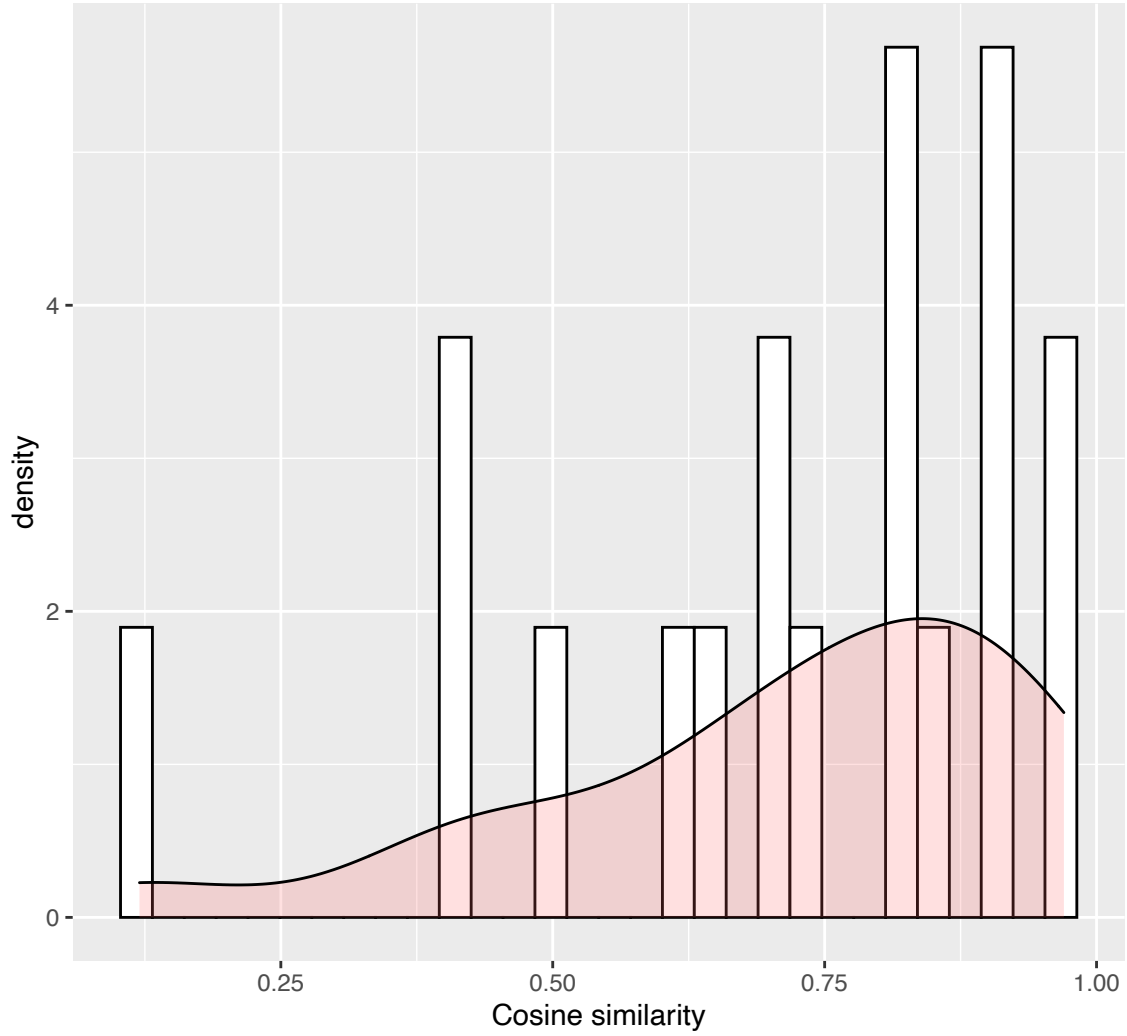


Figure 5.8: Trend of cosine similarity observed: Cosine similarity values between mutREAD derived signatures and their matched WGS were measured on a scale of 0 to 1 for the 23

pre(diagnostic) samples. 0= not matching ;1=100% matching. Number of cases falling in the range of 0 to 1 were plotted on this density histogram.

When the tumour cellularity (percentage of tumour content in whole section of the tumour block) was compared with the cosine similarity for WGS signatures from mutREAD data (Figure 5.9) no pattern was observed. Furthermore, mutREAD can be used to obtain signatures from samples with tumour content as low as 10-20%. (Table 5.3). Tumour cellularity estimated from pathology is quite crude and often an over-estimate, unless it is estimated from the sequencing data.

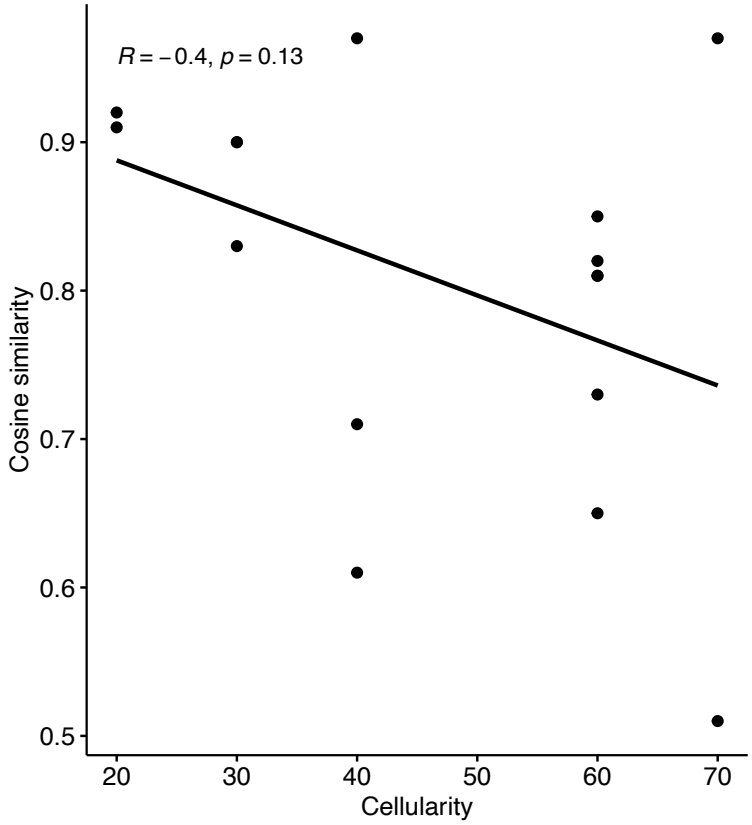


Figure 5.9: Tumour Cellularity v/s Cosine similarity

Scatter plot comparing the tumour cellularity of pre(diagnostic) samples with cosine similarity obtained for mutREAD signatures with WGS data. Pearson correlation was computed with *R*=-0.4, *p*=0.13

Sample ID	Pre(diagnostic) Blocks	Resection Blocks (Treated)
Case1	40%	20%
Case2	60%	20%
Case3	60%	60%
Case4	70%	10-20%
Case5	60%	30%
Case6	20%	30%
Case7	<5%	NA
Case8	10%	30%
Case9	30%	<5%
Case10	60%	10%
Case11	60%	10%
Case12	40%	30%
Case13	60%	20%
Case14	20%	20%
Case15	30%	20%
Case16	70%	60%
Case17	80%	30%
Case18	40%	50%
Case19	70%	20%
Case20	30%	<5%
Case21	NA	<10%
Case22	70%	NA
Case23	60%	60%
Case24	NA	60%
Case25	40%	NA

Table 5.3: Tumor cellularity of FFPE samples from Pre(diagnostic) and matched resection samples estimated by pathology

Estimated percent of tumour content for the twenty-five biopsy and matched resection samples as available, estimated by pathologist review of diagnostic slides.

5.5 Influence of therapy on proportions of mutREAD derived OAC signatures.

Having ascertained that the mutREAD data recapitulated the signatures the next task was to examine the differences in proportions of OAC mutational signatures between the pre(diagnostic) and chemo treated tumour samples. The standard triple therapy regimen (ECX or EOX) was planned for all the 25 cases of the cohort that preceded the recent switch to FLOT chemotherapy. The aim was to see how the treatment influences proportions of a type of signature in these samples. Proportions obtained for a particular signature across all samples in each subtype was compared. Three predominant OAC signatures (S1 (Aging), S17B and S18-like) showed differences in their proportions before and after therapy.(Figure 5.10). The Proportion of the S1 signature associated with aging tended to increase after treatment similar to the increasing proportions of S17B. On the other hand the S18-like proportion decreased after treatment. However, these were not statistically significant likely due to the sample size.

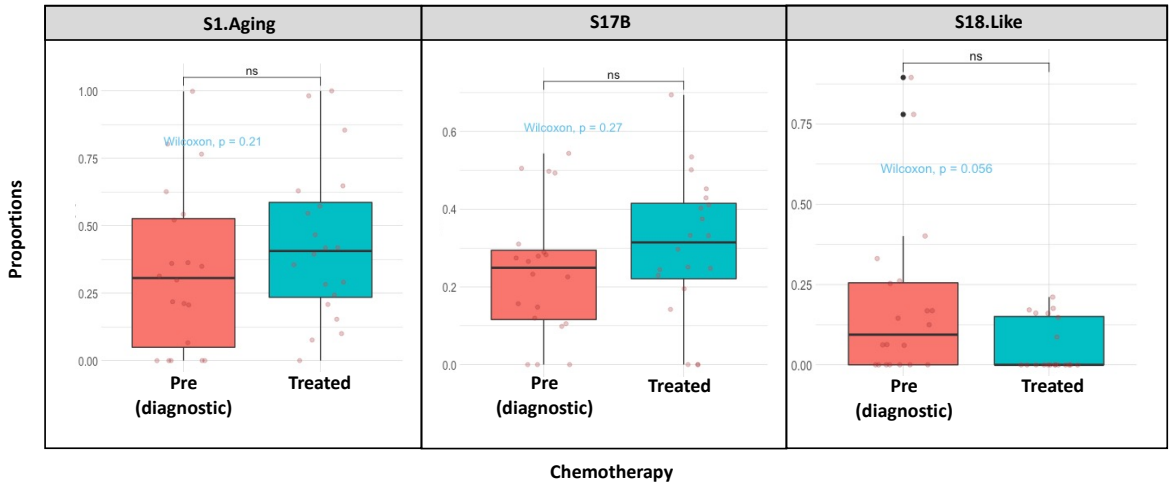


Figure 5.10: Influence of therapy on proportions of mutational signatures obtained from mutREAD data. Change in proportions of predominant OAC signatures between pre(diagnostic) biopsies and resection tumours after treatment are shown here. Three of six OAC signatures tend to vary between the sample type (S1.Aging- p=0.21; S17B- p=0.27; S18.like- p=0.05)

5.6 Summary

In this chapter, I have presented the validation of mutREAD in a new cohort of 25 OAC patients, using FFPE samples from 23 pre(diagnostic) biopsies and matched post treatment resection tumours for all 25 patients. The experimental protocol for the mutREAD DNA library preparation performed well with a relatively low input of FFPE DNA (250g). The quality check showed that mutREAD DNA libraries had minimal to no contamination of by-products of ligation and amplification (un-ligated adapters). The cohort had varied quality of FFPE DNA. For instance the post treatment samples were relatively degraded compared to pre(diagnostic samples), yet across all these parameters it was possible to generate signature data from these cases. This suggests that the mutREAD lab protocol is reproducible, consistent and can be adapted to account for changes in input DNA. The method therefore has potential to be used for large scale clinical screening after accounting for its limitations. There are limitations of this method, it will not perform well for signatures at low prevalence, for instance S2-APOBEC in this cohort. Also, the method is designed for known signatures in a given cancer type. As I have shown only predominant signatures in OAC were recapitulated (S17, S18-like , S1-Aging and S3-BRCA). This method is not recommended for *de novo* discovery of signatures in a cohort. For faithful recapitulation of signatures, a sequencing depth of 100x is recommended.

On sequencing, a range of SNVs were obtained (median 191), this can be in part explained by the cellularity, which was varied across blocks. DNA was extracted from whole block tumour sections to help ease the protocol and avoid time consuming macro-dissections. This contributed to varied cellularity however despite this mutREAD performed well, with 82.6% of samples recapitulating the WGS mutational signature profiles with a good cosine similarity

(median=0.8). This is an additional advantage of mutREAD for large scale applications, where tumour content and quality varies.

Using the mutREAD mutational signatures data I investigated the differences in the proportions of the mutational signatures in pre (diagnostic) and post treatment tumours. While the sample size was small to make definitive conclusions, I observed an increasing trend for signature 1 (aging) and S17B after treatment. All the patients in this cohort were treated with capecitabine which has previously been shown to be linked with the S17B mutational signature. This is also in line with reports in pan cancer analysis of treatment tumours for these signatures[24]. In contrast we found that S18, linked to ROS, decreases after treatment - this observation needs to be confirmed in large cohort.

Overall, mutREAD was shown to be a feasible method to generate data from low-input and poor-quality clinical samples. This method is promising for use as a cost-effective and easy protocol to ascertain the signature profiles in routine clinical specimens.

6. Discussion and Future Work

In this thesis I have presented a comprehensive characterization of dynamics of mutational processes during the evolution of OAC and development of a novel cost-effective DNA sequencing method from small amounts of archival DNA samples. This included investigation of the mutational processes during the course of OAC development from pre-cancerous stages to advanced metastatic spread. To facilitate this analysis, I therefore gathered data from a large and clinically annotated cohort of 997 samples from different stages of OAC development. The cohort also has 45 matched Barrett's samples from the same OAC patients. This extensive cohort is one of the strengths of this study. Also, tumours from 4 OAC patients with metastases. In total, I have 59 metastases samples and mostly these are from nodal spread. Building on existing knowledge of mutational signatures described in OAC tumours[35], I have explored the details of the temporal behaviour and clinical extrinsic risk factors affecting these mutational processes during the evolution of OAC. To my knowledge this study provides the first comprehensive description of the dynamics of mutational events during tumorigenesis in this disease. One of the limitations of the cohort is limited samples from metastases and their linked primaries. In future, enriching linked Barrett's, OAC and metastases trios from same patient would be ideal for further understanding of the dynamics of mutational processes during OAC development. Also, it would be ideal to enrich the cohort with multiple samples from the same tumour to help in heterogeneity analysis.

In this study, my focus was to study single base substitutions (SBS) type mutational signatures, as these are very informative and can be linked to deficiencies in cellular processes and have implications in disease progression.

I started with a comparative signature analysis, in order to rule out any bias or artefacts due to usage of various signature extraction methods. I evaluated three different methods, all of which were based on NMF but were slightly different in their statistical methods. I used tumour WGS data (416 samples) for ease of data handling. Overall, our analysis showed a coherent output across methods, I therefore extended the analysis to 997 samples.

I have characterized and compared the landscape of mutational processes at each stage of the disease and the larger number of samples enabled us to capture mutational processes operating even in low magnitude. These results are in line with our previous study[35]. In addition I have identified new mutational signatures for the first time in OAC development, such as the Colibactin associated SBS41 and other signatures with small proportions such as SBS30, associated with impairment in the BER pathway, and the platinum therapy linked SBS35.As expected, aging associated processes (SBS1 and SBS5) were running constantly in the background during tumour evolution, whereas SBS40 which has been correlated to age of the patients in some cancer types [45] was relatively more prevalent at late disease stages. I found that a few mutational signatures are activated early, such as SBS17a/b(Unknown), SBS8(DDRD), SBS40(Age?), SBS41(Colibactin), and SBS44 (MMRD), and these were relatively enriched in the late stages of cancer and therefore might have role in tumour evolution of this cancer type. In future, this study can be extended to indel and rearrangement type signatures. To shed light into the implications of complex cellular events during disease development.

I showed that OAC evolution is marked by frequent mutagenesis bottlenecks, whereby mutational signature dynamics change. In future, it would be helpful to look into the tumour heterogeneity and mutational processes using multiple samples from the same tumour.

The dominant SBS17b/a process appears to be triggered early in preneoplastic stages and increasingly accumulate during the later course of the disease. SBS17 processes were also reported to accumulate in late stages of breast cancers[104,105]. The spike in their proportions especially in metastatic samples might be partly contributed to by the mutagenic effects of therapies such as the nucleoside metabolic inhibitors (capecitabine and 5 Fluro Uracil), as majority of the cases studied were pre-treated resection samples after surgery. These findings are in agreement with a study on effects of therapies in pan cancer metastatic samples[24]. However, the prevalence of SBS17 processes in Barrett's and in chemo naive tumours[48]suggests that the trigger for this mutational event is a combined effect of impairment in endogenous processes as well as the influence of exogenous exposures during therapy[54,24]. It should also be noted that, SBS17 predominates especially in OAC when compared to other cancer types, suggesting the influence of tissue specific mutagenesis[35,45]

SBS17b was also accompanied by increased copy number instability (CIN), DDR and telomerase activity. Polyploidy representing CIN in OAC cell lines has previously been reported by collaborators and by our lab, caused by mitotic slippage due to impairment of chromosomal attachments[106]. SBS17b with a CIN background suggests that there maybe a clonal advantage for the clones harbouring these mutations for expansion[107]. It is possible that SBS17b activity may promote CIN, which will then elicit an immune response by activating inflammatory pathways via induction of genomic DNA into the cytosol[108]. Telomeric regions are also susceptible to oxidative DNA damage as they are enriched with guanine residues, SBS17b association with telomeric activity suggests partial involvement of external oxidative damage[109]. Also, we observed a reduction of regulatory T cells and monocyte infiltration, which favours tumour progression. This provides an avenue for a detailed study

of influence of tumour microenvironment determinants, especially with regards to SBS17b mediated tumour evolution. In future, a IHC study to look into the tumour tissue of patients with SBS17 predominance for these immune cell infiltration by staining for their respective markers can help validate the findings of the current study.

We also investigated positively selected genes in primary tumours with dominant SBS17 proportions. Interestingly, genes from chromatin remodelling and transcriptional control pathway such as SMARCA4, KMT2D and ARID2 were positively selected compared with the recurrent OAC drivers such as KRAS, PIK3CA, PTEN,ARID1A and APC. In future, to validate these finding in OAC organoid models, a CRISPR panel can be designed to probe their influence in tumour progression. Further, we observed a common nucleosome periodicity pattern across cancer stages and this may be linked with changes in chromatin remodelling genes which appear to be selected for in the presence of this signature, as observed in the pan cancer data for this signature[53]. In future, MNaseSeq data from the OAC cell lines/organoids can be generated and used to obtain the nucleosome positioning to address any tissue specific changes in nucleosome periodicity patterns and their influence in tissue specific mutational events such as SBS17 associated processes.

I also looked into the prognostic relevance of SBS17 and surprisingly patients with higher proportions of this signature showed a better outcome and survival. Further, I also looked into subset of patients with and without evidence of adjacent Barrett's, and a similar trend was observed. These findings are in contrast with a small study of untreated adenocarcinoma of gastro oesophageal junction tumours (n=124) from Chinese patients[110]. Since most of the tumour samples are obtained after treatment, it would be helpful to investigate outcome with SBS17 proportions in chemo naive and chemo treated samples. This will suggest the influence of treatment for a good outcome.

Mutagenesis linked to DDR deficiencies had small, but significant contributions. In particular, base excision repair (BER) impairment and APOBEC associated processes appeared to be distinctly active after OAC transformation (OAC) and tended to be less prevalent in advanced disease stages. APOBEC mutagenesis has been reported to be associated with the development resistance for hormone therapy in breast cancer[111]. During the course of OAC , I observed a decrease in its activity during metastasis. BER associated mutagenesis tends to decrease in late stages of OAC, which was also the case in breast cancer[104]

Importantly, while mutations arising due to BER deficiency were relatively few on average, they appeared predominantly in TP53 mutated cancers. This mutational process was associated with a significantly worse patient outcome, which is in line with studies in colorectal adenocarcinoma[112], Furthermore, we uncovered signatures of early and late impairment of DDR processes, mainly acting on the NHEJ, HR, MMR and BER pathways. These findings are in agreement with our previous study[35], and suggests that DDR deficiency mutagenesis may be an underappreciated prognostic and therapeutic opportunity in a subset of OAC patients.

We also observed the presence of a colibactin-linked mutational signature (SBS41) as an early event that expanded in advanced stages, this accounted for ~8% of total signatures proportions/sample in almost all samples (990/997). This signature has been predominately reported in colorectal cancers, where it was suggested to be linked with genotoxins originating from certain strains of *E.Coli* during tumour progression[113,114]. *E.Coli* has been reported to form part of the microbiota in Barrett's and OAC, and not in normal squamous oesophagus[115]. Our analysis strengthens this possible contribution of colibactin-induced stress with regards to OAC tumour development. For future directions, this can be further interrogated to identify traces of the bacterial DNA content in WGS data as well as matched

expression profiles. Such an analysis may be facilitated by recruiting patients with history of bacterial infection. This could also be interrogated experimentally by co-culturing E Coli genotoxins in oesophageal squamous or gastric and Barrett'sorganoids adapting *Manzano,C.P et al 2020* methodology. This will help better understand the role of E.Coli infections during OAC development similar to the role of H.pylori infections in gastric cancer and gastric mucosa-associated lymphoid tissue (MALT) lymphoma[116–118].

In this study, I observed certain mutational processes like APOBEC and colibactin were linked to primary tumours than Barrett's and advanced metastases, whereas SBS17b was linked to Barrett's. With collaboration we have looked into a machine learning based tissue classification algorithm and were able to classify tissue type based on mutational signatures. This can be further developed and trained using even more large data sets and validated on independent cohorts. This algorithm has a potential to be used in clinical diagnosis and early detection.

In addition to these processes, we found that exposure to certain risk factors may modulate mutagenesis in Barrett and OAC patients. In particular, we found novel associations between alcohol consumption or NSAID intake and DDR deficiencies in Barrett's. The use of NSAIDs was reported to be linked with a reduction in mutation rate, especially of SBS17[119]. In primary tumours, new associations were discovered between SBS17 and alcohol intake as well as smoking. Alcohol consumption has previously been linked to SBS16 in liver and oesophageal squamous cell carcinoma[20,120]; nevertheless, SBS16 was absent in human liver stem cells chronically exposed to alcohol[121], suggesting further investigations are needed into the potential mutagenic marks left by this risk factor in the genome. Also, among the tumour factors we studied, positive nodes were associated with SBS17b and SBS3 proportions, further confirming the increased prevalence of SBS17b in advanced stages[122].

By expanding the cohorts of analysed cancer genomes, it is becoming clear that the repertoire of uncovered mutational processes in OAC continues to expand. While the SBS17 process undoubtedly dominates across tumour development stages, and APOBEC/DDRD mutagenesis appear particularly important in shaping primary tumours, it is likely that a variety of mutational processes will continue to emerge as acting in a minority of OACs, much like the long tail of cancer drivers.

Despite the relatively large size of the cohort in the present study, the findings should be interpreted taking into account the uncertainty around the contribution of the less prevalent signatures. This is particularly true for pre-neoplastic stages since the size of our cohort was smaller and the mutation burden is smaller. In addition, our insights into metastatic disease are limited by the small number of metastatic and lymph node samples available for analysis. Though we gathered a well annotated clinical data, further data curation will help to validate these findings and perhaps will help in expanding on other risk exposures.

Future research should focus on experimentally validating and further elucidating the role of BER and SBS17 mutagenesis in the progression of OAC, from a genetic and environmental perspective. Some initial experiments have already been designed towards experimental characterization of SBS17 but time did not permit me to complete this work. The hypothesis underlying these experiments is that T>G/C mutations at CTT trinucleotide may be due to misincorporation of oxidised guanines from the altered nucleotide pool by trans-lesion DNA polymerases during replication. Model cell lines for these experiments would be the gastric cell line (HFE-145) as a columnar normal control, since our lab has shown origins of Barrett's from gastric cardia[123]. We would use Barrett's(CP-D) and OAC (FLO-1) to provide tissue type context, mutation data from untreated cell lines could be used for normalisation.

The experimental protocol would be to treat the cell lines with a near lethal concentration, already determined for these cell lines, of 750uM of 8-oxo-2-deoxyguanosine-5'-triphosphate in the growth media, for 48hrs. Single live cells would then be sorted by flow cytometry and propagated. The treatment cycle is repeated to ensure, enough mutations accumulate for downstream signature analysis. After a few passages the cells are harvested to extract DNA and perform sequencing using our in house mutREAD protocol.

Aside from the insights into cancer pathogenesis that maybe obtained from mutational signatures, they are also being suggested as clinical tools for early detection, prevention and patient stratification[23,35,66,124,125,126]. Despite their enormous clinical potential, the only standard method available for their identification to date has been WGS. WGS is relatively expensive method, that requires high quality and quantities of DNA. For application at scale and in the clinic, it is not feasible, as the samples are usually preserved in FFPE. DNA from FFPE material is low in yield and quality. So, WGS will not suit the purpose as a clinical tool. Another key component of my PhD was therefore to develop a low cost DNA sequencing method for study of mutational signature from low quantities of archival clinical samples.

mutREAD produces reproducible and highly specific reduced representation libraries and the derived mutational signatures mirror the WGS-derived signatures with good to high cosine similarity. Importantly, this method is also appliable when used with highly degraded DNA samples.

When applied to tumour samples from OAC patients, I have shown that mutREAD outperforms the previously proposed methods WES and 10x sWGS. OAC is characterized by abundant somatic mutations, which are most prevalent in intergenic and intronic regions which are also covered by the sequenced fragments[35,127]. The limitation of our method is the

choice of library preparation protocols to study mutational signatures in other cancer types, which will also depend on the overall mutation rate and the genomic distribution of the somatic mutations. Since our method sequences 1.5% of the whole genome, it can be applied for identification of predominant mutational signatures in a cancer type. *De novo* identification of mutational signatures is not recommended for our method, as this was developed based on already informed mutational signatures.

In terms of scalability and cost mutREAD outperforms other methods. In our hands, the cost associated with mutREAD libraries synthesis is 80% lower than for 10x sWGS and 96% lower than for WES libraries. Sequencing costs on the Illumina HiSeq 4000 are comparable for WES and mutREAD libraries, while sequencing 10x WGS libraries is at least three times more expensive. Further, due to its high multiplexing capabilities for sequencing and for library preparation mutREAD is highly scalable for studying larger cohorts.

I validated mutREAD on an additional archival FFPE samples from OAC patients. I slightly modified the protocol so as to use relatively low quantities of DNA for library preparation and the coverage of sequencing was also lowered. This is a challenging test for any method. With these changes, I was still able to recapitulate the OAC specific six mutational signatures from the mutREAD data. I used tumour samples pre and post chemotherapy and I found SBS1 and SBS17b tend to increase after treatment in line with the literature in pan cancer data[24].

Given its ease of use and low cost, we envision a wide range of applications for mutREAD to study mutational signatures in basic research and translational settings. For example, clinical trials using mutational signature-based patient stratification to assign optimal therapies become feasible. mutREAD could further improve the mutational signature-based prediction of homologous recombination deficiency in clinical samples[66,128]. Together with computational tools for coarse-grained copy alteration detection[129,130], mutREAD could

provide a detailed view of the role of mutational processes in cancer progression and evolution from archived material. Our method can also be extended to study other DNA alterations such as copy number changes, structural variations and indels. A collaboration with University of Dundee is ongoing, where mutREAD is employed to screen for mutational signature based homologous recombination deficiency (HRD) status in treatment responders from a clinical trial of advanced gastroesophageal cancer patients[131] It is hoped that this method will ultimately allow the study of mutational signatures in much larger cohorts and in clinical settings where FFPE-derived DNA samples are routinely collected[132].

Overall this study has made a contribution to characterisation of the mutational signatures across stages of OAC development to help in deepening our understanding of the active mutational processes. These mutational changes could help inform therapies in a stage-dependent manner. It is hoped that development of a new cost-effective DNA sequencing method will allow mutational signatures to be applied in the clinic.

Bibliography:

1. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300**, 149–152 (1982).
2. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
3. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
4. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
5. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
6. Mardis, E. R. *et al.* Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
7. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
8. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
9. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
10. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science (80-.).* **354**, 618–622 (2016).

11. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

12. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).

13. Stratton, M. R. Exploring the genomes of cancer cells: Progress and promise. *Science (80-.).* **331**, 1553–1558 (2011).

14. Nik-Zainal, S. & Morganella, S. Mutational signatures in breast cancer: The problem at the DNA level. *Clin. Cancer Res.* **23**, 2617–2629 (2017).

15. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

16. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 1–11 (2018).

17. COSMIC SBS mutational signature database.

18. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).

19. Macintyre, G. *et al.* Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).

20. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, (2017).

21. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282-1294.e20 (2019).

22. Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage.

Nat. Cancer **2**, 643–657 (2021).

23. Connor, A. A. *et al.* Association of Distinct Mutational Signatures With Correlates of Increased Immune Activity in Pancreatic Ductal Adenocarcinoma. *JAMA Oncol.* **3**, 774–783 (2017).

24. Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).

25. Tan, V. Y. F. in Nonnegative Matrix Factorization with the -Divergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1592–1605 (2013).

26. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, (2015).

27. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).

28. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 1–11 (2016).

29. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 1–12 (2018).

30. Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538–549 (2016).

31. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.* **4**, 1–9 (2013).

32. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on

regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).

33. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove file:///Users/shujath/Desktop/References/jessica zukermann 2017.pdfaround Nucleosomes. *Cell* **175**, 1074-1087.e18 (2018).

34. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e16 (2019).

35. Secrier, M. *et al.* Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.* **48**, 1131–1141 (2016).

36. Killcoyne, S. & Fitzgerald, R. C. Evolution and progression of Barrett's oesophagus to oesophageal cancer. *Nat. Rev. Cancer* **0123456789**, (2021).

37. Kamangar, F. *et al.* The global, regional, and national burden of oesophageal cancer and its attributable risk factors in 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol. Hepatol.* **5**, 582–597 (2020).

38. Ho, A. L. K. & Smyth, E. C. A global perspective on oesophageal cancer: two diseases in one. *Lancet Gastroenterol. Hepatol.* **5**, 521–522 (2020).

39. Cunningham, D. *et al.* Capecitabine and Oxaliplatin for Advanced Esophagogastric Cancer. *N. Engl. J. Med.* **358**, 36–46 (2008).

40. Shaheen, N. J., Falk, G. W., Iyer, P. G. & Gerson, L. B. ACG Clinical Guideline: Diagnosis and Management of Barrett's Esophagus. *Am. J. Gastroenterol.* **111**, 30–50 (2016).

41. Rustgi, A K; El-Serag, H. B. Esophageal carcinoma. *N. Engl. J. Med.* 2499–2509 (2014). doi:10.5694/j.1326-5377.1937.tb53511.x

42. Lordick, F. *et al.* Oesophageal cancer: ESMO clinical practice guidelines for diagnosis,

treatment and follow-up. *Ann. Oncol.* **27**, v50–v57 (2016).

43. van Nistelrooij, A. M. J. *et al.* Hereditary Factors in Esophageal Adenocarcinoma. *Gastrointest. Tumors* **1**, 93–98 (2014).

44. Frederik Hvid-Jensen, M.D., Lars Pedersen, Ph.D., Asbjørn Mohr Drewes, M.D., D. M. S. & Henrik Toft Sørensen, M.D., Dr. Med. Sci., and Peter Funch-Jensen, M.D., D. M. S. Incidence of Adenocarcinoma among Patients with Barrett's Esophagus. *N. Engl. J. Med.* (2011).

45. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

46. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).

47. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).

48. Ross-Innes, C. S. *et al.* Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat. Genet.* **47**, 1038–1046 (2015).

49. Greenman, C. D. *et al.* Estimation of rearrangement phylogeny for cancer genomes. *Genome Res.* **22**, 346–361 (2012).

50. Gerstung, M., Jolly, C., Leshchiner, I. & Dentro, S. C. The evolutionary history of 2 , 658 cancers. **578**, (2020).

51. Rubanova, Y. *et al.* Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. *Nat. Commun.* **11**, (2020).

52. Weaver, J. M. J. *et al.* Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat. Genet.* **46**, 837–843 (2014).

53. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* **175**, 1074-1087.e18 (2018).

54. Christensen, S. *et al.* 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. *Nat. Commun.* **10**, 1–11 (2019).

55. Erichsen, R. *et al.* Erosive Reflux Disease Increases Risk for Esophageal Adenocarcinoma, Compared With Nonerosive Reflux. *Clin. Gastroenterol. Hepatol.* **10**, 475-480.e1 (2012).

56. Fein, M., Maroske, J. & Fuchs, K. H. Importance of duodenogastric reflux in gastro-oesophageal reflux disease. *Br. J. Surg.* **93**, 1475–1482 (2006).

57. Räsänen, J. V. et al. The expression of 8-hydroxydeoxyguanosine in oesophageal tissues and tumours. *Eur. J. Surg. Oncol.* **Volume 33**, 1164–1168 (2007).

58. Jiménez Molinos, P. *et al.* Free radicals and antioxidant systems in reflux esophagitis and Barrett's esophagus. *World J. Gastroenterol.* **11**, 2697–2703 (2005).

59. Dvorak, K. *et al.* Bile acids in combination with low pH induce oxidative stress and oxidative DNA damage: Relevance to the pathogenesis of Barrett's oesophagus. *Gut* **56**, 763–771 (2007).

60. Inoue, M. *et al.* Induction of chromosomal gene mutations in Escherichia coli by direct incorporation of oxidatively damaged nucleotides: New evaluation method for mutagenesis by damaged dna precursors in vivo. *J. Biol. Chem.* **273**, 11069–11074 (1998).

61. Satou K, Kawai K, Kasai H, Harashima H, K. H. Mutagenic effects of 8-hydroxy-dGTP in live mammalian cells. *Free Radic Biol Med* **42**, 1552–60 (2007).

62. Satou K, Hori M, Kawai K, Kasai H, Harashima H, K. H. Involvement of specialized DNA polymerases in mutagenesis by 8-hydroxy-dGTP in human cells. *DNA Repair* **8**, 637–

42 (2009).

63. Kamiya, H. Mutations Induced by Oxidized DNA Precursors and Their Prevention by Nucleotide Pool Sanitization Enzymes. *Genes Environ.* **29**, 133–140 (2007).

64. Suzuki, T. & Kamiya, H. Mutations induced by 8-hydroxyguanine (8-oxo-7,8-dihydroguanine), a representative oxidized base, in mammalian cells. *Genes Environ.* **39**, 4–9 (2017).

65. Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).

66. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).

67. Gulhan, D. C., Lee, J. J. K., Melloni, G. E. M., Cortés-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* **51**, 912–919 (2019).

68. Franchini, P., Monné Parera, D., Kautt, A. F. & Meyer, A. quaddRAD: a new high-multiplexing and PCR duplicate removal ddRAD protocol produces novel evolutionary insights in a nonradiating cichlid lineage. *Mol. Ecol.* **26**, 2783–2795 (2017).

69. Peterson BK, Weber JN, Kay EH, Fisher HS, H. H. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS One* **7**, e37135. https://doi.org/10.1371/journal.pone.00371 (2012).

70. Robert J. Toonen, Jonathan B. Puritz, Zac H. Forsman, Jonathan L. Whitney, Iria Fernandez-Silva, Kimberly R. Andrews, C. E. B. ezRAD: a simplified method for genomic genotyping in non-model organisms. *Peer J* **1**, (2013).

71. Scaglione, D., Acquadro, A., Portis, E. et al. RAD tag sequencing as a source of SNP

markers in Cynara cardunculus L. *BMC Genomics* **13**, (2012).

72. Wang, N. *et al.* Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci. Rep.* **10**, 1–12 (2020).

73. Poland JA, Brown PJ, Sorrells ME, J. J. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One.* **7**, e32253. doi (2012).

74. Esposito, S. *et al.* ddRAD sequencing-based genotyping for population structure analysis in cultivated tomato provides new insights into the genomic diversity of Mediterranean 'da serbo' type long shelf-life germplasm. *Hortic. Res.* **7**, (2020).

75. Wang, S., Meyer, E., Mckay, J. K. & Matz, M. V. 2b-RAD: A simple and flexible method for genome-wide genotyping. *Nat. Methods* **9**, 808–810 (2012).

76. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *bioRxiv* (2020). doi:10.1101/2020.12.13.422570

77. Wang, S. *et al.* Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. *PLoS Genet.* **17**, 1–23 (2021).

78. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 1–11 (2016).

79. Gaffney, D. J. *et al.* Controls of Nucleosome Positioning in the Human Genome. *PLoS Genet.* **8**, 1–13 (2012).

80. Cui, F. & Zhurkin, V. B. Structure-based analysis of dna sequence patterns guiding nucleosome positioning in vitro. *J. Biomol. Struct. Dyn.* **27**, 821–841 (2010).

81. Kang, Y. J. & Noh, Y. Development of Hartigan's Dip Statistic with Bimodality Coefficient to Assess Multimodality of Distributions. *Math. Probl. Eng.* **2019**, (2019).

82. Theo A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* **23**, 239-254.e6 (2018).

83. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, (2010).

84. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

85. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

86. Geostatistics, M., Diggle, P. J. & 間違っている. . Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

87. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9**, 1–10 (2013).

88. Yuan, H. *et al.* CancerSEA: A cancer single-cell state atlas. *Nucleic Acids Res.* **47**, D900–D908 (2019).

89. Jimenez-Sanchez, A., Cast, O. & Miller, M. L. Comprehensive benchmarking and integration of tumor microenvironment cell estimation methods. *Cancer Res.* **79**, 6238–6246 (2019).

90. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).

91. Bro, R. & de Jong, S. A Fast Non-negative Constrained Linear Least Squares Algorithm. *J. Chemom.* **11**, 393–401 (1997).

92. Pandeya, N., Williams, G., Green, A. C., Webb, P. M. & Whiteman, D. C. Alcohol Consumption and the Risks of Adenocarcinoma and Squamous Cell Carcinoma of the Esophagus. *Gastroenterology* **136**, 1215-1224.e2 (2009).

93. Gao, Y. B. *et al.* Genetic landscape of esophageal squamous cell carcinoma. *Nat. Genet.* **46**, 1097–1102 (2014).

94. Kim, J. *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–174 (2017).

95. Abudureheman, A. *et al.* High MLL2 expression predicts poor prognosis and promotes tumor progression by inducing EMT in esophageal squamous cell carcinoma. *J. Cancer Res. Clin. Oncol.* **144**, 1025–1035 (2018).

96. Kantidakis, T. *et al.* Mutation of cancer driver MLL2 results in transcription stress and genome instability. *Genes Dev.* **30**, 408–420 (2016).

97. Traven, A., Hammet, A., Tenis, N., Denis, C. L. & Heierhorst, J. Ccr4-not complex mRNA deadenylase activity contributes to DNA damage responses in Saccharomyces cerevisiae. *Genetics* **169**, 65–75 (2005).

98. Mulder, K. W., Winkler, G. S. & Timmers, H. T. M. DNA damage and replication stress induced transcription of RNR genes is dependent on the Ccr4-Not complex. *Nucleic Acids Res.* **33**, 6384–6392 (2005).

99. Yuan, Z. M. *et al.* Regulation of DNA damage-induced apoptosis by the c-Abl tyrosine kinase. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 1437–1440 (1997).

100. Baba, R. A. *et al.* E3B1/ABI-1 isoforms are down-regulated in cancers of human gastrointestinal tract. *Dis. Markers* **32**, 273–279 (2012).

101. Illumina. Illumina Adapter Sequences Introduction 3 Sequences for Nextera Kits 3 Sequences for AmpliSeq for Illumina Panels 16 Sequences for TruSight Kits 18

Sequences for TruSeq Kits 24 Process Controls for TruSeq Kits 36 Legacy Kits 42 Revision History 48 Technic. (2019).

102. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213 (2013).

103. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

104. Bertucci, F. *et al.* Genomic characterization of metastatic breast cancers. *Nature* **569**, 560–564 (2019).

105. Angus, L. *et al.* The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. *Nat. Genet.* **51**, 1450–1458 (2019).

106. Scott, S. J. *et al.* Evidence that polyploidy in esophageal adenocarcinoma originates from mitotic slippage caused by defective chromosome attachments. *Cell Death Differ.* **28**, 2179–2193 (2021).

107. D P Cahill , K W Kinzler, B Vogelstein, C. L. Genetic instability and darwinian selection in tumours. *Trends Cell Biol* **9**, M57-60 (1999).

108. Bakhoum, S. F. *et al.* Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* **553**, 467–472 (2018).

109. Coluzzi, E. *et al.* Oxidative stress induces persistent telomeric DNA damage responsible for nuclear morphology change in mammalian cells. *PLoS One* **9**, (2014).

110. Lin, Y. *et al.* Genomic and transcriptomic alterations associated with drug vulnerabilities and prognosis in adenocarcinoma at the gastroesophageal junction. *Nat. Commun.* **11**, 1–14 (2020).

111. Emily K Law , Anieta M Sieuwerts, Kelly LaPara , Brandon Leonard , Gabriel J Starrett , Amy M Molan , Nuri A Temiz , Rachel Isaksson Vogel , Marion E Meijer-van Gelder ,

Fred C G J Sweep, Paul N Span , John A Foekens , John W M Martens , Douglas Yee, R. S. H. The DNA cytosine deaminase APOBEC3B promotes tamoxifen resistance in ER-positive breast cancer. *Sci Adv* **2**, (2016).

112. Azambuja, D. B. *et al.* Prognostic impact of changes in base excision repair machinery in sporadic colorectal cancer. *Pathol. Res. Pract.* **214**, 64–71 (2018).

113. Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks + E. coli. *Nature* **580**, 269–273 (2020).

114. Dziubańska-Kusibab, P. J. *et al.* Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nat. Med.* **26**, 1063–1069 (2020).

115. Zaidi, A. H. *et al.* Associations of microbiota and toll-like receptor signaling pathway in esophageal adenocarcinoma. *BMC Cancer* **16**, 1–10 (2016).

116. Bayerdörffer, E. *et al.* Regression of primary gastric lymphoma of mucosa-associated lymphoid tissue type after cure of Helicobacter pylori infection. *Lancet* **345**, 1591–1594 (1995).

117. Stolte, M. *et al.* Helicobacter and gastric MALT lymphoma. *Gut* **50**, 19–24 (2002).

118. NAOMI UEMURA, M.D., SHIRO OKAMOTO, M.D., SOICHIRO YAMAMOTO, M.D., NOBUTOSHI MATSUMURA, M. D., SHUJI YAMAGUCHI, M.D., MICHIO YAMAKIDO, M.D., KIYOMI TANIYAMA, M.D., NAOMI SASAKI, M. D. & AND RONALD J. SCHLEMPER, M. D. HELICOBACTER PYLORI INFECTION AND THE DEVELOPMENT OF GASTRIC CANCER. *N. Engl. J. Med.* **345**, 784–789 (2001).

119. Galipeau, P. C. *et al.* NSAID use and somatic exomic mutations in Barrett's esophagus. *Genome Med.* **10**, 1–14 (2018).

120. Li, X. C. *et al.* A mutational signature associated with alcohol consumption and prognostically significantly mutated driver genes in esophageal squamous cell

carcinoma. *Ann. Oncol.* **29**, 938–944 (2018).

121. Jager, M. *et al.* Mutational impact of chronic alcohol use on stem cells in cirrhotic liver. *bioRxiv* 0–30 (2019). doi:10.1101/698894

122. De Mattos-Arruda, L. *et al.* The Genomic and Immune Landscapes of Lethal Metastatic Breast Cancer. *Cell Rep.* **27**, 2690-2708.e10 (2019).

123. Nowicki-Osuch, K & Zhuang, L. et al. Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition. Science. (2021). doi:10.1126/science.abd1449

124. Staaf, J. *et al.* Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **25**, 1526–1533 (2019).

125. Polak, P. *et al.* A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476 (2017).

126. Sarah Moody et al. Mutational signatures in esophageal squamous cell carcinoma from eight countries of varying incidence. *medRxiv* (2021).

127. Dulak, A. M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478 (2013).

128. Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M., Cortés-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* **51**, 912–919 (2019).

129. Zheng, C. *et al.* Determination of genomic copy number alteration emphasizing a restriction site-based strategy of genome re-sequencing. *Bioinformatics* **29**, 2813–2821 (2013).

130. Perry, E. B. *et al.* Tumor diversity and evolution revealed through RADseq. *Oncotarget*

8, 41792–41805 (2017).

131. Hall, P. S. *et al.* Efficacy of Reduced-Intensity Chemotherapy with Oxaliplatin and Capecitabine on Quality of Life and Cancer Control among Older and Frail Patients with Advanced Gastroesophageal Cancer: The GO2 Phase 3 Randomized Clinical Trial. *JAMA Oncol.* **7**, 869–877 (2021).

132. Sujath Abbas,Juliane Perner,Karol Nowicki-osuch, R. F. mutREAD: An affordable method to capture the archaeology of DNA damaging events in cancer genomes. *Nat. Portf. Cancer Community* (2020).

Appendix:

List of Publications

Mutational processes unveil bottlenecks that shape the evolution of oesophageal adenocarcinoma.
Sujath Abbas, Oriol Pich, Ginny Devonshire, Shawn Zamani, Annalise Katz-Summercorn, Sarah Killcoyne, Calvin Cheah, Barbara Nutzinger, Nuria Lopez-Bigas, Rebecca C. Fitzgerald*, Maria Secrier*, OCCAMS (Under Submission)

The mutREAD method detects mutational signatures from low quantities of cancer DNA. Nat.Commun 11, 3166 (2020).
Perner, J*., **Abbas, S***., Nowicki-Osuch, K*. et al. https://doi.org/10.1038/s41467-020-16974-3 *shared first author

mutREAD: An affordable method to capture the archaeology of DNA damaging events in cancer genomes.Behind the paper; Nature Research Cancer Community Blog July27,2020 Author: **Sujath Abbas** Contributing authors: Juliane Perner, Karol Nowicki-Osuch, Rebecca Fitzgerald https://go.nature.com/32VKddz

Multi-omic cross-sectional cohort study of pre-malignant Barrett's esophagus reveals early structural variation and retrotransposon activity. Nat.Commun accepted (2022); NCOMMS-21-27519A
Katz-Summercorn, S, Jammula S ,Frangou A ,Peneva I ,Donovan M O ,Tripathi M ,Malhotra S ,Pietro M ,**Abbas S** ,Devonshire G ,Januszewicz W ,Blasko A ,Nowicki-Osuch K ,MacRae S , Northrop A ,Redmond A , Wedge D, Fitzgerald RC

Rearrangement processes and structural variations show evidence of selection in oesophageal adenocarcinomas.
Alvin Wei Tian Ng*, Gianmarco Contino*, Sarah Killcoyne, Ginny Devonshire, Ray Hsu, **Sujath Abbas**, Jing Su, Aisling M. Redmond, Jamie M.J. Simon Tavaré, Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium, Paul A.W. Edwards and Rebecca C. Fitzgerald. Communications Biology accepted (2022); COMMSBIO-21-2438B

Genomic copy number predicts esophageal cancer years before transformation. Killcoyne S, Gregson E, Wedge DC, Woodcock DJ, Eldridge MD, de la Rue R, Miremadi A, **Abbas S**, Blasko A, Kosmidou C, Januszewicz W, Jenkins AV, Gerstung M, Fitzgerald RC. Nat Med. 2020 Sep 7. Doi: 10.1038/s41591-020-1033-y. Online ahead of print. https://rdcu.be/b6Utl

Identification of Subtypes of Barrett's Esophagus and Esophageal Adenocarcinoma Based on DNA Methylation Profiles and Integration of Transcriptome and Genome Data. Jammula S, Katz-Summercorn A C, Li X, Linossi C, Smyth E, Killcoyne S, Biasci D, Subash V V, **Abbas S**, Blasko A, Devonshire G, Grantham A, Wronowski F,Donovan M O, Grehan N, Eldridge M D, Tavaré S, Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) consortium, Fitzgerald R C .Gastroenterology, 2020 Feb 4 https://www.gastrojournal.org/action/showPdf?pii=S0016-5085%2820%2930156-6

The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. Frankell AM, Jammula S, Li X, Contino G, Killcoyne S, **Abbas S**, Perner J, Bower L, Devonshire G, Ococks E, Grehan N, Mok J, O'Donovan M, MacRae S, Eldridge MD, Tavaré S; Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium, Fitzgerald RC. Nat Genet 51, 506–516 (2019). https://doi.org/10.1038/s41588-018-0331-5

Posters at Conferences

CRUK Early Detection of Cancer Conference (Virtual). Stanford, October 2020. "mutREAD:a cost effective method of detecting mutational signatures for early cancer detection and classification"
Cold Spring Harbor Laboratory (CSHL) The Biology of Genomes, (Virtual). New York USA, *May*, 2020. "Mutational processes active in esophageal adenocarcinoma and their associations with clinical factors"
CRUK Oesophageal Cancer Symposium, London, April 2019. "Signatures of mutational processes active in Oesophageal adenocarcinoma: Insights from the cohort of 416 patients"