Impact of genetic inter-tumour heterogeneity on oesophageal adenocarcinoma development

Sartini, Giulia

Awarding institution:
King's College London

Impact of genetic inter-tumour heterogeneity on oesophageal adenocarcinoma development

Giulia Sartini

King's College London

and

The Francis Crick Institute

MPhil Supervisor: Francesca D. Ciccarelli

A thesis submitted for the degree of

Master of Philosophy

King's College London

April 2022

Declaration

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

Abstract

Oesophageal adenocarcinoma (OAC) incidence in Western countries has increased in the last decades. In the UK, 70-80% of patients are diagnosed with OAC at an advanced stage with metastatic disease. The late diagnosis impacts the success of OAC treatment resulting in a five-year survival rate of ~15%. This poor patient survival is further compounded by a high level of genetic inter-tumour heterogeneity. Despite international consortia extensively investigating the driver landscape of OAC, a significant proportion of OACs are partially explained by too few cancer driver genes.

Consequently, the aim of this thesis is to complete the driver repertoire of a cohort of 675 OACs by employing a machine learning-based algorithm, sysSVM2, and to investigate the drivers' role in OAC development. sysSVM2 learns about the molecular and systems-level properties (SLPs) of well-known cancer drivers to score and predict new ones in individual patients. Molecular properties define the driver alterations of individual OACs. SLPs describe the genes' evolutionary origin and central role within the cell.

Given that sysSVM2 prioritises drivers in single patients using a scoring system, we used a functional approach to compare two models predicting varying number of drivers per tumour. We estimated that five drivers per sample fully explained the development of OAC. Therefore, we compiled a comprehensive list of cancer drivers for all samples and investigated their frequency and functionality. Newly predicted drivers, some of which were targetable by oncological drugs, were preferentially sample-specific and hit immune- and DNA repair-related pathways. Well-known drivers were recurrent across samples and preferentially perturbed proliferative and invasion pathways.

Finally, we found that OAC clinical stratification is sustained by different pathways that are active in a stage-specific fashion throughout the development of OAC.

Acknowledgement

I would like to thank my first supervisor Prof. Francesca D. Ciccarelli for her guidance and supervision throughout these years. I would also like to thank my second supervisor Prof. Jesper Lagergren, and my thesis committee members Prof. Peter Sasieni, Prof. Peter Van Loo, and Dr. Samra Turajlic for the discussions and feedback that ultimately shaped my research.

I would also like to thank the funding bodies, Cancer Research UK, the Francis Crick Institute, and King's College London, without whom this work would not have been possible.

I would like to thank all of the past and present members of the Ciccarelli lab. Specifically, I would like to personally thank Hrvoje Misetic, Kalum Clayton, Lisa Dressler, Amelia Acha-Sagredo, Lucia Montorsi, and Lorena Benedetti for the constant support and useful opinions they have offered throughout this experience.

I would like to thank my friends and family, who have been – and will always be – there to support me. Finally, I would like to thank SJ for helping me realise that I am responsible for my own fate, and I only have the power to limit myself, or to make myself limitless.

Table of Contents

Abstract 3
Acknowledgement 4
Table of Contents 5
Table of figures 7
List of tables 8
Abbreviations 9
Chapter 1. Introduction 11
1.1 Oesophageal adenocarcinoma 11
1.1.1 The epidemiology of OAC and its origin 11
1.1.2 OAC clinical treatment 13
1.2 OAC cancer driver genes 17
1.2.1 Next-generation sequencing studies in OAC 17
1.2.2 Cancer driver genes involved in the BO-to-OAC progression 20
1.2.3 A debated point: the number of driver events needed per tumour 23
1.3 Beyond cohort-based cancer driver detection methods 25
1.3.1 Limitations of cohort-based approaches 25
1.3.2 Systems-level properties of cancer driver genes 28
1.3.3 sysSVM2: a tool for patient-specific driver predictions 29
1.4 Aim of the thesis 33
Chapter 2. Materials & Methods 35
2.1 The Network of Cancer Genes 35
2.1.1 Literature curation 35
2.1.2 Systems-level properties 36
2.1.3 Pan-cancer cell line data 39
2.1.4 Driver functional annotation 39
2.1.5 Drug interactions 39
2.1.6 Database and website implementation 40
2.2 Clinical and molecular characterisation of the BO and OAC sample cohort 40
2.2.1 Annotation of damaged genes 40
2.2.2 Estimation of WGD 41
2.2.3 Clinical annotation 42
2.2.4 Curation of OAC-specific canonical cancer driver genes 42
2.3 Training of sysSVM2 and evaluation of its predictions 43
2.3.1 Prioritisation of cancer driver genes 43
2.3.2 Performance and stability metrics 47
2.4 Investigating cancer driver genes 48
2.4.1 Deriving the list of cancer driver genes in individual samples 48
2.4.2 Gene set enrichment analysis 49
2.4.3 Comparing the contribution of different sets of drivers to Reactome level 1 pathways 49
2.4.4 Reducing Reactome redundancy 50
Chapter 3. The Network of Cancer Genes 51
3.1 Motivation 51
3.2 Curated annotation of cancer and healthy drivers 52

3.3 SLPs define the central role of cancer and healthy drivers within the cell 54
3.4 Annotation of gene function and interactions with drugs 57
3.5 Conclusion 61
Chapter 4. Resolving OAC genetic inter-tumour heterogeneity 63
4.1 Motivation 63
4.2 Curation of a comprehensive cohort of BO and OAC cases 65
4.3 OAC genetic inter-tumour heterogeneity 70
4.4 OAC-specific training of sysSVM2 74
4.5 Conclusion 79
Chapter 5. Investigating the role and therapeutic vulnerabilities of OAC genetic drivers 82
5.1 Motivation 82
5.2 Drivers identified under the positive selection model perturb novel pathways 83
5.3 Frequency of cancer drivers across OAC samples 87
5.4 OAC-specific canonical drivers and sysSVM2 predictions perturb different processes 90
5.5 Newly discovered targetable drivers 94
5.6 Stage-specific driver perturbations 96
5.7 Conclusion 98
Chapter 6. Discussion 102
6.1 Summary 102
6.2 NCG: a manually curated repository of cancer and healthy driver genes 103
6.3 sysSVM2 optimisation on an OAC-specific setting 106
6.4 Role and clinical relevance of the driver genes contributing to OAC development 108
6.5 Concluding remarks and future trajectories 111
Chapter 7. Appendix 113
7.1 Supplementary figures 113
Reference List 114

Table of figures

Figure 1.1 Clinical progression from BO to OAC.....12
Figure 1.2 Clinical treatment of OAC 16
Figure 1.3 Genetic models of BO-OAC progression 22
Figure 1.4 Overview of sysSVM2 workflow.....30
Figure 3.1 Literature annotation and resulting list of cancer and healthy drivers 53
Figure 3.2 SLPs of cancer and healthy drivers 55
Figure 3.3 Functional and drug annotations of cancer driver genes 58
Figure 4.1 Annotation of damaged genes across sources and clinical stages..67
Figure 4.2 The driver gene landscape of OAC based on current knowledge....72
Figure 4.3 Evaluation of the three sysSVM2 settings.....76
Figure 4.4 sysSVM2 score of different gene categories.....78
Figure 5.1 OAC samples with not enough cancer drivers 84
Figure 5.2 Comparison of the age-incidence and positive selection models.....86
Figure 5.3 Frequency of OAC cancer drivers across samples 88
Figure 5.4 Pathways perturbed by OAC-specific canonical drivers and sysSVM2 predictions 91
Figure 5.5 OAC drivers targeted by oncological drugs.....95
Figure 5.6 OAC clinical stage-specific driver alteration and pathway perturbation 97
Figure 7.1 Pyramidal structure of the Reactome database 113

List of tables

Table 1.1 sysSVM2 features 31
Table 2.1 Prediction tools used to annotate the damaging effect of mutations 37
Table 2.2 Contribution of gene categories to top five drivers 47
Table 3.1 Proportion of enriched pathways per driver category 59
Table 4.1 Median number of damaged genes across cohorts 68
Table 4.2 Training and prediction sets for sysSVM2 under the three settings 74
Table 4.3 Best model parameters 79
Table 5.1 Cancer driver genes targeted by oncological drugs 96

Abbreviations

Abbreviation	Meaning
AJCC	American joint committee on cancer
AUROC	Area under the receiver operating characteristic
BFB	Breakage-fusion bridge
BO	Barrett's oesophagus
bp	Base pair
CCLE	Cancer cell line encyclopaedia
CGC	Cancer Gene Census
CIN	Chromosomal instability
CLP	Cosmic cancer cell line project
CN	Copy number
CNA	Copy number alteration
COSMIC	Catalogue of somatic mutations in cancer
CT	Computerised tomography
DBO	Dysplastic Barrett's oesophagus
EGFR	Epidermal growth factor receptor
ERBB4	Erb-b2 receptor tyrosine kinase 4
FDA	Food and Drug Administration
FDR	False discovery rate
FISH	Fluorescent in situ hybridisation
FPKM	Fragments per kilobase per million
GDSC	Genomics of drug sensitivity in cancer
GNE	Genentech
GoF	Gain-of-function
GORD	Gastro-oesophageal reflux disease
GSEA	Gene set enrichment analysis
HER2	Human epidermal growth factor receptor 2
HGD	High-grade dysplasia
HGF	Hepatocyte growth factor
ICGC	International cancer genome consortium
IGF	Insulin-like growth factor
IGF1R	Insulin-like growth factor 1 receptor
IHC	Immunohistochemistry
Indel	Small insertion or deletion
LOEUF	Loss-of-function observed/expected upper bound fraction
LoF	Loss-of-function
LOH	Loss of heterozygosity

MAPK	Mitogen-activated protein kinase
MET	Mesenchymal epithelial transition factor
MHC-I	Major histocompatibility complex class I
miRNA	microRNA
NCG	Network of Cancer Genes
NF-kB	Nuclear factor kappa B
NGS	Next-generation sequencing
OAC	Oesophageal adenocarcinoma
OCC	One-class classifier
OCCAMS	Oesophageal cancer clinical and molecular stratification
OG	Oncogene
PARP	Poly ADP-ribose polymerase
PD1	Programmed cell death protein 1
PDL1	Programmed death-ligand 1
PET	Positron emission tomography
PI3K	Phosphoinositide 3-kinase
PPIN	Protein-protein interaction network
RBO	Rank-biased overlap
RNAi	RNA interference
RPKM	Reads per kilobase million
RTK	Receptor tyrosine kinase
SCNA	Somatic copy number analysis
SEER	Surveillance Epidemiology and End Results
SLP	Systems-level property
SNV	Single nucleotide variant
SV	Structural variant
SVM	Support vector machine
TCGA	The cancer genome atlas
TGF-β	Transforming growth factor-β
TLR	Toll-like receptor
TNM	Tumour, lymph node, metastasis
TOGA	Trastuzumab for gastric cancer
TPM	Transcript per million
TSG	Tumour suppressor gene
UK	United Kingdom
VEGF	Vascular endothelial growth factor
VEGFR2	Vascular endothelial growth factor receptor 2
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Chapter 1. Introduction

1.1 Oesophageal adenocarcinoma

1.1.1 The epidemiology of OAC and its origin

Oesophageal adenocarcinoma (OAC) is a disease of the distal oesophagus (Smyth et al., 2017). OAC incidence in Western countries has rapidly increased in the last decades (Coleman et al., 2018) exceeding that of oesophageal squamous cell carcinoma (OSCC), which remains the predominant subtype of oesophageal cancer worldwide (Smyth et al., 2017). OAC incidences are the highest in the United Kingdom (UK) and the Netherlands where the age-standardised incidence rate for both countries in the male population was between 10.5-26.5 per 100000 people in 2012 (Rubenstein & Shaheen, 2015). There is a strong male predominance of OAC, with a 6:1 male-to-female ratio in Europe, reaching a 9:1 ratio in North America (Xie & Lagergren, 2016). Such high male predominance suggests that sex hormones may be involved in the development of OAC. A recent investigation has characterised the association between the genetic regulation of sex hormones and the risk of OAC (Xie et al., 2020). The authors found that, based on the predicted effect of genetic mutations on the production of sex hormones, high levels of follicle-stimulating and luteinizing hormones were associated with increased and decreased risk of OAC, respectively.

In addition to gender, a history of gastro-oesophageal reflux disease (GORD), obesity, age and tobacco smoking are other associated risk factors for OAC (Smyth et al., 2017). A 30-year history of GORD is associated with a 6.2-fold increased risk of developing OAC, in comparison tobacco smoking approximately doubles the risk of developing OAC compared with never smokers (Coleman et al., 2018).

Although more than 50% of patients when diagnosed with OAC have no evidence of any precursor lesion (Sawas et al., 2018), OAC is thought to arise from a precancerous lesion known as Barrett's oesophagus (BO) (Contino et al., 2017) (Figure 1.1). In BO the squamous epithelium of the distal oesophagus

is replaced with a crypt-like columnar epithelium resembling that of the intestine. This metaplastic state increases the risk of tumour development. As previously mentioned, one of the major risk factors for OAC is GORD, and BO is likely a reparative response to the damage induced by the refluxate (Peters et al., 2019). The incidence of BO in the general population is around 0.5–2% (Runge et al., 2015) and BO-to-OAC progression increases in the presence of dysplasia. While the annual progression rate of non-dysplastic BO (NDBO) into OAC is 0.1–0.3%, in the presence of dysplasia it increases to at least 10% (Contino et al., 2017).

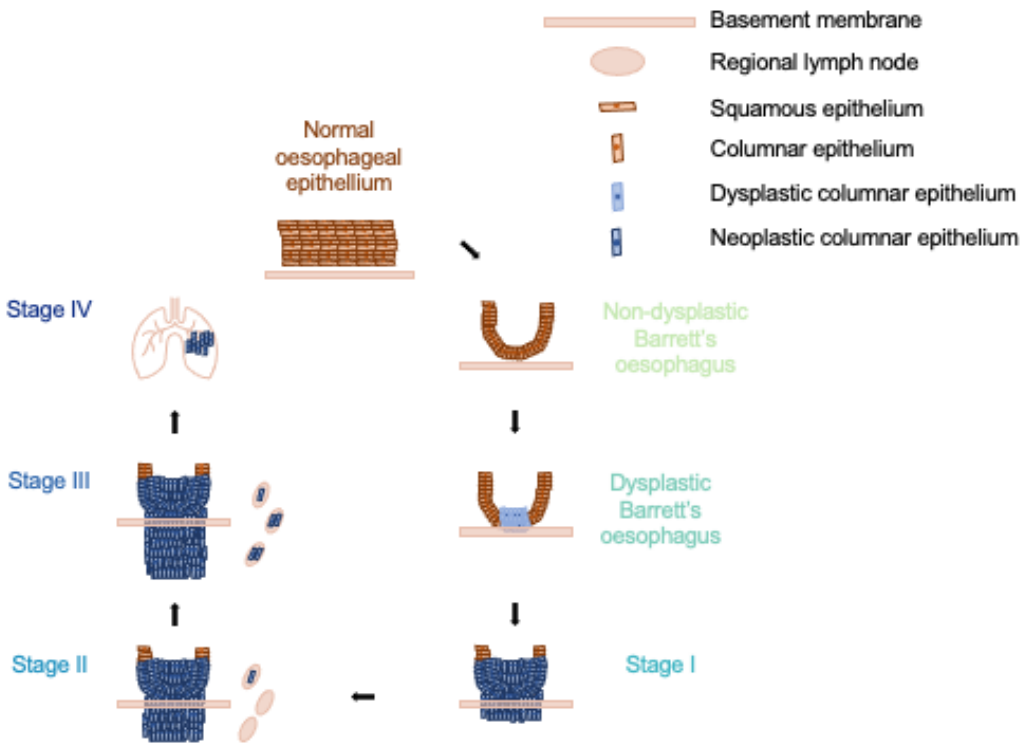

Figure 1.1 Clinical progression from BO to OAC BO clinically progresses to OAC in a stepwise fashion where the non-dysplastic lesion progresses through different grades of dysplasia and transforms into an invasive disease where neoplastic cells have invaded the underlying basement membrane (Stage I). The tumour increases in size and metastasises regional lymph nodes (Stages II and III) and distant metastatic sites (Stage IV). The classification of OAC into clinical stages is based upon the 8th edition of the American Joint Committee on Cancer (AJCC) guidelines for oesophageal and oesophagogastric tumours (Rice et al., 2017).

OAC is thought to originate from BO because of the cell of origin of the tumour. OAC is a disease of the columnar epithelium whereas the resident epithelium of the oesophagus is the squamous type. Genetic studies on matched BO and

OAC samples have identified that up to 80% of mutations overlap between the two samples in the same patient, further confirming the shared origin of BO and OAC (Agrawal et al., 2012; Ross-Innes et al., 2015). This is further supported by a genome-wide methylation analysis that described similar methylation profiles between BO and OAC (Xu et al., 2013).

Some level of debate exists around the cell of origin of BO, hence also of OAC. A trans-differentiation of the oesophageal epithelium has been proposed as the mechanism initiating BO (Minacapelli et al., 2017). Alternatively, gastric cells have been proposed to migrate to the distal oesophagus, seeding the BO lesion (Quante et al., 2012). Finally, transitional basal progenitor or residual embryonic cells at the gastro-oesophageal junction (Jiang et al., 2017; Wang et al., 2011) or the expansion of the oesophageal submucosal gland duct into the BO segment (Owen et al., 2018) have been proposed to originate BO. A recent investigation integrating single-cell transcriptomic profiling and *in silico* lineage tracing using open chromatin, methylation and somatic mutation analyses has shown how BO originated from migrating gastric cardia cells to the distal oesophagus (Nowicki-Osuch et al., 2021). Additionally, the authors showed that all OAC tumours, even those diagnosed without any evidence of BO, shared the same marker expression profile with undifferentiated BO confirming the unique BO origin of OAC (Nowicki-Osuch et al., 2021).

1.1.2 OAC clinical treatment

In the UK, 70–80% of OAC patients are diagnosed with an advanced-stage tumour, when the lesion has become large and local (lymph node) or distant (other organs) metastases are already present (source: https://www.cancerresearchuk.org). This has a profound impact on the success of OAC treatment. OAC treatment is determined by evaluating the clinical staging of the tumour, which is usually done by following the American Joint Committee on Cancer (AJCC) staging guidelines (Rice et al., 2017). The AJCC system reports in standardised terms the size and depth of invasion of the tumour mass. It describes the size of the primary tumour (T), the number of

lymph nodes invaded by neoplastic cells (N), and the presence of distant metastasis (M). T varies from one to four, with higher scores representing a tumour that has invaded structures adjacent to the original epithelium. N is scored on a scale of zero to three with the maximum score representing evidence of at least seven lymph nodes invaded by neoplastic cells. The presence or absence of distant metastasis (M) is determined by a binary score of one or zero (Rice et al., 2017).

Clinical staging is based upon an endoscopic examination and biopsy of the tumour for histological confirmation (Figure 1.2). Additionally, computerised tomography (CT) and positron emission tomography (PET) scans are carried out to exclude the presence of lymph node and distant metastases, respectively (Smyth et al., 2020). Only 19% of patients are treated with curatively-intended surgery as primary care. The remaining patients are treated, often palliatively, with radiotherapy or chemotherapy. The treatment is chosen based on the results of the clinical assessment. Early-stage tumours (T1-2 N0 M0) are usually resected either endoscopically or surgically. Locally advanced non-metastatic tumours (T3-4 N1-3 M0) are usually treated with sequential lines of neoadjuvant chemotherapy to reduce the tumour mass. Resection is only performed when the presence of distant metastases in the patient can be excluded (M0). Advanced metastatic tumours (T1-4 N0-3 M1) are usually treated with multiple lines of chemotherapy that can be coupled with targeted or immunotherapy to improve patient survival. Clinically approved targeted and immunotherapies for OAC include trastuzumab, ramucirumab, nivolumab, and pembrolizumab (Smyth et al., 2020).

Trastuzumab is an anti-human epidermal growth factor receptor 2 (HER2) monoclonal antibody. *HER2* encodes the receptor tyrosine kinase (RTK) ErbB-2 that is part of the epidermal growth factor receptor (EGFR) pathway. The involvement of ErbB receptors along with their ligands has been extensively reported in human cancers (Normanno et al., 2006). Given the central role of the EGFR pathway in multiple cellular signal transduction pathways, the ligand-receptor network is involved in tumour pathogenesis and progression by promoting tumour growth and survival. Trastuzumab is suitable for patients

whose tumour shows amplification and overexpression of *HER2*. In the phase III Trastuzumab for Gastric Cancer (TOGA) trial patients were evaluated for HER2 overexpression by means of immunohistochemistry (IHC) and fluorescence *in situ* hybridisation (FISH). The median overall survival was 13.8 months in patients treated with trastuzumab combined with chemotherapy compared with 11.1 months in patients treated with chemotherapy alone (Bang et al., 2010). Trastuzumab is currently used as first line treatment in *HER2*-positive advanced metastatic tumours in addition to chemotherapy (Figure 1.2).
Ramucirumab is an anti-vascular endothelial growth factor receptor 2 (VEGFR2) monoclonal antibody. Vascular endothelial growth factor (VEGF) is a key mediator of angiogenesis in cancer and binds two receptors: VEGFR1 and VEGFR2 (Carmeliet, 2005). The vasculature grows around the tumour to provide nutrients, and thus promoting sustained tumour growth. In addition, the VEGF pathway is involved in invasion and metastasis. The phase III RAINBOW trial evaluated the effect of ramucirumab in a cohort of patients with gastric and gastro-oesophageal junction adenocarcinoma. The investigators found that the median overall survival was 9.6 months in patients treated with ramucirumab combined with chemotherapy compared with 7.4 months in patients treated with chemotherapy alone (Wilke et al., 2014). Ramucirumab is currently used as second line treatment in combination with chemotherapy in advanced metastatic tumours (Figure 1.2).
Nivolumab and pembrolizumab are anti-programmed cell death protein 1 (PD1) monoclonal antibodies. The immune-checkpoint PD1 plays a central role in the inhibition of the immune system via modulating the activity of T cells (Han et al., 2020). The binding of PD1 to its ligand programmed death-ligand 1 (PDL1) results in T cell inactivation and mechanisms of immune escape. Consequently, the blockade of PD1 or PDL1 results in the release of mechanisms for anti-tumour response. This inhibition has proven successful in the treatment of different solid tumours and haematological malignancies (Han et al., 2020). In the phase II KEYNOTE-059 study patients with gastric and gastro-oesophageal junction adenocarcinoma treated with pembrolizumab showed a median response duration of 8.4 months supporting further investigation into this drug

for advanced tumours (Fuchs et al., 2018). These two immunotherapy drugs are currently used in the clinic for patients whose advanced tumours have progressed after two or more lines of treatment and are thus defined as chemo-refractory (Figure 1.2).

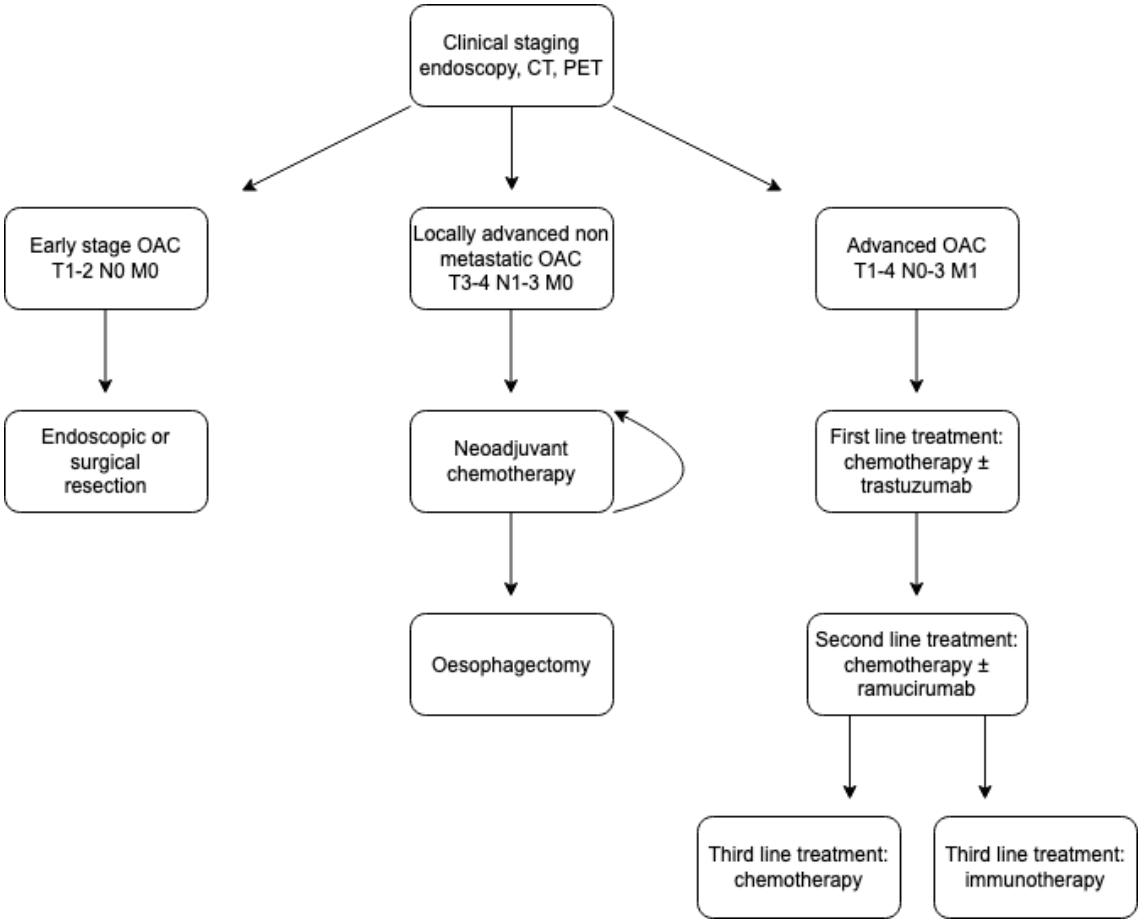

Figure 1.2 Clinical treatment of OAC

The staging of the tumour is usually done at diagnosis, termed clinical staging. This step entails an endoscopy of the primary tumour along with CT and PET scans to investigate the presence of lymph node or distant metastases. Tumours whose size is limited (T1-2) and show no evidence of lymph nodes (N0) and distant (M0) metastases are usually resected without any neoadjuvant treatment. Locally advanced non-metastatic tumours are larger in size (T3-4) and show evidence of lymph node metastases (N1-3). These tumours are treated with one or more lines of neoadjuvant chemotherapy in order to reduce the tumour size. If the neoadjuvant treatment is successful and there is no evidence of distant metastases (M0), the surgery is done. Advanced tumours where the tumour has metastasised (M1) are treated with multiple lines of chemo-treatment combined with targeted or immunotherapy if the corresponding biomarkers are present. Figure adapted from (Smyth et al., 2017, 2020).

The above-mentioned drugs in combination with chemotherapy and radiotherapy represent the current standard of care for OAC treatment. However, other targeted therapies have been tested over the years in the context of OAC. Targeting other components of the EGFR pathway, such as HER3 and HER4, via monoclonal antibodies and small molecule tyrosine kinase inhibitors has not proven beneficial in the clinical setting (Woo et al., 2015).
Similarly, the inhibition of the mesenchymal epithelial transition factor (MET) receptor, which binds the hepatocyte growth factor (HGF) and is associated with proliferation and invasion, has not shown any increase in patient survival (Young & Chau, 2016). In addition, *FGFR2* controls mitogenesis and differentiation. Although it is amplified in 5-10% of advanced gastric cancers, the inhibition of it has not shown any impact on improving patient progression-free survival (Bang et al., 2015).
Despite all efforts to improve the clinical treatment of OAC, the overall five-year survival rate for this cancer remains relatively poor, between 15% and 20% (Ferlay et al., 2015).

1.2 OAC cancer driver genes

1.2.1 Next-generation sequencing studies in OAC

The advancement of the field of targeted therapies relies heavily on molecular studies to elucidate the genetic mechanisms responsible for OAC development. Recent years have seen a surge in cancer genomics studies aimed at the identification of genes that, upon acquiring somatic alterations, contribute to the growth of the tumour, thus termed cancer driver genes. This surge has resulted in the sequencing of tens of thousands of tumours (Dressler et al., 2022). Further, decreased sequencing costs have created a large pool of cancer genome data available for the investigation of cancer driver genes.
Specifically, due to the increase of OAC incidence in Western countries, most of the recent analyses and resulting literature represent the foundation stone for the clinical treatment of OAC. The first study analysing the genome of a small cohort

of OAC patients (e.g., n=11) using next-generation sequencing (NGS) was published in 2012 (Agrawal et al., 2012). In this seminal paper the authors compared the recurrently altered genetic coding drivers in OAC and OSCC. They identified differences by means of different frequencies of altered genes: OSCC was characterised by frequent coding driver alterations in *TP53*, *NOTCH1* and *NOTCH3* (>20% of samples), whereas OAC was characterised by frequent driver alterations in *TP53* alone (>70% of samples) (Agrawal et al., 2012). Recent screenings have focused on larger cohorts and on the integration of different data types such as somatic copy number analysis (SCNA), DNA methylation, and mRNA expression. They showed that the molecular phenotype of OAC more closely resembles that of gastric cancer than that of OSCC (Kim et al., 2017; Liu et al., 2018). Based on the integration of multiple sources of molecular data, the authors argued against clinical trials combining the two subtypes of oesophageal cancers (Kim et al., 2017), supporting instead the grouping of OAC with gastric cancers for clinical investigations on neoadjuvant, adjuvant and systemic therapies (Smyth et al., 2020).

Both OAC and its precursor lesion, BO, have a high mutational frequency comparable to that of cancer types known to be caused by environmental mutagens, such as lung and melanoma (Dulak et al., 2013; Stachler et al., 2015). This raises the question as to whether OAC can be also attributed to environmental stimuli. Different groups working with different OAC cohorts have observed a common mutational signature that comprises T>G substitutions in a CTT context (Dulak et al., 2013; Secrier et al., 2016; Stachler et al., 2015). Though the signature is of unknown aetiology, it has been suggested to result from the exposure to bile acids that continuously invade the distal oesophagus in patients with GORD (Contino et al., 2017).

Most of the mutations identified by mutational screenings on OAC resulted in a loss-of-function (LoF) effect on the genes harbouring them (Agrawal et al., 2012; Dulak et al., 2013). Genes that drive tumour progression due to their lost or inactivation are commonly defined as tumour suppressor genes (TSGs) (Vogelstein et al., 2013). Under normal conditions TSGs control cell proliferation by arresting the cell cycle and inducing cell death upon DNA damage. In contrast,

oncogenes (OGs) promote neoplastic growth when activated or present in multiple copies within the cell (Vogelstein et al., 2013).

Shortly after the first seminal studies on mutational coding drivers (Agrawal et al., 2012; Dulak et al., 2013), it became evident that the major drivers of OAC development were structural rearrangements (Nones et al., 2014). Nones and colleagues observed that 33% of OACs showed evidence of catastrophic chromosomal events. Evidence of large and clustered genomic rearrangements (chromothripsis) and breakage-fusion-bridges (BFBs) were among the most recurrent events of chromosomal aberrations in OAC (Nones et al., 2014). Such events have been proposed to underlie mechanisms of oncogenic activation given the overlap of these aberrant regions with potent OGs. The acquisition of such catastrophic events might also explain why BO patients under routine surveillance quickly develop the tumour without linearly progressing through the clinical stages of OAC development.

Further genomic instability in OAC is acquired through events of whole genome doubling (WGD), during which the entire genome duplicates resulting in twice the number of chromosomes of a diploid cell. Evidence of at least one event of WGD was observed in up to 62.5% of sequenced OAC cases, confirming the role of different forms of chromosomal instability (CIN) in the development of OAC (Stachler et al., 2015).

Recent whole genome sequencing (WGS) and whole exome sequencing (WES) studies have investigated the frequency of cancer driver genes in large cohorts of samples (Frankell et al., 2019; Liu et al., 2018; Secrier et al., 2016). The largest of these comprised 551 OAC cases (Frankell et al., 2019). These studies confirmed that the most frequent genetic alterations (via mutations or copy number alterations, CNAs) affected *TP53* in more than 70% of samples. The second most frequently altered gene (via mutations or CNAs) was *CDKN2A* with a frequency of less than 20%. Interestingly, *CDKN2A* was found silenced through epigenetic mechanisms in 75% of OACs (Liu et al., 2018), suggesting the importance of non-genetic mechanisms to explain the presence of OAC.

Developing tailored strategies based on the pool of alterations present in OAC is challenging given the heterogeneous nature of the disease. Most of the tools

used so far for driver detection mainly rely on the identification of genes that are frequently altered within sample cohorts. Such an approach is limited by the driver heterogeneity present in OAC.

The application of a machine learning tool, developed by the Ciccarelli lab, for patient-specific driver predictions to a cohort of 261 OACs allowed for the first time to complete the driver repertoire of all the OAC samples under investigation (Mourikis et al., 2019). This resulted in almost 1000 genes predicted as drivers across the sample cohort with the majority of them being rare or patient-specific. Mourikis and colleagues showed how tools for patient-specific predictions can help filling in the gap and understanding the specific evolutionary history of individual OACs. Identifying the genetic drivers responsible for tumour development is the first step towards improving the poor patient survival that is pervasive of this cancer type.

1.2.2 Cancer driver genes involved in the BO-to-OAC progression

Given the low progression rate of BO to OAC (see Chapter 1.1.1), a key aim of current research is to identify biomarkers of progression able to predict the group of BO patients that are most at risk of progressing to OAC, thus requiring additional surveillance. This will shed light on the mechanisms responsible for OAC initiation and open possibilities for early intervention in the treatment of OAC. In this context, some studies have investigated the driver landscape of both BO and OAC in matched and unmatched cases (Ross-Innes et al., 2015; Stachler et al., 2015; Weaver et al., 2014).

Weaver and co-authors studied a cohort of samples composed of 112 OACs, 66 NDBO cases, and 43 patients with high-grade dysplasia (HGD, the stage preceding the development of OAC) (Weaver et al., 2014). They found that the majority of recurrently mutated genes in OAC were also mutated in NDBO and HGD cases at similar frequencies. However, two such genes acquired mutations in a stepwise manner and defined stage boundaries of the disease. They found *TP53* mutated in 72% of HGD cases and 69% of OACs but only 2.5% of NDBO samples. *SMAD4* was altered, although at a low frequency (13%), only in OACs.

Although informative, the study did not investigate matched BO and OAC cases and, given the low progression rate of BO to OAC, it is difficult to understand which characteristics apply to progressors and which apply to non-progressor BO patients (Weaver et al., 2014).

To better address the study of the clonal organisation of progressor BO and the genetic mechanisms underlying malignant transformation, two studies focused on matched BO and OAC pairs (Ross-Innes et al., 2015; Stachler et al., 2015). In 57% of the cases, fewer than 20% of the single nucleotide variants (SNVs) overlapped between BO and OAC pairs, with dysplastic cases sharing more SNVs with the corresponding OAC than NDBO cases. Such little overlap might be the result of the long BO evolutionary history before seeding the tumour. An in-depth spatiotemporal study of a single BO patient showed that BO originated from an individual founder cell. The initiating cell quickly gave rise to multiple clones within the BO segment with varying ability to expand and seed further clones (Ross-Innes et al., 2015).

Previous analysis on BO already suggested it to be a multiclonal condition (Galipeau et al., 1999). In BO, clones with either one or both of 9p (harbouring *CDKN2A*) and 17p (harbouring *TP53*) loss of heterozygosity (LOH) would expand and invade a large proportion of the oesophagus. This seminal analysis also showed that 9p LOH usually occurred before 17p LOH (Galipeau et al., 1999). Further analyses identified a sequential acquisition of driver alterations whereby *CDKN2A* inactivation occurred in the early stages of BO and the clone bearing such alterations usually expanded and invaded the whole segment (Maley et al., 2004). The early inactivation of *CDKN2A* was supported by the evidence that *CDKN2A* alterations underwent fixation more frequently than any other alteration in BO. For example, *TP53* modifications occurred later in the evolutionary history of the disease and they required an existing *CDKN2A* mutant clone to expand on (Maley et al., 2004). In a later paper the same group identified *TP53* as the gene responsible for BO-to-OAC progression (Maley et al., 2006). *TP53* favoured progression through genomic instability and patients with a more unstable and clonally diverse BO were more likely to progress to OAC (Maley et al., 2006). Based on these observations, the very first model of

BO progressing to OAC postulated that BO acquired alterations in *CDKN2A* via LOH, point mutations or changes in the methylation status. These were followed by alterations in *TP53,* via LOH or point mutations, whereas aneuploidy marked the final malignant transformation (Maley, 2007) (Figure 1.3, Model 1). This mechanism was additionally supported by the observation that alterations in *TP53* and *CDKN2A* gene loci were common in both BO and OAC, suggesting early acquisition of such mutations in the evolutionary history of the disease even before the tumour developed (Barrett et al., 1999).

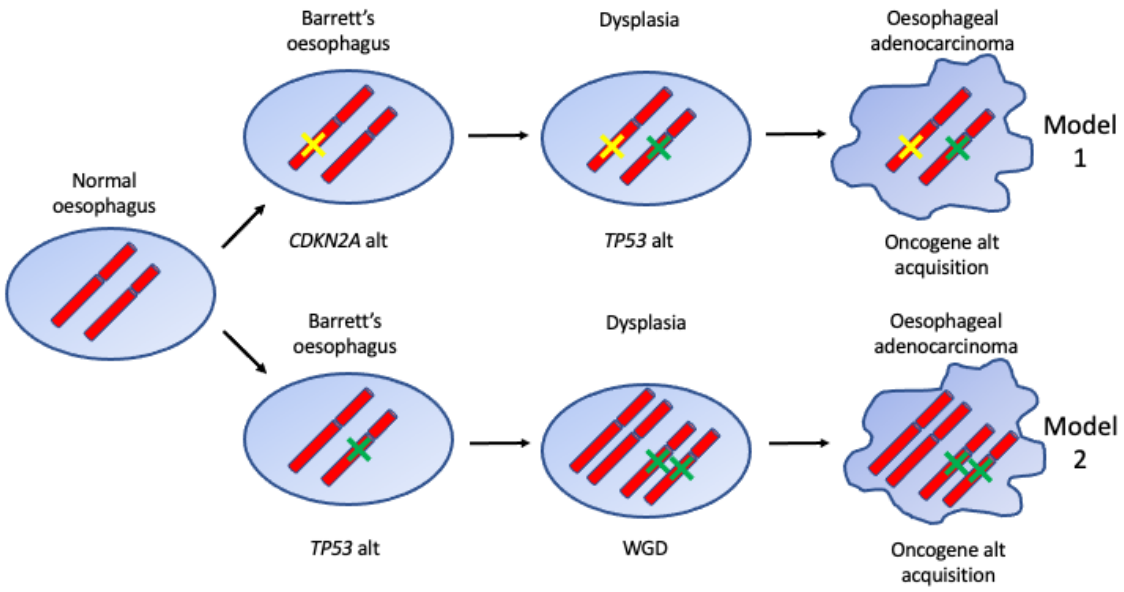

Figure 1.3 Genetic models of BO-OAC progression

The first model postulates that alterations in *CDKN2A* are acquired very early on, usually at the level of NDBO. When dysplasia occurs, the lesion usually acquires alterations in *TP53*. Finally, the cell becomes malignant after acquiring additional aneuploidy (Maley, 2007; Maley et al., 2004; Maley et al., 2004). The second model postulates that alterations in *TP53* are acquired when BO develops before the appearance of dysplasia. When dysplasia develops the cell usually doubles its genetic material (WGD). Finally, the tumour develops by acquiring additional alterations affecting mainly oncogenes (Stachler et al., 2015). WGD: whole-genome doubling, alt: alteration.

Stachler and colleagues presented evidence in OAC of a fast-track mechanism for tumour development (Stachler et al., 2015). According to their observation, *TP53* acquired driver alterations in the early, pre-dysplastic stages of BO. These alterations were then followed by a WGD event at the level of dysplastic BO

(DBO) (Stachler et al., 2015). The final malignant conversion was characterised by increased aneuploidy through the acquisition of alterations in oncogenes mainly via gene amplifications (Stachler et al., 2015) (Figure 1.3, Model 2). Although the *TP53* and WGD model was most frequent, the authors also found evidence of the *CDKN2A* and *TP53* model in their cohort of matched cases (Stachler et al., 2015). Their finding argued against the evidence that the acquisition of mutations in *TP53* marked the boundary between NDBO and DBO (Weaver et al., 2014).

The investigation of the driver events responsible for BO-to-OAC progression can help inform strategies for the early detection and treatment of patients who are more likely to progress. However, given the heterogeneous driver landscape of OAC, it is tempting to speculate that these two proposed models are not sufficient to explain all OAC cases and further investigation into these mechanisms is likely needed in the future.

1.2.3 A debated point: the number of driver events needed per tumour

A long-standing question in the cancer genomics field concerns the number of driver genes needed for malignant transformation. This number is highly heterogeneous and varies across cancer types, primary sites, and theoretical approaches used to derive it.

The first attempt to calculate the number of drivers predated the advent of cancer genomics and relied on cancer age incidence data (Armitage & Doll, 1954). The approach was based on the evidence that the incidence of cancer in the population increased with age. The incidence of common cancers, such as colorectal adenocarcinoma, breast carcinoma, and pancreatic ductal adenocarcinoma, increased to the power of four to six as a function of age. This led to the hypothesis that cancer could be the result of four to six rate-limiting steps, namely driver events, that accumulated randomly at a constant rate throughout life (Armitage & Doll, 1954). The fundamental idea behind this approach is to use mathematical approximation to fit the observed trend of tumour incidence.

This method has been applied to multiple cancer types (Martincorena & Campbell, 2015; Tomasetti et al., 2015), including OAC where it predicted the occurrence of three drivers per sample (Jeon et al., 2006). The authors obtained age-specific incidence data for 4483 white males and 746 white females diagnosed with OAC between 1973 and 2000 from the Surveillance Epidemiology and End Results (SEER) registry in the United States. The most parsimonious model fitting the data suggested that OAC developed from an initial step of tissue conversion (BO) followed by a multi-stage process characterised by three rate-limiting steps. According to the model, the first two steps resulted in an initiated cell that expanded clonally into a premalignant lesion followed by a final conversion to malignancy (Jeon et al., 2006). The mathematical model of OAC development well overlapped with the evidence available at the time according to which in BO there was evidence of clonal expansion (Maley et al., 2004) and of a sequential acquisition of genetic changes that resulted in tumour formation (Maley, 2007) (Figure 1.3, Model 1).

More recently, the number of drivers has been measured as the number of genes whose mutations are under positive selection because of the selective advantage they provide to tumour cells (Martincorena et al., 2017). In the case of OAC, these resulted in a median of five driver events per sample. The authors applied dNdScv, a method widely used to predict driver genes, on a pan-cancer cohort of 7664 samples from 29 cancer types including OAC. The tool is based on the rationale that genes could accumulate mutations under positive, neutral or negative selection. These different evolutionary trajectories are obtained by calculating the rate of non-synonymous to synonymous mutations across the coding region of the genome.

Genes under neutral selection represent the vast majority of the human coding genes (97-98%). They represent those genes with a mutation rate comparable to that expected by chance, the background mutation rate. Genes under positive selection represent 1-3.9% of all the human coding genes and they accumulate an excess of mutations that are likely to act as drivers in cancer patients. Finally, genes under negative selection represent a small portion of the human coding genome (0.02-0.5%) and are the pool of genes in which mutations are selected

against since their mutation rate is below the background mutation rate (Martincorena et al., 2017).

In the context of oesophageal cancer, the authors found that five drivers are needed per tumour (Martincorena et al., 2017). One limitation of this study was that, due to the low number of samples, OSCC and OAC were treated as a single homogenous group. Interestingly, the study showed that half of the driver events occurred in yet-to-be-discovered cancer driver genes (Martincorena et al., 2017).

Both the age-incidence and the positive selection method assume a constant number of drivers throughout the evolutionary history of an individual cancer. For example, Martincorena and colleagues compared the number of driver genes between early-stage and late-stage tumours and found no statistically significant difference at the pan-cancer level (Martincorena et al., 2017). This is in line with the hypothesis according to which the transformed cell already has the capacity to metastasise and no additional drivers must be acquired to gain the invasive phenotype after the malignant transformation (Vogelstein & Kinzler, 2015). However, genomic instability increases as OAC progresses (Newell et al., 2019; Nones et al., 2014) and some of these newly acquired mutations might have functional implications in the progression of the disease.

Based on the literature and as outlined above, two models explain the development of OAC in terms of the number of driver genes (Jeon et al., 2006; Martincorena et al., 2017); however, evidence shows that there is heterogeneity regarding the definitive number of cancer driver genes needed to explain the presence of the disease.

1.3 Beyond cohort-based cancer driver detection methods

1.3.1 Limitations of cohort-based approaches

The identification of cancer driver genes along with the number of drivers required in individual tumours enables the unravelling of the processes that sustain tumour initiation and growth. This mechanistic understanding can then be exploited to

develop targeted therapies based on the genetic profiling of cancer patients. The main challenge for identifying driver genes is to differentiate between genes that harbour driver mutations and genes acquiring alterations that do not contribute to the disease (passenger alterations).

Most current methods for driver identification rely on the recurrence of mutations across patient cohorts. The rationale is to identify the mutated genes that confer selective advantage to tumour cells, causing their mutations to be selected for and fixed across patients. Such tools inevitably rely on the size of the cohort under study. The larger the cohort, the higher the number of driver genes identified and the more comprehensive our understanding of cancer genes becomes (Repana et al., 2019). Some tools, such as MutSigCV (Lawrence et al., 2013) and MuSiC (Dees et al., 2012), focus on genes whose mutation rate is above the background mutation rate in the cohort under investigation. dNdScv (Martincorena et al., 2017) identifies driver genes by deriving the ratio of non-synonymous to synonymous mutations in order to characterise genes that are more likely to contribute to cancer given the accumulation of excess mutations (see Chapter 1.2.3). Such excess mutations are more likely to be drivers given their selection across multiple samples. OncodriveCLUST (Tamborero et al., 2013a) identifies mutations that cluster in specific amino acids, while ActiveDriver (Reimand & Bader, 2013) and OncodriveFM (Gonzalez-Perez & Lopez-Bigas, 2012) identify mutations with a functional impact on the encoded protein.

For the identification of CNAs, GISTIC2 is the most widely used tool for driver detection (Mermel et al., 2011). However, the tool identifies recurrently deleted or amplified regions with little information as to which gene within the region is acting as driver.

Despite all these efforts, even when several methods for driver detection are combined (Tamborero et al., 2013b; Bailey et al., 2018), a sizeable number of cancer patients are still left with very few driver genes. In a screening of 3205 tumours from 12 different cancer types the authors applied five tools for driver identification based on the recurrence and functional impact of mutations. Despite identifying 291 cancer driver genes overall, almost 35% of the samples had less than three drivers per tumour (Tamborero et al., 2013b). The high number of

patients with too few cancer driver genes is particularly evident for cancer types that have a high degree of inter-tumour heterogeneity, such as OAC (Frankell et al., 2019).

The limitations of cohort-based approaches have already been encountered and acknowledged in the context of OAC (Dulak et al., 2013; Weaver et al., 2014). Dulak and co-authors applied MutSigCV on a cohort of 149 samples and identified 26 significantly mutated (false discovery rate (FDR) <0.1) genes. However, they rescued 22 additional genes whose mutation frequency did not reach statistical significance but whose involvement in cancer had been extensively reported (Dulak et al., 2013). The inclusion of biological knowledge in order to compile a comprehensive list of driver genes has also been applied in other studies (Weaver et al., 2014). Such an approach, however valid, is limited by the current knowledge on cancer drivers and does not allow the discovery of new genes previously unrelated to cancer.

One further limitation of cohort-based tools is their bias to identify frequently mutated driver genes, while neglecting rare driver events that are active in individual patients. To overcome this limitation, other tools aim to detect driver genes at the patient level rather than at the cohort level. This is more challenging because it is not possible to rely on an underlying population of tumours to identify signals of positive selection but predictions are performed on the individual tumour sample. DriverNet (Bashashati et al., 2012) and OncoIMPACT (Bertrand et al., 2015) combine several types of omics data to define perturbed networks of gene expression from which to infer driver genes at the patient level. PHIAL (Van Allen et al., 2014) and iCAGES (Dong et al., 2016), in addition to characterising patient-specific driver genes, return a clinical interpretation of somatic variants. sysSVM, developed and maintained by the Ciccarelli lab, predicts cancer driver genes in individual patients (Mourikis et al., 2019). The tool is based on supervised machine learning and the rationale behind it is that somatic alterations promoting cancer affect genes with distinct properties (D'Antonio & Ciccarelli, 2013). Such properties are a combination of molecular and systems-level properties (SLPs). Molecular properties include somatic alterations that individual tumours acquire throughout their evolutionary history.

SLPs are evolutionary, genomic, gene expression, and network properties that describe the cellular role and evolutionary origin of human genes. They include gene duplicability (Rambaldi et al., 2008), gene evolutionary origin (Syed et al., 2010), breadth of gene and protein expression in human tissues (An et al., 2016), gene essentiality (Repana et al., 2019), connections and position of the protein in the protein-protein interaction network (PPIN), and number of associated regulatory microRNAs (miRNAs) (D'Antonio et al., 2012).

1.3.2 Systems-level properties of cancer driver genes

Cancer sequencing screens predict cancer driver genes based on statistical frameworks (Raphael et al., 2014). Given the variability in cancer driver detection methods and the corresponding output it is important to maintain an up-to-date and consistent overview of driver genes involved in cancer. To this end, the Network of Cancer Genes (NCG) was first established in 2010 (Syed et al., 2010). NCG is a resource developed and curated by the Ciccarelli lab and available to the community online at http://network-cancer-genes.org/. The goal of NCG is to maintain an up-to-date comprehensive list of cancer driver genes from the literature and to annotate their SLPs.

Since cancer driver genes show a distinct SLP profile that is different from that of the rest of human genes (Repana et al., 2019), this characteristic was later exploited to develop a patient-specific cancer driver prediction tool (Mourikis et al., 2019).

Specifically, cancer genes are more present as singletons in the human genome (Rambaldi et al., 2008) and originate earlier in evolution (Syed et al., 2010) compared to the rest of human genes. The central role of cancer genes across tissues is exemplified by their ubiquitous expression both at the gene (An et al., 2016) and protein (Repana et al., 2019) level. Cancer genes are targeted by many more miRNAs than the rest of human genes (D'Antonio et al., 2012). They encode proteins that are highly-interconnected hubs in the PPIN (D'Antonio et al., 2012) and that are involved in more complexes than the rest of human genes

(An et al., 2016). Finally, cancer genes are essential in a higher fraction of human cell lines than the rest of human genes (Repana et al., 2019).
SLPs also highlight different cellular roles and evolutionary paths of TSGs and OGs. TSGs are older, less duplicated, more ubiquitously expressed, and more essential than OGs (Repana et al., 2019). These traits underlie the specific role of these two sets of genes in the cell. TSGs are involved in the maintenance of basic cellular functions such as cell cycle and gene expression. OGs, on the other hand, are preferentially involved in regulatory functions, which tend to be more tissue-specific and less ubiquitous, such as signal transduction and the control of the immune system (Dressler et al., 2022).
Overall, NCG provides the cancer research community with the crucial collection of an up-to-date list of cancer driver genes and builds the foundation of the cancer prediction tool, sysSVM (Mourikis et al., 2019), and its implementation, sysSVM2 (Nulsen et al., 2021).

1.3.3 sysSVM2: a tool for patient-specific driver predictions

sysSVM2, the recent implementation of sysSVM (Mourikis et al., 2019), is a tool for the identification of driver genes at the individual patient level (Nulsen et al., 2021). The tool is based on a supervised machine learning algorithm that scores genes based on their molecular and SLPs.
sysSVM2 prioritises genes with features similar to those of genes that have been experimentally validated to be involved in tumorigenesis (namely, canonical cancer driver genes). Canonical drivers differ from the rest of human genes by SLPs that define these genes as a group as previously described (see Chapter 1.3.2). Additionally, canonical drivers are described using molecular properties. Molecular properties are defined as mutations and CNAs, affecting the gene expression or impacting on the protein functionality (hence termed damaging), that occur in the individual cancer sample.
sysSVM2 leverages these two sets of properties to rank damaged genes in individual cancer patients. The more similar the gene properties are to those of canonical drivers, the higher rank the gene will be assigned. Highly ranked genes,

given their similar property profile to that of canonical drivers, will then be considered the cancer drivers for that patient.

The sysSVM2 algorithm uses a one-class classifier (OCC) that consists of four support vector machines (SVMs). An OCC has been selected as the most appropriate classification system due to limitations in the current knowledge of cancer driver genes (Nulsen et al., 2021). While it is possible to confidently define the true set of cancer driver genes (Dressler et al., 2022; Saito et al., 2020; Sondka et al., 2018; Vogelstein et al., 2013), it is more challenging to describe and identify non-cancer genes, i.e., genes that categorically do not contribute to cancer.

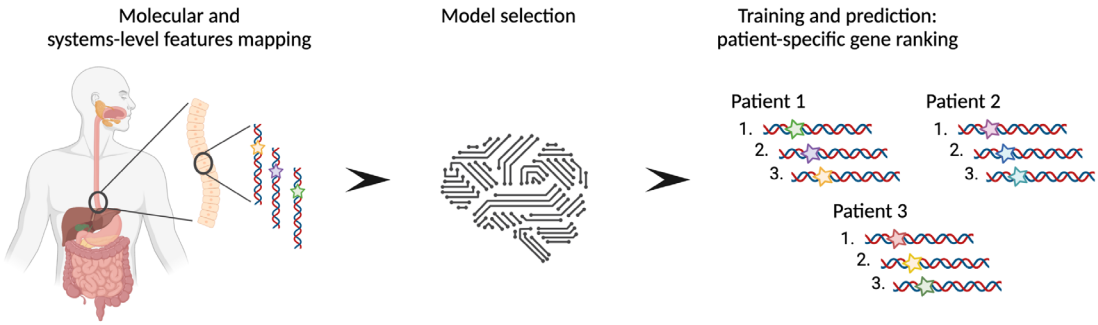

Figure 1.4 Overview of sysSVM2 workflow

The three steps that constitute sysSVM2 pipeline and its final output are shown: in the first 'feature mapping' step, molecular and systems-level features are mapped to each damaged gene in each patient. The second 'model selection' step consists of multiple iterations to tune the four SVMs' parameters. After defining the parameter values the tool performs the actual 'training and prediction' phase and returns a ranked list of patient-specific predictions. The figure was created using BioRender.

The four SVMs are trained on canonical cancer driver genes. sysSVM2 returns a list of damaged genes ranked in each patient accordingly to their similarity to the canonical drivers used for training. Specifically, sysSVM2 consists of three steps (Figure 1.4):

- Feature mapping;
- Model selection;
- Training and prediction.

In the first step, seven molecular and 19 systems-level features (Table 1.1) are mapped to all the damaged genes in individual patients. The genes are then

divided into training and prediction datasets. The training set consists of canonical drivers damaged in the cohort under study that will be used to optimise the parameters of the four SVMs. This is done through iterations of cross-validation. The training set is randomly split at every iteration, and 2/3 of the genes are used to train the models with the remaining 1/3 used as a test set. The choice of the optimal model parameters is based on the sensitivity to retrieve canonical cancer drivers calculated at every iteration. Predictions are performed on the group of genes used as test set and the sensitivity to predict canonical driver genes is computed for each combination of parameters. The best models for the four SVMs are those with the highest sensitivity to retrieve canonical drivers and the lowest variance across multiple iterations (Nulsen et al., 2021).

Table 1.1 sysSVM2 features

The 26 features used by sysSVM2 for prioritising cancer driver genes are listed. For each feature the corresponding molecular or systems-level property, the type (whether it is binary or continuous) and the category are reported.

Property	Feature	Feature type	Category
Mutation	Exonic mutations (n)	Continuous	Molecular
Mutation	Non-truncating damaging mutations (n)	Continuous	Molecular
Mutation	Truncating mutations (n)	Continuous	Molecular
Mutation	GoF mutations (n)	Continuous	Molecular
CNA	Gene copy number (n)	Continuous	Molecular
CNA	Gene amplification	Binary	Molecular
CNA	Gene deletion	Binary	Molecular
Conservation	Pre-metazoan origin	Binary	Systems-level
Conservation	Metazoan origin	Binary	Systems-level
Conservation	Vertebrate origin	Binary	Systems-level
Conservation	Post-vertebrate origin	Binary	Systems-level
Duplication	Gene duplication	Binary	Systems-level
Duplication	Gene ohnolog	Binary	Systems-level
Expression	Tissues expressing gene (n)	Continuous	Systems-level
Expression	Gene expressed in 0 tissues	Binary	Systems-level
Expression	Gene expressed in 7≤tissues≤36	Binary	Systems-level
Expression	Tissues expressing protein (n)	Continuous	Systems-level
Expression	Protein expressed in ≥41 tissues	Binary	Systems-level

Protein interactions	Complexes the protein is part of (n)	Continuous	Systems-level
Protein interactions	PPIN degree	Continuous	Systems-level
Protein interactions	Protein: hub in PPIN	Binary	Systems-level
Protein interactions	PPIN betweenness	Continuous	Systems-level
Protein interactions	PPIN clustering coefficient	Continuous	Systems-level
miRNA interactions	miRNAs targeting the gene (n)	Continuous	Systems-level
Essentiality	Cell lines in which the gene is essential (%)	Continuous	Systems-level
Essentiality	Essential gene	Binary	Systems-level

Once the best models are identified, the whole training set is used to train the four SVMs and the trained models are finally employed for predictions. A score that combines the predictions from the four SVMs is computed and used to rank genes in single samples (Figure 1.4). In each sample the genes that more closely resemble canonical cancer drivers used for training the model will rank higher and are therefore more likely to contribute to cancer in the corresponding sample. One of the strengths of sysSVM2 is that it prioritises cancer driver genes over false positives and non-cancer genes regardless of the size of the training cohort. Even in small cohorts of samples (e.g., n=10), the recall of cancer driver genes over false positives and the rest of human genes is comparable to that of large cohorts of up to 1000 samples (Nulsen et al., 2021).

When benchmarked against other patient-specific driver detection methods, such as DriverNet (Bashashati et al., 2012) and OncoIMPACT (Bertrand et al., 2015), sysSVM2 showed a lower recall of false positives and outperformed the other methods in predicting cancer driver genes in individual patients (Nulsen et al., 2021).

sysSVM2 has been developed and implemented with the ultimate goal of prioritising cancer genes in individual patients by exploiting the similarity of genes SLPs to that of canonical drivers. This rationale has proven successful in

identifying the complete cancer driver gene repertoire in a pan-cancer cohort of 7646 tumour samples (Nulsen et al., 2021).

1.4 Aim of the thesis

This thesis aims to investigate the role of genetic inter-tumour heterogeneity in shaping oesophageal adenocarcinoma (OAC) initiation and progression. The high grade of genetic inter-tumour heterogeneity hampers the efficacy of large-scale cancer genomic studies for the identification of targetable genes in OAC. Furthermore, traditional cohort-based approaches hamper the identification of cancer driver genes in individual samples. Some patients are left with too few or no driver genes to explain the presence of the disease. In the context of precision oncology, it is important to identify the complete pool of driver genes responsible for the initiation and development of the tumour. Being able to identify the cancer driver repertoire in each patient will enable the employment of strategies tailored to the individual molecular landscape of each patient.

In order to investigate inter-tumour heterogeneity in OAC, I will first introduce the Network of Cancer Genes (NCG). NCG is a curated repository of cancer driver genes and their systems-level properties (SLPs). NCG represents the foundation of the rest of the work presented in this thesis.

In the second result chapter I will describe the process used to compile a comprehensive list of cancer driver genes in individual patients. Given the genetic heterogeneity of OAC, I will apply sysSVM2 to a large cohort of OAC samples that I curated and annotated. sysSVM2 ranks putative driver genes in individual samples based on molecular and SLPs. Since there is no hard cut-off that defines what is or is not a cancer driver gene, it is fundamental to quantify the number of drivers needed in individual patients.

In the third result chapter, I will describe the comparison of two models on the number of cancer drivers in OAC: the age-incidence model and the positive selection model. The two models differ by the number of driver genes that they estimate are needed in each OAC. The age-incidence model predicts that three driver genes are enough to explain the presence of OAC in single patients. The

positive selection model, on the other hand, postulates that five driver genes are needed to explain the disease in each patient. I will compare the two models using pathway perturbations to inspect whether the additional drivers, predicted only under the positive selection model, are functionally involved in the development and progression of the tumour.

After defining how many drivers are needed per OAC I will investigate the frequency of cancer driver genes. I will divide the identified drivers in two groups: those that have been experimentally validated to be involved in cancer (canonical cancer drivers) and the new candidate drivers (sysSVM2 predictions). I will next investigate whether the two sets of drivers (canonical drivers and sysSVM2 predictions) perturb similar pathways. I will inspect sysSVM2 predictions for new potential drug targets by using the recently curated list of antineoplastic and immunomodulating drugs, published as part of the latest update of NCG.

Finally, taking advantage of the clinical stratification of OAC samples I will look at the frequency of drivers across stages and investigate whether different pathways are perturbed in a stage-specific fashion during the development of OAC.

Chapter 2. Materials & Methods

2.1 The Network of Cancer Genes

2.1.1 Literature curation

A literature search integrating PubMed, The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) and the International Cancer Genome Consortium (ICGC, https://dcc.icgc.org/) was carried out to retrieve cancer screens published between 2018 and 2020. The literature search resulted in 135 coding and 154 non-coding cancer screens, of which 37 were retained after examining abstracts and full texts. Exclusion criteria included the absence of driver genes or driver detection methods and the impossibility to map non-coding driver alterations to protein-coding genes. The 37 new cancer screens were added to the 273 publications previously curated by the Ciccarelli lab (Repana et al., 2019), totalling 310 cancer publications. A similar literature search retrieved 24 sequencing screens of non-cancer and healthy tissues published before 2020, 18 of which were retained after applying the same criteria as above. Each paper was reviewed independently by two experts and further discussed if any mismatch was found in their annotations, including the list of driver genes, the number of donors, the type of screen (whole-genome, whole-exome, target gene sequencing), the cancer or healthy tissues, and the driver detection method.

Canonical cancer drivers were extracted from two publications (Saito et al., 2020; Vogelstein et al., 2013) and the Cancer Gene Census (CGC) v.91 (Tate et al., 2019). From CGC, all tier 1 and 2 genes were retained, except those derived from gene fusions. Canonical cancer driver genes were further stratified into TSG, OG, and unclassified if having a dual role or having conflicting or unavailable annotation.

Drivers from cancer screens and canonical sources underwent additional filtering. They were intersected with a list of 148 possible false positives derived from two sources (Lawrence et al., 2013; Saito et al., 2020). After manual checks of the supporting evidence, two drivers were retained as canonical, five were retained as candidates, and 41 were removed.

The three resulting lists (canonical drivers, drivers from cancer screens, and healthy drivers) were intersected to annotate:

- canonical drivers in cancer screens,
- remaining drivers in cancer screens (candidate cancer drivers),
- canonical healthy drivers,
- candidate healthy drivers,
- remaining healthy drivers.

2.1.2 Systems-level properties

Protein sequences from RefSeq v.99 (O'Leary et al., 2016) were aligned to the human genome assembly GRCh38 using BLAT (Kent, 2002). Unique genomic loci were identified for 19756 genes based on gene coverage, span, score, and identity (Bhagwat et al., 2012). Genes sharing at least 60% of their protein sequence were considered as duplicates (Rambaldi et al., 2008).

Evolutionary conservation was obtained for 18922 human genes using their orthologs in EggNOG v.5.0 (Huerta-Cepas et al., 2019). Genes were considered to have a pre-metazoan origin (and therefore conserved in evolution) if they had orthologs in prokaryotes, eukaryotes, or opisthokonts (Matteo D'Antonio & Ciccarelli, 2011).

Gene expression for 19231 genes in 49 healthy tissues was derived from the union of Protein Atlas v.19.3 (Uhlén et al., 2015) and GTEx v.8 (Aguet et al., 2020). Genes were considered expressed in a tissue if their expression value was ≥1 transcript per million (TPM). Protein expression for 13229 proteins in 45 healthy tissues was derived from Protein Atlas v.19.3 (Uhlén et al., 2015). When multiple expression values were available the highest one was retained.

A total of 542397 non-redundant binary interactions between 17883 proteins were gathered from the integration of five sources (BioGRID v.3.5.185 (Oughtred et al., 2019), IntAct v.4.2.14 (Orchard et al., 2014), DIP (February 2018) (Salwinski et al., 2004), HPRD v.9 (Keshava Prasad et al., 2009) and Bioplex v.3.0 (Huttlin et al., 2021)). Data on 9476 protein complexes involving 8504 proteins were derived from CORUM v.3.0 (Giurgiu et al., 2019), HPRD v.9

(Keshava Prasad et al., 2009), and Reactome v.72 (Jassal et al., 2020). Experimentally-supported interactions between 14747 genes and 1758 miRNAs were obtained from the integration of miRTarBase v.8.0 (Huang et al., 2020) and miRecords v.4.0 (Xiao et al., 2009). Degree, betweenness, and clustering coefficient were calculated for protein and miRNA networks using the R package igraph v.1.2.6 (Csardi & Nepusz, 2006).

The loss-of-function observed/expected upper bound fraction (LOEUF) score for 18392 genes was obtained from gnomAD v.2.1.1 (Karczewski et al., 2020). Germline mutations (SNVs and small insertions and deletions, indels) were obtained from the union of 2504 samples from the 1000 Genomes Project Phase 3 v.5a (Auton et al., 2015) and 125748 samples from gnomAD v.2.1.1 (Karczewski et al., 2020). Mutations were annotated with ANNOVAR (April 2018) (Wang et al., 2010) and dbNSFP v3.0 (Liu et al., 2016) and only those identified as exonic or splicing were retained. Mutations were annotated as damaging if at least one of the following conditions was true:

- truncating (stopgain, stoploss, frameshift) mutations,
- missense mutations predicted by at least five out of the seven function-specific methods and by at least two out of the three conservation-specific methods (Table 2.1),
- splicing mutations predicted by at least one of two splicing-specific methods (Table 2.1),
- hotspot mutations identified with OncodriveCLUST v1.0.0 (Tamborero et al., 2013).

A total of 18812 genes were retained as damaged. A total of 32558 germline structural variants (SVs) for 14158 genes were derived using 15708 samples from gnomAD v.2.1.1 (Karczewski et al., 2020). The numbers of damaging mutations and SVs per base pair (bp) were calculated for each gene.

Table 2.1 Prediction tools used to annotate the damaging effect of mutations

Tool	Method	Reference
SIFT	Function	(P. C. Ng & Henikoff, 2003)

PolyPhen-2 HDIV	Function	(Adzhubei et al., 2013)
PolyPhen-2 HVAR	Function	(Adzhubei et al., 2013)
MutationTaster	Function	(Schwarz et al., 2010)
MutationAssessor	Function	(Reva et al., 2011)
LRT	Function	(Chun & Fay, 2009)
FATHMM	Function	(Shihab et al., 2013)
PhyloP	Conservation	(Pollard et al., 2010)
GERP++RS	Conservation	(Davydov et al., 2010)
SiPhy	Conservation	(Garber et al., 2009)
ADA	Splicing	(Liu et al., 2016)
RF	Splicing	(Liu et al., 2016)

Essentiality data for 19013 genes in 1122 cell lines were obtained by integrating three RNA interference (RNAi) knockdown and six CRISPR/Cas9 knockout screens (Behan et al., 2019; Dempster et al., 2019; Lenoir et al., 2018; McFarland et al., 2018; Meyers et al., 2017; Tsherniak et al., 2017). Genes with CERES (Meyers et al., 2017) or DEMETER (Tsherniak et al., 2017) scores <− 1 or Bayes score (Hart & Moffat, 2016) > 5 were considered as essential.

The proportions of duplicated genes, pre-metazoan genes, essential genes, and proteins engaging in complexes were compared between gene groups using a two-sided Fisher's exact test. Distributions of healthy tissues expressing genes or proteins, protein and miRNA network properties, LOEUF scores, damaging mutations and SVs per bp were compared between gene groups using a two-sided Wilcoxon rank-sum test. Multiple comparisons within each property were corrected using the Benjamini-Hochberg procedure.

For each SLP in each driver group (d), a normalised property score was calculated as:

$$\text{Normalised property score} = sgn(\Delta_d) \times \frac{|\Delta_d| - min_t|\Delta_t|}{max_t|\Delta_t| - min_t|\Delta_t|}$$

where t represents the ten gene groups (canonical cancer drivers, candidate cancer drivers, TSGs, OGs, drivers with coding alterations, drivers with non-coding alterations, canonical healthy drivers, candidate healthy drivers,

remaining healthy drivers, and the rest of human genes); $sgn(\Delta_d)$ is the sign of the difference; and Δ_d indicates the difference of medians (for continuous properties) or proportions (for categorical properties) between each driver group and the rest of human genes. Minima and maxima were taken over all ten gene groups for each property.

2.1.3 Pan-cancer cell line data

Gene expression data for cancer genes in 2443 cancer cell lines were taken from the Cancer Cell Line Encyclopaedia (CCLE, May 2020) (Ghandi et al., 2019), the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Cell Line Project (CLP) v.91 (Futreal et al., 2004), and a Genentech study (GNE, June 2014) (Klijn et al., 2015). Gene expression levels were derived directly from the original sources, namely reads per kilobase million (RPKM) values for CCLE and GNE, and microarray z-scores for CLP. Genes were categorized as expressed if their expression value was ≥1 RPKM in CCLE or GNE and were annotated as over, under, or normally expressed in CLP, as determined by COSMIC.

2.1.4 Driver functional annotation

Gene functions were collected for 11778 proteins from Reactome v.72 (Jassal et al., 2020) and KEGG v.94.1 (Kanehisa et al., 2017) (levels 1 and 2). Driver enrichment in Reactome pathways (levels 2–8) compared to the rest of human genes was calculated using a one-sided Fisher's exact test and corrected for multiple testing with the Benjamini-Hochberg method. Enriched pathways were then mapped to the corresponding Reactome level 1.

2.1.5 Drug interactions

A total of 247 Food and Drug Administration (FDA)-approved, antineoplastic, and immunomodulating drugs targeting 212 human genes were downloaded from DrugBank v.5.1.8 (Wishart et al., 2018). Genetic biomarkers of response and

resistance to drugs in cancer cell lines were obtained from Genomics of Drug Sensitivity in Cancer (GDSC) v.8.2 (Iorio et al., 2016). Of those, only 467 associations with FDR ≤0.25 involving 129 drugs and 106 genes were retained. Genetic biomarkers of response and resistance in clinical studies were obtained from the Variant Interpretation for Cancer Consortium Meta-Knowledgebase v.1 (Wagner et al., 2020). A total of 868 associations between drugs and genomic features involving 64 anti-cancer drugs and drug combinations and 24 human genes were retained (Wagner et al., 2020).

2.1.6 Database and website implementation

All annotations of driver genes were stored into a relational database based on MySQL v.8.0.21 (source: https://dev.mysql.com/doc/refman/8.0/en/) connected to a web interface enabling interactive retrieval of information through gene identifiers. The frontend was developed with PHP v.7.4.15 (Bakken et al., 2020). The interactive displays of miRNA-gene and protein-protein interactions were implemented using the R packages Shiny v.1.6.0 (Chang et al., 2017) and igraph v.1.2.6 (Csardi & Nepusz, 2006) and ran on a Shiny Server v.1.5.16.958.

2.2 Clinical and molecular characterisation of the BO and OAC sample cohort

2.2.1 Annotation of damaged genes

Aligned genomic sequencing data in the form of BAM files were downloaded from public repositories for two cohorts (Dulak et al., 2013; Stachler et al., 2015). Mutation (SNVs and indels) and copy number (CN) data for the International Cancer Genome Consortium - Oesophageal Cancer Clinical and Molecular Stratification (ICGC-OCCAMS) samples were obtained from the OCCAMS consortium, of which the Ciccarelli lab is a part (source: https://www.occams.org.uk/). Mutation (SNVs and indels), CN, and gene

expression data for TCGA samples were obtained from the Genomic Data Commons portal I (source: https://gdc.cancer.gov/; Grossman et al., 2016).
For the two publication datasets (Dulak et al., 2013; Stachler et al., 2015), somatic mutations (SNVs and indels) were called using Strelka v.2.9.0 (Saunders et al., 2012). The absolute CN of genomic regions, the sample ploidy and the sample purity were obtained by running ASCAT v.2.5.2 (Loo et al., 2010). To obtain the gene CN, gene coordinates were intersected with genomic regions for which ASCAT calculated the CN. If at least 25% of the gene length was contained in one of the genomic regions, the CN of the region was assigned to the gene.
For all 748 samples, mutations were annotated as damaging as previously described (see Chapter 2.1.2). In addition to genes altered via damaging mutations, CNAs were considered. A gene was considered affected by damaging CNAs if it was either homozygously deleted (gene CN =0) or amplified (gene CN $\geq$2 times sample ploidy).
An additional filtering was carried out for TCGA samples. To remove possible false-positive CNAs and given the availability of gene expression data for the whole TCGA cohort, CNAs were corrected by RNA-seq. Homozygously deleted genes were confirmed if their expression was $<$1 fragments per kilobase per million (FPKM) over sample purity. Heterozygously deleted genes had CN =1 or CN =0 and FPKM $>$1 over sample purity. The expression of putatively amplified genes was compared between samples with and without each gene amplification using a two-sided Wilcoxon rank-sum test and corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure. Only amplified genes with an FDR $<$0.05 were retained.

2.2.2 Estimation of WGD

The CN of the genomic segments, obtained with ASCAT v.2.5.2 (Loo et al., 2010), were used to identify the presence of WGD in the OAC samples. An adaptation of the method used by Dentro et al. (Dentro et al., 2021) was used to estimate whether samples had undergone WGD.

For each major allele CN the length of the genome with the same CN value was calculated. If, throughout the whole genome, the longest total segment corresponded to major allele CN ≥2, the sample was predicted to have undergone WGD.

2.2.3 Clinical annotation

Clinical data were obtained from the same sources as the molecular data (Dulak et al., 2013; Stachler et al., 2015; https://www.occams.org.uk/; https://gdc.cancer.gov/).

BO cases progressing to OAC were retained even if the paired OAC had not been sequenced. BO clinical information included whether dysplasia was present, which resulted in the stratification of samples into two major groups: NDBO and DBO.

OAC clinical information included the patient gender, the age at diagnosis, the size of the primary tumour (T), the number of metastasised lymph nodes (N), the presence of distant metastasis (M), whether the patient received neoadjuvant chemotherapy, and whether the sample had been collected at the clinical or at the pathologic staging.

The information on TNM stages was used to classify OAC samples into stage I, II, III and IV by applying the guidelines as described in the 8th edition AJCC staging of tumours of the oesophagus and oesophagogastric junction (Rice et al., 2017).

BO and OAC samples with both clinical and molecular information were retained for downstream analyses.

2.2.4 Curation of OAC-specific canonical cancer driver genes

The list of OAC-specific canonical cancer driver genes was derived from the integration of OAC-specific canonical drivers reported in NCG6 (Repana et al., 2019) and the manual curation of five additional OAC-specific screens (Dulak et

al., 2012; Frankel et al., 2014; Frankell et al., 2019; Murugaesu et al., 2015; Nones et al., 2014).

The manual annotation of the five additional screens resulted in 311 putative cancer drivers. This list was then intersected with the 711 canonical cancer drivers from NCG6 (Repana et al., 2019), and 78 canonical cancer drivers were retained.

Given the presence of canonical cancer driver genes identified through the study of CNAs, only genes for which there was agreement between their role in tumorigenesis and their driver event were retained (i.e., deleted TSGs and amplified OGs). This filtering resulted in the removal of three genes that were reported as TSGs but were found amplified in the corresponding screen.

The final list of OAC-specific canonical cancer driver genes was composed of 77 genes.

2.3 Training of sysSVM2 and evaluation of its predictions

2.3.1 Prioritisation of cancer driver genes

In order to complete the list of cancer driver genes in individual samples, sysSVM2 (Nulsen et al., 2021), was used. sysSVM2 consists of four one-class SVMs trained on the molecular and systems-level properties of the canonical cancer driver genes damaged in the BO and OAC cohort. The tool ranks damaged genes outside the training set based on how closely their properties resemble those of canonical cancer drivers used for training.

The algorithm consists of three stages: feature mapping, model selection, and training and prediction.

FEATURE MAPPING

In the feature mapping stage, molecular and systems-level properties are mapped to all the damaged genes in the cohort. The pool of damaged genes across the whole cohort is then divided into training and prediction sets.

TSGs with LoF alterations (truncating mutations, missense or splicing damaging mutations, homozygous deletions, or double hits) and OGs with gain-of-function (GoF) alterations (hotspot mutations, missense or splicing damaging mutations, or gene amplifications) were retained. For *TP53* both LoF and GoF alterations were retained as driver events. Somatic alterations in TSGs and OGs that were of the opposite types (i.e. GoF in TSGs and LoF in OGs) were discarded.

Molecular properties were derived from the annotation of damaged genes in individual samples as previously described (see Chapter 2.2.1). From these properties, seven molecular features were derived and used for feature mapping. SLPs of human genes were obtained from NCG6 (Repana et al., 2019) which collected information from different sources. Briefly, duplicated gene loci (genes with 60% protein sequence shared with another locus) were identified as previously described (Repana et al., 2019). The gene ohnolog status (i.e. whether gene duplicates appeared as a result of whole-genome duplication events that occurred at the basis of vertebrates) was obtained from Nakatani et al. (Nakatani et al., 2007). Data on gene essentiality in cell lines were obtained from PICKLES (September 2017) (Lenoir et al., 2018) and OGEE v.2 (Chen et al., 2017). Genes in PICKLES were considered to be essential in a cancer cell line if they had Bayes factor (Hart & Moffat, 2016) >3, while original annotations of essentiality from OGEE were retained. mRNA expression data for healthy human tissues were obtained from GTEx v.7 (Lonsdale et al., 2013) and Protein Atlas v.18 (Uhlén et al., 2015). A gene was considered expressed in a tissue if its median expression was ≥1 TPM in both databases (or in one database if said tissue was not included in both). Protein expression data for healthy human tissues was downloaded from Protein Atlas v.18 (Uhlén et al., 2015). The PPIN was built as previously described (Repana et al., 2019), from the union of BioGRID v.3.4.157 (Chatr-Aryamontri et al., 2017), MIntAct v.4.2.10 (Orchard et al., 2014), DIP (February 2018) (Salwinski et al., 2004) and HPRD v.9 (Keshava Prasad et al., 2009). Network properties (PPIN degree, betweenness and centrality) were calculated using custom scripts as previously described (see Chapter 2.1.2). Genes were identified as participating in complexes using data downloaded from CORUM (July 2017) (Ruepp et al., 2009), HPRD v.9 (Keshava Prasad et al., 2009) and

Reactome v.63 (Fabregat et al., 2018). The number of miRNAs regulating a gene was calculated using data from miRTarBase v.7 (Chou et al., 2018) and miRecords v.4 (Xiao et al., 2009). The evolutionary origin of genes was identified as previously described (Matteo D'Antonio & Ciccarelli, 2011), using data from EggNOG v.4.5.1 (Huerta-Cepas et al., 2016). From these properties, 19 systems-level features were derived and used for feature mapping.

For each systems-level feature, missing values were imputed using the median or mode (for continuous and categorical features, respectively) of available data for canonical drivers and the rest of human genes separately. All features used in the model significantly differentiated cancer drivers from other human genes or subgroups of cancer drivers among each other (i.e. TSGs and OGs).

MODEL SELECTION

The kernels used in sysSVM2 are linear, polynomial, radial, and sigmoid, and each of them is controlled by parameters whose values can vary. For this reason, a grid search was run to select the best parameters for each kernel separately that defined the best models. The best models were defined as those with a high sensitivity to retrieve canonical cancer drivers and high stability (low standard deviation of sensitivity).

These parameters and their default grid ranges are:

- Nu (ν, all kernels): values ranging from 0.05 to 0.35 in incremental steps of 0.05;
- Gamma (γ, radial and sigmoid kernels): values assessed were $\gamma = 2^x$, where $x \in \{-7, -6, \dots, 4\}$;
- Degree (d, polynomial kernel): chosen from the set $\{3, 4, 5\}$.

A parameter grid search was carried out for each kernel separately, for a total of 196 kernel-parameter combinations which were the result of 7 parameter combinations for the linear kernel, $7 \times 12 = 84$ combinations each for the radial and sigmoid kernels, and $7 \times 3 = 21$ combinations for the polynomial kernel.

To identify the best models, cross-validation iterations were performed on the training set. At each iteration the training set was randomly split with 2/3 of the genes used to train the models and the remaining 1/3 used as a test set. The

sensitivity of each parameter combination to retrieve canonical cancer drivers was calculated at every iteration. For each kernel k and parameter combination i, the mean μ_{ki} and standard deviation σ_{ki} of the sensitivity were calculated across the cross-validation iterations. These were then converted into z-scores, $z_{ki}^{(\mu)}$ and $z_{ki}^{(\sigma)}$, which measured the relative values of mean and standard deviation between the different parameter combinations such that:

$$\sum_i z_{ki}^{(\mu)} = \sum_i z_{ki}^{(\sigma)} = 0$$

and

$$\text{Variance}_i\left(z_{ki}^{(\mu)}\right) = \text{Variance}_i\left(z_{ki}^{(\sigma)}\right) = 1.$$

Finally, the Δ_z score was defined as:

$$\Delta_z = z_{ki}^{(\mu)} - z_{ki}^{(\sigma)}$$

High Δ_z scores corresponded to parameter combinations that had high mean sensitivity and low standard deviation relative to the other combinations for that kernel. The four parameter combinations (one per kernel) with the highest Δ_z scores were selected and used to train the four kernels on the entire training set.

TRAINING AND PREDICTION

Once the parameters for each kernel were selected, the four SVMs were trained using the entire training set of canonical drivers. The trained sysSVM2 models were, thus, used for prediction in individual samples. To combine the outputs of the four kernels, a combined score S_{gs}, was calculated for each gene g in sample s. S_{gs} measured the similarity of the features of gene g to those of the training set. It combined the rank of g in sample s according to each of the four kernels. This resulted in a normalised final score between 0 and 1. High ranks in each kernel were given exponential weighting and the kernels were weighted according to their sensitivity, with more sensitive kernels contributing more to the score. The formula used to calculate the combined score, S_{gs}, was:

$$S_{gs} = \frac{\sum_{k=1}^{4}\left(-\log_{10}\left(\frac{R_{kgs}}{N_s}\right) \times \mu_k\right)}{4 \times \log_{10}(N_s)}$$

where R_{kgs} represents the rank of gene g in sample s according to the decision value of kernel k; N_s is the total number of damaged genes in sample s; and μ_k is the mean sensitivity of kernel k as assessed by cross-validation iterations.

2.3.2 Performance and stability metrics

The results of the three driver settings tested in sysSVM2 were measured using two performance and two stability metrics: Area Under the Receiver Operating Characteristic (AUROC) curve, composition score, Rank-Biased Overlap (RBO) score, and overlap of the top five predictions between models.

AUROC curves were derived for each sample individually by comparing the ranks of canonical drivers not used for training and of candidate cancer drivers to false positives and to the rest of human genes. The median AUROC curve was then measured across samples.

The composition score assessed the identity of the top five predictions in each sample and measured the prevalence and ranks of different types of genes. The score S was calculated as a weighted sum according to the following formula:

$$S = \sum_{g=1}^{5} w_g \times t_g.$$

The weight w_g of each gene g in the top five was such that higher-ranked genes were assigned greater weight; specifically, $w_g = 6 - r_g$ where r_g was the rank of gene g (with 1 being the highest). The contribution t_g of gene g was defined for different gene categories as reported in Table 2.2.

Table 2.2 Contribution of gene categories to top five drivers

Gene (g)	Gene contribution (t_g)
OAC-specific canonical driver	3
Non OAC-specific canonical driver	2
OAC-specific candidate driver	1.5

Non OAC-specific candidate driver	1
Rest of human genes	0
False positive	-1

The RBO score (Webber et al., 2010) was used to assess the similarity of the top five predictions from pairs of models. It measured the overlap of ranked lists at incrementally increasing depths using a convergent series. It was calculated according to the formula:

$$\mathrm{RBO} = \frac{1-p}{1-p^5} \times \sum_{d=1}^{5} p^{d-1} \times A_d$$

where p determines how steep the decline in weights is (the smaller p, the more top-weighted is the metric; the closer it is to 1, the flatter the weights are) (Webber et al., 2010) and was set to 0.9 so that the five rankings were weighted similarly, d indicates the depth in the rankings (starting from the top-ranked elements), and A_d is the overlap of the two lists, restricted to depth d.

2.4 Investigating cancer driver genes

2.4.1 Deriving the list of cancer driver genes in individual samples

The list of cancer driver genes in individual OAC samples was derived using the following procedure. OAC-specific canonical cancer drivers damaged in each sample were considered as cancer drivers for that sample. If more than three or five canonical drivers, based on the model under investigation (Jeon et al., 2006; Martincorena et al., 2017), were present in each sample, then OAC-specific canonical cancer drivers were prioritised as follows. *CDKN2A*, *TP53* and OAC-specific OGs were prioritised according to previous literature on the acquisition of driver alterations in the BO-to-OAC progression (Maley, 2007; Stachler et al., 2015). If either *CDKN2A* or *TP53* were not damaged, OAC-specific TSGs were prioritised.

If less than three or five OAC-specific canonical drivers were damaged, the highest-ranked genes from sysSVM2 were added in order to obtain three or five drivers in total per individual sample.

2.4.2 Gene set enrichment analysis

Human pathways for gene set enrichment analysis (GSEA) were obtained from Reactome v.72 (Jassal et al., 2020) and from MSigDB v.7.4 (Liberzon et al., 2015). Reactome was used to investigate differences between driver models (see Chapter 5.2) and between sysSVM2 predictions and OAC-specific canonical cancer drivers (see Chapter 5.4). MSigDB was used to investigate differences across OAC clinical stages (see Chapter 5.6).

Before testing, Reactome pathways were restricted to those of level 2 or higher, and with between 10 and 500 genes (total of 1303 pathways containing 10178 unique genes). All 50 MSigDB hallmarks pathways were used comprising a total of 4378 unique genes.

Across all comparisons, as well as in individual OAC clinical stages separately, the unique set of cancer genes was tested for enrichment in pathways containing at least one gene against the rest of human genes using one-sided Fisher's exact tests. The resulting p-values within each clinical stage were corrected for FDR using the Benjamini-Hochberg method.

2.4.3 Comparing the contribution of different sets of drivers to Reactome level 1 pathways

After identifying the enriched Reactome pathways by applying the filtering on the pathway size as explained above (see Chapter 2.4.2), each enriched pathway was mapped to their corresponding level 1. Differences in the proportion of enriched pathways mapping to each level 1 were compared between the sets of drivers in each comparison and within each OAC clinical stage. The proportions were compared using a chi-squared test and the resulting p-values were

corrected using the Benjamini-Hochberg method. Proportions with an FDR <0.1 were considered as significantly different.

2.4.4 Reducing Reactome redundancy

Given the pyramidal structure of the Reactome database (Figure 7.1), the genes mapping to lower level pathways (e.g. level 6 or level 7) are fully contained in higher level pathways (e.g. level 2 or level 3).

The pathways enriched in the positive selection-specific drivers (see Chapter 5.2) were tested in order to investigate whether they were, in terms of genes mapping to them, fully contained in higher level pathways enriched in the age-incidence drivers. If all the genes from an enriched pathway in the positive-selection-specific drivers were contained in any of the pathways enriched in the age-incidence drivers, the pathway was identified as redundant, hence removed from further analyses.

Of the total 899 enriched pathways in the positive selection-specific drivers, 343 were retained as non-redundant and investigated further.

Chapter 3. The Network of Cancer Genes

3.1 Motivation

NCG is a curated database of cancer driver genes and their SLPs. SLPs are global properties of genes, that, although they are not directly related to cancer, differentiate cancer driver genes from the rest of human genes. SLPs includes properties that describe the evolutionary history of the gene, such as the first appearance of the gene in the species phylogenetic tree (Syed et al., 2010) and the presence of duplicated copies of the gene in the human genome (Rambaldi et al., 2008). Moreover, SLPs define the pervasive role of the gene in terms of its expression at the gene and protein level (An et al., 2016; Repana et al., 2019), the essentiality in cancer cell lines (Repana et al., 2019), the level of miRNA regulation (D'Antonio et al., 2012), the participation of the encoded protein in protein complexes (An et al., 2016), and its position in the PPIN (D'Antonio et al., 2012).

The seventh release of NCG (NCG7) is available online at http://network-cancer-genes.org/. NCG7 (Dressler et al., 2022) contains some novelties with respect to previous releases (An et al., 2016, 2014; D'Antonio et al., 2012; Repana et al., 2019; Syed et al., 2010). In addition to cancer drivers, it now reports information relative to genes that, upon acquiring somatic mutations, drive the expansion of clones in non-cancer tissues (namely, healthy drivers). Another novelty of NCG7 is that it now also annotates cancer drivers implicated in disease progression through the accumulation of mutations in their non-coding regions (namely, drivers with non-coding alterations). Finally, NCG7 includes some new SLPs and gene annotations. These newly introduced SLPs describe the tolerance of the gene towards the accumulation of germline variation and are used as further evidence of the essential role of the gene within the cell. Furthermore, we have introduced the annotation of interactions between human genes and anti-cancer drugs.

The seventh release of NCG has been a collaborative effort of the Ciccarelli lab and my contribution to it involves the:

- Curation of gene and protein expression in human healthy tissues;

- Curation of gene expression in cancer cell lines;
- Curation of gene function;
- Annotation of drug interactions (specifically, drug targets and biomarkers of response and resistance to antineoplastic drugs);
- Co-curation of the database;
- Review of the literature curation performed by other colleagues.

In the rest of the chapter, I will acknowledge and discuss the contributions of the other group members who curated the remaining information.

3.2 Curated annotation of cancer and healthy drivers

In order to gather a comprehensive understanding of cancer and healthy driver genes we collected information from 331 scientific papers published between 2008 and 2020. My colleagues Amelia Acha-Sagredo, Lucia Montorsi, Dimitra Repana and Neshika Wijewardhane carried out and curated the literature search, which was reviewed by the rest of the group, including myself.

We integrated three sources of canonical cancer driver genes (Saito et al., 2020; Sondka et al., 2018; Vogelstein et al., 2013) (Figure 3.1A). As additional sources of cancer driver genes we included 310 publications, of which 291 reported mutations in coding regions of the human genes, five investigated mutations in non-coding regions of the human genes, and 14 annotated mutations in both coding and non-coding regions of the human genes (Figure 3.1A). In addition, we selected 18 publications investigating healthy drivers in non-cancer tissues.

The curation of the 331 publications present in NCG7 resulted in the identification of 3347 cancer driver and 95 healthy driver genes (Figure 3.1A). We further divided the 3347 cancer drivers into canonical and candidate cancer driver genes. While 591 cancer driver genes had robust experimental evidence supporting their implication in tumorigenesis, and thus were defined as canonical, we labelled the remaining 2756 as candidate drivers given that their involvement in cancer is based exclusively on statistical methods for the prediction of driver genes from cancer patients. We then curated nine SLPs and three annotations for the pool of drivers that we collected (Figure 3.1A).

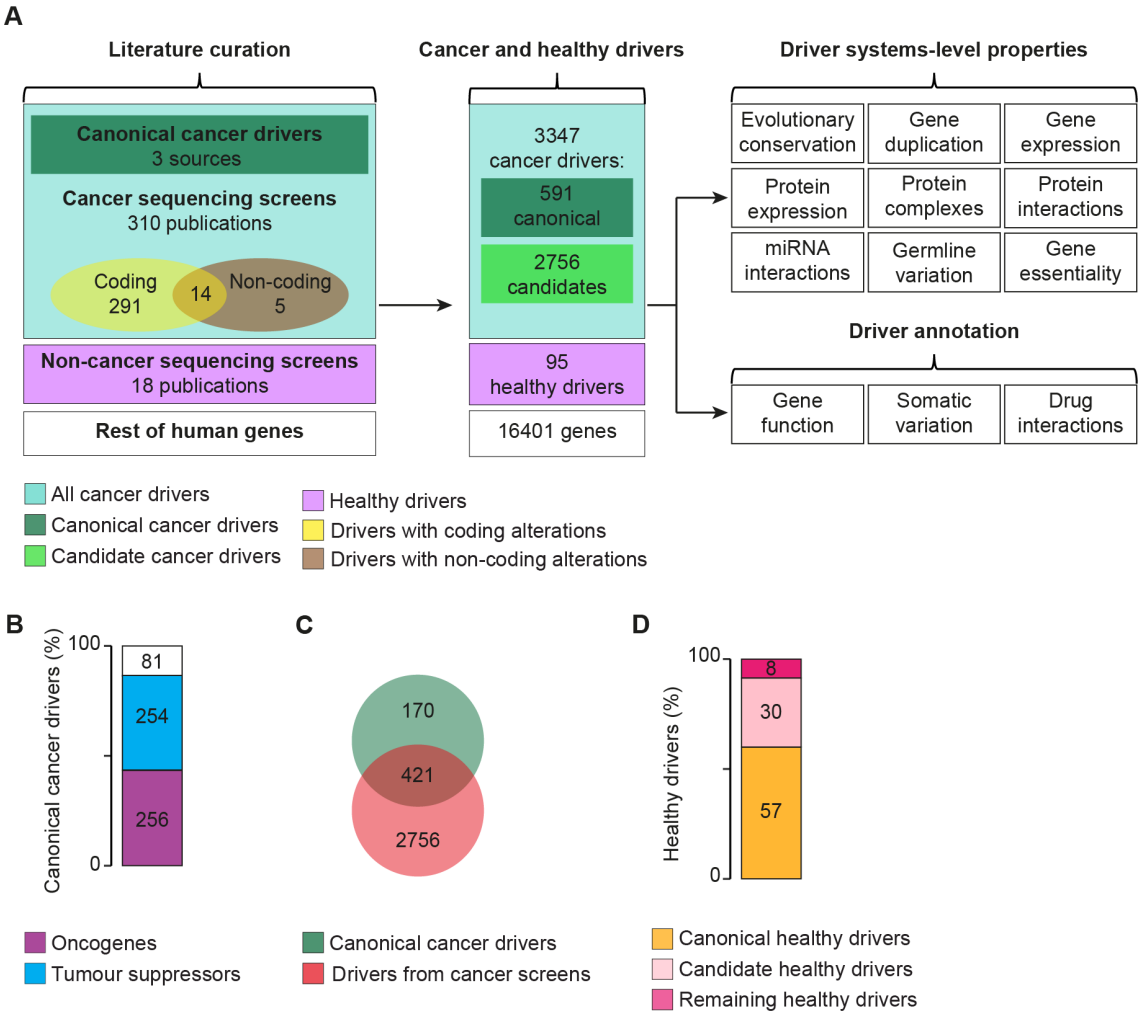

Figure 3.1 Literature annotation and resulting list of cancer and healthy drivers

A) Pipeline overview of the literature curation and annotation of SLPs and additional driver information (e.g., gene function, somatic variation and gene interactions with anti-cancer drugs). **B)** Proportion of canonical cancer driver genes that are TSGs, OGs and unclassified. **C)** Overlap between canonical cancer drivers (derived from the three sources of canonical drivers (Saito et al., 2020; Sondka et al., 2018; Vogelstein et al., 2013)) and the cancer drivers derived from the curation of 310 cancer publications. **D)** Proportion of healthy drivers that are also canonical cancer drivers, candidate cancer drivers and only healthy. Figure adapted from (Dressler et al., 2022).

We divided the 591 canonical drivers into subgroups based on their mechanism of action in promoting and sustaining tumorigenesis. We identified 254 TSGs, 256 OGs and 81 canonical drivers with unclear or no evidence for their mechanism of action in the context of tumorigenesis (Figure 3.1B). We did not identify 29% (170/591) of the canonical drivers in any of the 310 cancer

sequencing screens (Figure 3.1C). This may suggest that these 170 canonical cancer drivers contribute to carcinogenesis via non-mutational mechanisms, such as CNAs or epigenetic deregulation.

Given the novelty of the field and the limited number of studies available (Wijewardhane et al., 2021), it came as no surprise that the number of healthy drivers that we identified was smaller than the overall number of cancer drivers. Interestingly, 92% of the healthy drivers were also cancer drivers (57 canonical healthy and 30 candidate healthy drivers) and only 8% (8/95) were involved exclusively in the expansion of mutated clones in non-cancer tissues (Figure 3.1D).

3.3 SLPs define the central role of cancer and healthy drivers within the cell

We compiled annotations of SLPs for all human genes, including evolutionary origin (curated by Michele Bortolomeazzi), gene duplication (curated by Reda Keddar), gene and protein expression in healthy human tissues (curated by me), protein-protein (curated by Lisa Dressler) and miRNA-gene (curated by Hrvoje Misetic) interactions, germline variation (curated by Joel Nulsen and Hrvoje Misetic), and gene essentiality (curated by Lisa Dressler). We evaluated the SLPs of canonical and candidate cancer drivers against the rest of human genes. We then investigated how specific subtypes of these two main categories differed from each other (e.g., TSGs versus OGs and candidate drivers with only coding alterations versus candidate drivers with only non-coding alterations). Finally, we investigated how the three categories of healthy drivers (Figure 3.1D) compared with the rest of human genes.

Canonical and candidate cancer driver genes were older and more broadly expressed in healthy human tissues than the rest of human genes (Figure 3.2A). This highlights their evolutionary conserved and ubiquitous role. Both cancer driver categories encoded central hubs of the PPIN (Figure 3.2A). Specifically, such proteins formed significantly more connections (higher degree), were more central and clustered (higher betweenness and clustering coefficient,

respectively) in the PPIN and took part in significantly more protein complexes than the rest of human genes. Moreover, canonical and candidate cancer drivers showed a higher level of miRNA regulation than the rest of human genes (Figure 3.2A).

As one of the novelties of NCG7, we investigated the tendency of genes to accumulate germline alterations by including three new metrics. We measured the number of damaging mutations and SVs per coding bp using germline data from healthy individuals (Karczewski et al., 2020) and introduced the LOEUF score which quantifies the tendency to accumulate germline LoF alterations (Karczewski et al., 2020). Further supporting the central role of these drivers in the cell, we observed that canonical and candidate cancer drivers accumulated fewer germline alterations than the rest of human genes (Figure 3.2A), pointing towards a negative selection against potentially damaging germline alterations in these sets of genes. In line with that, we observed that both driver categories were also significantly more essential in cancer cell lines than the rest of human genes (Figure 3.2A).

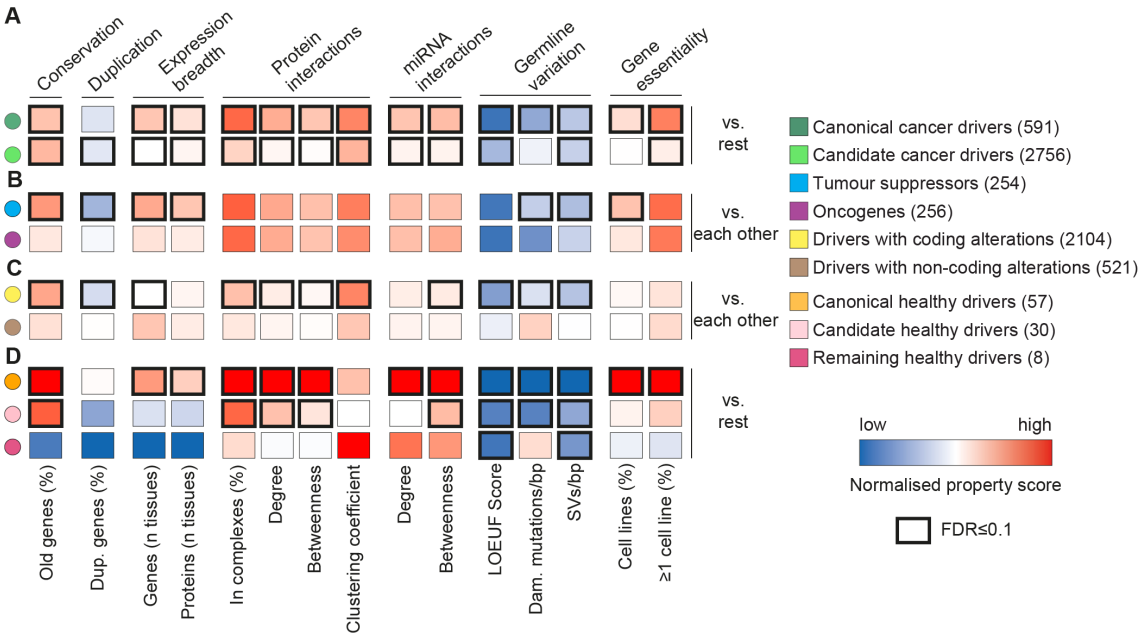

Figure 3.2 SLPs of cancer and healthy drivers

Comparison of SLPs between **A)** canonical or candidate cancer drivers and the rest of human genes; **B)** TSGs and OGs; **C)** candidate cancer drivers with only coding alterations and candidate cancer drivers with only non-coding alterations; **D)** canonical or candidate or remaining healthy drivers and the rest of human genes.

The normalised property score was calculated as the normalised difference between the median or proportion values in each driver category and the rest of human genes (see Chapter 2.1.2). Proportions of pre-metazoan (old), duplicated, proteins involved in complexes and essential genes were compared using a two-sided Fisher's exact test. Distributions of gene and protein expression, protein-protein, miRNA-gene interactions and germline variation were compared using a two-sided Wilcoxon test. The resulting p-values were corrected within each SLP using the Benjamini-Hochberg method. Figure adapted from (Dressler et al., 2022).

TSGs were older, more enriched in single-copy genes and more broadly expressed across human healthy tissues than OGs (Figure 3.2B). These differences were due to the different roles of these two groups of genes within the cell, as previously suggested (D'Antonio & Ciccarelli, 2011; Domazet-Lošo & Tautz, 2010; Michor et al., 2004). TSGs accumulated fewer germline alterations than OGs and this was further supported by the higher proportion of cancer cell lines in which TSGs were essential (Figure 3.2B).

Candidate cancer drivers with non-coding alterations showed a weaker SLP profile than candidates with coding alterations (Figure 3.2C). Particularly, candidate drivers with non-coding alterations were younger, more often present as duplicates in the human genome, less central in the PPIN and the miRNA-gene network, and accumulated more germline alterations than candidate driver genes with coding alterations. These properties suggest that candidate drivers with non-coding alterations play different roles within the cell than those accumulating alterations in coding regions.

Finally, we investigated the differences within the heterogeneous group of healthy drivers (Figure 3.2D). We used the rest of human genes as a reference for these comparisons to better characterise how healthy drivers behaved with respect to genes that were not involved in cancer or clonal expansion in healthy tissues. Strikingly, we observed that canonical healthy drivers localised at the extreme spectrum for most of the SLPs (e.g., conservation, PPIN, miRNA interactions, germline variation and essentiality) showing a much more pronounced profile than all canonical cancer drivers (Figure 3.2D). Candidate healthy drivers showed an intermediate SLP profile between that of canonical healthy and the remaining healthy drivers (Figure 3.2D). From a property perspective the remaining healthy drivers had a completely different SLP profile from that of all other driver

categories (Figure 3.2D). Since we only identified eight genes that belonged to this category, the differences we observed will have to be confirmed on a larger and more robust group of remaining healthy drivers.

3.4 Annotation of gene function and interactions with drugs

We investigated the function of the different groups of driver genes through GSEA (see Chapter 2.1.4). We found that at least 60% of enriched pathways (FDR <0.05) across all driver groups converged to five main cellular processes (i.e., signal transduction, gene expression, immune system, cell cycle, and DNA repair) (Figure 3.3A, Table 3.1). TSGs showed a convergence to cell cycle and DNA repair pathways, while oncogenes were enriched in signal transduction and immune system-related pathways (Figure 3.3A, Table 3.1). The SLPs along with the functional enrichment (Figure 3.2B, Figure 3.3) suggest that TSGs are involved in the maintenance of the basic cellular machinery, such as mismatch repair mechanisms and cell cycle checkpoints. Oncogenes, on the other hand, are preferentially involved in regulatory functions such as signal transduction and the regulation of the immune system.

Candidate cancer drivers showed a functional profile closely resembling that of canonical cancer drivers with few exceptions such as extracellular matrix organisation (Figure 3.3A).

We investigated the functional enrichment of the two categories of candidate drivers individually. Interestingly, the candidate drivers with non-coding alterations were not enriched in any pathway (Table 3.1). The candidate drivers with coding alterations showed functional enrichment that closely resembled that of candidate drivers altogether (Figure 3.3A). However, the number of pathways enriched in candidate cancer drivers was lower than the number of pathways enriched in candidate drivers with coding alterations (Figure 3.3A). Given the lack of functional enrichment in candidate drivers with non-coding alterations (Table 3.1), the lower numbers of enriched pathways in all candidate drivers may be explained by the fact that drivers with non-coding alterations are contributing to the functional enrichment of all candidate cancer drivers by adding noise.

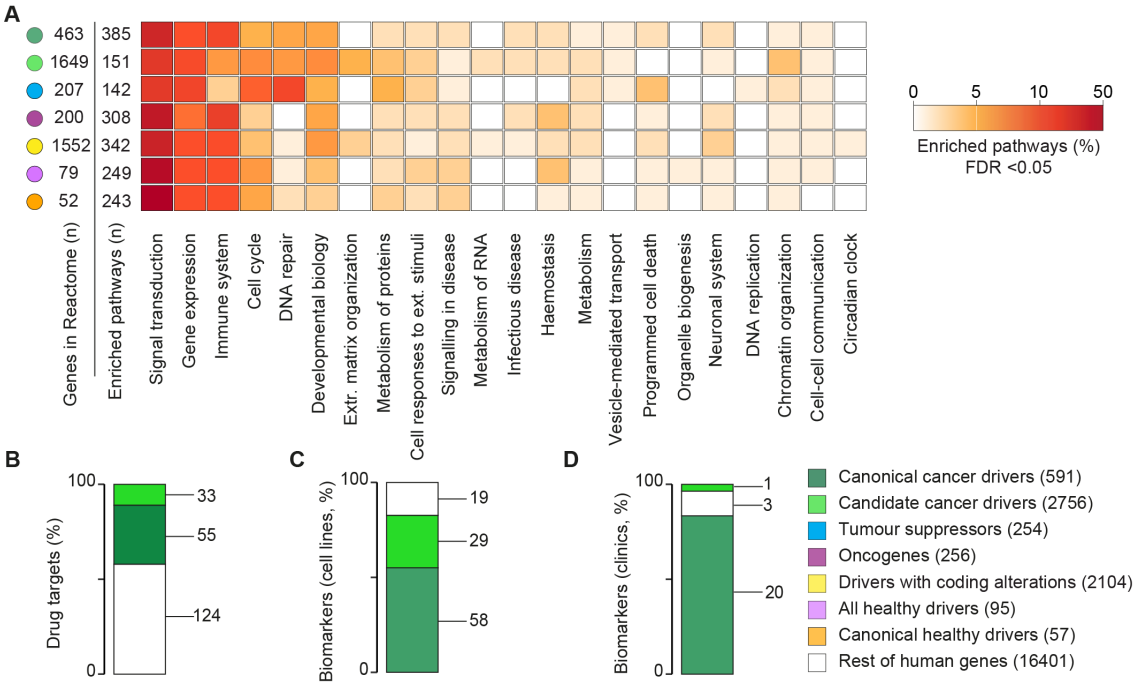

Figure 3.3 Functional and drug annotations of cancer driver genes

A) Proportion of enriched Reactome levels 2-8 pathways mapping to the corresponding level 1 in each driver category. Enrichment was measured comparing the proportion of drivers in each pathway against that of the rest of human genes using a one-sided Fisher's exact test. FDR was calculated by applying Benjamini-Hochberg correction. Proportion of canonical, candidate cancer drivers and the rest of human genes that are **B)** targets of FDA-approved antineoplastic drugs or biomarkers of response or resistance to immunomodulating and oncological drugs in **C)** cancer cell lines and **D)** clinical trials. The corresponding numbers for each group is reported. This figure was adapted from Dressler et al. (Dressler et al., 2022).

As a consequence of the high degree of overlap, the functional profile of healthy drivers closely resembled that of cancer drivers (Figure 3.3A). The candidate healthy drivers showed enrichment in only six pathways overall mapping to four level 1 Reactome pathways, namely cell cycle, haemostasis, organelle biogenesis and transport of small molecules (Table 3.1) These enriched pathways included transport of plasma lipoproteins, injury-induced platelet activation, and formation of sensory cilia. There was no functional enrichment in the remaining healthy drivers (Table 3.1), likely due to the small number of genes in this category (Figure 3.1D).

Table 3.1 Proportion of enriched pathways per driver category

For each driver category the proportion of enriched level 2-8 pathways is reported. Enriched pathways are mapped to the corresponding Reactome level 1 pathway. The enrichment was tested for all driver categories against the rest of human genes.

Reactome pathway level 1	Canonical cancer drivers (%)	Candidate cancer drivers (%)	TSGs (%)	OGs (%)	Drivers with coding alterations (%)	Drivers with non-coding alterations (%)	All healthy drivers (%)	Canonical healthy drivers (%)	Candidate healthy drivers (%)	Remaining healthy drivers (%)
Signal Transduction	35	26	25	40	37	0	46	50	0	0
Gene expression	12	16	19	10	12	0	12	12	0	0
Immune system	18	7	3	22	15	0	12	14	0	0
Cell cycle	5	8	10	3	4	0	7	6	16.7	0
DNA repair	6	7	18	0	2	0	1	2	0	0
Developmental biology	6	8	6	5	7	0	4	3	0	0
Extr. matrix organisation	0	5	0	0	3	0	0	0	0	0
Metabolism of proteins	2	4	5	2	2	0	2	3	0	0
Cell responses to external stimuli	2	3	3	2	2	0	3	2	0	0
Signalling in disease	2	1	1	2	2	0	3	3	0	0
Metabolism of RNA	0	2	0	0	1	0	0	0	0	0

Infectious disease	2	2	0	2	1	0	0	0	0	0
Haemostasis	2	2	0	4	2	0	4	1	50	0
Metabolism	1	2	2	2	2	0	1	1	0	0
Vesicle-mediated transport	1	1	0	1	0	0	0	0	0	0
Programmed cell death	2	0	4	1	2	0	1	1	0	0
Organelle biogenesis	0	0	0	0	0	0	1	0	16.7	0
Neuronal system	2	1	0	2	3	0	1	1	0	0
DNA replication	0	0	1	0	0	0	0	0	0	0
Chromatin organisation	1	4	2	1	1	0	1	1	0	0
Cell-cell communication	1	1	1	1	1	0	1	0	0	0
Circadian clock	0	0	0	0	1	0	0	0	0	0
Transport of small molecules	0	0	0	0	0	0	0	0	16.7	0

Finally, we studied the interaction between genes and oncological drugs. Specifically, we investigated the gene categories that were frequently targeted by anti-cancer drugs (Figure 3.3B). We did not differentiate between targeted therapy and more traditional treatment such as chemotherapy, and as a result we observed that the majority (58%) of drug targets fell within the category of the rest of human genes (Figure 3.3B). On the other hand, most of the biomarkers of resistance or response to oncological treatment in cancer cell lines (82%, Figure 3.3C) or clinical trials (88%, Figure 3.3D) were cancer driver genes with a large prevalence of canonical cancer drivers.

All SLPs, along with the gene functional annotation and interactions with drugs, are available online to the cancer research community at http://network-cancer-genes.org/. Reda Keddar, Hrvoje Misetic, Michele Bortolomeazzi, and Lisa Dressler curated the implementation of the website.

3.5 Conclusion

The SLP profile shows that driver genes involved in tumorigenesis have distinctive characteristics. Based on such properties we were able to distinguish cancer driver genes from the rest of human genes. Within the cancer driver repertoire we observed heterogeneity between TSGs and OGs. The evolutionary and functional analyses showed that TSGs and OGs represent two distinct groups of canonical cancer drivers. Not only do they contribute to cancer in different ways, they also evolved through different paths and play distinctive roles within human cells.

Interestingly, canonical healthy drivers show the most extreme property profile within the driver categories we analysed. This could support the evidence that not all driver genes have the same effect on promoting tumorigenesis. It has been previously suggested that cancer driver genes can be distributed across a gradient of oncogenic potential (Davoli et al., 2013; Grossmann et al., 2020). At one end of the gradient, driver genes have a very strong potential for promoting tumorigenesis and, hence, are termed super-drivers (Grossmann et al., 2020). Those at the opposite end have a lower oncogenic potential (Davoli et al., 2013).

Based on this reasoning, it is tempting to speculate that the 57 canonical healthy driver genes represent a subset of super-drivers given their marked property profile. The remaining healthy drivers are represented by such a small number of genes on which currently, any conclusion regarding their evolutionary history and cellular role would be premature.

From a functional perspective, we observed three top pathways enriched across all categories of drivers at similar proportions: signal transduction, gene expression and immune system. However, within some groups of drivers, the functional profile might be diluted by the presence of genes whose role within the cell has not yet been clearly defined, such as candidate drivers with non-coding alterations.

From a treatment perspective, many targets of anti-cancer drugs are non-cancer genes. This is mainly explained by the fact that some traditional treatments, such as chemotherapy, do not rely on the presence of alterations within their targets, even though they target different aspects of the cellular machinery. On the other hand, most known mechanisms of resistance or response to oncological drugs are mediated by cancer driver genes because this group of genes has been investigated more thoroughly in the context of carcinogenesis.

The advantage of SLPs is that they allow us to distinguish genes involved in the initiation and development of cancer from the pool of genes that are not involved in cancer. This advantage can be employed in cancer prediction tools for the prioritisation of cancer driver genes in patient cohorts. Results are particularly insightful when applied to cancer types that are genetically heterogenous and whose driver repertoire is not fully explained by acquired alterations in canonical cancer drivers. The use of SLPs in prioritising cancer driver genes can be applied in order to derive a comprehensive list of cancer driver genes in individual patients regardless of the size of the cohort under investigation (Mourikis et al., 2019; Nulsen et al., 2021).

Chapter 4. Resolving OAC genetic inter-tumour heterogeneity

4.1 Motivation

OAC is genetically heterogeneous (Contino et al., 2017). This means that across patients only a handful of genes are commonly predicted as cancer drivers and many tumours are left with no drivers to explain the presence of the disease in the corresponding patient. However, identifying cancer driver genes is key for the clinical treatment of patients. The field of precision oncology relies heavily on the identification of cancer driver genes in order to select the most appropriate therapy to treat individual patients (Malone et al., 2020).

Cancer driver genes are those genes that, upon acquiring driver alterations, sustain tumour growth, promote local invasion and distant metastasis. Driver alterations confer a selective advantage to cells that harbour these alterations over cells that do not. The resulting selective advantage enables tumour cells to produce more daughter cells than their neighbours through mechanisms that result in resistance to apoptosis or accelerated proliferation. Additionally, driver alterations can favour cancer cells by reprogramming their cellular metabolism and avoiding immune destruction (Hanahan, 2022).

The limitation in identifying the complete repertoire of cancer driver genes in the sample cohort under investigation is undoubtedly due to the cohort size and the tools used to identify them (see Chapter 1.3.1). One common approach to prioritise cancer driver genes is to identify recurrent mutations within their genetic sequence.

The study of tumour evolution and the identification of cancer driver genes have inherited fundamental concepts from the field of evolutionary biology. Mutations are the source of new variation within a population and they exist in three different forms: neutral, deleterious, and beneficial (Gregory, 2009). Beneficial mutations can be rare and provide only a minor advantage. But, over time, the proportion of these beneficial and heritable traits within the population increases due to a process known as natural selection. Mutations that confer a fitness advantage

are thus selected for and become predominant over time, whereas the deleterious ones disappear due to their negative outcome.

In cancer this process of selection occurs much faster than in the evolution of species for two reasons. Firstly, the replication time of cells is much faster than the reproduction of entire individuals. Secondly, somatic mutations represent the main driving force in cancer. They occur frequently, especially when cells are exposed to mutagens, and can become fixed in a few replication cycles via clonal expansion (Nowell, 1976).

Recurrence-based methods rely on this idea: if a mutation gives a selective advantage to tumour cells it will be selected for and will appear more often than usual in sample cohorts (Martincorena et al., 2017). The larger the cohort is, the greater the statistical power available to identify cancer genes harbouring less frequent driver mutations. These approaches have proved particularly useful in defining our current knowledge of genes that drive cancer (Dressler et al., 2022; Sondka et al., 2018; Vogelstein et al., 2013). However, most of the genes reported in these repositories are those that accumulate driver mutations most frequently.

Higher mutation rate, after correcting for confounding factors such as the level of gene expression, replication time and the mutation rate of the genomic region, means higher selective advantage (Davoli et al., 2013). However, it has long been acknowledged that the genomic landscape of most tumours is dominated by cancer driver genes altered in 5% or less of samples (Wood et al., 2007). Driver mutations in such genes result in a lower selective advantage but are key to understanding tumour initiation, development, and expansion to other organs. Additionally, in some cancer types the mutational landscape is further hampered by extended driver heterogeneity.

In the context of OAC, even the most comprehensive genetic study so far, which comprised 551 OAC samples and the employment of multiple driver detection methods, failed to identify cancer driver genes across all patients (Frankell et al., 2019). In the study, the authors combined seven driver detection methods that resulted in the prioritisation of 77 driver genes affected by coding and non-coding mutations and CNAs. Despite the large cohort size and the combination of

multiple cancer driver detection methods, 50% of the samples had less than five driver events (Frankell et al., 2019), a commonly accepted number of driver events to explain the development of the disease (Martincorena et al., 2017).
For this reason, we compiled a large cohort of OAC and BO samples and investigated their heterogeneity by applying sysSVM2 (see Chapter 1.3.3). The tool, instead of relying on the frequency of alterations, uses SLPs to prioritise cancer driver genes. As shown above (see Chapter 3.3), we can exploit these properties to differentiate between cancer driver genes and the rest of human genes. This approach allows the prioritisation of frequently altered cancer drivers and of genes that rarely accumulate driver alterations and contribute to cancer in very few or individual tumours.

4.2 Curation of a comprehensive cohort of BO and OAC cases

We assembled a cohort of 748 BO and OAC samples derived from 671 patients, combining various sources (Figure 4.1A):

- 489 samples obtained from ICGC-OCCAMS (source: https://www.occams.org.uk/);
- 73 samples obtained from TCGA (source: https://gdc.cancer.gov/; Grossman et al., 2016);
- 186 samples obtained from two NGS screens on OAC (Dulak et al., 2013; Stachler et al., 2015).

The dataset consisted of 73 BO cases (all of which progressed to cancer) and 675 OACs (Figure 4.1A). Despite all 73 BO patients in the cohort progressing to OAC, genomic data of the paired OAC were available for 70 BO cases. The average age at cancer diagnosis was 67 years and 85% of patients were male, in accordance with the strong male predominance of the disease (Xie & Lagergren, 2016).
We stratified the 748 samples into clinical stages (Figure 4.1B). We divided the 73 BO cases based on the presence of dysplasia and found 73% of them to be NDBO. We decided not to include the grade of dysplasia due to the high inter-observer variability in evaluating Barrett's dysplasia (Fitzgerald et al., 2014).

This resulted in a single stage, namely DBO, that included both low-grade and high-grade dysplasia samples. We annotated OAC samples using the 8th edition of the AJCC staging of cancers of the oesophagus and oesophagogastric junction (Rice et al., 2017). The staging system describes the size of the tumour using the TNM notation (see Chapter 1.1.2). We found that 75% of the OAC samples in our cohort were classified as Stage III or IV tumours. This is in line with the late diagnosis of the disease where 70-80% of patients are diagnosed when lymph node or distant metastases are already present (source: https://www.cancerresearchuk.org).

Given the different original sources of samples we set out to treat and analyse them consistently using the same bioinformatic pipeline (see Chapter 2.2.1; Figure 4.1C). The tools for variant calling applied to the bulk of our data, namely ICGC-OCCAMS, were benchmarked against various other available methods and had among the best sensitivity and specificity for variant calling (Ding et al., 2015). Hence, we decided to apply the same tools to the remaining sources of samples.

For the samples obtained from the NGS screens on OAC (Dulak et al., 2013; Stachler et al., 2015), we downloaded the BAM files and called in house both the CN segments and the mutations (Figure 4.1C). For the TGCA (curated by Hrvoje Misetic) and ICGC-OCCAMS samples we obtained the called CN segment and mutations and moved further to identify the genes affected by damaging alterations (see Chapter 2.2.1; Figure 4.1C). We defined as damaging those alterations (mutations or CNAs) that resulted in the activation or inactivation of the gene. Specifically, we considered four categories of damaging alterations:

- Damaging SNVs and indels;
- Homozygous deletions;
- Gene amplifications;
- Double hits, in which one of the two alleles was lost and the other harboured a damaging SNV or indel.

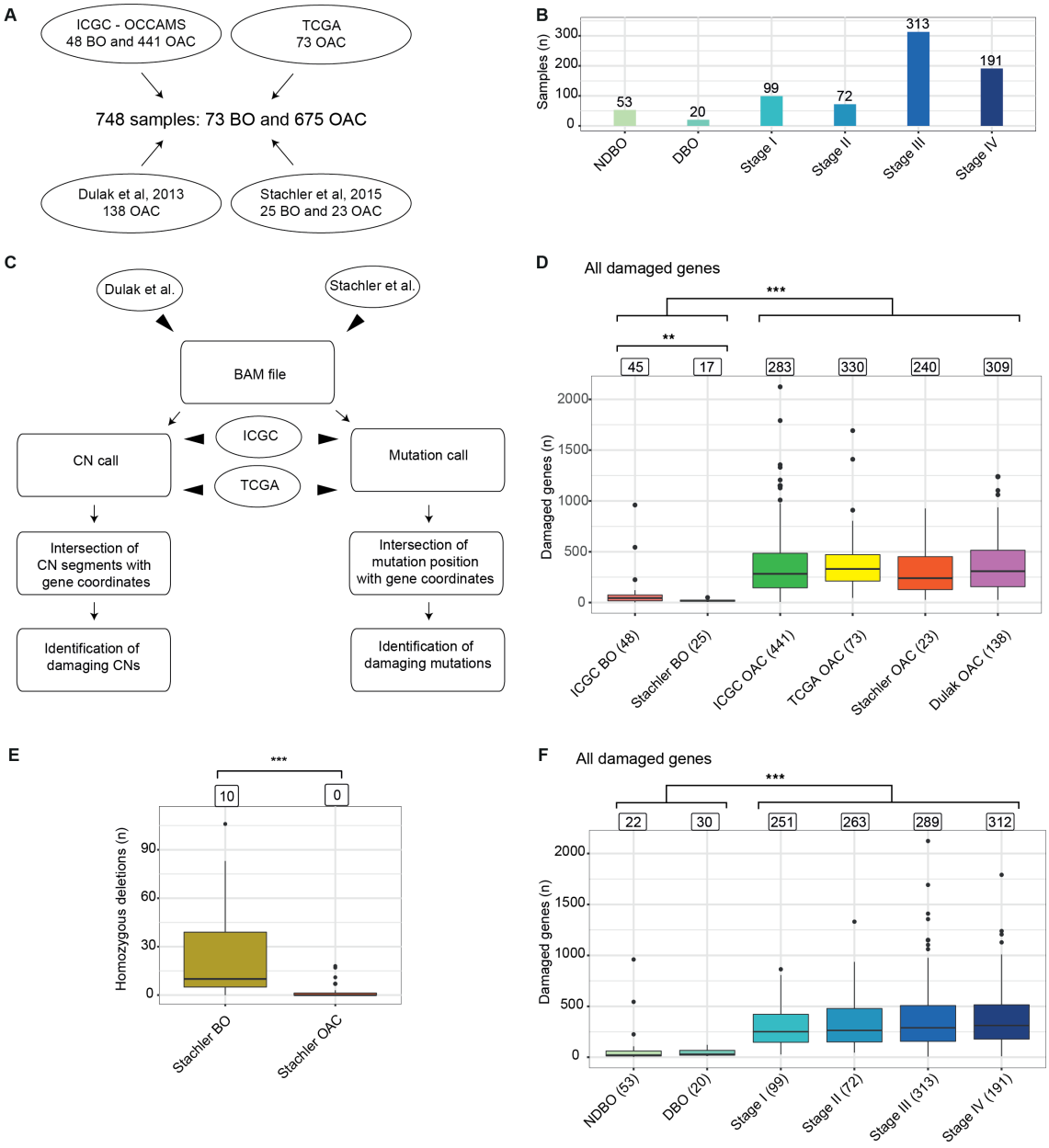

Figure 4.1 Annotation of damaged genes across sources and clinical stages

A) Original sources of BO and OAC samples. **B)** Clinical stage of the samples included in this study. The number of samples in each clinical stage is reported on top of each bar. **C)** Pipeline overview for the annotation of damaged genes. The pipeline is split into two parts that run in parallel: on one side the annotation of genes affected by damaging CNAs, while on the other side the annotation of genes affected by damaging mutations. Reported is the entry point into the pipeline for each source of samples. **D)** Distribution of damaged genes per sample across original sources. Reported on the horizontal axis is the source and the number of samples obtained from it. The numbers on top of the graphs represent the median value for the corresponding distribution. **E)** Comparison of the distribution of homozygous deletions between BO and OAC samples obtained from (Stachler et al., 2015) **F)** Distribution of damaged genes per sample across clinical stages. Reported on the horizontal axis is the clinical stage and the number of samples mapping to it. The

numbers on top of the graphs represent the median value for the corresponding distribution. In **D)** – **E)** – **F)** each pairwise comparison was done using a two-sided Wilcoxon test. The resulting p-values were corrected using the Benjamini-Hochberg method. *** FDR <0.0001, ** FDR <0.001.

We compared the number of damaged genes across original sources and found that BO samples had significantly fewer genes with damaging alterations than OAC samples (Figure 4.1D). We also observed that within the four OAC sources the number of damaged genes per sample was comparable (Figure 4.1D). However, within BO samples, the ICGC source had significantly more damaged genes than the Stachler source. This was caused by a significantly higher number of homozygous deletions per sample in the BO cases from the Stachler cohort (Stachler et al., 2015) than in the paired OAC samples (Figure 4.1E). Since the transition from BO to OAC is commonly characterised by increased aneuploidy (Nones et al., 2014; Stachler et al., 2015), we decided not to call CNAs in the BO samples from the Stachler cohort (Stachler et al., 2015). In line with what was previously reported in the literature, which showed a stable genome with very few or no copy number changes in BO cases that progressed to OAC (Ross-Innes et al., 2015), we assumed that these BO samples had a normal diploid genome without any evidence of aneuploidy.

Finally, we compared the distributions of damaged genes per sample across clinical stages. We observed that samples accumulated more damaged genes across the four OAC stages than within BO stages (Figure 4.1F). This was in line with previous evidence of increased aneuploidy in the progression from BO to OAC (Ross-Innes et al., 2015; Stachler et al., 2015) mainly due to an increase in gene amplifications (Table 4.1). Interestingly, we did not observe any increase in the distribution of the number of damaged genes per sample between early and advanced OAC stages (Figure 4.1F).

Table 4.1 Median number of damaged genes across cohorts

For each category of damaged genes (all damaging alterations, damaging mutations, homozygous deletions, amplifications, double hits) reported is the median number of damaged genes within the cohort.

Cohort	Median number of damaged genes	Damaging category
ICGC BO (48)	45	All damaging
Stachler BO (25)	17	All damaging
ICGC OAC (441)	283	All damaging
TCGA OAC (73)	330	All damaging
Stachler OAC (23)	240	All damaging
Dulak OAC (138)	309	All damaging
ICGC BO (48)	17	Damaging mutations
Stachler BO (25)	17	Damaging mutations
ICGC OAC (441)	31	Damaging mutations
TCGA OAC (73)	35	Damaging mutations
Stachler OAC (23)	20	Damaging mutations
Dulak OAC (138)	22	Damaging mutations
ICGC BO (48)	2	Homozygous deletions
Stachler BO (25)	0	Homozygous deletions
ICGC OAC (441)	4	Homozygous deletions
TCGA OAC (73)	1	Homozygous deletions
Stachler OAC (23)	2	Homozygous deletions
Dulak OAC (138)	4	Homozygous deletions
ICGC BO (48)	1	Amplifications
Stachler BO (25)	0	Amplifications
ICGC OAC (441)	223	Amplifications
TCGA OAC (73)	280	Amplifications
Stachler OAC (23)	215	Amplifications
Dulak OAC (138)	250	Amplifications
ICGC BO (48)	0	Double hits
Stachler BO (25)	0	Double hits
ICGC OAC (441)	0	Double hits
TCGA OAC (73)	0	Double hits
Stachler OAC (23)	2	Double hits

Dulak OAC (138)	1	Double hits

4.3 OAC genetic inter-tumour heterogeneity

We then set out to investigate the current knowledge on OAC-specific cancer drivers and analyse how frequently these drivers were altered in the 748-sample cohort that we assembled and curated.

Since this project was started prior to the release of the seventh update of the NCG database, we used the, then current, version of the database, NCG6, as a reference (Repana et al., 2019). The NCG database, in addition to reporting cancer driver genes and their SLPs, contains information regarding the list of cancer drivers associated with individual cancer types. We additionally used NCG6 to derive the pan-cancer list of canonical cancer driver genes and SLPs for all human genes.

From NCG6, we retrieved nine canonical cancer driver genes that had been previously reported to be involved in OAC tumorigenesis and whose role in promoting cancer was experimentally validated (Figure 4.2A). NCG6 contains information on the mutational drivers predicted in six publications on OAC published between 2012 and 2017 (Agrawal et al., 2012; Dulak et al., 2013; Fels Elliott et al., 2017; Kim et al., 2017; Secrier et al., 2016; Weaver et al., 2014). Shortly after we obtained this list, the most comprehensive study on OAC driver genes was published (Frankell et al., 2019) and we decided to include it in order to have a thorough understanding of the driver landscape in OAC. The curation of the drivers predicted overall in these seven screenings resulted in 48 OAC-specific canonical cancer drivers (Figure 4.2A). Given the role of aneuploidy in driving OAC progression (Stachler et al., 2015) and the fact that NCG6 reports only mutational cancer drivers, we decided to include OAC-specific cancer driver genes that are altered via CNAs such as amplifications or homozygous deletions. We curated four additional screenings that focused specifically on the study of CNAs in OAC (Dulak et al., 2012; Frankel et al., 2014; Murugaesu et al., 2015; Nones et al., 2014). This resulted in 29 new canonical cancer driver genes that were identified by exclusively focusing on drivers affected by CNAs

(Figure 4.2A).

The final and, to our knowledge, most comprehensive list of OAC-specific canonical cancer driver genes we obtained resulted in 77 genes that were shown to be involved in OAC tumorigenesis. Although Frankell et al. reported the same number of OAC driver genes (Frankell et al., 2019), the two lists differ from each other in that we only included drivers with robust experimental evidence supporting their role in cancer, namely canonical.

The largest proportion of the 77 OAC-specific canonical cancer drivers were represented by TSGs (49%), followed by 35% OGs, with the remaining 16% being undefined canonical drivers with no clear understanding of how they mechanistically promote tumorigenesis (Figure 4.2B). We noticed that 66% of the 77 drivers were predicted as drivers in only one of the 11 screenings that we annotated (Figure 4.2C). The most frequently predicted genes were *CDKN2A* and *SMAD4,* reported as drivers in nine and eight OAC-specific screenings respectively (Figure 4.2). Conversely, *TP53*, the most frequently altered gene in OAC (Contino et al., 2017), was predicted as a cancer driver gene in six screenings.

We mapped the 77 OAC-specific canonical cancer driver genes to our cohort of 748 samples in order to investigate the prevalence and frequency of their damaging alterations. We found these genes to be damaged 2690 times in 715 samples. Interestingly, three OAC-specific canonical cancer drivers were never damaged in our cohort. These three genes were all reported as TSGs, namely *FAM46C*, *KLF6* and *SDHB*, and were identified exclusively as drivers affected by CNAs. Due to the limitations in the current methods for the identification of drivers with CNAs (see Chapter 1.3.1), it is possible that, even though they were reported as the putative driver genes affected by the recurrent genomic loss (Murugaesu et al., 2015), they were not the driver genes targeted by the loss of the genomic locus.

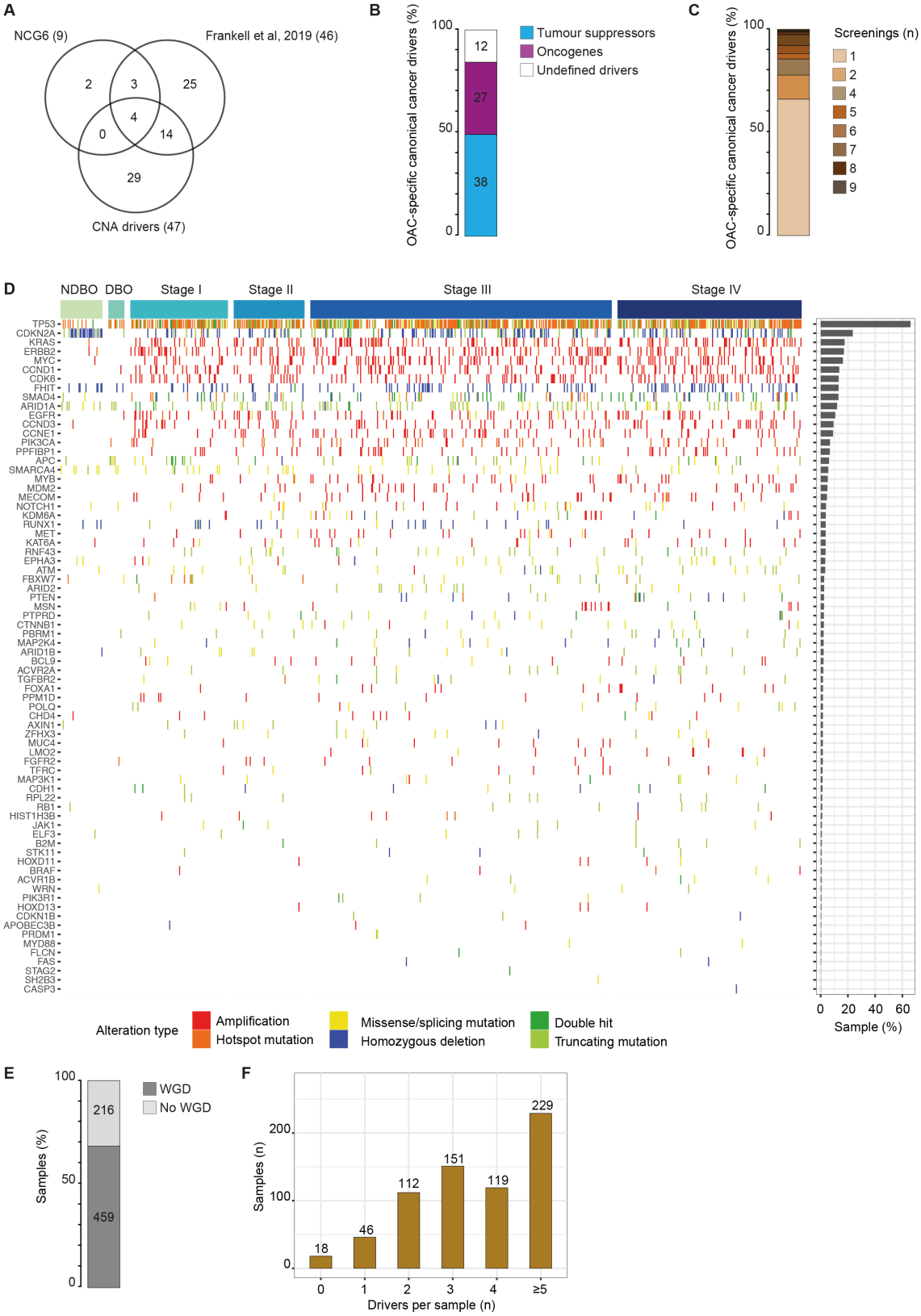

Figure 4.2 The driver gene landscape of OAC based on current knowledge

A) Sources of OAC-specific canonical cancer driver genes with the corresponding number derived from each of them. **B)** Classification of OAC-specific canonical

cancer drivers in TSGs, OGs and unclassified. Unclassified drivers are genes with dual roles in cancer or no clear evidence regarding their mechanism of action in the context of tumorigenesis. **C)** Level of support for the OAC-specific canonical cancer drivers. The levels refer to the number of OAC-specific screenings that predicted the gene as cancer driver. **D)** List of the alterations present in the 77 OAC-specific canonical drivers in individual samples. The colour of each cell represents the type of alteration affecting the corresponding gene in the corresponding sample. On the right, the percentage of total samples in which the gene is altered is shown. The 748 samples are divided into clinical stages. **E)** Proportion of OAC samples with and without whole-genome doubling (WGD). **F)** Distribution of OAC-specific canonical cancer driver genes in the 675 OAC samples.

Unsurprisingly, the most frequently damaged gene was *TP53*, damaged in 66% of our cohort (494/748) (Figure 4.2D). *CDKN2A* was the second most damaged gene with alterations present in 25% of samples. All of the remaining canonical cancer driver genes were damaged in less than 20% of samples, with 56 genes damaged in 5% or less of samples (Figure 4.2D). Interestingly, *CDKN2A* alterations were more common in precancerous stages than across OAC clinical stages. Specifically, we found *CDKN2A* altered in 45% of NDBO cases (24/53) whereas across OAC clinical stages we found it altered in 16% to 24% of the cases.

We observed that the vast majority of the most frequently altered drivers were OGs, affected by amplifications (Figure 4.2D), although proportionally OGs were less represented than TSGs (Figure 4.2B). The high frequency of amplifications in OAC-specific canonical drivers might be explained by the common presence of WGD in OAC. We found that 68% of our 675 OAC samples underwent at least one WGD event (Figure 4.2E). In this context, it is probably easier for a cancer lesion to accumulate further aneuploidy in the form of amplifications rather than losing genes given the duplicated number of gene copies present in WGD samples. Additionally, mutations in *TP53* were reported as favouring the proliferation of cancer cells with WGD (Quinton et al., 2021). The high prevalence of *TP53*-inactivating mutations in our cohort is likely to explain the high incidence of WGD, hence further aneuploidy in the form of gene amplifications.

Finally, we looked at the distribution of OAC-specific canonical cancer driver genes across the 675 OAC cases. We observed some level of variability in

terms of the number of canonical drivers damaged in individual samples (Figure 4.2F). The median number of damaged OAC-specific canonical drivers per sample was four, with one sample accumulating 14 damaged OAC-specific canonical driver genes. We observed that 26% of OAC cases had less than three drivers (Figure 4.2F), demonstrating an urgent need to complete the list of drivers in a significant proportion of OAC samples.

4.4 OAC-specific training of sysSVM2

In order to complete the list of driver genes in each sample, we used sysSVM2 (Nulsen et al., 2021). sysSVM2 is trained on canonical cancer drivers but the exact composition of the training set is dependent on the goal of the experiment. In order to find the optimal SVM models for OAC, we tested three settings of sysSVM2. The settings corresponded to three different lists of canonical drivers that we used to optimise the SVM parameters through cross-validation.

We aimed to investigate how large, non OAC-specific lists of canonical drivers performed as compared to small, OAC-specific canonical drivers in the prioritisation of cancer driver genes in OAC. We used 711 canonical cancer drivers divided into 239 TSGs, 239 OGs and 233 drivers with an undefined role as derived from NCG6 (Repana et al., 2019). In addition to this, we included a list of OAC-specific drivers which did not include CNA drivers (Figure 4.2A). Different lists of canonical drivers resulted in different training and prediction sets across the three settings.

We tested three groups of driver genes as training set (Table 4.2):

- 48 OAC-specific canonical cancer drivers, derived from the union of NCG6 and Frankell et al. (Frankell et al., 2019) (setting 1),
- 239 TSGs and 239 OGs from NCG6 (setting 2),
- All 711 canonical cancer drivers from NCG6 (setting 3).

Table 4.2 Training and prediction sets for sysSVM2 under the three settings

	Training set	Prediction set

	# samples	# genes (redundant)	# genes (unique)	# samples	# genes (redundant)	# genes (unique)
Setting 1	712	2428	48	748	266421	18883
Setting 2	738	7358	455	748	261491	18476
Setting 3	741	11525	683	748	257324	18248

sysSVM2 used the molecular and systems-level features of these three sets of genes to train the SVMs based on the four kernels. To select the best models, which define the four kernels, we ran a cross-validation with 10000 iterations for each of the three settings. During the cross-validation, we evaluated the sensitivity of each kernel to predict canonical drivers. We identified the best four models (one for each of the four SVMs) as those with the highest average and lowest standard deviation of sensitivity to retrieve canonical drivers in multiple iterations of cross-validation (see Chapter 2.3.1).

Using these best models, we used the whole training sets to train and predict on the corresponding prediction sets (Table 4.2). To select the best setting of the three, we investigated their predictions in terms of performance and stability. In order to investigate the performance of the three settings, we used the AUROC curve and the composition score (Nulsen et al., 2021) (see Chapter 2.3.2). These two metrics enabled us to evaluate the prevalence of different gene categories in top ranked positions. To evaluate the stability of the predictions across the three settings, we measured the degree of overlap of top ranked genes in pairwise comparisons and the RBO score (Nulsen et al., 2021) (see Chapter 2.3.2). The RBO score evaluates the overlap between two lists keeping track of the position where it occurs.

Settings 1 and 2 were comparably able to distinguish canonical cancer drivers not included in the training set from non-cancer genes and false positives (Figure 4.3A). We could not evaluate this metric in Setting 3 as the training set included all canonical cancer drivers. Although the performance decreased, sysSVM2 was also able to separate candidate cancer genes from false positives across all three settings we tested (Figure 4.3A).

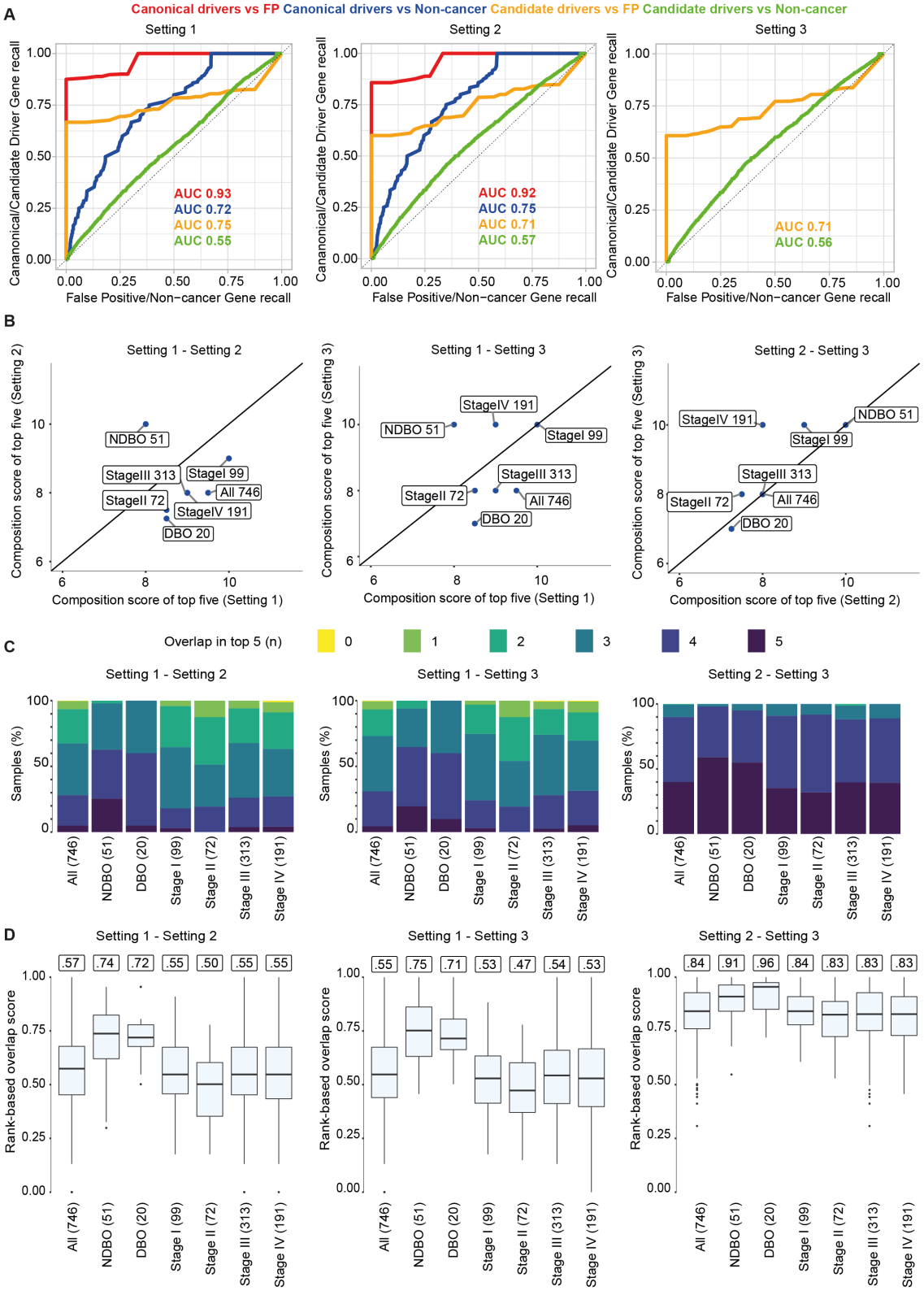

Figure 4.3 Evaluation of the three sysSVM2 settings

A) AUROC curve comparing canonical drivers to false positive (red), canonical drivers to non-cancer genes (blue), candidate drivers to false positive (yellow), candidate drivers to non-cancer genes (green). Recall rates were calculated for each

sample within each setting separately and the median AUROC curve across samples is plotted. Median AUROC curve for all comparisons is indicated. **B)** Median composition scores of the top five predictions in terms of canonical drivers, candidate drivers, non-cancer genes and false positives in the pairwise comparisons of the three settings. **C)** Distribution of the number of top five predictions shared between settings. The overlap was calculated between each pair of predictions in each sample. The colours reflect the number of genes overlapping between settings. **D)** RBO score of the top five predictions in each sample. For each distribution the median value is reported on top of the graph. **B) – C) – D)** Samples were divided into clinical stages and also treated as a single group. The number of samples per stage is reported.

We then investigated the composition of the top five predictions in order to understand which gene categories were prioritised across the three settings. Top-ranking predictions in setting 1 had a higher composition score overall than those in settings 2 and 3 (Figure 4.3B). By contrast, settings 2 and 3 had similar composition scores across stages and altogether. Since canonical and candidate cancer genes were weighted more than non-cancer genes and false positives, this meant that cancer-related genes ranked higher than non-cancer genes in setting 1 over the other two (Figure 4.3B).

Regarding stability, we compared the top five predictions of each set for each sample. Although setting 2 and 3 shared more genes scoring among the top five predictions with each other than with setting 1 (Figure 4.3C), 90% of samples overall shared at least two predictions between setting 1 and setting 2 or 3 (Figure 4.3C). Furthermore, we observed that across settings 2 and 3 most of the drivers were predicted in the same order (Figure 4.3D). Instead, setting 1 showed more variability when compared to the other two settings with 50% of drivers in common and predicted in the same order across stages (Figure 4.3D).

Based on these comparisons, we chose the OAC-specific training set (setting 1) because, although it had less predictions shared with the other two (Figure 4.3C-D), it showed the highest performance in terms of prioritising cancer-related genes among the top five predictions (Figure 4.3B). Additionally, the choice of using OAC-specific canonical drivers resulted in the training of the sysSVM2 bearing a closer resemblance to the original driver landscape of OAC.

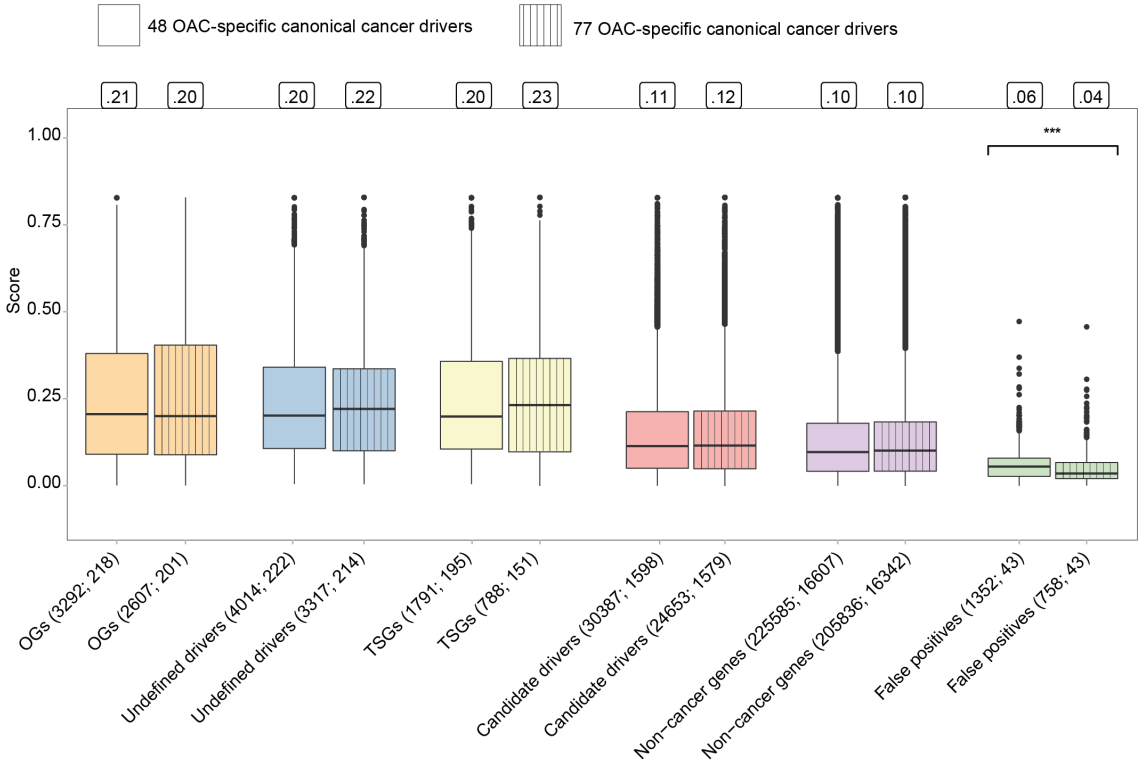

Figure 4.4 sysSVM2 score of different gene categories

Shown are the distributions of sysSVM2 scores assigned to the six gene categories present in the prediction sets. For each category the score distribution is reported for the OAC-specific training done on the 48 canonical cancer drivers (no fill pattern) and on the 77 canonical cancer drivers (square fill pattern). The genes are divided into OGs, canonical cancer drivers with undefined role, TSGs, candidate cancer drivers, non-cancer genes and false positives. The parentheses indicate the number of redundant and unique genes, respectively, for each category present in the prediction set. The median value is reported above each plot. Each pairwise comparison was done using a two-sided Wilcoxon test. *** p-value <0.0001.

We later compiled a more comprehensive list of OAC-specific canonical cancer drivers that included genes altered via CNAs and resulted in 77 genes in total (Figure 4.2A). We trained sysSVM2 using this more comprehensive knowledge of OAC and evaluated it with respect to the previously chosen OAC-specific setting based on the initial 48 canonical drivers. The parameters that defined the best four models converged to the same values in the two settings (see Chapter 2.3.1; Table 4.3). Additionally, we found that under both settings canonical cancer drivers that were part of the prediction sets were scored higher than candidate cancer drivers, non-cancer genes, and false positives (Figure 4.4). The 77 OAC-specific canonical drivers setting assigned a lower score to false positives than

the 48 OAC-specific canonical drivers setting. Based on this evidence, we continued our analysis using the training of sysSVM2 done on the comprehensive list of 77 OAC-specific canonical drivers.

Table 4.3 Best model parameters

Reported are the values describing the best models for each SVM-parameter combination for which an optimisation was done. The values reported show the parameters selected when training sysSVM2 on the 48 and on the 77 OAC-specific canonical cancer drivers.

SVM	Parameter	Value under 48 genes	Value under 77 genes
Linear	ν	0.05	0.05
Polynomial	ν	0.05	0.05
	d	2	2
Radial	ν	0.05	0.05
	γ	0.0078125	0.0078125
Sigmoid	ν	0.05	0.05
	γ	0.25	0.25

4.5 Conclusion

To our knowledge, we compiled the most comprehensive cohort of BO and OAC samples to date and annotated their damaged genes. Given the variety of sources from which we obtained these samples, we annotated their damaged genes consistently using the same bioinformatic pipeline. This resulted in comparable distributions of the number of damaged genes per sample within OAC sources. Within BO sources, we observed some level of variability due to the inability of calling reliable CNAs in the samples obtained from Stachler et al. (Stachler et al., 2015). We found a higher number of homozygous deletions in the Stachler cohort of BO cases than in the paired cohort of OACs. Since the literature reported evidence of low CNAs in BO and high CNAs in OAC (Ross-Innes et al., 2015), we decided to assume that these BO cases were diploid with no CNAs in their genome.

We stratified BO and OAC samples into clinical stages, classifying most of the tumours as advanced due to the presence of lymph node or distant metastasis. Indeed, the late diagnosis of the disease is one of the causes that hampers the successful treatment of OAC (Contino et al., 2017). We observed that OAC samples accumulated more damaged genes compared to BO samples. This higher level of instability was mainly driven by the acquisition of gene amplifications as the tumour develops (Nones et al., 2014; Ross-Innes et al., 2015). Interestingly, we did not observe any differences in the level of instability across OAC stages, suggesting that no new genetic drivers are involved in acquiring the ability to metastasise. However, sub-clonal drivers in the primary lesion may become clonal and promote metastasis, especially in treated tumours (Hu et al., 2020).

We compiled a comprehensive list of canonical cancer driver genes that were previously reported to be involved in OAC tumorigenesis by integrating 11 OAC-specific screenings. By mapping this comprehensive knowledge to our cohort of 748 BO and OAC samples we confirmed that OAC is indeed heterogeneous. *TP53* was the most frequent driver with alterations in 66% of the samples. It was followed by *CDKN2A* altered in 25% of the samples. Interestingly, *CDKN2A* showed some level of stage-specificity in that it was more frequently altered in the precancerous lesion, NDBO, than across OAC clinical stages. A previous study investigated the acquisition of driver alterations in a stage-specific fashion (Weaver et al., 2014). However, this study did not report *CDKN2A* as one of the genes differentiating the pre-cancer from the cancer lesion. The authors observed a similar rate of *CDKN2A* alterations across BO-to-OAC progression (Weaver et al., 2014). Given that this study focused exclusively on mutation rates, this discrepancy most likely results from the missing CNA analysis.

We found that the driver repertoire across the 675 OAC samples was quite heterogeneous. The median number of driver genes per sample was four. Five drivers per tumours have been proposed to be needed to explain the presence of OAC (Martincorena et al., 2017). Under the five-driver model, 66% of our samples would require additional cancer driver genes to be predicted. Even under more conservative estimates of three cancer driver genes needed per

tumour (Tomasetti et al., 2015; Vogelstein & Kinzler, 2015), 26% of our samples did not have enough cancer driver genes to explain the presence of the disease in the corresponding patient.

Given the heterogeneity of the OAC driver gene landscape, we decided to use a cancer driver prediction tool that allows us to prioritise cancer driver genes at the individual patient level (Nulsen et al., 2021). sysSVM2 learns the molecular (damaging somatic alterations) and systems-level (SLPs) features of canonical cancer drivers and predicts as drivers the damaged genes in single patients that more closely resemble these features. We tested three settings of sysSVM2 and compared their performance and stability. Specifically, we tested three training sets, namely the OAC-specific, TSGs and OGs, and all canonical drivers settings. We selected the OAC-specific setting given its higher ability in prioritising cancer-related genes in top positions compared to the other two settings.

Chapter 5. Investigating the role and therapeutic vulnerabilities of OAC genetic drivers

5.1 Motivation

After training sysSVM2 on an OAC-specific setting of 77 canonical cancer drivers, we were able to complete the list of cancer driver genes in all 675 OAC samples we annotated. The sample-specific lists were the result of combining OAC-specific canonical cancer drivers with drivers prioritised by sysSVM2, henceforth termed sysSVM2 predictions.

Firstly, we investigated how many cancer driver genes were needed in individual OACs, aiming for the search of a clearly defined number of driver events among the altered genes. A reliable estimate on the number of driver genes needed per tumour, and in this work per OAC, is crucial for the identification of targeted therapies in individual patients and will help the cancer research community to understand how the disease initiates and develops.

We analysed two published models that described the number of drivers needed in OAC (Jeon et al., 2006; Martincorena et al., 2017). Jeon et al. took advantage of age incidence data on OAC to derive the number of rate-limiting steps, hence drivers, that described the relation between age and tumour incidence. This model estimated that three cancer driver genes were needed per OAC. By contrast, Martincorena et al. relied on identifying signals of positive selection across cancer genomes and to derive the number of events needed to explain cancer evolution. As a result, the authors found that five driver events were needed per OAC.

sysSVM2 ranks the damaged genes that are part of the prediction set in individual samples using a combined score that weights the four kernels based on their sensitivity (see Chapter 2.3.1). The score is a proxy of how closely the genes' properties resemble those of OAC-specific canonical cancer drivers used for training. Highly ranked genes have the most similar property profile to that of OAC-specific canonical drivers and will then be prioritised as cancer drivers for that patient.

Given the ranking nature of sysSVM2 predictions, we compared the two driver models by selecting different numbers of cancer driver genes per sample. We studied whether the two additional drivers prioritised exclusively under the five-driver-per-sample (namely positive selection) model were predicted to be functionally involved in OAC tumorigenesis. We looked at the pathways they perturbed to see if they converged to the pathways already altered by the three-driver-per-sample (namely age-incidence) model or were involved in novel molecular processes.

After selecting the positive selection model as the most thorough estimate on the number of drivers needed per OAC, we investigated the role of OAC-specific canonical drivers and sysSVM2 predictions in the context of OAC tumorigenesis. Furthermore, we searched for drivers in the list of sysSVM2 predictions that could be exploited as drug targets. Finally, we investigated whether there were any differences in the frequency of drivers across OAC clinical stages suggesting that some cancer driver genes are acquired at specific stages, hence are able to define clinical stage boundaries, during OAC development.

5.2 Drivers identified under the positive selection model perturb novel pathways

Based on the number of drivers needed per individual tumour according to the age-incidence and the positive selection model, the proportion of samples that required sysSVM2 predictions varied. Under the age-incidence model, which assumed three drivers per OAC (Jeon et al., 2006), 26% (176/675) of the samples required additional drivers to be found (Figure 5.1A). Within these 176 samples, 64% (112/176) required one additional driver to be predicted whereas the remaining 36% required either two or three drivers to be identified (Figure 5.1A). Under the positive selection model predicting five drivers per tumour (Martincorena et al., 2017), 66% (446/675) of the OACs required further investigation in order to complete their list of cancer driver genes (Figure 5.1B). In this context, 61% of the samples (270/446) required one or two drivers to complete their list of cancer driver genes, whereas the remaining 39% of the

samples required between three and five driver genes to be further predicted (Figure 5.1B).

We predicted 234 and 521 cancer driver genes in total under the age-incidence and the positive selection model, respectively. As a consequence of the larger number of drivers needed under the positive selection model and of the fact that we derived the drivers from the same ranked list of damaged genes in each sample (see Chapter 2.4.1), all of the 234 age-incidence drivers represented a subset of the 521 positive selection drivers (Figure 5.1C).

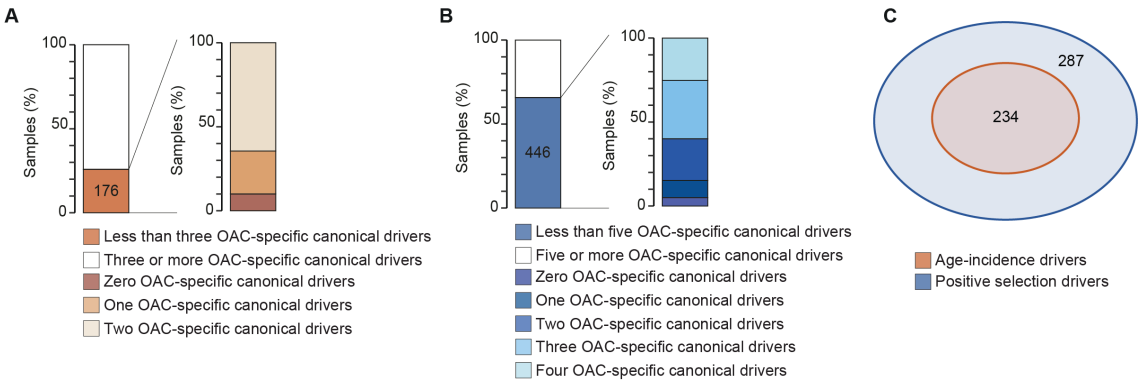

Figure 5.1 OAC samples with not enough cancer drivers

A) Proportion of OAC samples that have an incomplete repertoire of cancer driver genes under the age-incidence model. Highlighted in different brown shades are the proportions of samples that require between one and three additional predictions. **B)** Proportion of OAC samples that have an incomplete repertoire of cancer driver genes under the positive selection model. Highlighted in different blue shades are the proportions of samples that require between one and five additional predictions. **C)** Overlap between the unique sets of driver genes predicted under the age-incidence and positive selection models.

Cancer driver alterations in core oncogenic pathways have been reported to occur in a nearly mutually exclusive mode (McLendon et al., 2008). Once a gene involved in one of the hallmarks of cancer (Hanahan, 2022) is altered, the cell acquires a selective advantage. A second driver alteration, that results in the perturbation of the same pathway, is less likely to occur within the cell since it is unlikely to provide additional advantage to the cancer cell (Ciriello et al., 2012).

We reasoned that, if the two additional drivers per sample predicted exclusively under the positive selection model (termed positive selection-specific drivers) perturbed pathways already altered by the age-incidence drivers, then the

positive selection-specific drivers would be more likely to be redundant and not functionally involved in OAC initiation and development. If, on the other hand, they perturbed novel pathways that were not directly affected by the age-incidence drivers, the positive selection-specific drivers would be more likely to be functionally relevant and needed to explain the presence of the tumour in individual patients. To investigate this, we carried out a cohort-level GSEA on the three drivers per sample predicted by the age-incidence model and on the two positive selection-specific drivers separately and compared their results (see Chapters 2.4.2 and 2.4.3).

We observed minimal differences in terms of pathway perturbations between the age-incidence drivers and the positive selection-specific drivers (Figure 5.2A). Specifically, metabolism in stage I OACs and DNA repair in stage II OACs were predominantly perturbed by positive selection-specific drivers. By contrast, signal transduction in stage IV OACs and immune system in stage I OACs were preferentially affected by alterations in age-incidence drivers.

Given the redundancy of Reactome, which we used as a reference database for cellular pathways, we decided to inspect the level of overlap between the pathways enriched in the positive selection-specific drivers and those enriched in the age-incidence drivers (Figure 5.2B, see Chapter 2.4.4). We found a high level of pathway overlap between the two models. However, between 17% and 58% of the pathways enriched in the positive selection-specific drivers across OAC clinical stages were novel and independent from the age-incidence driver enrichment.

When inspecting these novel pathways, we noticed that the positive selection-specific drivers preferentially perturbed pathways related to DNA activities and replication, such as transcription, cell cycle, and DNA repair (Figure 5.2C). Specifically, we found the positive selection-specific drivers enriched in processes like chromosome maintenance, nucleotide and base excision repair, and the regulation of gene expression via non-coding RNAs and epigenetic mechanisms.

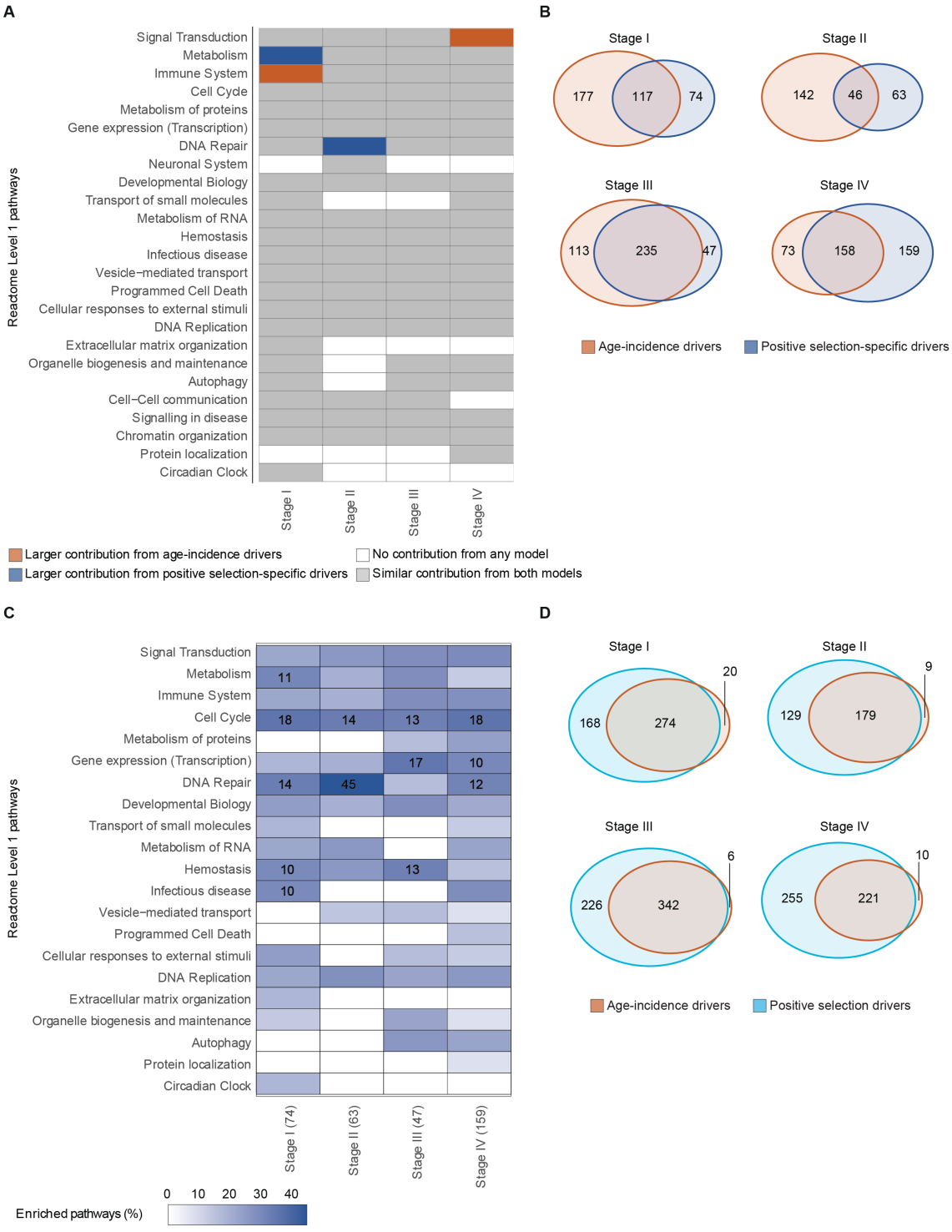

Figure 5.2 Comparison of the age-incidence and positive selection models

A) Comparison of the proportion of level 1 Reactome pathways between the age-incidence drivers (brown) and the positive selection-specific drivers (blue). Highlighted in brown and blue are the pathways in which the proportion is larger in the corresponding model (FDR <0.1, one-sided Fisher's exact test). FDR was calculated by applying the Benjamini-Hochberg correction. **B)** Overlap of the enriched level 2-8 Reactome pathways in the age-incidence drivers (brown) and those enriched in the positive selection-specific drivers (blue). **C)** Proportion of the

positive selection-specific enriched level 2-8 Reactome pathways mapping to the corresponding level 1. **D)** Overlap of the enriched level 2-8 Reactome pathways in the age-incidence drivers (brown) and those enriched in the positive selection drivers (cyan).

Finally, we wondered whether all of the pathways perturbed by the 234 age-incidence drivers were contained in the list of pathways perturbed by the 521 positive selection drivers. In order to determine this, we tested the enrichment of the two sets of drivers separately and compared their results. Across all OAC clinical stages at least 94% of the enriched pathways in the age-incidence drivers were part of the enriched pathways under the positive selection driver model (Figure 5.2D). Unsurprisingly, the vast majority of pathways perturbed by the 234 drivers predicted under the age-incidence model were part of the pathways perturbed by the 521 drivers under the positive selection model.

Based on these results, we selected the positive selection model as a thorough estimate on the number of drivers needed per OAC and discarded the age-incidence model. We decided to drop the small driver model (three driver genes per OAC) because the age-incidence model was fully contained in terms of pathway perturbations (Figure 5.2D) and driver genes (Figure 5.1C) in the larger positive selection model. The drivers predicted exclusively under the positive selection model added perturbations to novel pathways that were not affected by the age-incidence drivers, pointing towards a putative functional implication of the positive selection-specific drivers in the perturbation of cellular pathways involved in OAC development.

5.3 Frequency of cancer drivers across OAC samples

Next we investigated the identity and the frequency of the drivers across samples. Under the positive selection model we identified in total 521 unique cancer driver genes divided into 73 OAC-specific canonical cancer drivers and 448 sysSVM2 predictions.

We found OAC-specific canonical drivers prioritised 2344 times in the whole cohort making up 69% of the total drivers (Figure 5.3A). sysSVM2 predictions

made up altogether 31% of the total driver repertoire and were prioritised 1031 times (Figure 5.3A). sysSVM2 predictions were a combination of canonical drivers that had not been previously reported to be involved in OAC (18%), candidate cancer driver genes (17%), and genes that had not been previously reported to be involved in cancer (65%) (Figure 5.3A).

We then investigated the number of samples in which each OAC-specific canonical driver and sysSVM2 prediction was predicted as a driver. OAC-specific canonical drivers were more frequent across samples than sysSVM2 predictions (Figure 5.3B). More than 60% of sysSVM2 predictions were sample-specific.

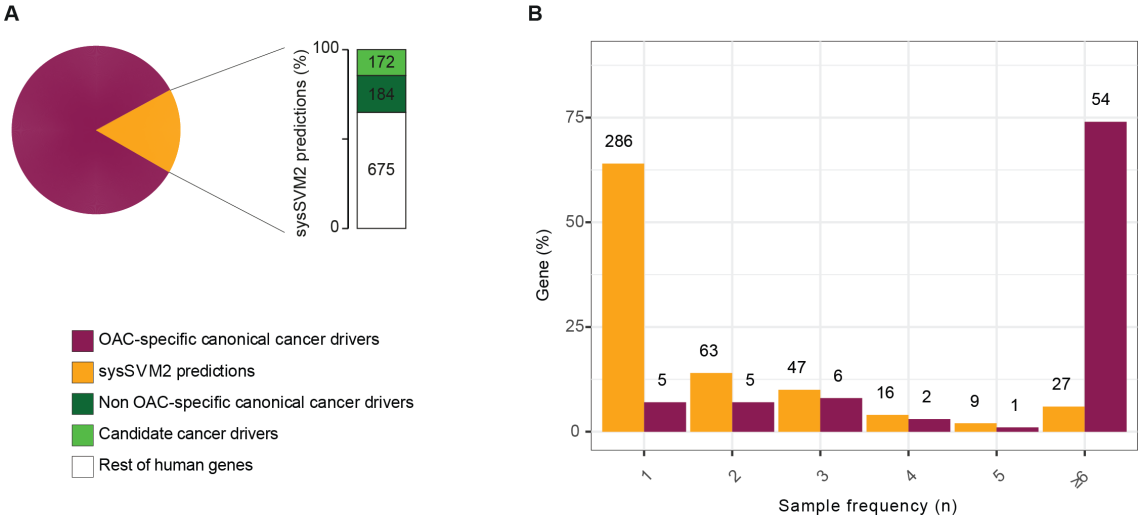

Figure 5.3 Frequency of OAC cancer drivers across samples

A) Proportion of OAC cancer drivers that are OAC-specific canonical cancer drivers (pink) and sysSVM2 predictions (orange). Highlighted in the bar plot is the proportion of sysSVM2 predictions that are non OAC-specific canonical, candidate cancer drivers and rest of human genes. **B)** Distribution of the frequency of cancer driver genes across samples. The driver genes are divided into OAC-specific canonical drivers and sysSVM2 predictions. On top of each bar the actual number is reported.

We observed some exceptions to the main trend. Some OAC-specific canonical cancer drivers were prioritised in individual samples. These comprised the TSGs *FAS*, *FLCN*, *PRDM1*, *SH2B3* and *STAG2*.

Interestingly, we found two sysSVM2 predictions prioritised as drivers in at least 5% of the samples: *SNRPD1* and *YWHAB*, predicted as drivers in 33 and 41 OAC samples respectively. Both of these genes were not previously reported to be involved in OAC tumorigenesis and were not described as cancer driver genes

before (Dressler et al., 2022; Repana et al., 2019). Both genes were altered via gene amplifications in all samples in which we predicted them as cancer drivers. *SNRPD1* encodes a spliceosome-associated protein. The spliceosome is a large cellular machinery that controls the splicing of the nuclear precursor mRNA into mature mRNA (Bonnal et al., 2012). Specifically, the spliceosome is responsible for removing the introns from the precursor mRNA and stitching together the exons into mRNA before transferring the mature mRNA into the cytoplasm for translation. Misregulation of splicing contributes to cancer progression in many ways, such as control of cell proliferation, cell death, angiogenesis, and metastasis (Bonnal et al., 2012). High *SNRPD1* expression was associated with unfavourable prognosis in breast cancer (Dai et al., 2021). The authors found that *SNRPD1* knockdown resulted in reduced cell viability and cell cycle arrest, and proposed that overexpression of the gene was needed to sustain cell cycle progression in cancer cells. A second study observed a marked reduction in cell viability in breast, lung and melanoma cancer cell lines as a consequence of *SNRPD1* depletion (Quidville et al., 2013). The authors also found that depletion of the gene resulted in cancer cell death through autophagy.

YWHAB encodes the protein 14-3-3β. The 14-3-3 family proteins are involved in many signal transduction pathways including those that regulate cell division (Fu et al., 2000). Takihara and colleagues observed that overexpression of 14-3-3β promotes cancer cell growth and tumour formation *in vivo* (Takihara et al., 2000). They also found increased activation of the mitogen-activated protein kinase (MAPK) cascade as a result of overexpression of 14-3-3β, probably due to the interaction of the protein with Raf-1 which mediates MAPK signalling. MAPK signalling is an evolutionarily conserved cascade responsible for processing extracellular signals that control multiple cellular responses, such as proliferation, apoptosis, and migration (Dhillon et al., 2007). Similarly, Sugiyama and co-authors found that reducing *YWAHB* expression in cancer cell lines endogenously expressing high levels of the protein led to reduced proliferation and the seeded tumours were smaller and histologically more benign (Sugiyama et al., 2003).

Overall, a large proportion of sysSVM2 predictions were rare or sample-specific, further highlighting the importance of tools that enable the prioritisation of cancer driver genes active in individual samples. However, we predicted two new driver genes, *SNRPD1* and *YWAHB*, which had never been reported to be involved in OAC before, in more than 5% of our cohort. Although experimental validation in OAC is needed, the driver alterations affecting these genes suggest they act as putative tumour-promoting genes.

5.4 OAC-specific canonical drivers and sysSVM2 predictions perturb different processes

We then investigated whether OAC-specific canonical drivers and sysSVM2 predictions perturbed similar or different pathways during OAC development. We tested the enrichment of drivers in individual clinical stages and divided the drivers into two sets: OAC-specific canonical drivers and sysSVM2 predictions. Across all clinical stages OAC-specific canonical drivers were preferentially perturbing pathways involved in signalling (Figure 5.4A). sysSVM2 predictions instead preferentially contributed to the alteration of immune-related and DNA repair-related pathways (Figure 5.4A). However, whilst OAC-specific canonical drivers preferentially affected signalling-related pathways across all stages, sysSVM2 predictions showed more variability. sysSVM2 predictions preferentially affected immune-related pathways in stage I, II, and IV tumours, whilst DNA repair was influenced by sysSVM2 predictions across all OAC clinical stages (Figure 5.4A).

Among the signalling-related pathways perturbed by OAC-specific canonical drivers we found signalling by activin, signalling by Erb-b2 receptor tyrosine kinase 4 (ERBB4), signalling by erythropoietin, and signalling by insulin-like growth factor 1 receptor (IGF1R) (Figure 5.4B).

Overall, these pathways contribute to processes like cancer cell proliferation and apoptosis, such as activin (Chen et al., 2006) and ERBB4 (Segers et al., 2020) signalling. IGF1R allows anchorage-independent growth of cancer cells (Maki,

2010). By contrast, erythropoietin stimulates tumour growth by favouring angiogenesis and promoting lymph node metastasis (Kimáková et al., 2017).

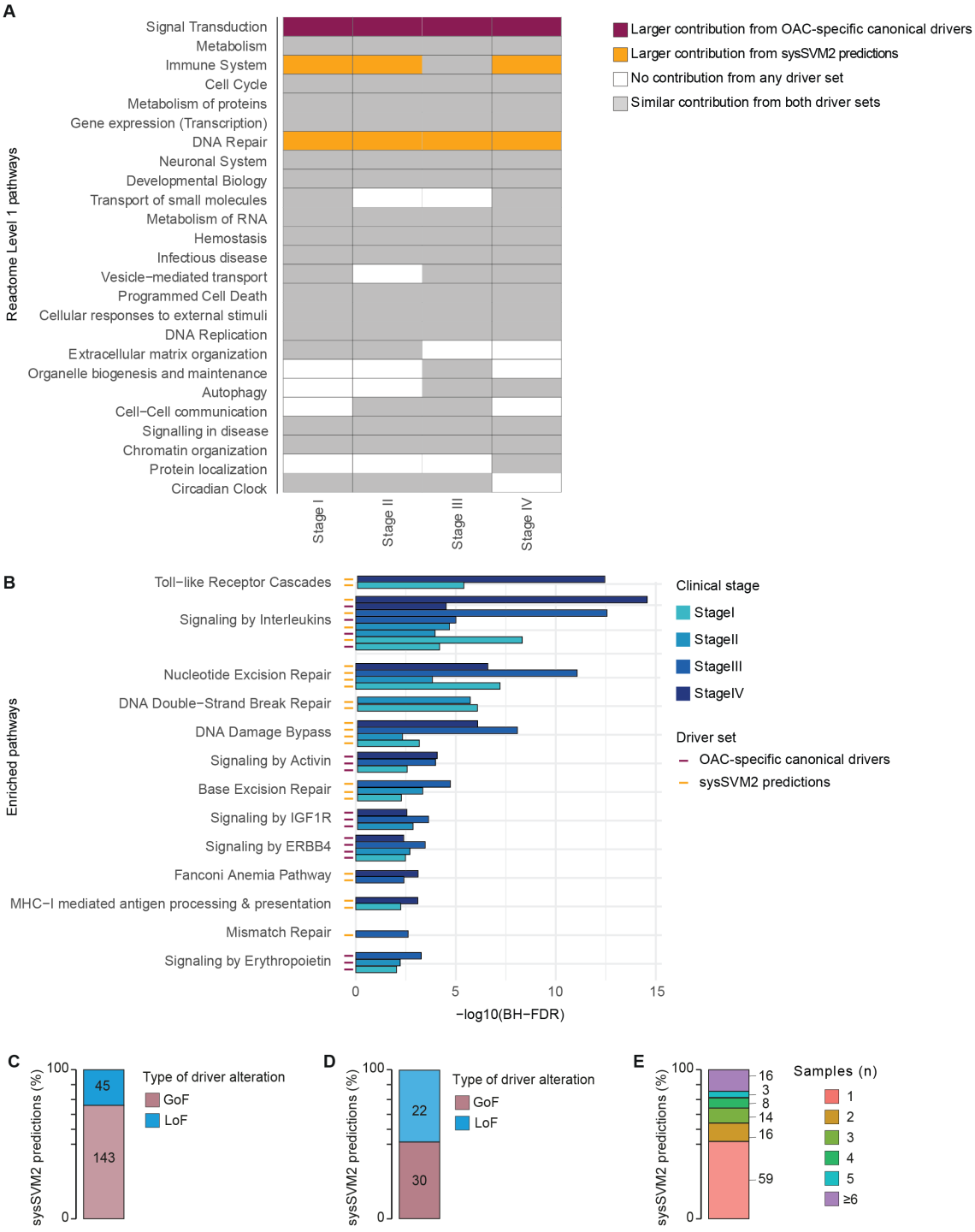

Figure 5.4 Pathways perturbed by OAC-specific canonical drivers and sysSVM2 predictions

A) Comparison of the proportion of Reactome level 1 pathways between the OAC-specific canonical cancer drivers (pink) and the sysSVM2 predictions (orange).

Highlighted in pink and orange are the pathways in which the proportion is larger in the corresponding set of drivers (FDR <0.1, one-sided Fisher's exact test). FDR was calculated by applying Benjamini-Hochberg correction. **B)** List of enriched pathways. Enrichment was measured comparing the proportion of drivers in each pathway against that of the non-driver genes. (FDR <0.1, one-sided Fisher's exact test). FDR was calculated by applying Benjamini-Hochberg correction. The colours of the bars describe the four OAC clinical stages. The colours under each bar describe the driver set, OAC-specific canonical drivers or sysSVM2 predictions, enriched in the corresponding pathway. **C)** Proportion of sysSVM2 predictions involved in inflammation affected by GoF and LoF alterations. **D)** Proportion of sysSVM2 predictions involved in MHC-I mediated antigen processing and presentation affected by GoF and LoF alterations. **E)** Proportion of immune-related sysSVM2 predictions predicted in the corresponding number of samples.

We found four DNA repair pathways perturbed exclusively by sysSVM2 predictions (Figure 5.4B). Three of these were involved in the repair of damage induced by mutagens such as oxidative reactive species or as a result of replication errors, namely base and nucleotide excision, and mismatch repair processes. sysSVM2 predictions, however, were also involved in mechanisms that tolerate unrepaired damage during genome replications, such as DNA damage bypass.

Thirty-four driver genes, predicted in 76 samples, were involved in DNA damage repair. These included one canonical cancer driver, seven candidate cancer genes, and 26 genes not reported to be involved in cancer before (Repana et al., 2019). Among them, *EP300* is a well-known TSG with experimental evidence that supports its inactivating role in promoting tumorigenesis (Iyer et al., 2004) and in contributing to genomic instability (Tini et al., 2002). Recent studies showed how cancers with mutations in *EP300* had a higher tumour mutational burden than cancers with wild-type *EP300* (Chen et al., 2021). We found that the only OAC of the 76 samples with the hypermutated phenotype, as defined by Bailey et al. (Bailey et al., 2018), had indeed *EP300* as a driver.

Lastly, we saw two different mechanisms perturbed by sysSVM2 predictions that impact the immune microenvironment surrounding the tumour (Figure 5.4B). On one hand, sysSVM2 predictions were involved in mechanisms of immune escape such as the major histocompatibility complex class I (MHC-I) mediated antigen processing and presentation. On the other hand, sysSVM2 predictions were enriched in pro-inflammatory mechanisms such as Toll-like receptor (TLR)

cascade and signalling by interleukins, with the latter pathway also affected by OAC-specific canonical drivers.

In line with evidence reported in other cancer types (Mantovani et al., 2008), we found that 76% of the sysSVM2 predictions involved in pro-inflammatory mechanisms were altered via GoF alterations underlying potential oncogenic mechanisms of these genes (Figure 5.4C). By contrast, 42% of the sysSVM2 predictions enriched in MHC-I mediated antigen processing and presentation were damaged via LoF alterations, such as *HLA-B*, suggesting a higher involvement of genes with tumour-suppressive roles in mechanisms of immune escape (Figure 5.4C). Although cancer cells preferentially inactivate TSGs in order to evade the immune system (Martin et al., 2021), 58% of the sysSVM2 predictions involved in immune escape were damaged through GoF alterations underlying a tumour-promoting rather than a tumour-suppressive role. Upon closer inspection we noticed that GoF alterations tended to affect genes that exert a suppressive effect on antigen processing and presentation, thus contributing to the inactivation of this process. Among these, we found *CDC27* and genes involved in protein degradation, hence antigen processing, such as *PSMD3*, *RPS27A*, and *SMURF1*.

For instance, Song and co-authors showed that *CDC27* had a regulatory effect on *CD274*, the gene encoding PDL1 (Song et al., 2020). CDC27 overexpression was associated with high expression of PDL1, a well-known inhibitor of antigen presentation to T cells (Han et al., 2020). In contrast, overexpression of proteasomal subunits such as PSME3 and PSME4 have been shown to induce immune escape by restricting proteasome activity, thus inhibiting the processing and presentation of the antigen (Boulpicante et al., 2020; Javitt et al., 2021).

Finally, we wondered whether sysSVM2 predictions, especially those involved in immune-related pathways, were recurrent across the sample cohort or rather sample-specific. We found that 77% of the 116 sysSVM2 predictions enriched in immune-related pathways were rare or patient-specific, predicted as cancer drivers in less than three samples (Figure 5.4B).

These analyses suggest that OAC-specific canonical cancer drivers and sysSVM2 predictions play different roles that together contribute to OAC

tumorigenesis by specialising on the perturbation of different hallmarks of cancer (Hanahan, 2022). OAC-specific canonical drivers preferentially perturbed signalling-related pathways that influence tumour proliferation, cell death, and the ability to metastasise. By contrast, sysSVM2 predictions were mainly involved in mechanisms of immune escape and of DNA damage repair. Interestingly, we found that the pathways involved in immune escape were preferentially altered by sample-specific drivers.

5.5 Newly discovered targetable drivers

We then investigated whether any drivers, especially sysSVM2 predictions, were targets of FDA-approved, antineoplastic, and immunomodulating drugs collected in NCG7 (Dressler et al., 2022).

Of the total 521 unique cancer driver genes that we identified, 16 can currently be targeted by oncological drugs (Figure 5.5A). Ten gene targets were OAC-specific canonical drivers, while the remaining six were part of sysSVM2 predictions (Table 5.1).

The six sysSVM2 predictions targeted by drugs were three oncogenes, one candidate cancer gene, and two non-cancer genes (Table 5.1). The three oncogenes - *ABL1, FGFR1* and *SRC* - were not reported previously as OAC-specific cancer drivers.

As part of the drugs that target OAC-specific canonical drivers, we found some that are already being used in the clinic, such as trastuzumab (see Chapter 1.1.2). Other drugs, such as afatinib targeting *ERBB2* and *EGFR*, are currently tested in clinical trials as they showed initial promising results (Janjigian et al., 2015). Other drugs, such as cetuximab targeting *EGFR*, provided no additional benefit to patients whose tumours harboured the corresponding driver alteration (Lordick et al., 2013).

Among the drugs that preferentially target sysSVM2 predictions we found tyrosine kinase inhibitors such as dasatinib and regorafenib, and poly ADP-ribose polymerase (PARP) inhibitors such as olaparib, niraparib, and rucaparib.

Dasatinib is a tyrosine kinase inhibitor with multiple targets including *ABL1*, *EPHA2*, *SRC*, *YES1* (Table 5.1). All four genes are non-receptor tyrosine kinases. With the exception of *EPHA2*, they were all amplified in those samples where we prioritised them as drivers. Dasatinib is a potent agent for the treatment of chronic myeloid leukaemia, and some preliminary studies showed evidence of its efficacy in gastric cancer cell lines, too (Choi et al., 2020). Regorafenib is also a multitarget tyrosine kinase inhibitor that targets *ABL1*, *EPHA2* and *FGFR1* (Table 5.1). Its use is approved for the treatment of locally advanced or metastatic tumours. Indeed, it was shown to increase progression-free survival in a cohort of refractory advanced oesophagogastric adenocarcinomas (Pavlakis et al., 2016). *PARP1* is involved in the DNA damage response and, more specifically, in the repair of DNA single-strand breaks and is the target of PARP inhibitors. Most PARP inhibitors are approved for the treatment of ovarian and breast cancer in patients with damaging germline *BRCA* mutations. Some clinical trials are in the process of evaluating the response to PARP inhibitors alone or in combination with chemotherapy and immunotherapy in gastric cancer (Wang et al., 2021). These initial studies are promising as they show an acceptable safety profile along with preliminary antitumour activity of these drugs.

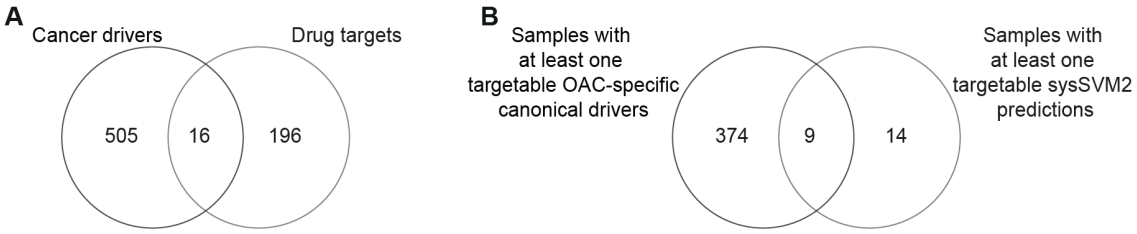

Figure 5.5 OAC drivers targeted by oncological drugs

A) Overlap between OAC cancer drivers and genes targeted by FDA-approved immunomodulating and antineoplastic drugs. **B)** Overlap of OAC samples that have at least one targetable OAC-specific canonical cancer driver and OAC samples that have at least one targetable sysSVM2 prediction.

In total, 59% of our sample cohort (397/675) had a driver alteration in at least one of these 16 targetable cancer genes with some of the samples having multiple drivers targetable by oncological drugs. The OAC-specific canonical drivers alone covered 374 samples; nine samples had at least one targetable OAC-specific

canonical driver and one targetable sysSVM2 predictions. Interestingly, 14 samples had only targetable sysSVM2 predictions (Figure 5.5B).

We found potential drug targets within the pool of rare or patient-specific drivers. Alterations harboured within these drivers should be considered in the context of precision medicine to develop therapies tailored on the driver landscape of individual tumours.

Table 5.1 Cancer driver genes targeted by oncological drugs

For each of the 16 cancer driver genes the gene name, whether it is a OAC-specific canonical cancer driver, its driver role according to NCG6 (Repana et al., 2019), the number of samples in which it is predicted as a driver and the number of oncological drugs that target it are reported. The genes are shown in ascending order based on the number of samples in which they are predicted as drivers.

Gene	OAC-specific driver	Driver role	# of samples	# of drugs
ABL1	N	OG	1	5
EPHA2	N	Candidate cancer	1	2
SRC	N	OG	2	1
BRAF	Y	OG	3	6
FGFR1	N	OG	3	6
PARP1	N	Non-cancer	6	4
JAK1	Y	Canonical cancer	7	2
MAP3K1	Y	Canonical cancer	7	1
FGFR2	Y	OG	10	5
YES1	N	Non-cancer	10	1
MET	Y	OG	22	2
PIK3CA	Y	OG	49	1
EGFR	Y	OG	79	11
CCND1	Y	OG	98	1
CDK6	Y	OG	98	3
ERBB2	Y	OG	122	6

5.6 Stage-specific driver perturbations

Next we investigated whether the four OAC clinical stages were characterised by the acquisition of specific cancer driver genes that could help defining stage boundaries (e.g., the transition from clinical stage I to clinical stage II). In addition,

we studied whether OAC clinical stratification was reflected at the level of pathway perturbation that could further help stratifying patients into subgroups.

In order to investigate the frequency of driver alterations across stages we focused on cancer driver genes predicted in at least 5% of our whole OAC sample cohort. We used this threshold as too few OACs are unlikely to be a representative sample of the entire OAC population.

We found 18 unique cancer driver genes altered 1811 times in 647 samples. Sixteen of these were OAC-specific canonical drivers and the other two were the sysSVM2 predictions discussed above (see Chapter 5.3). None of the 18 drivers were stage-specific or acquired driver alterations at specific stage boundaries (Figure 5.6A).

Although not significant, we observed that three genes showed a trend of stage-specific driver event acquisition. *ERBB2* acted as a driver in 23% of stage I OAC tumours; the percentage progressively decreased across stages, and in stage IV tumours *ERBB2* was predicted as a cancer driver in 16% of cases (Figure 5.6A). Conversely, *MYC* and *KRAS* showed the opposite trend: we predicted them as drivers in 11% and 12% of OAC stage I cases, respectively. In stage IV tumours we found *MYC* acting as a driver in 20% of samples and *KRAS* in 21% of samples.

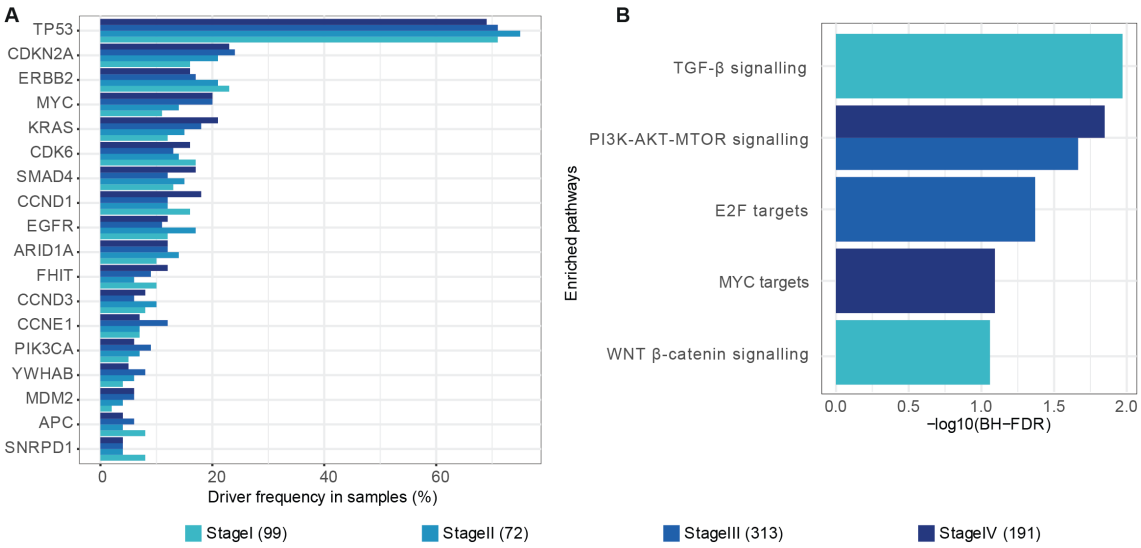

Figure 5.6 OAC clinical stage-specific driver alteration and pathway perturbation

A) Frequency of recurrent drivers in OAC samples. The samples were divided into clinical stages. Recurrence was defined as cancer driver genes predicted in at least 5% of the whole OAC cohort. **B)** List of enriched pathways across OAC clinical stages. Enrichment was measured comparing the proportion of drivers in each pathway against that of the non-driver genes. (FDR <0.1, one-sided Fisher's exact test). FDR was calculated by applying Benjamini-Hochberg correction. The colours identify the four OAC clinical stages. The numbers in brackets report the number of sample in the corresponding stage.

We finally tested the enrichment of all the drivers grouped into clinical stages. We identified differences in terms of pathways that the stage-specific drivers perturbed (Figure 5.6B). Stage I cancer drivers were enriched in TGF-β and WNT signalling. The drivers we predicted in stage II tumours were not enriched in any pathway, most likely because stage II OACs were the smallest group among the four OAC clinical stages (Figure 4.1B). The drivers predicted in the two advanced OAC stages were enriched in phosphoinositide 3-kinase (PI3K) signalling and in targets of two transcription factors, namely E2F and MYC. MYC targets enriched in stage IV drivers reflected the trend we observed for individual cancer driver genes (Figure 5.6A).

Recurrent cancer drivers in OAC were not enriched, at the single gene level, in any clinical stage of the disease suggesting that, based on our cohort size, cancer driver genes do not acquire alterations at specific stage boundaries. We observed, however, a few trends that suggest a sequential acquisition of driver alterations affecting specifically *ERBB2*, *MYC* and *KRAS*. When looking at the pathway perturbed by drivers instead of individual driver genes, we observed clear differences across stages. All of the pathways that we reported enriched across the four clinical stages had previously been discussed in the context of OAC (Mourikis et al., 2019), however we were able to show that they acquired alterations in a stage-specific fashion.

5.7 Conclusion

sysSVM2 ranks damaged genes in individual samples based on how much the gene properties resemble those of canonical cancer drivers used for training. In this work, the training set for sysSVM2 consisted of a list of 77 OAC-specific

canonical drivers derived from two sources; NCG6 (Repana et al., 2019) and the curation of additional OAC screenings obtained from a comprehensive literature search (Dulak et al., 2012; Frankel et al., 2014; Frankell et al., 2019; Murugaesu et al., 2015; Nones et al., 2014). The training of sysSVM2 resulted in ranked sample-specific lists of damaged genes based on the driver landscape, which was specific for OAC.

Given the ranking nature of sysSVM2 predictions and the fact that the tool does not impose a hard cut-off on what a cancer driver gene is, we wondered how many drivers were needed to explain OAC in single patients. This led us to compare two previously reported models on the number of driver genes needed in the context of OAC. The first model - age-incidence - reported that three drivers were needed per tumour, whereas the second model - positive selection - estimated five drivers per tumour.

We investigated whether the two positive selection-specific driver genes in each sample perturbed novel pathways. After removing redundant pathways, the positive selection-specific drivers indeed altered pathways that were not directly affected by the age-incidence drivers. Pathways involved in the regulation of gene expression, mechanisms of nucleotide and base repair, and chromosome maintenance were exclusively altered by the positive selection-specific drivers.

This led us to select the positive selection model, five drivers per sample, as the most accurate model available so far to explain OAC evolution.

We then investigated the frequency of drivers across the 675-sample cohort. We found that, in line with previous investigations (Mourikis et al., 2019), sysSVM2 predictions were preferentially sample-specific whilst OAC-specific canonical drivers were more often recurrent. Interestingly, two sysSVM2 predictions were altered at a similar rate to that of some OAC-specific canonical cancer drivers, specifically in 5% of our cohort. These two genes were *SNRPD1* and *YWHAB,* neither of which was reported as a driver before. The type of alterations affecting these two drivers in our cohort along with experimental evidence on their mechanism of action (Dai et al., 2021; Quidville et al., 2013; Sugiyama et al., 2003; Takihara et al., 2000) suggested that they are likely to act in a tumour-promoting mode during OAC development.

Given the heterogenous list of sysSVM2 predictions, composed of 448 unique driver genes, we compared their role to that of OAC-specific canonical cancer drivers. We found a high level of overlap in terms of pathways perturbed by both sets of drivers, as previously reported (Mourikis et al., 2019). However, in addition to the above, these two sets of drivers played distinct roles in the cell. OAC-specific canonical driver genes preferentially perturbed well-known signalling cascades involved in the control of cell fate such as activin signalling (Chen et al., 2006) and the promotion of angiogenesis such as erythropoietin activation (Kimáková et al., 2017). Interestingly, sysSVM2 predictions were preferentially involved in DNA repair and the regulation of the immune system. The enrichment of sysSVM2 predictions in DNA repair processes recapitulates the observed trend in which positive selection-specific drivers were mainly involved in the perturbation of similar processes. The two results are probably related, as models that predict more drivers per tumour require more sysSVM2 predictions to complete their list of driver genes.

sysSVM2 predictions were enriched in pathways involved in the recognition and presentation of antigens to cytotoxic $CD8^+$ T cells. It is tempting to speculate that alterations acquired in this set of drivers are thus responsible for the impairment of the detection of cancer cells by the immune system. Additionally, sysSVM2 predictions perturbed pathways involved in the creation of a pro-inflammatory environment around the tumour mass such as TLR cascade and interleukin signalling. Our group and others already described the involvement of TLR cascade in the context of OAC (Fels Elliott et al., 2017; Mourikis et al., 2019). The TLR signalling has been suggested to be key in creating a pro-inflammatory immune environment, which favours and sustains tumour progression, in the distal oesophagus through the recruitment of NF-kB. Overall, we found that most of the immune-related pathways were perturbed by sample-specific sysSVM2 predictions, highlighting the important role of these driver genes during OAC development.

Next we inspected the list of OAC drivers, searching for targetable genes. We found 16 driver genes as potential targets of anticancer treatment. Ten of these were OAC-specific canonical drivers and the remaining six were sysSVM2

predictions, mostly predicted in single patients. Among the six sysSVM2 predictions we found non-receptor tyrosine kinases targeted by multiple drugs (Huang et al., 2020) and targets of PARP inhibitors which are showing promising preliminary results in the treatment of gastric cancer (Wang et al., 2021).

Finally, we inspected the identity and role of cancer driver genes from a clinical perspective. We wondered whether cancer drivers were acquired at specific stage boundaries that could help us in stratifying OAC samples into clinical groups. No driver gene acquired alterations at specific stage boundaries. However, some pathways were perturbed in a stage-specific fashion. TGF-β and WNT signalling were altered in stage I tumours, whereas PI3K signalling and transcription factor networks were altered in late stage tumours.

In the context of BO-to-OAC progression it has been reported that TSGs acquire driver alterations before OGs, whose alterations usually characterise the acquisition of the definitive neoplastic phenotype (Maley, 2007; Stachler et al., 2015). Pan-cancer studies have additionally showed that increased genomic instability is usually a hallmark of late stage OAC tumours (Gerstung et al., 2020). Given this, it is tempting to speculate that driver events in early stage tumours preferentially perturb pathways that involve well-known TSGs, such as *SMAD4* and *APC* in TGF-β and WNT signalling, respectively. By contrast, late stage tumours, characterised by higher genomic instability and chromosomal rearrangements, which usually underlie OG activation (Nones et al., 2014), show an increased dysregulation of transcriptional networks regulated by well-know OGs such as *MYC*.

Chapter 6. Discussion

6.1 Summary

The work presented in this thesis aims to inspect the role of genetic inter-tumour heterogeneity in the initiation and progression of OAC and to identify the cellular and molecular processes that the driver genes perturb during progression from early to late tumour stages. Furthermore, whether the drivers converge to disrupt similar cellular pathways is a key aim of this work alongside understanding which molecular processes underlie the clinical stages of OAC.

In the first results chapter I describe the role along with the evolutionary, genomic, expression, and network properties of cancer and healthy driver genes whose extensive annotation and SLPs have been published as part of the seventh release of the NCG database, available online at http://network-cancer-genes.org/. SLPs are intrinsic properties of human protein-coding genes that describe their role within the cell and the evolutionary path across the species phylogenetic tree. This chapter lays the foundation for all of the subsequent analysis.

In order to investigate how heterogeneity favours OAC development, it is critical to first define and complete the list of drivers in each OAC sample. Given the inter-tumour genetic heterogeneity present in OAC, in the second results chapter I derive single-patient driver predictions from the large cohort of samples using sysSVM2, a machine learning tool for driver identification. The tool relies on similarities at the level of molecular and SLPs between the canonical cancer drivers used for training and the remaining damaged genes. The molecular properties summarise the somatic alterations present in each tumour and are specific for each sample as they describe the genetic changes acquired by the cancer cells throughout the individual tumour evolutionary history. SLPs, on the other hand, are derived from NCG.

In the final results chapter the training and prediction of sysSVM2 is used to derive a comprehensive list of drivers for each OAC sample. I present and compare two lists of drivers predicted under two models that estimate different numbers of drivers needed in OAC. Finally, I investigate the role of the identified

drivers, their potential as targets of anti-cancer treatment, and the cellular processes that these drivers perturb across OAC clinical stages in order to understand their translational potential.

6.2 NCG: a manually curated repository of cancer and healthy driver genes

Cancer genomes acquire hundreds of somatic alterations throughout their evolutionary history. Some of these alterations affect genes that, upon acquiring a mutated phenotype, promote cancer growth by providing affected cells with a selective advantage over their neighbours. Indeed, the goal of cancer genomics is to identify which genes are responsible for driving cancer initiation and progression (Campbell et al., 2020).

Different resources exist, that are available to the cancer research community, for annotating the specific alterations (Ainscough et al., 2016; Cerami et al., 2012; Tamborero et al., 2018) or the genes (Futreal et al., 2004; Sondka et al., 2018) that drive cancer. Similarly, NCG curates an up-to-date overview of driver genes that are well-known or predicted to be involved in tumorigenesis. NCG differs from similar repositories in that it focuses on cancer genes, rather than alterations, classifies drivers into canonical (namely those with experimental validation) and candidate (namely those predicted by driver identification tools), and annotates their SLPs (An et al., 2016; Matteo D'Antonio et al., 2012; Dressler et al., 2022; Repana et al., 2019; Syed et al., 2010). Such properties are a proxy for the central role of the genes within the cell and across species.

NCG7 now annotates driver genes that promote tumorigenesis by accumulating mutations in their non-coding regions and genes that are involved in the expansion of mutated clones within non-cancer tissues (healthy drivers). The database also integrates a larger number of sources on gene essentiality, contains new properties that describe the genomic accumulation of germline variation, and annotates the interaction between human genes and anti-cancer drugs. Altogether, these provide the novelties of the latest release and iteration of the NCG repository (Dressler et al., 2022).

We confirmed that cancer driver genes appeared earlier in evolution (Syed et al., 2010), are more present as singletons in the human genome (Rambaldi et al., 2008), are more broadly expressed across human healthy tissues (An et al., 2016; Repana et al., 2019), are more finely regulated by miRNA (D'Antonio et al., 2012), and encode proteins that are more central, connected and clustered in the PPIN (D'Antonio et al., 2012) than the rest of human genes. Finally, cancer drivers are more essential in cancer cell lines as confirmed by the tendency to accumulate fewer germline alterations than the rest of human genes (Dressler et al., 2022).

Within canonical cancer drivers we observed some level of heterogeneity since TSGs and OGs specialised in different cellular roles during their evolution. We confirmed that TSGs are preferentially involved in the maintenance of basic cellular processes such as the control of the cell cycle and the repair of DNA damage. This shows evidence of an enrichment of caretaker genes within TSGs (Domazet-Lošo & Tautz, 2010; Kinzler & Vogelstein, 1997; Michor et al., 2004; Negrini et al., 2010). Their role in maintaining a functional and stable genome is of fundamental importance for cells, hence shared across many species. Their evolutionary-conserved maintenance role is indeed supported by the older age, lower duplicability, broader expression in human tissues, and higher essentiality of TSGs with respect to OGs.

OGs, on the other hand, are preferentially involved in regulatory functions such as signalling and the immune system. Such processes tend to vary more across cells, and likely reflect the lower level of expression in human healthy tissues and the lower grade of essentiality across cancer cell lines of this group of drivers compared to TSGs.

Canonical healthy drivers (n=57) represent the subgroup of genes with the most extreme SLP profile. With respect to their evolutionary conservation, PPIN, miRNA-gene interactions, germline variation, and gene essentiality, these 57 drivers display a profile that is even stronger than that of canonical cancer drivers altogether.

This observation raises a few questions about the initiation of tumorigenesis and the expansion of mutated clones in healthy tissues. Based on their extreme SLP

profile, it is tempting to speculate that these 57 genes are representative of a subset of drivers that show a very strong oncogenic potential. If this is the case, it is intriguing to understand why clones with somatic mutations in these drivers do not always show a malignant phenotype. Related to this, it becomes particularly interesting to identify the genomic and environmental changes that represent the final tipping point responsible for malignant transformation.

Mutations affecting the same gene result in different phenotypes between cancer and normal tissues (Wijewardhane et al., 2021). This has been a topic of interest with some compelling mechanisms proposed to explain such differences. Firstly, a driver alteration in only one gene is probably not sufficient to cause malignant transformation. Even the most conservative estimates report that at least two or three driver events must be acquired by cells to transform into a tumour (Martincorena & Campbell, 2015; Tomasetti et al., 2015). A second important aspect to consider is the underlying genomic context in which these changes are acquired. Specifically, since cancer is characterised by a high grade of genetic instability (Hanahan, 2022; Hanahan & Weinberg, 2011), the absence of this instability is probably a major brake on the acquisition of an invasive phenotype. Similarly, it is becoming increasingly evident that, in some cases, pre-cancer lesions accumulating high levels of CIN are more likely to progress to cancer compared to lesions that are CN-neutral (Killcoyne et al., 2020, 2021). In this context, an interesting point is to understand how mutations in these 57 drivers are able to activate changes that result in the acquisition of CIN and in the surrounding microenvironment.

Finally, understanding the role of drivers involved only in the clonal expansion within healthy tissues and whether they have a protective role against carcinogenesis, or are involved in processes not affecting cancer, will be critical to fully elucidating the acquisition of healthy drivers.

New studies are being continuously published on the acquisition of somatic mutations in non-cancer tissues that result in the expansion of mutated clones (Fowler et al., 2021; Li et al., 2021; Moore et al., 2021; Ng et al., 2021; Robinson et al., 2021; Saini et al., 2021). New investigations will help in understanding what

the role of mutated clones in healthy tissues is that will ultimately result in further insights into the initial stages of tumour development.

6.3 sysSVM2 optimisation on an OAC-specific setting

We curated one of the largest cohorts of BO and OAC samples comprising 748 samples. We characterised these samples from a clinical and a molecular perspective using a consistent bioinformatic pipeline for the annotation of damaged genes in individual samples. We stratified samples into clinical stages and found that, as the tumour developed, the samples acquired significantly more CNAs underlying an increased level of CIN as previously reported (Ross-Innes et al., 2015; Stachler et al., 2015).

The higher grade of genomic instability we observed in OAC samples was driven by gene amplifications rather than homozygous deletions and double hits whose numbers were comparable between pre-cancer and cancer lesions. The high level of amplifications in OAC samples is likely due to structural rearrangements acquired when the tumour develops, as suggested by other groups (Nones et al., 2014).

We then compiled a comprehensive list of canonical cancer driver genes identified in OAC, including all NGS screenings focusing on this cancer type (Agrawal et al., 2012; Dulak et al., 2012, 2013; Fels Elliott et al., 2017; Frankel et al., 2014; Frankell et al., 2019; Kim et al., 2017; Murugaesu et al., 2015; Nones et al., 2014; Secrier et al., 2016; Weaver et al., 2014). This list represents a *bona fide* snapshot of the current knowledge of cancer driver genes involved in OAC tumorigenesis. It resulted in 77 driver genes whose alterations have robust experimental evidence of their involvement in cancer development.

We found that, with the only exception of *TP53* and *CDKN2A*, damaged in 66% and 25% of samples respectively, the remaining OAC-specific canonical cancer drivers were damaged in less than 20% of samples. This confirmed the high grade of heterogeneity in OAC as previously reported (Agrawal et al., 2012; Dulak et al., 2013). This was also confirmed by the vast majority (66%) of OAC samples

with not enough cancer driver genes to explain the presence of the disease in the corresponding patient.

Interestingly, we noticed that *CDKN2A* acquired more alterations in the pre-cancer lesion with 45% of NDBO cases having the gene in their list of drivers. The gene was mainly altered via homozygous deletions rather than damaging mutations. A previous study investigating the sequential acquisition of driver events in the progression from BO to OAC did not identify *CDKN2A* as one of the genes more commonly altered in BO than in OAC (Weaver et al., 2014). This was likely due to the fact that the authors only inspected mutations and discarded the role of CNAs in the study of BO-to-OAC progression.

Other groups found *CDKN2A* to be key in the progression from BO to OAC (Maley, 2007; Stachler et al., 2015). However, our observation argues against this thesis given the lower percentage of OAC samples with driver *CDKN2A* alterations. If *CDKN2A* is an early driver of progression whose alterations are acquired at the level of BO, and BO is the lesion that seeds OAC, it is reasonable to expect that the alteration is maintained in OAC at a similar rate to that observed in BO. If this is not the case, two possible explanations exist. In the first case, given the polyclonality of the BO lesion (Ross-Innes et al., 2015), it is possible that the clone harbouring the *CDKN2A* alteration was not the one that seeded the tumour. Alternatively, *CDKN2A* is a key driver involved in the evolution of the BO lesion but has little to do with the progression to OAC.

Given OAC genetic inter-tumour heterogeneity, which has been extensively reported and discussed by us and other groups (Contino et al., 2017; Mourikis et al., 2019), we decided to apply a tool for driver detection, sysSVM2, that allows the making of single-patient predictions (Nulsen et al., 2021). sysSVM2 prioritises as cancer drivers the damaged genes that more closely resemble the canonical drivers used as training set.

The Ciccarelli lab applied a previous version of the sysSVM2 tool to a smaller cohort of OAC samples that enabled the completion of the list of cancer driver genes in individual samples (Mourikis et al., 2019). In this work, we applied an optimised version of the tool to a much larger cohort of BO and OAC samples than previously. In addition, we performed further investigation in order to select

an optimised OAC-specific setting for training sysSVM2 that recapitulated the driver repertoire specific for OAC carcinogenesis.

6.4 Role and clinical relevance of the driver genes contributing to OAC development

A long-standing question in cancer biology concerns the number of driver events that are needed for a normal cell to transform into a cancer cell. In this context, multiple methods have been applied and different cancer types have been inspected (Anandakrishnan et al., 2019; Makohon-Moore et al., 2018; Martincorena et al., 2017; Tomasetti et al., 2015; Vogelstein & Kinzler, 2015). The final number is variable and has been estimated to be between two and eleven, based on the cancer type and the methodology applied to derive it.

In OAC, two of said models have been previously applied in an attempt to explain the mechanisms of tumorigenesis (Jeon et al., 2006; Martincorena et al., 2017). Jeon et al. used an approach based on age incidence data that predicted three driver events per tumour, whereas Martincorena et al. studied the genes that accumulate more non-synonymous mutations than the rest of the genome and, from there, derived that five driver events are needed per tumour.

The ranking nature of sysSVM2 predictions allowed us to investigate and compare the two models in terms of the functional implication of the two sample-specific positive selection-specific drivers that were not predicted under the age-incidence model. We reasoned that, in order to investigate whether the positive selection-specific drivers were needed to explain the presence of OAC, we would look at their role within the cancer cell. If they converged to perturb pathways that were not already altered by the age-incidence drivers, they would more likely be functionally implicated in OAC carcinogenesis, hence needed to explain the presence of the disease.

Based on this reasoning, we found that the two positive selection-specific drivers added perturbations to new pathways. They contributed to the perturbations of novel pathways such as DNA damage repair, gene expression, and cell cycle

pathways, therefore suggesting that they are likely functionally involved in OAC development.

It is reasonable to assume that, as the number of drivers estimated to explain the disease increases, more pathways will be perturbed by the additional drivers. However, similar analysis done within the group on a pan-cancer dataset showed that, in some cancer types, models predicting a larger number of drivers added no or very little contribution to the perturbation of new pathways (unpublished data, Hrvoje Misetic). This suggests that additional drivers do not always contribute to the perturbation of novel pathways that sustain tumorigenesis.

After selecting the number of drivers needed per OAC, we were able to complete the list of drivers in individual samples and investigate their role in the development of OAC. We identified two distinct behaviours for the recurrence across samples of OAC-specific canonical drivers and sysSVM2 predictions. Specifically, OAC-specific canonical drivers were more often predicted across samples, whereas sysSVM2 predictions were preferentially rare or sample-specific. Interestingly, we found two genes within the pool of sysSVM2 predictions prioritised as drivers in 5% of our OAC cohort which were not previously reported to be involved in cancer. In all samples we found these altered via gene amplifications. This confirmed the suggested role of these two genes in the context of carcinogenesis as potential OGs (Dai et al., 2021; Quidville et al., 2013; Sugiyama et al., 2003; Takihara et al., 2000).

Moreover, we confirmed that, overall, the two sets of cancer drivers (OAC-specific canonical driver genes and sysSVM2 predictions) perturbed almost all of the processes reported to sustain cancer development (Hanahan, 2022; Hanahan & Weinberg, 2011) across all four OAC clinical stages. We also showed that most of these processes were similarly perturbed by both OAC-specific canonical drivers and sysSVM2 predictions. We observed, however, some interesting exceptions to this general trend. The alterations affecting OAC-specific canonical drivers preferentially perturbed pathways involved in signalling. Alterations affecting these pathways are likely to result in proliferative growth, resistance to cell death, evasion of growth suppressors, induction of angiogenesis, and promotion of metastasis. By contrast, alterations in sysSVM2

predictions preferentially contributed to the perturbation of other aspects of tumour development. Specifically, we saw a larger contribution from sysSVM2 predictions in creating genomic instability, escaping immune destruction, and creating a microenvironment that promotes inflammation and cancer growth.

We also inspected the list of drivers for putative targets of anti-cancer treatment. Within sysSVM2 predictions we found some interesting targets of PARP inhibitor therapy that are currently under investigation in the context of gastric cancer (Wang et al., 2021).

These results suggest that, within the pool of patient-specific driver genes, putative clinical targets reside. Such targets can only be identified by patient-specific driver prediction tools and can contribute to the development of strategies for precision medicine. Additionally, we showed that patient-specific drivers alone are involved in important mechanisms that sustain tumour growth, such as immune escape.

Finally, we looked at possible biomarkers in the form of drivers and cellular pathways that can help in further characterising OAC clinical stages. In terms of individual drivers, we did not see any difference in the frequency of these across stages. When inspecting the pathways we observed that some processes, such as TGF-β and WNT signalling, were altered at early stages, whereas others, such as transcription network controlled by MYC and E2F, were perturbed at more advanced stages of disease.

The stage-specific enriched pathways might recapitulate the evolutionary history of OAC. In OAC, TSG inactivation, which often underlies TGF-β and WNT signalling, is usually an early event acquired in the first stages of pre-cancer development (Gerstung et al., 2020; Maley et al., 2004; Maley et al., 2004). Later on when the tumour develops OG amplifications become more frequent (Stachler et al., 2015). Even though in this work we investigated only the acquisition of driver perturbations across the tumour clinical stages, OAC evolutionary history could be reflective of mechanisms in which tumorigenesis is driven by LoF alterations in early tumour stages and followed, at later stages, by an increase in GoF alterations.

6.5 Concluding remarks and future trajectories

Understanding the mechanisms of tumour development is extremely important for the development of strategies that are able to inhibit or counteract it. In this work, we showed that two groups of drivers exist within OAC. Well-known OAC drivers, given their higher frequency rate, are commonly perturbed and easily identified when investigating sample cohorts. The second group of drivers is represented by genes that, instead, are preferentially sample-specific and can only be identified by applying statistical tools for single patient driver prediction.
This concept has important implications in the context of OAC tumorigenesis. The recurrent drivers preferentially perturb processes that sustain tumour growth, prevent cancer cell death, and promote invasion. The sample-specific drivers, on the other hand, preferentially interfere with processes involved in immune escape and DNA damage repair creating vulnerabilities that can be exploited by anticancer treatment.
When identifying therapies tailored to the specific landscape of individual tumours, it is pivotal to identify the precise number of driver genes present in the patient. We showed that five drivers per tumour is a comprehensive estimate on the number of driver events needed in individual OACs. Once the driver genes are identified it is easier to tailor strategies that consider the driver landscape of single tumours. We demonstrated how sample-specific driver genes can be useful targets of anti-cancer treatment, especially in those tumours that have no targetable driver alteration within their pool of canonical cancer drivers.
In the future, it will be of interest to follow up on some of the remaining open questions regarding:

- The role of non-coding drivers in OAC development. It would be informative to integrate, within the sysSVM2 framework, features that will enable the prioritisation of non-coding drivers, such as long non-coding RNAs, in order to obtain a broader and more comprehensive view of driver events in OAC.
- The temporal acquisition of driver genes. Previous work has already investigated the acquisition of recurrent events in OAC (Gerstung et al.,

2020). It would be interesting to see if differences in the acquisition of OAC-specific canonical drivers and sysSVM2 predictions exist and, if so, in which direction.

- The interplay between cancer drivers and the tumour microenvironment in the BO-to-OAC progression. Additionally, it will be particularly useful for the clinic to investigate how the tumour microenvironment of progressors and non-progressors differs in order to identify potential biomarkers of progression and further refine the early detection of this deadly disease.

Chapter 7. Appendix

7.1 Supplementary figures

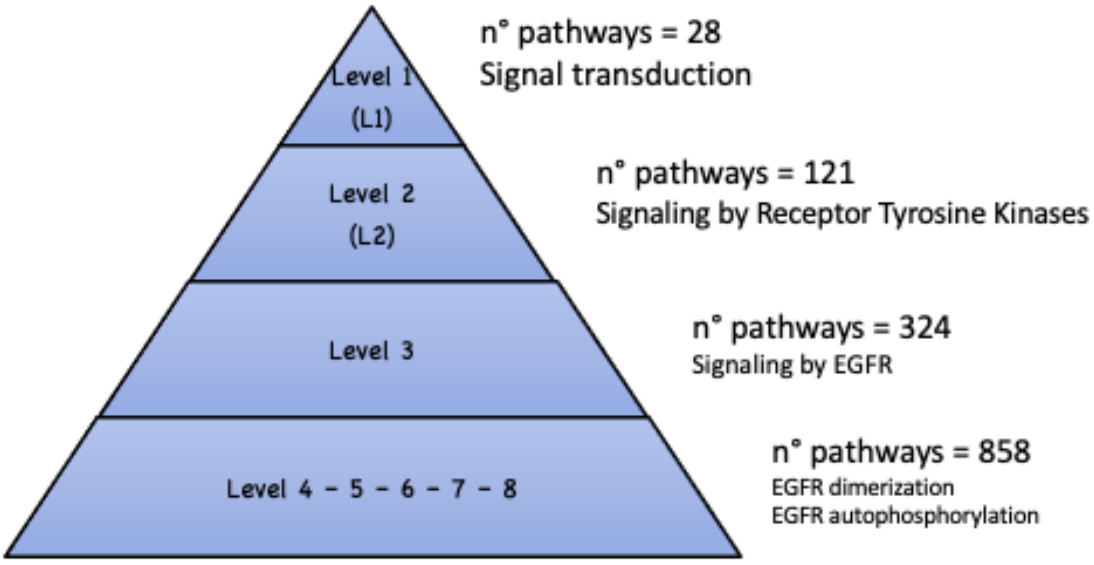

Figure 7.1 Pyramidal structure of the Reactome database

The Reactome database is organised into levels. Level 1 pathways are 28 in total and broad. They describe general cellular processes. As the level increases, so does the granularity of the information present in each level. For each level the number of pathways corresponding to that level is reported, along with an example of at least one pathway.

Reference List

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, *76*(7), 7.20.1-7.20.41. https://doi.org/10.1002/0471142905.hg0720s76

Agrawal, N., Jiao, Y., Bettegowda, C., Hutess, S. M., Wang, Y., David, S., Cheng, Y., Twaddell, W. S., Latt, N. L., Shin, E. J., Wang, L. D., Wang, L., Yang, W., Velculescu, V. E., Vogelstein, B., Papadopoulos, N., Kinzler, K. W., & Meltzer, S. J. (2012). Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discovery*, *2*(10), 899–905. https://doi.org/10.1158/2159-8290.CD-12-0189

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., Flynn, E. D., Parsana, P., Fresard, L., Gamazon, E. R., Hamel, A. R., He, Y., Hormozdiari, F., Mohammadi, P., Muñoz-Aguirre, M., … Volpi, S. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330. https://doi.org/10.1126/SCIENCE.AAZ1776

Ainscough, B. J., Griffith, M., Coffman, A. C., Wagner, A. H., Kunisaki, J., Choudhary, M. N. K., McMichael, J. F., Fulton, R. S., Wilson, R. K., Griffith, O. L., & Mardis, E. R. (2016). DoCM: A database of curated mutations in cancer. In *Nature Methods* (Vol. 13, Issue 10, pp. 806–807). Nature Publishing Group. https://doi.org/10.1038/nmeth.4000

An, O., Dall'Olio, G. M., Mourikis, T. P., & Ciccarelli, F. D. (2016). NCG 5.0: Updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic Acids Research*, *44*(D1), D992–D999. https://doi.org/10.1093/nar/gkv1123

An, O., Pendino, V., D'Antonio, M., Ratti, E., Gentilini, M., & Ciccarelli, F. D. (2014). NCG 4.0: The network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database*, *2014*, 1–10. https://doi.org/10.1093/database/bau015

Anandakrishnan, R., Varghese, R. T., Kinney, N. A., & Garner, H. R. (2019).

Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations. *PLOS Computational Biology*, *15*(3), e1006881. https://doi.org/10.1371/journal.pcbi.1006881

Armitage, P., & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, *8*(1), 1–12. https://doi.org/10.1038/bjc.1954.1

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. In *Nature* (Vol. 526, Issue 7571, pp. 68–74). Nature Publishing Group. https://doi.org/10.1038/nature15393

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K. S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Karchin, R. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *173*(2), 371–385. https://doi.org/10.1016/j.cell.2018.02.060

Bakken, S., Suraski, Z., & Schmid, E. (2020). *PHP Manual*. https://scholar.google.com/scholar_lookup?title=PHP manual&publication_year=2020&author=Bakken%2CS&author=Suraski%2CZ&author=Schmid%2CE

Bang, Y.-J., Van Cutsem, E., Mansoor, W., Petty, R. D., Chao, Y., Cunningham, D., Ferry, D., Landers, D., Stockman, P., Smith, N. R., Geh, C., & Kilgour, E. (2015). A randomized, open-label phase II study of AZD4547 (AZD) versus Paclitaxel (P) in previously treated patients with advanced gastric cancer (AGC) with Fibroblast Growth Factor Receptor 2 (FGFR2) polysomy or gene amplification (amp): SHINE study. *Journal of Clinical Oncology*, *33*(15_suppl), 4014–4014. https://doi.org/10.1200/jco.2015.33.15_suppl.4014

Bang, Y. J., Van Cutsem, E., Feyereislova, A., Chung, H. C., Shen, L., Sawaki,

A., Lordick, F., Ohtsu, A., Omuro, Y., Satoh, T., Aprile, G., Kulikov, E., Hill, J., Lehle, M., Rüschoff, J., & Kang, Y. K. (2010). Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): A phase 3, open-label, randomised controlled trial. *The Lancet*, *376*(9742), 687–697. https://doi.org/10.1016/S0140-6736(10)61121-X

Barrett, M. T., Sanchez, C. A., Prevo, L. J., Wong, D. J., Galipeau, P. C., Paulson, T. G., Rabinovitch, P. S., & Reid, B. J. (1999). Evolution of neoplastic cell lineages in Barrett oesophagus. *Nature Genetics*, *22*(1), 106–109. https://doi.org/10.1038/8816

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., Huntsman, D. G., Caldas, C., Aparicio, S. A., & Shah, S. P. (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biology*, *13*(12), R124. https://doi.org/10.1186/gb-2012-13-12-r124

Behan, F. M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C. M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., Ansari, R., Harper, S., Jackson, D. A., McRae, R., Pooley, R., Wilkinson, P., van der Meer, D., Dow, D., Buser-Doepner, C., … Garnett, M. J. (2019). Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature*, *568*(7753), 511–516. https://doi.org/10.1038/s41586-019-1103-9

Bertrand, D., Chng, K. R., Sherbaf, F. G., Kiesel, A., Chia, B. K. H., Sia, Y. Y., Huang, S. K., Hoon, D. S. B., Liu, E. T., Hillmer, A., & Nagarajan, N. (2015). Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Research*, *43*(7), e44. https://doi.org/10.1093/nar/gku1393

Bhagwat, M., Young, L., & Robison, R. R. (2012). Using BLAT to Find Sequence Similarity in Closely Related Genomes. *Current Protocols in Bioinformatics*, *37*(1), 10.8.1-10.8.24. https://doi.org/10.1002/0471250953.bi1008s37

Bonnal, S., Vigevani, L., & Valcárcel, J. (2012). The spliceosome as a target of novel antitumour drugs. In *Nature Reviews Drug Discovery* (Vol. 11, Issue 11, pp. 847–859). Nature Publishing Group. https://doi.org/10.1038/nrd3823

Boulpicante, M., Darrigrand, R., Pierson, A., Salgues, V., Rouillon, M.,

Gaudineau, B., Khaled, M., Cattaneo, A., Bachi, A., Cascio, P., & Apcher, S. (2020). Tumors escape immunosurveillance by overexpressing the proteasome activator PSME3. *OncoImmunology*, *9*(1), 1–16. https://doi.org/10.1080/2162402X.2020.1761205

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H. K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C. L., Faquin, W. C., Deshpande, V., Boutros, P. C., Lazar, A. J., Hoadley, K. A., … Zhang, J. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

Carmeliet, P. (2005). VEGF as a key mediator of angiogenesis in cancer. *Oncology*, *69*(SUPPL. 3), 4–10. https://doi.org/10.1159/000088478

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, *2*(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2017). *Shiny: web application framework for R*. https://scholar.google.com/scholar_lookup?title=shiny%3A web application framework for R&publication_year=2021&author=Chang%2CW&author=Cheng%2CJ&author=Allaire%2CJ&author=Sievert%2CC&author=Schloerke%2CB&author=Xie%2CY

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B. J., Dolinski, K., & Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, *45*(D1), D369–D379. https://doi.org/10.1093/nar/gkw1102

Chen, W. H., Lu, G., Chen, X., Zhao, X. M., & Bork, P. (2017). OGEE v2: An update of the online gene essentiality database with special focus on

differentially essential genes in human cancer cell lines. *Nucleic Acids Research*, *45*(D1), D940–D944. https://doi.org/10.1093/nar/gkw1013

Chen, Y. G., Wang, Q., Lin, S. L., Chang, C. D., Chung, J., & Ying, S. Y. (2006). Activin signaling and its role in regulation of cell proliferation, apoptosis, and carcinogenesis. *Experimental Biology and Medicine*, *231*(5), 534–544. https://doi.org/10.1177/153537020623100507

Chen, Z., Chen, C., Li, L., Zhang, T., & Wang, X. (2021). Pan-Cancer Analysis Reveals That E1A Binding Protein p300 Mutations Increase Genome Instability and Antitumor Immunity. *Frontiers in Cell and Developmental Biology*, *9*(729927), 1–11. https://doi.org/10.3389/fcell.2021.729927

Choi, K.-M., Cho, E., Bang, G., Lee, S.-J., Kim, B., Kim, J.-H., Park, S.-G., Han, E. H., Chung, Y.-H., Kim, J. Y., Kim, E., & Kim, J.-Y. (2020). Activity-Based Protein Profiling Reveals Potential Dasatinib Targets in Gastric Cancer. *International Journal of Molecular Sciences*, *21*(9276), 1–14. https://doi.org/10.3390/ijms21239276

Chou, C. H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., Huang, W. C., Sun, T. H., Tu, S. J., Lee, W. H., Chiew, M. Y., Tai, C. S., Wei, T. Y., Tsai, T. R., Huang, H. T., Wang, C. Y., Wu, H. Y., Ho, S. Y., Chen, P. R., … Huang, H. Da. (2018). MiRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, *46*(D1), D296–D302. https://doi.org/10.1093/nar/gkx1067

Chun, S., & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research*, *19*(9), 1553–1561. https://doi.org/10.1101/gr.092619.109

Ciriello, G., Cerami, E., Sander, C., & Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, *22*(2), 398–406. https://doi.org/10.1101/gr.125567.111

Coleman, H. G., Xie, S. H., & Lagergren, J. (2018). The Epidemiology of Esophageal Adenocarcinoma. *Gastroenterology*, *154*(2), 390–405. https://doi.org/10.1053/j.gastro.2017.07.046

Contino, G., Vaughan, T. L., Whiteman, D., & Fitzgerald, R. C. (2017). The evolving genomic landscape of Barrett's esophagus and esophageal

adenocarcinoma. *Gastroenterology.*, *153*(3), 657–673. https://doi.org/10.1016/j.physbeh.2017.03.040

Csardi, G., & Nepusz, T. (2006). The Igraph Software Package for Complex Network Research. *Inter J Complex Syst*, *1695*, 1–9. https://www.researchgate.net/publication/221995787

D'Antonio, M., & Ciccarelli, F. (2013). Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biology*, *14*(R52). https://doi.org/10.1186/gb-2013-14-5-r52

D'Antonio, Matteo, & Ciccarelli, F. D. (2011). Modification of gene duplicability during the evolution of protein interaction network. *PLoS Computational Biology*, *7*(4). https://doi.org/10.1371/journal.pcbi.1002029

D'Antonio, Matteo, Pendino, V., Shruti, S., & Ciccarelli, F. D. (2012). Network of Cancer Genes (NCG 3.0): Integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Research*, *40*(D1), 978–983. https://doi.org/10.1093/nar/gkr952

Dai, X., Yu, L., Chen, X., & Zhang, J. (2021). SNRPD1 confers diagnostic and therapeutic values on breast cancers through cell cycle regulation. *Cancer Cell International*, *21*(229), 1–16. https://doi.org/10.1186/s12935-021-01932-w

Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., & Elledge, S. J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, *155*(4), 948–962. https://doi.org/10.1016/j.cell.2013.10.011

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology*, *25*(12), e1001025. https://doi.org/10.1371/journal.pcbi.1001025

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., Wilson, R. K., & Ding, L. (2012). MuSiC: Identifying mutational significance in cancer genomes. *Genome Research*, *22*(8), 1589–1598. https://doi.org/10.1101/gr.134635.111

Dempster, J. M., Rossen, J., Kazachkova, M., Pan, J., Kugener, G., Root, D. E., & Tsherniak, A. (2019). Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. *BioRxiv*, 720243. https://doi.org/10.1101/720243

Dentro, S. C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A. G., Yu, K., Rubanova, Y., Macintyre, G., Demeulemeester, J., Vázquez-García, I., Kleinheinz, K., Livitz, D. G., Malikic, S., Donmez, N., Sengupta, S., Anur, P., Jolly, C., Cmero, M., … Yang, T. P. (2021). Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*, *184*(8), 2239–2254. https://doi.org/10.1016/j.cell.2021.03.009

Dhillon, A. S., Hagan, S., Rath, O., & Kolch, W. (2007). MAP kinase signalling pathways in cancer. In *Oncogene* (Vol. 26, Issue 22, pp. 3279–3290). Nature Publishing Group. https://doi.org/10.1038/sj.onc.1210421

Ding, J., McConechy, M. K., Horlings, H. M., Ha, G., Chun Chan, F., Funnell, T., Mullaly, S. C., Reimand, J., Bashashati, A., Bader, G. D., Huntsman, D., Aparicio, S., Condon, A., & Shah, S. P. (2015). Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nature Communications*, *6*(1), 1–13. https://doi.org/10.1038/ncomms9554

Domazet-Lošo, T., & Tautz, D. (2010). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biology*, *8*(66), 1–10. https://doi.org/10.1186/1741-7007-8-66

Dong, C., Guo, Y., Yang, H., He, Z., Liu, X., & Wang, K. (2016). iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes. *Genome Medicine*, *8*(1), 135. https://doi.org/10.1186/s13073-016-0390-0

Dressler, L., Bortolomeazzi, M., Keddar, M. R., Misetic, H., Sartini, G., Acha-sagredo, A., Montorsi, L., Wijewardhane, N., Repana, D., Nulsen, J., Goldman, J., Pollitt, M., Davis, P., Strange, A., Ambrose, K., & Ciccarelli, F. D. (2022). Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues : an update of the Network of Cancer Genes (NCG) resource. *Genome Biology*, *23*(35), 1–22.

Dulak, A. M., Schumacher, S. E., Van Lieshout, J., Imamura, Y., Fox, C., Shim,

B., Ramos, A. H., Saksena, G., Baca, S. C., Baselga, J., Tabernero, J., Barretina, J., Enzinger, P. C., Corso, G., Roviello, F., Lin, L., Bandla, S., Luketich, J. D., Pennathur, A., … Bass, A. J. (2012). Gastrointestinal adenocarcinomas of the esophagus, stomach, and colon exhibit distinct patterns of genome instability and oncogenesis. *Cancer Research*, *72*(17), 4383–4393. https://doi.org/10.1158/0008-5472.CAN-11-3893

Dulak, A. M., Stojanov, P., Peng, S., Lawrence, M. S., Fox, C., Stewart, C., Bandla, S., Imamura, Y., Schumacher, S. E., Shefler, E., McKenna, A., Carter, S. L., Cibulskis, K., Sivachenko, A., Saksena, G., Voet, D., Ramos, A. H., Auclair, D., Thompson, K., … Bass, A. J. (2013). Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature Genetics*, *45*(5), 478–486. https://doi.org/10.1038/ng.2591

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., … D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *46*(D1), D649–D655. https://doi.org/10.1093/nar/gkx1132

Fels Elliott, D. R., Perner, J., Li, X., Symmons, M. F., Verstak, B., Eldridge, M., Bower, L., O'Donovan, M., Gay, N. J., & Fitzgerald, R. C. (2017). Impact of mutations in Toll-like receptor pathway genes on esophageal carcinogenesis. *PLoS Genetics*, *13*(5), 1–21. https://doi.org/10.1371/journal.pgen.1006808

Ferlay, J., Soerjomataram, I., Dikshit, R., & et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, *136*, E359–E386. https://doi.org/10.1002/ijc.29210

Fitzgerald, R. C., Di Pietro, M., Ragunath, K., Ang, Y., Kang, J. Y., Watson, P., Trudgill, N., Patel, P., Kaye, P. V., Sanders, S., O'Donovan, M., Bird-Lieberman, E., Bhandari, P., Jankowski, J. A., Attwood, S., Parsons, S. L., Loft, D., Lagergren, J., Moayyedi, P., … De Caestecker, J. (2014). British

Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. *Gut*, *63*(1), 7–42. https://doi.org/10.1136/gutjnl-2013-305372

Fowler, J. C., King, C., Bryant, C., Hall, M. W. J., Sood, R., Ong, S. H., Earp, E., Fernandez-Antoran, D., Koeppel, J., Dentro, S. C., Shorthouse, D., Durrani, A., Fife, K., Rytina, E., Milne, D., Roshan, A., Mahububani, K., Saeb-Parsy, K., Hall, B. A., … Jones, P. H. (2021). Selection of oncogenic mutant clones in normal human skin varies with body site. *Cancer Discovery*, *11*(2), 340–361. https://doi.org/10.1158/2159-8290.CD-20-1092

Frankel, A., Armour, N., Nancarrow, D., Krause, L., Hayward, N., Lampe, G., Mark Smithers, B., & Barbour, A. (2014). Genome-wide analysis of esophageal adenocarcinoma yields specific copy number aberrations that correlate with prognosis. *Genes, Chromosomes & Cancer*, *53*, 324–338. https://doi.org/10.1002/gcc

Frankell, A. M., Jammula, S. G., 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, M. D., Tavaré, S., Fitzgerald, R. C., Noorani, A., Edwards, P. A. W., … Fitzgerald, R. C. (2019). The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nature Genetics*, *51*(3), 506–516. https://doi.org/10.1038/s41588-018-0331-5

Fu, H., Subramanian, R. R., & Masters, S. C. (2000). 14-3-3 Proteins: Structure, function, and regulation. In *Annual Review of Pharmacology and Toxicology* (Vol. 40, pp. 617–647). Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA . https://doi.org/10.1146/annurev.pharmtox.40.1.617

Fuchs, C. S., Doi, T., Jang, R. W., Muro, K., Satoh, T., Machado, M., Sun, W., Jalal, S. I., Shah, M. A., Metges, J. P., Garrido, M., Golan, T., Mandala, M., Wainberg, Z. A., Catenacci, D. V., Ohtsu, A., Shitara, K., Geva, R., Bleeker, J., … Yoon, H. H. (2018). Safety and efficacy of pembrolizumab monotherapy in patients with previously treated advanced gastric and gastroesophageal junction cancer: Phase 2 clinical KEYNOTE-059 trial.

JAMA Oncology, *4*(5), 2–9. https://doi.org/10.1001/jamaoncol.2018.0013

Futreal, P. A., Coin, L., Marshall, M., & Al., E. (2004). A census of human cancer genes. *Nature Reviews Cancer*, *4*(3), 177–183. https://doi.org/10.1038/nrc1299.A

Galipeau, P. C., Prevo, L. J., Sanchez, C. A., Longton, G. M., & Reid, B. J. (1999). Clonal expansion and loss of heterozygosity at chromosomes 9p and 17p in premalignant esophageal (Barrett's) tissue. *Journal of the National Cancer Institute*, *91*(24), 2087–2095. https://doi.org/10.1093/jnci/91.24.2087

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., & Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, *25*(12), i54–i62. https://doi.org/10.1093/bioinformatics/btp190

gdc.cancer.gov. (n.d.). *Home | NCI Genomic Data Commons*. Retrieved June 12, 2020, from https://gdc.cancer.gov/

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., Mitchell, T. J., Rubanova, Y., Anur, P., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vázquez-García, I., Haase, K., Jerman, L., Sengupta, S., Macintyre, G., … Wedge, D. C. (2020). The evolutionary history of 2,658 cancers. *Nature*, *578*(7793), 122–128. https://doi.org/10.1038/s41586-019-1907-7

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. V., Paolella, B. R., … Sellers, W. R. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, *569*(7757), 503–508. https://doi.org/10.1038/s41586-019-1186-3

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., & Ruepp, A. (2019). CORUM: The comprehensive resource of mammalian protein complexes - 2019. *Nucleic Acids Research*, *47*(D1), D559–D563. https://doi.org/10.1093/nar/gky973

Gonzalez-Perez, A., & Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, *40*(21), 1–10.

https://doi.org/10.1093/nar/gks743

Gregory, T. R. (2009). Understanding Natural Selection: Essential Concepts and Common Misconceptions. *Evolution: Education and Outreach*, *2*(2), 156–175. https://doi.org/10.1007/s12052-009-0128-1

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, *375*(12), 1109–1112. https://doi.org/10.1056/NEJMp1607591

Grossmann, P., Cristea, S., & Beerenwinkel, N. (2020). Clonal evolution driven by superdriver mutations. *BMC Evolutionary Biology*, *20*(89), 1–11. https://doi.org/10.1186/s12862-020-01647-y

Han, Y., Liu, D., & Li, L. (2020). PD-1/PD-L1 pathway: current researches in cancer. *American Journal of Cancer Research*, *10*(3), 727–742. http://www.ncbi.nlm.nih.gov/pubmed/32266087

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hart, T., & Moffat, J. (2016). BAGEL: A computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics*, *17*(164), 1–7. https://doi.org/10.1186/s12859-016-1015-8

Hu, Z., Li, Z., Ma, Z., & Curtis, C. (2020). Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. *Nature Genetics*, *52*(7), 701–708. https://doi.org/10.1038/s41588-020-0628-z

Huang, H. Y., Lin, Y. C. D., Li, J., Huang, K. Y., Shrestha, S., Hong, H. C., Tang, Y., Chen, Y. G., Jin, C. N., Yu, Y., Xu, J. T., Li, Y. M., Cai, X. X., Zhou, Z. Y., Chen, X. H., Pei, Y. Y., Hu, L., Su, J. J., Cui, S. D., … Huang, H. Da. (2020). MiRTarBase 2020: Updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Research*, *48*(D1), D148–D154. https://doi.org/10.1093/nar/gkz896

Huang, L., Jiang, S., & Shi, Y. (2020). Tyrosine kinase inhibitors for solid tumors

in the past 20 years (2001–2020). In *Journal of Hematology and Oncology* (Vol. 13, Issue 1, pp. 1–23). BioMed Central Ltd. https://doi.org/10.1186/s13045-020-00977-0

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., Von Mering, C., & Bork, P. (2016). EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–D293. https://doi.org/10.1093/nar/gkv1248

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., Von Mering, C., & Bork, P. (2019). EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., Gygi, M. P., Thornock, A., Zarraga, G., Tam, S., Szpyt, J., Gassaway, B. M., Panov, A., Parzen, H., Fu, S., Golbazi, A., Maenpaa, E., Stricker, K., Guha Thakurta, S., … Gygi, S. P. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, *184*(11), 3022-3040.e28. https://doi.org/10.1016/j.cell.2021.04.011

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., van Dyk, E., Chang, H., de Silva, H., Heyn, H., Deng, X., Egan, R. K., Liu, Q., … Garnett, M. J. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, *166*(3), 740–754. https://doi.org/10.1016/j.cell.2016.06.017

Iyer, N. G., Özdag, H., & Caldas, C. (2004). p300/CBP and cancer. In *Oncogene* (Vol. 23, Issue 24, pp. 4225–4231). Nature Publishing Group. https://doi.org/10.1038/sj.onc.1207118

Janjigian, Y. Y., Ku, G. Y., Ilson, D. H., Boyar, M. S., Capanu, M., Chou, J. F., Kelsen, D. P., Imtiaz, T., Berger, M. F., & Vakiani, E. (2015). A phase II study

of afatinib in patients (pts) with metastatic human epidermal growth factor receptor (HER2)-positive trastuzumab refractory esophagogastric (EG) cancer. *Journal of Clinical Oncology*, *33*(3_suppl), 59–59. https://doi.org/10.1200/jco.2015.33.3_suppl.59

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., … D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503. https://doi.org/10.1093/nar/gkz1031

Javitt, A., Shmueli, M. D., Kramer, M. P., Kolodziejczyk, A. A., Cohen, I. J., Kamer, I., Litchfield, K., Bab-Dinitz, E., Zadok, O., Neiens, V., Ulman, A., Radomir, L., Wolf-Levy, H., Eisenberg-Lerner, A., Kacen, A., Alon, M., Rêgo, A. T., Stacher-Priehse, E., Lindner, M., … Merbl, Y. (2021). The proteasome regulator PSME4 drives immune evasion and abrogates anti-tumor immunity in NSCLC. *BioRxiv*, *10*(24), 464690. https://doi.org/10.1101/2021.10.24.464690

Jeon, J., Luebeck, E. G., & Moolgavkar, S. H. (2006). Age effects and temporal trends in adenocarcinoma of the esophagus and gastric cardia (United States). *Cancer Causes and Control*, *17*(7), 971–981. https://doi.org/10.1007/s10552-006-0037-3

Jiang, M., Li, H., Zhang, Y., Yang, Y., Lu, R., Liu, K., Lin, S., Lan, X., Wang, H., Wu, H., Zhu, J., Zhou, Z., Xu, J., Lee, D. K., Zhang, L., Lee, Y. C., Yuan, J., Abrams, J. A., Wang, T. C., … Que, J. (2017). Transitional basal cells at the squamous-columnar junction generate Barrett's oesophagus. *Nature*, *550*(7677), 529–533. https://doi.org/10.1038/nature24269

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L.

D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kent, W. J. (2002). BLAT — The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., … Pandey, A. (2009). Human Protein Reference Database - 2009 update. *Nucleic Acids Research*, *37*(SUPPL. 1), 767–772. https://doi.org/10.1093/nar/gkn892

Killcoyne, S., Gregson, E., Wedge, D. C., Woodcock, D. J., Eldridge, M. D., de la Rue, R., Miremadi, A., Abbas, S., Blasko, A., Kosmidou, C., Januszewicz, W., Jenkins, A. V., Gerstung, M., & Fitzgerald, R. C. (2020). Genomic copy number predicts esophageal cancer years before transformation. *Nature Medicine*, *26*(11), 1726–1732. https://doi.org/10.1038/s41591-020-1033-y

Killcoyne, S., Yusuf, A., & Fitzgerald, R. C. (2021). Genomic instability signals offer diagnostic possibility in early cancer detection. *Trends in Genetics*, *37*(11), 966–972. https://doi.org/10.1016/j.tig.2021.06.009

Kim, J., Bowlby, R., Mungall, A. J., Robertson, A. G., Odze, R. D., Cherniack, A. D., Shih, J., Pedamallu, C. S., Cibulskis, C., Dunford, A., Meier, S. R., Kim, J., Raphael, J., Wu, H. T., Wong, A. M., Willis, J. E., Bass, A. J., Derks, S., Garman, K., … Zhang, J. (2017). Integrated genomic characterization of oesophageal carcinoma. *Nature*, *541*(7636), 169–174. https://doi.org/10.1038/nature20805

Kimáková, P., Solár, P., Solárová, Z., Komel, R., & Debeljak, N. (2017). Erythropoietin and its angiogenic activity. *International Journal of Molecular Sciences*, *18*(7), 1–14. https://doi.org/10.3390/ijms18071519

Kinzler, K. W., & Vogelstein, B. (1997). Gatekeepers and caretakers. *Nature*, *386*(18), 761–762. https://doi.org/10.1001/jama.287.18.2353-jwm20005-3-1

Klijn, C., Durinck, S., Stawiski, E. W., Haverty, P. M., Jiang, Z., Liu, H., Degenhardt, J., Mayba, O., Gnad, F., Liu, J., Pau, G., Reeder, J., Cao, Y., Mukhyala, K., Selvaraj, S. K., Yu, M., Zynda, G. J., Brauer, M. J., Wu, T. D., … Zhang, Z. (2015). A comprehensive transcriptional portrait of human cancer cell lines. *Nature Biotechnology*, *33*(3), 306–312. https://doi.org/10.1038/nbt.3080

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218. https://doi.org/10.1038/nature12213

Lenoir, W. F., Lim, T. L., & Hart, T. (2018). PICKLES: The database of pooled in-vitro CRISPR knockout library essentiality screens. *Nucleic Acids Research*, *46*(D1), D776–D780. https://doi.org/10.1093/nar/gkx993

Li, R., Di, L., Li, J., Fan, W., Liu, Y., Guo, W., Liu, W., Liu, L., Li, Q., Chen, L., Chen, Y., Miao, C., Liu, H., Wang, Y., Ma, Y., Xu, D., Lin, D., Huang, Y., Wang, J., … Wu, C. (2021). A body map of somatic mutagenesis in morphologically normal human tissues. *Nature*, *597*(7876), 398–403. https://doi.org/10.1038/s41586-021-03836-1

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, *1*(6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, *37*(3), 235–241. https://doi.org/10.1002/humu.22932

Liu, Y., Sethi, N. S., Hinoue, T., Schneider, B. G., Cherniack, A. D., Sanchez-Vega, F., Seoane, J. A., Farshidfar, F., Bowlby, R., Islam, M., Kim, J., Chatila, W., Akbani, R., Kanchi, R. S., Rabkin, C. S., Willis, J. E., Wang, K. K., McCall, S. J., Mishra, L., … Laird, P. W. (2018). Comparative Molecular

Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell*, *33*(4), 721–735. https://doi.org/10.1016/j.ccell.2018.03.010

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. In *Nature Genetics* (Vol. 45, Issue 6, pp. 580–585). Nature Publishing Group. https://doi.org/10.1038/ng.2653

Loo, P. Van, Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A.-L., & Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, *107*(39), 16910–16915. https://doi.org/10.1073/PNAS.1009843107

Lordick, F., Kang, Y. K., Chung, H. C., Salman, P., Oh, S. C., Bodoky, G., Kurteva, G., Volovat, C., Moiseyenko, V. M., Gorbunova, V., Park, J. O., Sawaki, A., Celik, I., Götte, H., Melezínková, H., & Moehler, M. (2013). Capecitabine and cisplatin with or without cetuximab for patients with previously untreated advanced gastric cancer (EXPAND): A randomised, open-label phase 3 trial. *The Lancet Oncology*, *14*(6), 490–499. https://doi.org/10.1016/S1470-2045(13)70102-5

Maki, R. G. (2010). Small is beautiful: Insulin-like growth factors and their role in growth, development, and cancer. *Journal of Clinical Oncology*, *28*(33), 4985–4995. https://doi.org/10.1200/JCO.2009.27.5040

Makohon-Moore, A. P., Matsukuma, K., Zhang, M., Reiter, J. G., Gerold, J. M., Jiao, Y., Sikkema, L., Attiyeh, M. A., Yachida, S., Sandone, C., Hruban, R. H., Klimstra, D. S., Papadopoulos, N., Nowak, M. A., Kinzler, K. W., Vogelstein, B., & Iacobuzio-Donahue, C. A. (2018). Precancerous neoplastic cells can move through the pancreatic ductal system. *Nature*, *561*(7722), 201–205. https://doi.org/10.1038/s41586-018-0481-8

Maley, C. C. (2007). Multistage carcinogenesis in Barrett's esophagus. *Cancer Letters*, *245*(1–2), 22–32. https://doi.org/10.1016/j.canlet.2006.03.018

Maley, C. C., Galipeau, P. C., Finley, J. C., Wongsurawat, V. J., Li, X., Sanchez, C. A., Paulson, T. G., Blount, P. L., Risques, R. A., Rabinovitch, P. S., & Reid, B. J. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genetics*, *38*(4), 468–473. https://doi.org/10.1038/ng1768

Maley, C. C., Galipeau, P. C., Li, X., Sanchez, C. A., Paulson, T. G., Blount, P. L., & Reid, B. J. (2004). The combination of genetic instability and clonal expansion predicts progression to esophageal adenocarcinoma. *Cancer Research*, *64*(20), 7629–7633. https://doi.org/10.1158/0008-5472.CAN-04-1738

Maley, C. C., Galipeau, P. C., Li, X., Sanchez, C. A., Paulson, T. G., & Reid, B. J. (2004). Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. *Cancer Research*, *64*(10), 3414–3427. https://doi.org/10.1158/0008-5472.CAN-03-3249

Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L., & Siu, L. L. (2020). Molecular profiling for precision cancer therapies. In *Genome Medicine* (Vol. 12, Issue 1, pp. 1–19). BioMed Central. https://doi.org/10.1186/s13073-019-0703-1

Mantovani, A., Allavena, P., Sica, A., & Balkwill, F. (2008). Cancer-related inflammation. In *Nature* (Vol. 454, Issue 7203, pp. 436–444). Nature Publishing Group. https://doi.org/10.1038/nature07205

Martin, T. D., Patel, R. S., Cook, D. R., Choi, M. Y., Patil, A., Liang, A. C., Li, M. Z., Haigis, K. M., & Elledge, S. J. (2021). The adaptive immune system is a major driver of selection for tumor suppressor gene inactivation. *Science*, *373*(6561), 1327–1335. https://doi.org/10.1126/science.abg5784

Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. In *Science* (Vol. 349, Issue 6255, pp. 1483–1489). American Association for the Advancement of Science. https://doi.org/10.1126/science.aab4082

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., & Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, *171*(5), 1029–1041.

https://doi.org/10.1016/j.cell.2017.09.042

McFarland, J. M., Ho, Z. V., Kugener, G., Dempster, J. M., Montgomery, P. G., Bryan, J. G., Krill-Burger, J. M., Green, T. M., Vazquez, F., Boehm, J. S., Golub, T. R., Hahn, W. C., Root, D. E., & Tsherniak, A. (2018). Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nature Communications*, *9*(4610), 1–13. https://doi.org/10.1038/s41467-018-06916-5

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., Yung, W. K. A., Bogler, O., Weinstein, J. N., VandenBerg, S., Berger, M., Prados, M., Muzny, D., Morgan, M., Scherer, S., … Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068. https://doi.org/10.1038/nature07385

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, *12*(4), R41. https://doi.org/10.1186/gb-2011-12-4-r41

Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Dharia, N. V., Montgomery, P. G., Cowley, G. S., Pantel, S., Goodale, A., Lee, Y., Ali, L. D., Jiang, G., Lubonja, R., Harrington, W. F., Strickland, M., Wu, T., Hawes, D. C., … Tsherniak, A. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nature Genetics*, *49*(12), 1779–1784. https://doi.org/10.1038/ng.3984

Michor, F., Iwasa, Y., & Nowak, M. A. (2004). Dynamics of cancer progression. *Nature Reviews Cancer*, *4*(3), 197–205. https://doi.org/10.1038/nrc1295

Minacapelli, C. D., Bajpai, M., Geng, X., Cheng, C. L., Chouthai, A. A., Souza, R., Spechler, S. J., & Das, K. M. (2017). Barrett's metaplasia develops from cellular reprograming of esophageal squamous epithelium due to gastroesophageal reflux. *American Journal of Physiology - Gastrointestinal and Liver Physiology*, *312*(6), G615–G622.

https://doi.org/10.1152/ajpgi.00268.2016

Moore, L., Cagan, A., Coorens, T. H. H., Neville, M. D. C., Sanghvi, R., Sanders, M. A., Oliver, T. R. W., Leongamornlert, D., Ellis, P., Noorani, A., Mitchell, T. J., Butler, T. M., Hooks, Y., Warren, A. Y., Jorgensen, M., Dawson, K. J., Menzies, A., O'Neill, L., Latimer, C., … Rahbari, R. (2021). The mutational landscape of human somatic and germline cells. *Nature*, *597*(7876), 381–386. https://doi.org/10.1038/s41586-021-03822-7

Mourikis, T. P., Benedetti, L., Foxall, E., Temelkovski, D., Nulsen, J., Perner, J., Cereda, M., Lagergren, J., Howell, M., Yau, C., Fitzgerald, R. C., Scaffidi, P., & Ciccarelli, F. D. (2019). Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. *Nature Communications*, *10*(1), 1–17. https://doi.org/10.1038/s41467-019-10898-3

Murugaesu, N., Wilson, G. A., Birkbak, N. J., Watkins, T. B. K., McGranahan, N., Kumar, S., Abbassi-Ghadi, N., Salm, M., Mitter, R., Horswell, S., Rowan, A., Phillimore, B., Biggs, J., Begum, S., Matthews, N., Hochhauser, D., Hanna, G. B., & Swanton, C. (2015). Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discovery*, *5*(8), 821–832. https://doi.org/10.1158/2159-8290.CD-15-0412

MySQL :: MySQL 8.0 Reference Manual. (n.d.). Retrieved March 21, 2022, from https://dev.mysql.com/doc/refman/8.0/en/

Nakatani, Y., Takeda, H., Kohara, Y., & Morishita, S. (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Research*, *17*(9), 1254–1265. https://doi.org/10.1101/gr.6316407

Negrini, S., Gorgoulis, V. G., & Halazonetis, T. D. (2010). Genomic instability - an evolving hallmark of cancer. *Nature Reviews Molecular Cell Biology*, *11*(3), 220–228. https://doi.org/10.1038/nrm2858

Newell, F., Patel, K., Gartside, M., Krause, L., Brosda, S., Aoude, L. G., Loffler, K. A., Bonazzi, V. F., Patch, A. M., Kazakoff, S. H., Holmes, O., Xu, Q., Wood, S., Leonard, C., Lampe, G., Lord, R. V., Whiteman, D. C., Pearson, J. V., Nones, K., … Barbour, A. P. (2019). Complex structural

rearrangements are present in high-grade dysplastic Barrett's oesophagus samples. *BMC Medical Genomics*, *12*(31), 1–14. https://doi.org/10.1186/s12920-019-0476-9

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

Ng, S. W. K., Rouhani, F. J., Brunner, S. F., Brzozowska, N., Aitken, S. J., Yang, M., Abascal, F., Moore, L., Nikitopoulou, E., Chappell, L., Leongamornlert, D., Ivovic, A., Robinson, P., Butler, T., Sanders, M. A., Williams, N., Coorens, T. H. H., Teague, J., Raine, K., … Campbell, P. J. (2021). Convergent somatic mutations in metabolism genes in chronic liver disease. *Nature*, *598*(7881), 473–478. https://doi.org/10.1038/s41586-021-03974-6

Nones, K., Waddell, N., Wayte, N., Patch, A. M., Bailey, P., Newell, F., Holmes, O., Fink, J. L., Quinn, M. C. J., Tang, Y. H., Lampe, G., Quek, K., Loffler, K. A., Manning, S., Idrisoglu, S., Miller, D., Xu, Q., Waddell, N., Wilson, P. J., … Barbour, A. P. (2014). Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nature Communications*, *5*, 5224. https://doi.org/10.1038/ncomms6224

Normanno, N., De Luca, A., Bianco, C., Strizzi, L., Mancino, M., Maiello, M. R., Carotenuto, A., De Feo, G., Caponigro, F., & Salomon, D. S. (2006). Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene*, *366*(1), 2–16. https://doi.org/10.1016/j.gene.2005.10.018

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, *194*(4260), 23–28. https://doi.org/10.1126/science.959840

Nowicki-Osuch, K., Zhuang, L., Jammula, S., Bleaney, C. W., Mahbubani, K. T., Devonshire, G., Katz-Summercorn, A., Eling, N., Wilbrey-Clark, A., Madissoon, E., Gamble, J., Di Pietro, M., O'Donovan, M., Meyer, K. B., Saeb-Parsy, K., Sharrocks, A. D., Teichmann, S. A., Marioni, J. C., & Fitzgerald, R. C. (2021). Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition. *Science*, *373*(6556), 760–767. https://doi.org/10.1126/science.abd1449

Nulsen, J., Misetic, H., Yau, C., & Ciccarelli, F. D. (2021). Pan-cancer detection

of driver genes at the single-patient resolution. *Genome Medicine*, *13*(12), 1–14. https://doi.org/10.1186/s13073-021-00830-0

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

OCCAMS | Oesophageal Cancer Clinical and Molecular Stratification. (n.d.). Retrieved March 22, 2022, from https://www.occams.org.uk/

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., … Hermjakob, H. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, *42*(D1), D358–D363. https://doi.org/10.1093/nar/gkt1115

Oughtred, R., Stark, C., Breitkreutz, B. J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., & Tyers, M. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, *47*(D1), D529–D541. https://doi.org/10.1093/nar/gky1079

Owen, R. P., White, M. J., Severson, D. T., Braden, B., Bailey, A., Goldin, R., Wang, L. M., Ruiz-Puig, C., Maynard, N. D., Green, A., Piazza, P., Buck, D., Middleton, M. R., Ponting, C. P., Schuster-Böckler, B., & Lu, X. (2018). Single cell RNA-seq reveals profound transcriptional similarity between Barrett's oesophagus and oesophageal submucosal glands. *Nature Communications*, *9*(4261), 1–12. https://doi.org/10.1038/s41467-018-06796-9

Pavlakis, N., Sjoquist, K. M., Martin, A. J., Tsobanis, E., Yip, S., Kang, Y. K., Bang, Y. J., Alcindor, T., O'Callaghan, C. J., Burnell, M. J., Tebbutt, N. C.,

Rha, S. Y., Lee, J., Cho, J. Y., Lipton, L. R., Wong, M., Strickland, A., Kim, J. W., Zalcberg, J. R., … Goldstein, D. (2016). Regorafenib for the treatment of advanced gastric cancer (INTEGRATE): A multinational placebo-controlled phase II Trial. *Journal of Clinical Oncology*, *34*(23), 2728–2735. https://doi.org/10.1200/JCO.2015.65.1901

Peters, Y., Al-Kaabi, A., Shaheen, N. J., & Al, E. (2019). Barrett oesophagus. *Disease Primers*, *5*(35). https://doi.org/10.1038/s41572-019-0086-z

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121. https://doi.org/10.1101/gr.097857.109

Quante, M., Bhagat, G., Abrams, J. A., Marache, F., Good, P., Lee, M. D., Lee, Y., Friedman, R., Asfaha, S., Dubeykovskaya, Z., Mahmood, U., Figueiredo, J. L., Kitajewski, J., Shawber, C., Lightdale, C. J., Rustgi, A. K., & Wang, T. C. (2012). Bile acid and inflammation activate gastric cardia stem cells in a mouse model of Barrett-like metaplasia. *Cancer Cell*, *21*(1), 36–51. https://doi.org/10.1016/j.ccr.2011.12.004

Quidville, V., Alsafadi, S., Goubar, A., Commo, F., Scott, V., Pioche-Durieu, C., Girault, I., Baconnais, S., Le Cam, E., Lazar, V., Delaloge, S., Saghatchian, M., Pautier, P., Morice, P., Dessen, P., Vagner, S., & Andre, F. (2013). Targeting the deregulated spliceosome core machinery in cancer cells triggers mTOR blockade and autophagy. *Cancer Research*, *73*(7), 2247–2258. https://doi.org/10.1158/0008-5472.CAN-12-2501

Quinton, R. J., DiDomizio, A., Vittoria, M. A., Kotýnková, K., Ticas, C. J., Patel, S., Koga, Y., Vakhshoorzadeh, J., Hermance, N., Kuroda, T. S., Parulekar, N., Taylor, A. M., Manning, A. L., Campbell, J. D., & Ganem, N. J. (2021). Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature*, *590*(7846), 492–497. https://doi.org/10.1038/s41586-020-03133-3

Rambaldi, D., Giorgi, F., & Al., E. (2008). Low Duplicability and Network Fragility of Cancer Genes. *Trends Genet*, *24*(June), 427–430.

Raphael, B. J., Dobson, J. R., Oesper, L., & Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: Computational approaches to

enable precision medicine. In *Genome Medicine* (Vol. 6, Issue 5, pp. 1–17). BioMed Central. https://doi.org/10.1186/gm524

Reimand, J., & Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular Systems Biology*, *9*(1), 637. https://doi.org/10.1038/msb.2012.68

Repana, D., Nulsen, J., Dressler, L., Bortolomeazzi, M., Venkata, S. K., Tourna, A., Yakovleva, A., Palmieri, T., & Ciccarelli, F. D. (2019). The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biology*, *20*(1), 1–12. https://doi.org/10.1186/s13059-018-1612-0

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, *39*(17), e118. https://doi.org/10.1093/nar/gkr407

Rice, T. W., Patil, D. T., & Blackstone, E. H. (2017). 8th edition AJCC/UICC staging of cancers of the esophagus and esophagogastric junction: Application to clinical practice. *Annals of Cardiothoracic Surgery*, *6*(2), 119–130. https://doi.org/10.21037/acs.2017.03.14

Robinson, P. S., Coorens, T. H. H., Palles, C., Mitchell, E., Abascal, F., Olafsson, S., Lee, B. C. H., Lawson, A. R. J., Lee-Six, H., Moore, L., Sanders, M. A., Hewinson, J., Martin, L., Pinna, C. M. A., Galavotti, S., Rahbari, R., Campbell, P. J., Martincorena, I., Tomlinson, I., & Stratton, M. R. (2021). Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nature Genetics*, *53*(10), 1434–1442. https://doi.org/10.1038/s41588-021-00930-y

Ross-Innes, C. S., Becq, J., Warren, A., Cheetham, R. K., Northen, H., O'Donovan, M., Malhotra, S., Di Pietro, M., Ivakhno, S., He, M., Weaver, J. M. J., Lynch, A. G., Kingsbury, Z., Ross, M., Humphray, S., Bentley, D., Fitzgerald, R. C., Hayes, S. J., Ang, Y., … Dawson, S. (2015). Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nature Genetics*, *47*(9), 1038–1046. https://doi.org/10.1038/ng.3357

Rubenstein, J. H., & Shaheen, N. J. (2015). Epidemiology, diagnosis, and

management of esophageal adenocarcinoma. *Gastroenterology*, *149*(2), 302–317. https://doi.org/10.1053/j.gastro.2015.04.053

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., & Mewes, H. W. (2009). CORUM: The comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research*, *38*(SUPPL.1), D497–D501. https://doi.org/10.1093/nar/gkp914

Runge, T. M., Abrams, J. A., Shaheen, N. J. (2015). Epidemiology of Barrett's esophagus and esophageal adenocarcinoma. *Gastroenterology Clin North Am*, *44*(2), 203–231. https://doi.org/10.1016/j.physbeh.2017.03.040

Saini, N., Giacobone, C. K., Klimczak, L. J., Papas, B. N., Burkholder, A. B., Li, J. L., Fargo, D. C., Bai, R., Gerrish, K., Innes, C. L., Schurman, S. H., & Gordenin, D. A. (2021). UV-exposure, endogenous DNA damage, and DNA replication errors shape the spectra of genome changes in human skin. *PLoS Genetics*, *17*(1), e1009302. https://doi.org/10.1371/journal.pgen.1009302

Saito, Y., Koya, J., Araki, M., Kogure, Y., Shingaki, S., Tabata, M., McClure, M. B., Yoshifuji, K., Matsumoto, S., Isaka, Y., Tanaka, H., Kanai, T., Miyano, S., Shiraishi, Y., Okuno, Y., & Kataoka, K. (2020). Landscape and function of multiple mutations within individual oncogenes. *Nature*, *582*(7810), 95–99. https://doi.org/10.1038/s41586-020-2175-2

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, *32*(90001), D449–D451. https://doi.org/10.1093/nar/gkh086

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, *28*(14), 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

Sawas, T., Killcoyne, S., Iyer, P. G., Wang, K. K., Smyrk, T. C., Kisiel, J. B., Qin, Y., Ahlquist, D. A., Rustgi, A. K., Costa, R. J., Gerstung, M., Fitzgerald, R. C., Katzka, D. A., Noorani, A., Edwards, P. A. W., Grehan, N., Nutzinger, B., Hughes, C., Fidziukiewicz, E., … Turkington, R. (2018). Identification of

Prognostic Phenotypes of Esophageal Adenocarcinoma in 2 Independent Cohorts. *Gastroenterology*, *155*(6), 1720–1728. https://doi.org/10.1053/j.gastro.2018.08.036

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. In *Nature Methods* (Vol. 7, Issue 8, pp. 575–576). Nature Publishing Group. https://doi.org/10.1038/nmeth0810-575

Secrier, M., Li, X., De Silva, N., Eldridge, M. D., Contino, G., Bornschein, J., Macrae, S., Grehan, N., O'Donovan, M., Miremadi, A., Yang, T. P., Bower, L., Chettouh, H., Crawte, J., Galeano-Dalmau, N., Grabowska, A., Saunders, J., Underwood, T., Waddell, N., … Grimmond, S. M. (2016). Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nature Genetics*, *48*(10), 1131–1141. https://doi.org/10.1038/ng.3659

Segers, V. F. M., Dugaucquier, L., Feyen, E., Shakeri, H., & De Keulenaer, G. W. (2020). The role of ErbB4 in cancer. *Cellular Oncology*, *43*(3), 335–352. https://doi.org/10.1007/s13402-020-00499-4

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M., & Gaunt, T. R. (2013). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, *34*(1), 57–65. https://doi.org/10.1002/humu.22225

Smyth, E. C., Lagergren, J., Fitzgerald, R. C., Lordick, F., Shah, M. A., Lagergren, P., & Cunningham, D. (2017). Oesophageal Cancer. *Nat. Rev. Dis. Primers*, *3*, 17048. https://doi.org/10.1038/nrdp.2017.48

Smyth, E. C., Nilsson, M., Grabsch, H. I., van Grieken, N. C., & Lordick, F. (2020). Gastric cancer. *The Lancet*, *396*(10251), 635–648. https://doi.org/10.1016/S0140-6736(20)31288-5

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. In *Nature Reviews Cancer* (Vol. 18, Issue 11, pp. 696–705). Nature Publishing Group. https://doi.org/10.1038/s41568-018-

0060-1

Song, Y., Song, W., Li, Z., Song, W., Wen, Y., Li, J., Xia, Q., & Zhang, M. (2020). CDC27 Promotes Tumor Progression and Affects PD-L1 Expression in T-Cell Lymphoblastic Lymphoma. *Frontiers in Oncology*, *10*(488), 1–13. https://doi.org/10.3389/fonc.2020.00488

Stachler, M. D., Taylor-Weiner, A., Peng, S., McKenna, A., Agoston, A. T., Odze, R. D., Davison, J. M., Nason, K. S., Loda, M., Leshchiner, I., Stewart, C., Stojanov, P., Seepo, S., Lawrence, M. S., Ferrer-Torres, D., Lin, J., Chang, A. C., Gabriel, S. B., Lander, E. S., … Bass, A. J. (2015). Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nature Genetics*, *47*(9), 1047–1055. https://doi.org/10.1038/ng.3343

Sugiyama, A., Miyagi, Y., Komiya, Y., Kurabe, N., Kitanaka, C., Kato, N., Nagashima, Y., Kuchino, Y., & Tashiro, F. (2003). Forced expression of antisense 14-3-3 beta RNA suppresses tumor cell growth in vitro and in vivo. *Carcinogenesis*, *24*(9), 1549–1559. https://doi.org/10.1093/carcin/bgg113

Syed, A. S., D'Antonio, M., & Ciccarelli, F. D. (2010). Network Of Cancer Genes: A web resource to analyze duplicability, orthology and network properties of cancer genes. *Nucleic Acids Research*, *38*(SUPPL.1), 670–675. https://doi.org/10.1093/nar/gkp957

Takihara, Y., Matsuda, Y., & Hara, J. (2000). Role of the beta isoform of 14-3-3 proteins in cellular proliferation and oncogenic transformation. *Carcinogenesis*, *21*(11), 2073–2077. https://doi.org/10.1093/carcin/21.11.2073

Tamborero, D., Gonzalez-Perez, A., & Lopez-Bigas, N. (2013). OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, *29*(18), 2238–2244. https://doi.org/10.1093/bioinformatics/btt395

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M. S., Getz, G., Bader, G. D., Ding, L., & Lopez-Bigas, N. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports*, *3*, 2650. https://doi.org/10.1038/srep02650

Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M. P., Vivancos, A., Rovira, A., Tusquets, I., Albanell, J., Rodon, J., Tabernero, J., de Torres, C., Dienstmann, R., Gonzalez-Perez, A., & Lopez-Bigas, N. (2018). Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, *10*(1), 25. https://doi.org/10.1186/s13073-018-0531-8

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Tini, M., Benecke, A., Um, S. J., Torchia, J., Evans, R. M., & Chambon, P. (2002). Association of CBP/p300 acetylase and thymine DNA glycosylase links DNA repair and transcription. *Molecular Cell*, *9*(2), 265–277. https://doi.org/10.1016/S1097-2765(02)00453-7

Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G., & Vogelstein, B. (2015). Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(1), 118–123. https://doi.org/10.1073/pnas.1421839112

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., Meyers, R. M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W. F. J., Howell, S., Merkel, E., … Hahn, W. C. (2017). Defining a Cancer Dependency Map. *Cell*, *170*(3), 564-576.e16. https://doi.org/10.1016/j.cell.2017.06.010

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I. M., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A. K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. https://doi.org/10.1126/science.1260419

Van Allen, E. M., Wagle, N., Stojanov, P., Perrin, D. L., Cibulskis, K., Marlow, S., Jane-Valbuena, J., Friedrich, D. C., Kryukov, G., Carter, S. L., McKenna, A., Sivachenko, A., Rosenberg, M., Kiezun, A., Voet, D., Lawrence, M., Lichtenstein, L. T., Gentry, J. G., Huang, F. W., … Garraway, L. A. (2014). Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature Medicine*, *20*(6), 682–688. https://doi.org/10.1038/nm.3559

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., & Al., E. (2013). Cancer genome landscapes. *Science*, *339*(6127), 1546:1558. https://doi.org/10.1126/science.233.4770.1246-b

Vogelstein, Bert, & Kinzler, K. W. (2015). The Path to Cancer — Three Strikes and You’re Out. *New England Journal of Medicine*, *373*(20), 1895–1898. https://doi.org/10.1056/NEJMp1508811

Wagner, A. H., Walsh, B., Mayfield, G., Tamborero, D., Sonkin, D., Krysiak, K., Deu-Pons, J., Duren, R. P., Gao, J., McMurry, J., Patterson, S., del Vecchio Fitz, C., Pitel, B. A., Sezerman, O. U., Ellrott, K., Warner, J. L., Rieke, D. T., Aittokallio, T., Cerami, E., … Margolin, A. A. (2020). A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nature Genetics*, *52*(4), 448–457. https://doi.org/10.1038/s41588-020-0603-8

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), 1–7. https://doi.org/10.1093/nar/gkq603

Wang, X., Ouyang, H., Yamamoto, Y., Kumar, P. A., Wei, T. S., Dagher, R., Vincent, M., Lu, X., Bellizzi, A. M., Ho, K. Y., Crum, C. P., Xian, W., & McKeon, F. (2011). Residual embryonic cells as precursors of a Barrett’s-like metaplasia. *Cell*, *145*(7), 1023–1035. https://doi.org/10.1016/j.cell.2011.05.026

Wang, Y., Zheng, K., Huang, Y., Xiong, H., Su, J., Chen, R., & Zou, Y. (2021). PARP inhibitors in gastric cancer: beacon of hope. In *Journal of Experimental and Clinical Cancer Research* (Vol. 40, Issue 211, pp. 1–13). BioMed Central Ltd. https://doi.org/10.1186/s13046-021-02005-6

Weaver, J. M. J., Ross-Innes, C. S., Shannon, N., Lynch, A. G., Forshew, T., Barbera, M., Murtaza, M., Ong, C. A. J., Lao-Sirieix, P., Dunning, M. J., Smith, L., Smith, M. L., Anderson, C. L., Carvalho, B., O'donovan, M., Underwood, T. J., May, A. P., Grehan, N., Hardwick, R., … O'Neil, J. R. (2014). Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nature Genetics*, *46*(8), 837–843. https://doi.org/10.1038/ng.3013

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, *28*(4), 1–20.

Wijewardhane, N., Dressler, L., & Ciccarelli, F. D. (2021). Normal Somatic Mutations in Cancer Transformation. *Cancer Cell*, *39*(2), 125–129. https://doi.org/10.1016/j.ccell.2020.11.002

Wilke, H., Muro, K., Van Cutsem, E., Oh, S. C., Bodoky, G., Shimada, Y., Hironaka, S., Sugimoto, N., Lipatov, O., Kim, T. Y., Cunningham, D., Rougier, P., Komatsu, Y., Ajani, J., Emig, M., Carlesi, R., Ferry, D., Chandrawansa, K., Schwartz, J. D., & Ohtsu, A. (2014). Ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (RAINBOW): A double-blind, randomised phase 3 trial. *The Lancet Oncology*, *15*(11), 1224–1235. https://doi.org/10.1016/S1470-2045(14)70420-6

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maclejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, Di., … Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037

Woo, J., Cohen, S. A., & Grim, J. E. (2015). Targeted therapy in gastroesophageal cancers: Past, present and future. *Gastroenterology Report*, *3*(4), 316–329. https://doi.org/10.1093/gastro/gov052

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z.,

Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., … Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*, *318*(5853), 1108–1113. https://doi.org/10.1126/science.1145720

www.cancerresearchuk.org. (n.d.). *Oesophageal cancer statistics | Cancer Research UK*. Retrieved May 3, 2020, from https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer#heading-Zero

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., & Li, T. (2009). miRecords: An integrated resource for microRNA-target interactions. *Nucleic Acids Research*, *37*(SUPPL. 1), D105–D110. https://doi.org/10.1093/nar/gkn851

Xie, S. H., Fang, R., Huang, M., Dai, J., Thrift, A. P., Anderson, L. A., Chow, W. H., Bernstein, L., Gammon, M. D., Risch, H. A., Shaheen, N. J., Reid, B. J., Wu, A. H., Iyer, P. G., Liu, G., Corley, D. A., Whiteman, D. C., Caldas, C., Pharoah, P. D., … Lagergren, J. (2020). Association Between Levels of Sex Hormones and Risk of Esophageal Adenocarcinoma and Barrett's Esophagus. *Clinical Gastroenterology and Hepatology*, *18*(12), 2701-2709.e3. https://doi.org/10.1016/j.cgh.2019.11.030

Xie, S. H., & Lagergren, J. (2016). A global assessment of the male predominance in esophageal adenocarcinoma. *Oncotarget*, *7*(25), 38876–38883. https://doi.org/10.18632/oncotarget.9113

Xu, E., Gu, J., Hawk, E. T., Wang, K. K., Lai, M., Huang, M., Ajani, J., & Wu, X. (2013). Genome-wide methylation analysis shows similar patterns in Barrett's esophagus and esophageal adenocarcinoma. *Carcinogenesis*, *34*(12), 2750–2756. https://doi.org/10.1093/carcin/bgt286

Young, K., & Chau, I. (2016). Targeted Therapies for Advanced Oesophagogastric Cancer: Recent Progress and Future Directions. *Drugs*, *76*(1), 13–26. https://doi.org/10.1007/s40265-015-0510-y