

# A Compendium of Genetic Drivers for Oesophageal Adenocarcinoma defines Prognostic and Therapeutic Biomarkers for use in the Clinic

Mr. Alexander M Frankell

MRC Cancer Unit, Hutchison/MRC Research Centre, University of Cambridge

## **Supervisor:**

Prof. Rebecca C. Fitzgerald

Group Leader, MRC Cancer Unit, Hutchison/MRC Research Centre, University of Cambridge

Honorary Consultant Gastroenterologist Addenbrooke's Hospital

## **Secondary Supervisor:**

Dr. Saraki Vanharanta

Group Leader, MRC Cancer Unit, Hutchison/MRC Research Centre, University of Cambridge

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

Corpus Christi College, University of Cambridge

Submitted March 2019

-- This page is deliberately left blank --

# A Compendium of Genetic Drivers for Oesophageal Adenocarcinoma defines Prognostic and Therapeutic Biomarkers for use in the Clinic

Mr. Alexander M Frankell

**Background:** Oesophageal adenocarcinoma (OAC) is a poor-prognosis cancer type with rapidly rising incidence. Understanding of the genetic events driving OAC development is limited, and there are few molecular biomarkers for prognostication or therapeutics. This study aimed to use a large cohort of genomically characterised OACs to determine the landscape of genetic driver events in OAC and their possible clinic uses.

**Methods:** We have collated a cohort of 551 genomically characterized OACs (398 whole genome sequenced and 153 whole exome sequenced) with a sub-cohort matched to RNA sequencing data (116 cases). Strelka, Manta and ASCAT were used to call SNVs and indels, structural variants and copy number aberrations respectively. A suite of published tools was used to detect regions of the genome under positive selection for mutations in OAC including dNdScv, Mutsigcv, OncodriveFM and others. Copy number drivers were identified using GISTIC and correlations with matched expression data. Univariate and Multivariate cox regressions were used to identify prognostic biomarkers. Treatment of *In vitro* cultures of human OAC cell lines and organoids with a range of targeted therapeutics was used to calculate AUCs and GI50s for various drugs in OAC models. These sensitivities were matched to the genomic background in these cell lines, provided by whole genome sequencing.

**Results:** We discovered 77 putative OAC driver genes and 21 putative noncoding driver elements for OAC. We identified a mean of 4.4 driver events per tumour, which were derived more commonly from mutations than copy number alterations and compared the prevalence of these mutations to the exome-wide mutational excess calculated using non-synonymous to synonymous mutation ratios (dN/dS). We observed mutual exclusivity or co-occurrence of events within and between several dysregulated OAC pathways, a result suggestive of strong functional relationships. Indicators of poor prognosis (*SMAD4* and *GATA4*) were verified in independent cohorts with significant predictive value. Over 50% of OACs contained sensitizing events for CDK4 and CDK6 inhibitors, which were highly correlated with clinically relevant sensitivity in a panel of OAC cell lines and organoids. In a smaller panel of OACs we also saw evidence for specificity of BET inhibitor efficacy to *MYC* amplified OACs, however did not observe responses to EZH2 inhibitors, designed to target SWI/SNF mutated cancers, even upon induction of these mutations using CRISPR-Cas9.

**Discussion:** We have compiled the most comprehensive analysis to date of positively selected genomic elements in OAC, significantly improving upon previous analyses. We use this to identify prognostic and therapeutic biomarkers with considerable potential clinical value. Limitation to this study include a lack of RNA sequencing on all samples, making it difficult to assess selection for low-frequency copy number events. Future directions include functional investigation of many of the novel driver genes identified and prospective validation of clinical biomarkers in the Oelixir trial, including use of CDK4/6 inhibitors in OAC patients.

## **Disclaimer**

I hereby declare that my thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text

This thesis does not exceed the prescribed 60,000 word limit set by the Clinical Medicine and Clinical Veterinary Medicine Degree Committee excluding figures, photographs, tables and bibliography

Alexander M Frankell

## Acknowledgements

I would like to take this opportunity to thank the many people that have supported me throughout my doctoral training.

I am extremely grateful to Prof. Fitzgerald for her guidance and supervision. Over the past four years of her supervision I feel I have grown hugely as a scientist and am now able to tackle difficult problems independently, which is the greatest gift a teacher can give. I am also thankful for her compassion during difficult moments both relating to my work and during times of personal hardship.

There are also many other members of the lab that have been generous with their time and expertise to help me progress during my doctoral training. In particular, Dr Gianmarco Contino (Clinical Scientist, Fitzgerald Group) guided me through my first steps as a research student, training me in certain techniques and helping me grapple with complex research questions. Dr SriGanesh Jammula (Post-doctoral Informatician, Fitzgerald & Tavare Groups) was instrumental in the later stages of my Doctoral training when my focus switched to informatics, a field in which I did not have previous experience before arriving in the lab. He had great generosity with his time and has also be a great friend in the lab.

I would also like to thank my family and friends for their support during these years where I have commonly been unavailable due to commitments in the lab. Their understanding and support for me despite this has been unconditional.

Lastly, I would like to dedicate this thesis to my Grandmother, Dr Wendy Smaridge. She was a successful scientist in the early 1950s when she worked at Queen Mary College, University of London receiving her PhD on elucidating mechanisms of Hemoglobin degradation in *Daphnia* (Water flea). Her subsequent post-doctoral work on the same subject was published in the prestigious journal *Nature* and is still be found referred to in the latest textbook on *Daphnia* physiology. She was unable to continue her career in research for long due to family commitments was took great pleasure in hearing about my progress during my time in research. She even visited the lab in 2015 and shadowed me while performing western blot experiments. Wendy passed away in 2017, but her memory will continue to inspire and drive me for the rest of my career.

## Abbreviations

SNV - Single Nucleotide Variant  
Indel - short insertion or deletion  
CNA - Copy Number Aberration  
SV - Structural variant  
PCAWG - Pan-Cancer Analysis of Whole Genomes  
OAC - Oesophageal Adenocarcinoma  
BO - Barrett's Oesophagus  
GOJ - Gastro-Oesophageal Junction  
IM - Intestinal Metaplasia  
GORD - Gastro-Oesophageal Reflux Disease  
LGD - Low Grade Dysplasia  
HGD - High Grade Dysplasia  
SNP - Single Nucleotide Polymorphism  
NDBO - Non-Dysplastic Barrett's Oesophagus  
TSG - Tumour Suppressor Gene  
TNM - Tumour Node Metastasis  
FISH - Fluorescence In-Situ Hybridisation  
IHC - Immunohistochemistry  
MSI - Microsatellite Instability  
ICGC - The International Cancer Genome Consortium  
OCCAMS - The Oesophageal Cancer Classification and Molecular Stratification consortium  
FDR - False Discovery Rate  
HR - Hazard Ratio  
TPM - Transcripts Per Million  
CN - Copy Number  
LOH - Loss of Heterozygosity  
WT - Wild type  
BH - Benjamini & Hochberg  
TCGA - The Cancer Genome Atlas  
WGS - Whole Genome Sequencing  
GI50 - Concentration corresponding to 50% growth inhibition  
AUC - Area Under the Curve  
FCS - Fetal Calf Serum  
PAGE - Polyacrylamide Gel Electrophoresis  
HRP - Horse Radish Peroxidase  
PBS - Phosphate Buffered Saline  
EV - Empty Vector  
FPKM - Fragments Per Kilobase of transcripts per Million  
MHC - Major Histocompatibility Complex

IQR - Inter-Quartile Range

TGF $\beta$  - Transforming Growth Factor  $\beta$

RTK - Receptor Tyrosine Kinase

BET - Bromodomain and Extra-Terminal domain

OSCC - Oesophageal Squamous Cell Carcinoma

## Collaborations

The type of work undertaken in this thesis is not possible without the support of a very large network of individuals with a very wide variety of expertise in a number of fields. Sample collection was undertaken by a number of surgeons, endoscopists and a team of research nurses including; Dr. Gianmarco Contino, Dr. Annalise Katz-Summercorn, Dr. Ayesha Noorani, Dr. Massimiliano Di'Pietro, Rachel de la Rue, Irene Debiram-Beecham, Nicola Grehan and Tara Nuckcheddy-Grant as well as many other individuals in the OCCAMS consortium. Samples were then managed by these individuals and other administrators including; Dr. Shona Macrae, Jason Crawte, Alex Northrop and others. All samples underwent pathological quality control by a team of pathologists lead by Dr. Maria O'Donovan. DNA and RNA extraction was performed by Jason Crawte and Alex Northrop. DNA samples were then sent to Illumina for library preparation and whole genomic sequencing. Library preparation and sequencing of RNA was undertaken by Sujath Abbas. Pipelines for quality control and primary analysis (alignment, SNV & Indel calling, copy number calling, Structural variant calling, Transcript read counts) of the sequencing data were constructed and managed by a team of Bioinformaticians including Lawrence Bower, Ginny Devonshire, Dr. Juliane Perner, Dr. Maria Secier, Dr. SriGanesh Jammula, Dr. Gianmarco Contino, Dr. Matthew Eldridge and others. These outputs files were then used by the PhD candidate to produce the analysis set out in this work. Areas in which collaborations aided such secondary analysis are listed below:

Dr. SriGanesh Jammula performed a number of analyses for detection and characterisation of Driver genes in close conjunction with the PhD candidate. Dr. Jammula ran a minority of the tools for detection of positive selection; MutsigCV, GISTIC and ActiveDriverWGS. He also implemented filtering approaches for baseline expression in the final driver list. Finally Dr. Jammula constructed a workflow for concurrent visualization of copy number and structural variant events to assess for signature of extrachromosomal DNA.

Dr. Gianmarco Contino worked closely with the PhD candidate to construct a Circos plot describing recurrently amplified and deletion regions of the genome and genes in which overexpression was occurring concurrently with these copy number changes. The analysis describing these genes or regions and their relationship to expression was performed solely by the PhD candidate, however the final visualization step was mostly undertaken by Dr. Contino.

While genome sequencing of oesophageal adenocarcinoma cell lines was also aided by Emma Ococks, who managed the sequencing of 2 of the cell lines used, Dr. Rachel Fels Elliot who managed sequencing of 7 of the cell lines used and Dr. Gianmarco Contino who constructed the original pipelines for filtering of germline SNPs in these cell lines, adapted by the PhD Candidate.

Organoid work was performed by Dr. Xiaodun Li, although the experiments were designed in conjunction with the PhD candidate.

James Mok, a pre-doctoral student under the supervision of the PhD candidate aided some of the repeat experiments for the cell line drug responses for CDK4/6 inhibitors.

Over the course of this PhD two predoctoral students in the Fitzgerald group were placed under the supervision of the PhD candidate; James Mok and Kleovolous Kofonicolas. These students both undertook projects under close supervision, both in terms of project direction and training in experimental and informatic techniques, where activity of additional drug targets was investigated. These drugs were chosen based on the genomic driver landscape of oesophageal adenocarcinoma and were in addition to the originally investigated agents, CDK4/6 inhibitors, by the PhD candidate.

It should also be noted that sections of this Thesis are partly copied from the candidate's Journal Article Frankell *et. al.* 2019 published in *Nature Genetics*. However, all of the copied sections were directly written by the candidate.

# Table of Contents

<b>Abstract .....</b>	<b>13</b>
<b>Introduction .....</b>	<b>15</b>
<b>Principals of cancer evolution .....</b>	<b>15</b>
<b>Detecting positive selection in cancer genomes .....</b>	<b>17</b>
Principals and coding mutations .....	17
Structural variation.....	20
Non-coding mutations.....	23
<b>Oesophageal adenocarcinoma.....</b>	<b>24</b>
Pathology and Aetiology .....	24
Driver events in oesophageal adenocarcinoma .....	26
Genetics of OAC evolution from BO.....	30
<b>Clinical pathway for OAC .....</b>	<b>33</b>
<b>Biomarkers.....</b>	<b>33</b>
Biomarkers for prognosis in oesophageal adenocarcinoma .....	34
Biomarkers for therapeutic intervention in oesophageal adenocarcinoma .....	36
<b>Hypothesis &amp; Aims .....</b>	<b>38</b>
<b>Methods .....</b>	<b>39</b>
<b>Calling genomic aberrations in OAC.....</b>	<b>39</b>
<b>RNA sequencing .....</b>	<b>40</b>
<b>Analysing OAC mutations for selection .....</b>	<b>40</b>
<b>Detecting selection in CNAs.....</b>	<b>43</b>
<b>Pathways and relative distributions of genomic events .....</b>	<b>46</b>
<b>Correlating genomics with the clinical phenotype.....</b>	<b>47</b>
<b>Therapeutics .....</b>	<b>48</b>
<b>Cell Passage.....</b>	<b>49</b>
<b>Protein Extraction .....</b>	<b>50</b>
<b>SDS PAGE and Western Blotting .....</b>	<b>50</b>
<b>DNA extraction.....</b>	<b>51</b>
<b>PCR.....</b>	<b>52</b>
<b>Sanger sequencing.....</b>	<b>52</b>
<b>Plasmid Manipulation in E.coli.....</b>	<b>53</b>
<b>Qiagen Midipreps.....</b>	<b>53</b>
<b>Designing and Cloning sgRNA expressing plasmid.....</b>	<b>54</b>

Phenol-Chloroform DNA precipitation .....	56
Transfection, infection and selection to perform CRISPR-Cas9 .....	56
<b>Results Chapter 1: Detecting novel genetics drivers for OAC.....</b>	<b>59</b>
OAC driver detection using published tools and 551 cases .....	59
Selection in copy number aberrations .....	63
<b>Results chapter 2: Characterising OAC Drivers .....</b>	<b>72</b>
Landscape of driver Events in OAC.....	72
Dysregulation of specific pathways and processes in OAC .....	81
<b>Results chapter 3: Using OAC drivers as clinical biomarkers.....</b>	<b>84</b>
Clinical significance of driver variants .....	85
Targeted therapeutics using OAC driver events .....	90
Investigation of other targeted therapeutic options for OAC .....	94
MEK inhibitors .....	96
EZH2/HDAC inhibitors .....	98
BET inhibitors .....	101
<b>Discussion .....</b>	<b>74</b>
<b>Identification and characterisation of OAC drivers.....</b>	<b>74</b>
Limitations to driver gene detection in this study.....	74
Comparing detected drivers to those known in the literature.....	75
Opportunities for further investigations .....	77
<b>Using novel OAC drivers and clinical biomarkers.....</b>	<b>79</b>
Genotype-clinical phenotype correlations .....	79
Targeted therapeutics for OAC.....	82
Future directions .....	83
<b>Bibliography.....</b>	<b>118</b>

# Abstract

**Background:** Oesophageal adenocarcinoma (OAC) is a poor-prognosis cancer type with rapidly rising incidence. Understanding of the genetic events driving OAC development is limited, and there are few molecular biomarkers for prognostication or therapeutics. This study aimed to use a large cohort of genomically characterised OACs to determine the landscape of genetic driver events in OAC and their possible clinic uses.

**Methods:** We have collated a cohort of 551 genomically characterized OACs (398 whole genome sequenced and 153 whole exome sequenced) with a sub-cohort matched to RNA sequencing data (116 cases). Strelka, Manta and ASCAT were used to call SNVs and indels, structural variants and copy number aberrations respectively. A suite of published tools was used to detect regions of the genome under positive selection for mutations in OAC including dNdScv, Mutsigcv, OncodriveFM and others. Copy number drivers were identified using GISTIC and correlations with matched expression data. Univariate and Multivariate cox regressions were used to identify prognostic biomarkers. Treatment of *In vitro* cultures of human OAC cell lines and organoids with a range of targeted therapeutics was used to calculate AUCs and GI50s for various drugs in OAC models. These sensitivities were matched to the genomic background in these cell lines, provided by whole genome sequencing.

**Results:** We discovered 77 putative OAC driver genes and 21 putative noncoding driver elements for OAC. We identified a mean of 4.4 driver events per tumour, which were derived more commonly from mutations than copy number alterations and compared the prevalence of these mutations to the exome-wide mutational excess calculated using non-synonymous

to synonymous mutation ratios (dN/dS). We observed mutual exclusivity or co-occurrence of events within and between several dysregulated OAC pathways, a result suggestive of strong functional relationships. Indicators of poor prognosis (*SMAD4* and *GATA4*) were verified in independent cohorts with significant predictive value. Over 50% of OACs contained sensitizing events for CDK4 and CDK6 inhibitors, which were highly correlated with clinically relevant sensitivity in a panel of OAC cell lines and organoids. In a smaller panel of OACs we also saw evidence for specificity of BET inhibitor efficacy to *MYC* amplified OACs, however did not observe responses to EZH2 inhibitors, designed to target SWI/SNF mutated cancers, even upon induction of these mutations using CRISPR-Cas9.

**Discussion:** We have compiled the most comprehensive analysis to date of positively selected genomic elements in OAC, significantly improving upon previous analyses. We use this to identify prognostic and therapeutic biomarkers with considerable potential clinical value. Limitation to this study include a lack of RNA sequencing on all samples, making it difficult to assess selection for low-frequency copy number events. Future directions include functional investigation of many of the novel driver genes identified and prospective validation of clinical biomarkers in the Oelixir trial, including use of CDK4/6 inhibitors in OAC patients.

# Introduction

## Principals of cancer evolution

Cancer is an acquired genetic disease, as described in the somatic mutation theory, whereby gain of genetic alterations in tumour-suppressor or oncogenes leads to tumour development and contributes to the cancer phenotype, summarised famously in the 'Hallmarks of Cancer' Cell papers<sup>1,2</sup>. As a tumour expands from an initiating cell, genetic alterations become heterogeneous as new mutations are gained in different daughter cells. These daughter cells grow to become distinct clones and are subject to Darwinian selection as the population expands. Positive selection occurs in clones with a proliferative advantage over their neighbours due to gain of additional oncogenic gene aberrations, referred to as "drivers", and such clones may come to dominate the tumour in a selective sweep. As this happens, heterogeneity and further selection occurs within this clone itself and additional oncogenic genetic aberrations are accrued. As well as driver mutations these positively selected clones will stochastically carry other genetic alterations, not involved in carcinogenesis, which we call "passengers". Cells or clones with alterations deleterious to cell growth may also be negatively selected and removed from the population<sup>3</sup>. The number of selective sweeps and the strength of either positive or negative selection is the subject of considerable debate and seems to vary considerably between different tumour types<sup>4,5,6,7,8</sup>. In particular some investigators in the field have questioned the frequency of neutral evolution after the most recent common ancestor (*i.e.* within intra-tumour heterogeneity)<sup>5,7</sup>. Once a tumour is large enough to be detected clinically we can capture the heterogeneity which has accrued since the last selective sweep using clonal deconvolution of genome sequencing data from single

samples and from multi-region sampling across the tumour<sup>9,10</sup>. We can also decipher the ordering of mutations before the last selective sweep using clock-like mutational signatures and duplicated mutations at copy number amplification events<sup>11</sup>. This evolutionary process is thought to occur over a number of decades and in some tumours may commonly initiate in childhood despite being usually detected >50 years later<sup>12</sup>. Although all tumours seem to undergo this process of evolution to a greater or lesser extent, the events which are accrued in this process vary very significantly. Some tumours, such as acute myeloid leukaemia, seem to be highly dependent on single events such as gene fusions but do not accrue many other mutations at all, whereas many carcinomas will accrue a very large number of point mutations and structural events through the genome<sup>13</sup>. Carcinomas seem to be classifiable on a continuum between two classes depending on the types of alterations which they tend to accrue<sup>14</sup>. Some tumours seem to accumulate large numbers of structural changes but have relatively few recurrently mutated genes (*i.e.* genes under strong selection for single nucleotide variants (SNVs) and small insertions and deletions (Indels) – C class tumours) whereas for others the converse is the case (M class tumours). Tumours from particular tissues seem prone towards M or C classes, although there is still considerable heterogeneity within cancer types in this regard, for instance colorectal carcinomas are mostly M class tumours whereas ovarian carcinomas are usually C class. These classes also correlate with specific positively selected events such as *TP53* mutations in C class tumours. These differences may be due to exposure to different types of mutagens at different anatomical locations, differences in epigenetic modifications in the cell of origin or specifically selected genetic events which may enable cells to tolerate different types of further mutation. Not only does the overall landscape of genetic alterations differ between tumours but specifically the events under strong positive selection also seem to vary greatly between and within

tumour types<sup>13</sup>. Events which are selected in almost every case in a given cancer type do occur (*KRAS* mutations in Pancreatic, *TP53* mutations in ovarian for instance) but are the exception rather than the rule. Most driver alterations discovered thus far occur in less than 10% of any given tumour type and form what is known as a 'long tail' distribution of driver genes mutated in only a small fraction of cases. This heterogeneity poses considerable challenges when distinguishing which genetic alterations have been selected for during cancer evolution.

## Detecting positive selection in cancer genomes

### Principals and coding mutations

To understand the pathogenesis of a particular tumour it is important to determine which genetic aberrations have been positively selected through evolution, to differentiate its malignant cells from the normal cells from which they derived. Various methodologies exist to define driver mutations, but they all rely on the same underlying principal that across of large cohort of tumours, genes (or other regions of the genome) will appear more frequently mutated than otherwise expected if under positive selection. More precisely these mutations will be more clonal, more often, than we expect by chance given the positive selection and clonal expansion which they undergo. Hence these more clonal alterations will be more easily detected in our sequencing data. Luckily, most genes contain detectable variants in a very low frequency of cases by chance *i.e.* their background mutation rate is low, usually <1% of tumours, and hence we can differentiate genes which contain at least some

driver variants by looking for those which contain a significantly higher number of mutations than this expectation. However, this presents two significant challenges to overcome. The first is that to detect such an enrichment requires grouping of mutations, usually into those that fall in specific areas of the genome such as genes or other genetic elements, as this gives us information about how they may affect oncogenesis. However, there are an extremely large number of these elements in the genome (>20,000 genes and >100,000 other regulatory elements) making it only possible to detect very significant differences due to the high false discovery rate associated with such a large number of hypotheses. To detect significant differences a large number of cases is therefore required to increase statistical power, however even the largest cohorts in individual cancer types are mostly <1000 cases, significantly less than the number of genomic elements to be tested. The second significant challenge is accurately modelling the expected mutation rate against which the observed mutation rate is to be compared. The mutation rate is highly uneven across the genome, as is our ability to detect the true variants that occur. Much of the recent progress in the driver mutation detection field has come from discovering and accounting for the genomic features which modulate this background rate. The seminal paper in this regard was published in 2013<sup>15</sup> and showed that replication timing, *i.e.* the time taken for a region of the genome to replicate in S phase given its concentration of replication origins, transcriptional expression and chromatin organisation collectively determine a reasonable percentage of the variability in mutation rate across the genome. By taking these covariates into account several previously detected putative driver genes such as *TTN* (Titin) or Olfactory receptors, without likely involvement in cancer, are no longer found to be significantly mutated. An additional important contribution to the variation between background mutation rate in genes was the

trinucleotide context in which a mutation occurred. Certain trinucleotide contexts are more mutable due to preferences of specific endogenous or exogenous mutagens (these phenomena are known as “mutational signatures”) and hence you can more accurately model the expected mutation rate given the mutational context of the observed mutations<sup>4</sup>. Beyond these and other<sup>16</sup> general determinants of background mutation rate, we can use the observed rate of mutations which we expect to be under no or relatively little selection, to more effectively model the expected rate of the other mutations where strong selection may be acting. There are two ways of identifying such neutral mutations, the first is by using mutations in coding regions which do not alter the coding sequence of the specified gene due to codon redundancy (synonymous mutations) and the second is using functional impact scores based on evolutionary sequence conservation or mutation types to determine which mutations are likely to have the least impact on the function of the gene. The first using synonymous mutations is the most commonly used, although it is only possible for coding regions of the genome, and allows us to adjust the background model of mutation rate based on the other mutation rate covariates we have discussed<sup>4</sup>. The second is used to assess genes for an enrichment of high impact vs low impact mutations, essentially modelling the expected number of high impact mutations, more likely to be under selection, based on the number of low impact mutations<sup>17</sup>. The impact of a mutation can be estimated using either the type of mutation (*e.g.* truncating or non-truncating) or the conservation of the altered amino acid across species. The final commonly used tool to help us in determining which mutations drive a particular tumour is to change the scale at which mutations are tested<sup>18</sup>. By looking more specifically at smaller regions of genes or even single amino acids we massively decrease the expected mutation rate, however selection often focuses on specific regions or

amino acids to dysregulate oncogenes in a specific manner. Methods of detecting mutational 'hotspots' in genes can therefore be more sensitive to such oncogenes however less sensitive to genes in which driver mutations are spread evenly across the primary sequence, as is common in tumour suppressors, hence are complementary to previously described methods. However seemingly false positive hotspots can be common in highly powered datasets as this method is also more sensitive to occasions when factors unknown and not yet accounted for affect the background mutation rate, particular in non-coding regions. Hence hotspot methods have been excluded from non-coding analysis in recent multi-tool studies<sup>19</sup>. Collectively these various methods have been successfully employed on large data sets and confidently identify functionally validated driver mutations<sup>19,20</sup>.

## Structural variation

The field of detecting driver alterations in other mutation types such as structural variants or copy number alterations (CNAs) is not as advanced. This is partly due to the nature of these events which make them more difficult to interrogate. Firstly, they are relatively rare in comparison to small mutations across the genome, although each individual event can affect many more genes, weakening our statistical power. Secondly, we have a poor understanding of the neutral mutational processes that lead to CNAs, making it difficult to build an accurate background model against which we could test the observed data. This is partly because it difficult to define a set of definitely neutral CNAs from which we could build such models. Finally, because CNAs arise across large regions of the genome it can be difficult

to define exactly which genes within copy number amplified or deleted regions provide the selective advantage to the cell upon dysregulation. Despite these limitations, recurrently amplified or deleted regions of the genome can be detected by deconvolving the copy number distribution of each tumour into a series of amplification and deletion events, then looking for regions of the genome in which these events are enriched<sup>21</sup>. This can identify very specific regions containing a clear canonical oncogene (*e.g.* *ERBB2* or *KRAS*) but also identifies quite large regions as discussed, regions without clear drivers and regions which are recurrently copy number aberrant due to non-selection based mechanisms (*e.g.* fragile sites)<sup>19,22</sup>. Additional disruption at the RNA level is a sensitive marker for driver genes in CNAs however it is not very specific as expression modulation can also occur in passengers. Defining driver gene CNAs, has then ultimately relied on extracting known, canonical, CNA cancer genes from these regions or on thorough functional validation *in vivo*. This has limited our ability to discover novel CNA driver genes which have fallen significantly behind mutational drivers in recent years as shown by the relatively small number of CNA drivers currently in the cosmic cancer gene consensus<sup>23</sup>. CNAs are detected by calculating the Log ratio of coverage between tumour and normal (Log R) and the relative contribution of heterozygous SNPs (*i.e.* the maternal and paternal alleles) to this coverage (B-allele frequencies) which both shift distinctly in areas of the genome undergoing CNAs. However, these methods can miss small or allele specific balanced rearrangements and do not indicate the relative positions of variant segments of the genome, only their sequence original position in the reference to which they are aligned<sup>24</sup>. We can use paired end sequencing to define when two genomic regions have been brought together through a structural variant (SV) given that we know the expected fragment size in the prepared library. Hence, reads

that align to the reference genome a significantly different distance apart to the expected library fragment size, or in the incorrect orientation, must have been formed through structural variation<sup>25</sup>. We can precisely define the breakpoint point position in the genome by recovering split reads, covering the exact breakpoint, from the alignment procedure where both ends of a read can be correctly aligned to separate regions of the reference genome. This provides different opportunities to define positively selected structural variants beyond CNAs. This is a relatively new field because of the requirement for large whole genome sequenced data sets but the most recent work published by the PCAWG (Pan-Cancer Analysis of Whole Genomes)<sup>26</sup> consortium takes two approaches. In the first, recurrent SVs that form clusters in the genome are defined across the genome and corrected for SV mutation rate covariates in a similar manner for small scale mutations defining genes falling into these clusters as possible SV drivers. In the second, instances where two regions of the genome are recurrently brought together through a structural variant, for example fusion genes in coding regions, are defined. There are several difficulties in this new field of investigation, first is a lack of known true positive genes that are selectively dysregulated specifically by SVs rather than simply CNAs, hence assessing the quality of output for these methods is difficult and secondly our understanding of the background mutational processes which govern the expected distribution of these events is also not as well studied. There are not, as yet, any well-established computational tools for detection of SV drivers.

## Non-coding mutations

Another field of driver gene detection which has emerged recently, also dependent on availability of large whole genome sequenced cancer cohorts, is detection of driver mutations in non-coding regions of the genome. This is associated with a number of challenges not presented by analysis of the exome. Firstly, non-coding elements are far more difficult to define given the unknown functionality of many putative non-coding elements and their boundaries are often defined arbitrarily. The number of possible non-coding elements to assess is also high (greater than 100,000) presenting an even greater requirement for multiple hypothesis correction, and a weakening of statistical power, than with exonic genes alone. Most elements are also significantly smaller than genes, limiting the number of observations and again statistical power. Some regions of the non-coding genome sequence poorly, particularly promoters which are GC rich<sup>19</sup> and given that whole genomes are usually sequenced at a low depth (often 50X) this can prevent calling of even clonal homozygous mutations in low cellularity tumours. An additional complication is that many of the markers used to model expected mutation rate use codon structure to define neutral or low impact mutations (*i.e.* synonymous mutations or non-truncating mutations) which are hence unavailable to us in the non-coding genome. The PCAWG consortium however has used a suite of tools to assess positive selection in the non-coding genome<sup>19</sup>, calculating expected background mutation rate in a similar manner as discussed with coding genes. They have found a disappointing lack of strong signals for positive selection outside the previously established *TERT* promoter, in comparison to coding regions, however a number of

previously unidentified non-coding driver elements were identified and 50% of the analysed tumours contained at least 1 non-coding driver mutation.

## **Oesophageal adenocarcinoma**

Oesophageal carcinoma is the eighth most common cancer type in the world and the sixth leading cause of cancer death. Oesophageal adenocarcinoma (OAC) is the most common oesophageal cancer subtype in the western world and its incidence has been rapidly rising over the past four decades<sup>27</sup>. This cancer is highly aggressive with the majority of patients surviving less than 1 year after diagnosis<sup>27</sup>. The pre-malignant lesion of OAC, Barrett's Oesophagus (BO), is common but rarely progresses to an invasive lesion<sup>28</sup>. A greater understanding of OAC evolution from BO will be vital to both predict progression and treat patient afflicted with this condition.

### **Pathology and Aetiology**

Oesophageal adenocarcinomas arise in and above the gastroesophageal junction (GOJ) at the lower end of the oesophagus and comprise histologically columnar rather than squamous cells, which would normally be found in the oesophagus. Such cancers seem to derive from lesions of BO metaplasia which consist of columnar cells which usually also contain intestinal metaplasia (IM) where crypts contain goblet cells, resembling intestinal epithelium are

found<sup>29</sup>. BO development is strongly associated with gastro-oesophageal reflux disease (GORD), which may contribute to the proliferation of columnar cells above the GOJ via inducing cell death and inflammation of the native squamous epithelium<sup>30</sup>. OAC is predisposed not only by GORD but also by central adiposity and male sex and is associated modestly with Tobacco smoking<sup>29</sup>. The origin of BO columnar cells, and hence OAC, is the subject of intense debate. The best supported theories include expansion of a sub-population of Keratin 7 positive cells at the squamo-columnar junction, which is supported strongly by mouse models<sup>31</sup>, and expansion of cells in the submucosal glands embedded at the distal end of the squamous esophagus, which supported by pathological and genetic evidence in humans<sup>32</sup>. The cells that form BO appear to have acid-resistant properties which enable them to dominate the distal esophagus in the context of GORD<sup>30</sup>.

The transition from BO to OAC involves, at least in many cases, a series of dysplastic transitions. Dysplasia refers to an increase in histological disorganization of a tissue, in this case BO epithelium, including, but not exclusive to, loss of apical-basal polarity, changes in nuclear morphology and increases in cell volume<sup>33</sup>. BO can progress through low-grade dysplasia (LGD), the histological diagnosis of which is often inconsistent among pathologists, to high-grade dysplasia (HGD) which strongly predisposes to OAC development. Clinical intervention to remove such cells now occurs once dysplasia (low or high grade) develops given the significant risk of progression<sup>34</sup>. Despite this well-defined route of clinical progression and the opportunities for clinical intervention most OACs are diagnosed not only as malignant lesions, but at an advanced stage (78% are nodal positive at presentation<sup>35</sup>). This is due to our inability to capture all BO cases, estimated to occur in 5.6% of the US population<sup>36</sup>, and to properly risk stratify BO cases without dysplasia, given their low rate of overall progression (0.3% per year<sup>37</sup>). Minimally invasive screening devices such as the

Cytosponge™<sup>38</sup> along with molecular stratification biomarkers<sup>39</sup> may allow a reduction in the number of late stage diagnoses and increased survival rates in OAC. However, more effective stratification may require a greater understanding of the molecular determinants of BO progression to OAC. More effective therapeutic strategies are also required to improve survival in the short term and for those who will escape even the most effective screening programmes in the long term.

### Driver events in oesophageal adenocarcinoma

The work to describe genetic changes that occur in OAC development was pioneered in the late 80s and early 90s, mostly in the context of Barrett's oesophagus (BO). BO screening programs, which began in the 80s, provided unique samples for the study of cancer evolution over time<sup>40</sup>. These studies were focused, due to the available technologies, on accumulation of chromosomal abnormalities, a very common event type in OAC, including common amplification of *ERBB2* and *EGFR*, and mutation of a few well-known cancer genes such as *TP53* and *CDKN2A* mutated in approximately 70% and 10% of OAC case<sup>41-46</sup>. Through the late 90s and early 2000s focused studies identified *SMAD4* as recurrently mutated, again in around 10% of cases<sup>47</sup>, *PIK3CA* in 6% of cases<sup>48</sup> and also identified mutations in other known cancer genes at a much lower rate such as *KRAS*<sup>49</sup> and *APC*<sup>50</sup>.

While these focused studies were highly informative, they did not attempt to estimate expected mutation rates in these genes and hence mutations in genes which were not very recurrently mutated (*e.g.* <5%) could have been occurring by chance rather than

due to selection. They also did not assess most genes in the genome. In 2012 a small cohort of 11 OACs were whole exome sequenced, assessing all genes for mutation rate, discovering a fifth gene mutated in >5% of cases, *ARID1A*, however expected mutation rates were still poorly defined in such a small cohort<sup>51</sup>. The same year Dulak *et. al.* published the first large genome scale study focusing of CNAs detected by single nucleotide polymorphism (SNP) arrays on 186 OACs using GISTIC to define recurrently deleted or amplified regions of the genome beyond expectation<sup>52</sup>. Possible genes within or nearby these regions which may be driving the CNAs were annotated including amplified genes: *MET, FGFR2, MYC, CCNE1, CCND1, MDM2, PRKCI, MYB, CDK6, KRAS, GATA4, GATA6, EGFR, ERBB2, MCL1, VEGFA* and deleted genes: *CDKN2A, ATM, CASP3, RUNX1, PTPRD1, FHIT, FAMP190A, PDE4D, PARK2, WWOX, MACROD2* and *SMAD4*. An additional 19 regions were also significant but without a suggested driving gene. The following year the same group published a similarly sized, overlapping cohort comprising 149 exome sequenced OACs and 15 whole genome sequenced OACs<sup>53</sup>. The cohort size and recently developed statistical methods allowed assessment of the expected mutation rate for each gene, based on the mutational distribution of across most genes in the genome. This identified all five previously identified, frequently mutated genes (*TP53, CDKN2A, SMAD4, PIK3CA* and *ARID1A*), as mutated at a higher frequency than expected by chance along with an additional 21 genes, mostly mutated in <10% of OACs. Unfortunately, the methodology used in the paper was subsequently discredited and several of these genes identified are now thought of as classical 'false positives' from this era (*EYS, SYNE1, CNTNAP5* for example) however some genes have been subsequently verified as cancer genes in studies in OAC or other cancers (*ARID2* and *SMARCA4*). The next large study of OACs profiled genome wide was in 2016 where Secrier *et.*

*al.* used a now well established method to define OAC driver genes, MutsigCV, on 129 whole genome sequenced, chemo naïve cases to identify 7 recurrently mutated genes, recovering the 4 most frequently mutated, established OAC drivers (*TP53*, *CDKN2A*, *SMAD4* and *ARID1A*) along with three novel genes to cancer (*KCNQ3*, *CCDC102B* and *CYP7B1*) and many recurrently amplified or deleted regions as detected by GISTIC, without defining driver CNA genes within these regions<sup>54</sup>. In 2017 The cancer genome atlas published their oesophageal cancer manuscript comparing oesophageal squamous carcinomas to oesophageal adenocarcinoma cases<sup>55</sup>. They profiled 89 OACs with whole exome sequencing and RNA-seq and discovered the four established high frequency OAC drivers (*TP53*, *CDKN2A*, *SMAD4* and *ARID1A*) as well as *ERBB2*, which was mutated at a higher frequency than in other studies (13%). They also noted amplified or deleted regions using GISTIC with very similar results to those in Dulak *et. al.* 2012<sup>52</sup>, with a few differences including annotation of *TERC* and *YEATS4* as possible amplification drivers and *PTEN* and *SMARCA4* as possible deletion drivers. Lastly in 2017 a larger study using a large of number of publicly available OACs (446) from Dulak *et. al.* 2013<sup>53</sup>, Secrier *et. al.* 2016<sup>54</sup>, Noorani *et. al.* 2017<sup>56</sup> and TCGA 2017<sup>55</sup> used MutsigCV and discovered 17 OAC driver genes including established OAC drivers (*TP53*, *CDKN2A*, *SMAD4*, *ARID1A*, *PIK3CA* and *SMARCA4*) several known cancer genes not previously associated with OAC with high confidence (*KRAS*, *FBXW7*, *PBRM1* and *CTNNB1*) and several other genes (*PCDH18*, *C6orf114*, *CHRNA1*, *SEMA5A*, *EPHA2*, *PGCP* and *DOCK2*)<sup>57</sup>. Including known cancer genes from Dulak *et. al.* 2013, this brings the total number of thus far

**Introduction Table 1. Publications of oesophageal adenocarcinoma mutation drivers**

Year	Journal	Publication	Novel Driver(s) observed in OAC	Cohort size	Methodology (detection, background rate estimation)	Mutation frequency of Driver(s) in ICGC data
1994	Gastroenterology	Neshat et al	<i>TP53</i>	14	Sanger sequencing, NA	73%
1996	OncoGene	Barrett et al	<i>CDKN2A</i>	32	Sanger sequencing, NA	11%
1996	Cancer research	Barrett et al	<i>SMAD4</i>	35	Sanger sequencing, NA	11%
1996	Eur J Gastroenterol Hepatol.	Trautmann et al	<i>KRAS</i>	11	RFLA*, NA	4%
1997	J Clin Pathol	Gonzalez et al	<i>APC</i>	14	SSCP**, NA	8%
2006	Int J Cancer	Phillips et al	<i>PIK3CA</i>	95	DHPLC***, NA	6%
2012	Cancer discovery	Agrawal et al	<i>ARID1A</i>	11	WES, NA	13%
2013	Nature Genetics	Dulak et al	<i>ARID2, SMARCA4</i>	149	WES, Mutsig	6%, 7%
2016	Nature Genetics	Secrier et al	<i>KCNQ3</i>	126	WGS, MutsigCV	9%
2017	Nature	TCGA	<i>ERBB2</i>	89	WES, MutsigCV	3% (13%****)
2017	Gut	Lin et al	<i>FBXW7, PBRM1, CTNNB1</i>	446	WGS & WES, MutsigCV	3%, 4%, 3%

\*Restriction fragment length analysis, \*\*Single strand confirmation polymorphism, \*\*\*Denaturing high pressure liquid chromatography \*\*\*\*13% in TCGA

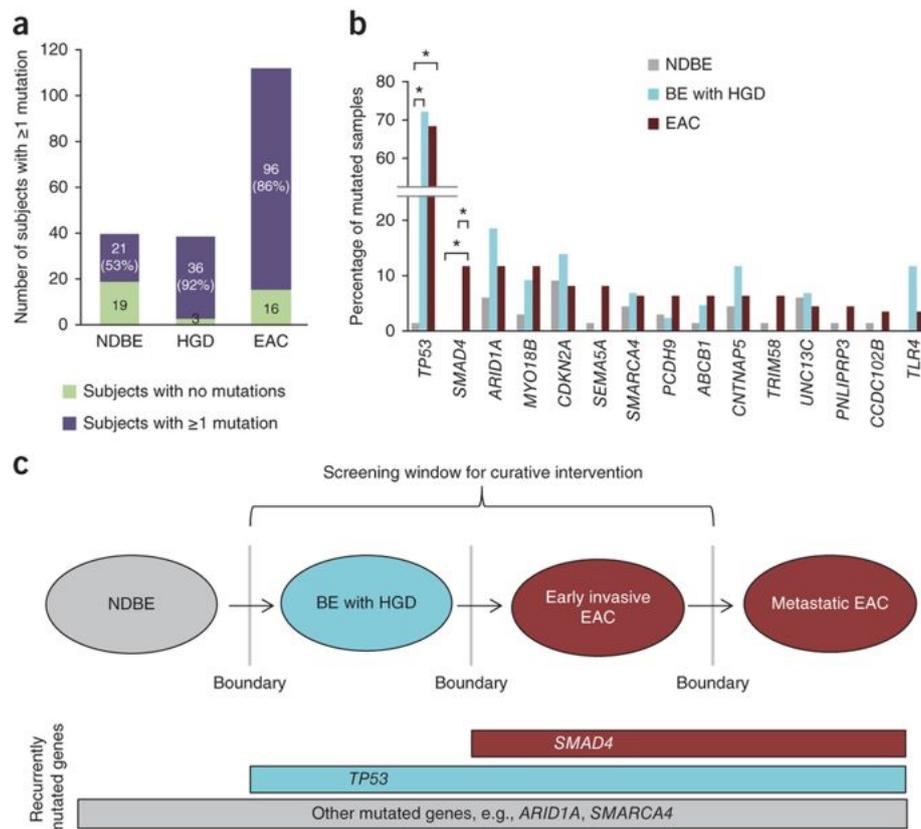
confidently identified OAC mutational driver genes to 22 including 12 known cancer genes (*TP53, CDKN2A, SMAD4, ARID1A, ARID2, PIK3CA, SMARCA4, ERBB2, KRAS, FBXW7, PBRM1* and *CTNNB1*) and 10 genes not previously associated with cancer (*PCDH18, C6orf114, CHRN1, SEMA5A, EPHA2, PGCP, DOCK2, KCNQ3, CCDC102B* and *CYP7B1*) and provides us with a long list of possible oncogenes and tumour suppressors in or nearby recurrently copy number aberrant regions. The most well functionally characterised of these are *ERBB2, EGFR, SMAD4* and *CDKN2A*. The PCAWG consortium also identified a single non-coding element under positive selection in OAC. This was an enhancer on chromosome 7 putatively linked to *TP53TG1*, a gene downstream of *TP53* which is required for *TP53* DNA damage response<sup>19</sup>. See Introduction table 1 for a summary of driver mutation publications in OAC.

## Genetics of OAC evolution from BO

BO lesions contain many genomic alterations which accumulate during progression through dysplasia to OAC<sup>58</sup>. In fact HGD lesions contain a similar number of mutations to many other carcinomas<sup>59</sup>. Many genes are recurrently mutated in Barrett's including known cancer driver genes such as *CDKN2A*, *SMARCA4*, *ARID1A*, *TP53* and others. In addition, known oncogenes are amplified including a plethora of receptor tyrosine kinases, transcription factors such as *MYC* and cell cycle components<sup>52,60</sup>. Whether such genomic alterations contribute to the abnormal non-dysplastic Barrett's oesophagus (NDBO) phenotype we have described is unknown, however they are likely to be important in driving the transition to OAC.

*TP53* mutations arise commonly in BO and seem to be an important driver in the transition from NDBO to HGD. *TP53* encodes a tetrameric transcription factor responsible for relaying stress signals to cause either cytostasis or if the damage is beyond repair, apoptosis. These signals include DNA damage and oxidative stress. It also appears to have roles in a large number of cellular processes, many of which are important in cancer including invasion, autophagy, stem-ness, angiogenesis and chromatin regulation<sup>61</sup>. Aberrant p53 expression has long been known to predict progression of BO to OAC, and sequencing data from the Fitzgerald lab on highly characterised patients with long-term follow-up showed that *TP53* mutations are found extremely rarely in non-dysplastic BO (NDBO) but in the majority of HGDs (72%) and OACs (68%)<sup>62</sup>. *TP53* mutation may cause dysplasia by allowing genomic instability and a build-up of tumorigenic genomic alterations, in particular copy number changes.

Copy number changes are also predictors of progression in BO as evidenced by a study in 2014 which showed copy number changes can precede a histological diagnosis of cancer by two years<sup>58</sup>. Two paths have been suggested for the HGD- OAC transition. Stachler *et. al.* 2015<sup>59</sup> found *TP53* mutations to be an early event in BO progression, relative to most tumour suppressor gene (TSG) mutations, and that whole genome doubling (WGD) was a common event in HGDs. WGD was inversely correlated with many TSG mutations in OAC including *SMAD4*, chromatin modifiers and cell cycle components. This suggests that those genomically unstable *TP53* mutant HGDs that undergo genome doubling gain the final drivers to transition to OAC most easily and quickly via oncogenic amplifications rather than via mutations. This may be because the doubling of WT gene copies makes loss of heterozygosity following TSG mutation more difficult or because amplification above a threshold number of copies, leading to over-expression, is much easier for the cell to achieve. Whereas it is proposed that those high-grade dysplasias that do not undergo genome doubling use a slower path to OAC, involving gain of further TSG mutations with some focal amplifications, to acquire these hallmarks. The specific amplifications and TSGs which are important in the HGD- OAC transition are largely unknown.



**Introduction figure 1. Figure from Weaver *et al.* Nature Genetics 2014 showing enrichment of TP53 mutation in High grade dysplasia (HGD) and SMAD4 mutation in Oesophageal adenocarcinoma (OAC) using targeted amplicon sequencing. A. A Stacked barplot showing the number of samples of each type and whether any mutations were found in each sample. B. A grouped barplot showing mutation rates of 15 putative OAC drivers across sample types C. Shows the proposed model of**

The best evidenced candidate HDG - OAC transition driver gene is *SMAD4*. *SMAD4* is a core member of the transforming-growth factor  $\beta$  pathway and is commonly mutated in other GI tumours, particularly in pancreatic cancer in >50% of tumours<sup>63</sup>. *SMAD4* mutations are the only alterations thus far found to be enriched in OAC over HGD<sup>62</sup> and copy number loss at the *SMAD4* is the only single copy number change found to predict progression from BO to OAC<sup>58</sup>. Finally, a number of recent reports have suggested that clonal heterogeneity, independent of which precise molecular features these clones possess, can predict progression to OAC<sup>64,65</sup>.

## Clinical pathway for OAC

Patients presenting with symptoms such as dysphagia (difficulty swallowing), GI bleeding, vomiting, weight loss and loss of appetite are referred for an endoscopy where abnormal regions of the esophagus are biopsied and studied histologically to diagnose malignancy<sup>66</sup>. These symptoms are usually associated with late stage disease and the lack of symptoms in earlier stages of disease likely contributes to the common late diagnosis of OAC.

Once patients are diagnosed clinicians use a variety of techniques to determine their TNM (Tumour - Node - Metastasis) stage<sup>66</sup>. In early disease (T1-2, N0, M0) surgical removal of the primary tumour is used, resulting in high cure rates. T1a tumours can be removed endoscopically without significant morbidity however larger tumours require esophagectomy, where at least part of the esophagus and stomach is removed and subsequently conjoined. This procedure is considered one of the most radical surgical interventions used and can be associated with significant morbidities such as anastomotic leakage. Patients with locally advanced disease (T3-4, N1-3, M0) are given more radical curative treatment including neoadjuvant, and sometimes additional adjuvant, chemotherapy or chemoradiotherapy. Metastatic OACs are usually considered palliative and can be treated with chemotherapy to extend lifespan.

## Biomarkers

Molecular phenotyping of cancers has led to improvements in patient management in a variety of areas. Distinct molecular subtypes have been identified in a variety of tumour types and even subtypes that span anatomical classifications. These subtypes often have distinct

outcomes and predictable disease courses that require specific management and has enabled researchers to focus on areas of unmet clinical need. Understanding the molecular drivers of cancer has also inspired novel therapeutics targeted towards specific patients, based on their subtype. Breast cancer provides the canonical example of such clinical and molecular subtypes where hormone receptor and *HER2* activation status defines prognosis and treatment opportunities<sup>67,68</sup>.

### **Biomarkers for prognosis in oesophageal adenocarcinoma**

Unlike in breast cancer, assessing prognosis of OACs in the clinic does not currently use molecular features of OAC tumours but relies solely on anatomical and histological features of the tumour. The TNM system is used across cancer types to try and quantify the disease burden and extent of spread across the body in a patient, hence their likely prognosis, and is used to inform treatment discussions. Specifically, it is used to choose a treatment strategy which maximises survival while minimising the associated morbidity of treatment, which can be very significant for this cancer type. To assess disease burden and spread we quantify: the extent of tumour invasion into the underlying tissue (T1-4), the number of cancer infiltrated lymph nodes (N0-4+) and the presence or absence of distant metastases (M0-1). A variety of different methodologies can be used to do this, with varying accuracy. In the latest TNM classification for oesophageal cancer released last year<sup>69</sup>, patients are advised to undergo histological examination of a biopsy specimen, a chest CT scan, an oesophageal ultrasound and a PET scan with fine needle aspiration or biopsy of possible lymph node or distant metastases to assess histology. Histological differentiation of the tumour is also assessed and together these factors are used to subgroup patients into eight prognostic groups (Stages IA,

IB, IC, IIA, IIB, IIIA, IIIB and IVA). The most commonly deterministic feature of prognosis is whether or not a tumour has spread beyond the primary site (*i.e.* lymph node positivity). This is probably in large part due to the success of surgical intervention at these early stages. Staging is also reassessed at several points after treatment to account for differences in treatment effectiveness and to detect disease progression. These stages, particularly when assessed after treatment, correlate well with prognosis, however there remains significant heterogeneity in prognosis within these groups that causes some patients to undergo inappropriate treatment pathways. This is likely to be partly due to incomplete assessment of disease spread, for example missing positive lymph nodes on a CT or PET scan, but also reflects that intrinsic tumour aggressiveness and rate of progression vary significantly and are not entirely captured by histological differentiation. Molecular phenotyping can aid identification of hyper-aggressive or indolent tumours, the likely disease course and hence the most appropriate treatment plan.

A large variety of molecular markers for prognostication have been investigated in OAC, most using protein level expression in Immunohistochemistry<sup>70</sup>, however the validation of some of these markers, either independent cohorts or comparing to current clinical prognostication in multivariate analysis, has been inconsistent and hence none have gone on to prospective biomarker studies. Markers of particular promise which have been suitably verified include a three gene IHC panel consisting of *EGFR*, *SIRT2* and *TRIM44* which improves prognostication above clinical factors but with a relatively small Hazard ratio of 1.2 (95% CI 1.03-1.40) for each additional positive marker. Another promising prognostic marker was identified this year in GLUT1<sup>71</sup>, a membrane transport protein involved in fluorodeoxyglucose (the imaging reagent) transport in PET scans. Bright PET imaging had also been shown to be poorly prognostic<sup>72</sup>. GLUT1 overexpression occurs in 20% of cases however while the hazard

ratio was encouraging in the discovery cohort, 2.08 (95% CI 1.1-3.94), this only remained statistically significant in the validation set in chemotherapy treated rather than naïve cases in a multivariate regression model. A poor understanding of the underlying biology in OACs has caused investigation of a large number of different markers in various studies without much knowledge of how they are involved in OAC oncogenesis. Unsurprisingly this has caused many of the markers found to be strongly affected by discovery bias and hence have not validated well. We believe a more promising strategy to predict OAC prognosis will be to investigate markers which have strong evidence as genes important in driving OAC biology.

### **Biomarkers for therapeutic intervention in oesophageal adenocarcinoma**

OAC patients are triaged into different treatment pathways depending on their disease stage<sup>66</sup>. The very earliest stage tumours (T1a N0 M0, and some T1b N0 M0) can be successfully treated using endoscopic resection, other early stage tumours which penetrate the submucosa require esophagectomy. Patients that present with later stage disease (T3 N1-3 M0) are treated with a neoadjuvant platinum-based chemotherapy regime sometimes including radiotherapy and then proceed to esophagectomy. Some cases will also receive adjuvant chemotherapy after surgery. Cases with metastatic disease are usually considered palliative but are still treated with chemotherapy to extend survival in cases which are fit and consent. Although these standard treatment strategies have evolved over the past years, including radiotherapy and neoadjuvant treatment to improve survival rates, they are still fundamentally based on therapies conceived over half a century ago.

Alternative strategies to these therapies are to use the molecular insights of recent decades to try and inform treatment based on the specific molecular features of each tumour. A small number of such targeted therapeutics are available in specific contexts already in OAC. The most well established is targeting of *HER2* (*ERBB2*) with the monoclonal antibody, Trastuzumab, in *HER2* amplified and/or overexpressing cases, as assessed by Fluorescence in-situ hybridisation (FISH) and Immunohistochemistry (IHC). However, this is only approved in the palliative setting and is not been shown to improve survival in the context of curative treatment<sup>73</sup>. The second treatment recently approved across gastroesophageal adenocarcinomas in the US is use of Immunotherapy (Pembrolizumab or Nivolumab) in microsatellite instable (MSI) cases<sup>74–76</sup> where the treatment provides a significant survival advantage and some evidence suggests it may also be effective in EBV positive cases<sup>77</sup>, however these markers, while relatively common in gastric cancer, are present in very few (<3%) of OACs. This leaves the vast majority of OACs without curative options beyond standard chemoradiotherapy and surgery. A better understanding of the molecular drivers in OAC may reveal novel drug-able targets or drug repurposing opportunities to improve patient survival.

## Hypothesis & Aims

Our understanding of the genetic events which drive each OAC tumour is currently poor because only 12 well characterised driver genes have thus far been associated with OAC and, excluding *TP53*, these are all infrequently mutated (<14% of cases) hence each tumour on average contains only one well characterised driver, usually *TP53*. The infrequent nature of these OAC drivers is likely caused by the nature of OAC as a C type cancer. The low number of OAC drivers is likely due to a lack of statistical power in smaller OAC cohorts and reliance of the OAC research community on a single methodology (MutsigCV) to quantify selection.

The central hypothesis of this thesis is that by identifying the genetics causes for each OAC tumour we could significantly increase our understanding of this disease and may be able to identify clinical biomarkers for prognosis and targeted therapeutic to improve the current management of OAC patients and increase survival rates in this poor prognosis tumour type.

To investigate this hypothesis, we aimed to use a large variety of methodologies and a large cohort of 551 OACs to reveal the landscape of selected events in OAC in unprecedented detail. We then used this compendium of OAC drivers to look for novel biomarkers of prognosis and for opportunities in drug repurposing, which we could validate *in vitro*.

# Methods

## Calling genomic aberrations in OAC

Our Oesophageal ICGC project has involved setting up a pipeline of sample collection, DNA extraction, quality control, library preparation, whole genome sequencing, read alignment and mutation calling to allow genomic characterisation of hundreds of OAC tumours. This has involved the work of a large group of people with a large variety of expertise over a number of years. In this pipeline samples with a pathology-based estimate of cellularity >70% were selected, whole genome sequenced by Illumina to at least x50 coverage, reads aligned by BWA-MEM to the reference human genome hg19. A series of mutation types were then called; SNV and indel mutations using Strelka<sup>78</sup>, copy number calls by ASCAT<sup>24</sup> and structural variant call by Manta<sup>25</sup>. Our methods were benchmarked against various other available methods and have among the best sensitivity and specificity for variant calling (ICGC benchmarking exercises<sup>79,80</sup>). In this analysis we've used 379 whole genome sequenced tumours from our ICGC pipeline. To enhance the analysis a larger cohort was constructed by adding 149 publicly available whole exome sequencing from Dulak *et. al.* 2013<sup>53</sup> performed at the Broad institute, Boston, US and 22 publicly available whole genomes from Nones *et. al.*<sup>81</sup>. The raw BAM files from these cases were requested and were ran through our alignment and mutation calling pipeline to ensure consistency with our calls. This gave us a total cohort of 551 exomes, the largest cohort of OAC analysed to our knowledge.

## RNA sequencing

Total RNA was extracted using All Prep DNA/RNA kit from Qiagen and the quality was checked on Agilent 2100 Bioanalyzer using RNA 6000 nano kit (Agilent). Qubit High sensitivity RNA assay kit from thermo fisher was used for quantification. Libraries were prepared from 250ng RNA, using TruSeq Stranded Total RNA Library Prep Gold (Ribo-zero) kit and ribosomal RNA (nuclear, cytoplasmic and mitochondrial rRNA) was depleted, whereby biotinylated probes selectively bind to ribosomal RNA molecules forming probe-rRNA hybrids. These hybrids were pulled down using magnetic beads and rRNA depleted total RNA was reverse transcribed. The libraries were prepared according to Illumina protocol<sup>82</sup>. Paired end 75bp sequencing on HiSeq4000 generated the paired end reads. For normal expression controls we chose gastric cardia tissue, from which some hypothesise Barrett's may arise, and duodenum which contains intestinal histology, including goblet cells, which mimics that of Barrett's. We did not use Barrett's tissue itself as a normal control given the heterogeneous and plentiful phenotypic and genomic changes which it undergoes early in its pathogenesis.

## Analysing OAC mutations for selection

To detect positively selected mutations in our OAC cohort, a multi-tool approach across various selection related 'Features' (recurrence, functional impact, clustering) was implemented in order to provide a comprehensive analysis. Method table 1 describes these tools and the advantages and limitations of each. This is broadly similar to several previous

approaches<sup>13,83</sup>; dNdScv<sup>4</sup>, MutsigCV<sup>15</sup>, e-Driver<sup>84</sup>, ActivedriverWGS<sup>85</sup> and e-Driver3D<sup>86</sup> were run using the default parameters. To run OncodriverFM<sup>87</sup>, Polyphen<sup>88</sup> and SIFT<sup>89</sup> were used to score the functional impact of each missense non-synonymous mutation (from 0, non-impactful to 1 highly impactful), synonymous mutations were given a score of 0 impact and truncating mutations (non-sense and frameshift mutations) were given a score of 1. Only genes with greater than 7 mutations, likely to contain detectable drivers using this method, were considered to decrease the requirement for multiple hypothesis correction and increase statistical power. OncodriveClust<sup>90</sup> was run using a minimum cluster distance of 3, minimum number of mutations for a gene to be considered of 7 and with a stringent probability cut off to find cluster seeds of  $p = \text{Ex}10^{-13}$  to prevent infiltration of large numbers of, likely, false positive genes. For all tool outputs we undertook quality control including Q-Q plots to ensure no tool produces inflated q-values and each tool produced at least 30% known cancer genes. Two tools were removed from the analysis due to failure for both of these parameters at quality control (Activedriver<sup>91</sup> and Hotspot<sup>92</sup>). For three of the QC-approved tools (dNdScv, OncodriveFM, MutsigCV), where this was possible, we also undertook an additional false discovery rate (fdr) reducing analysis by re-calculating q values based on analysis of known cancer genes only<sup>4,23,93</sup>, as has been previously implemented<sup>4,94</sup>. Significance cut offs were set at  $q < 0.1$  for coding genes. Tool outputs were then put through various filters to remove any further possible false positive genes. Specifically, genes where <50% of OAC cases had no expression (TPM (transcripts per million) < 0.1) in our matched RNA-seq cohort were removed and, using dNdScv, genes with no significant mutation excess (observed: expected ratio > 1.5:1) of any single mutation type were also removed. We also removed two (MT-MD2, MT-MD4) mitochondrial genes which were highly enriched for

Methods Table 1. A Description of Driver detection tools used

Tool Name	Reference	Method	Advantages	Limitations
MutSigCV	Lawrence et al 2013	Measuring recurrence of all mutations in a gene, estimating background using mutation covariates in the genome and synonymous mutations	Good sensitivity for commonly mutated genes	Poor sensitivity for infrequently mutated genes or driver genes where a large % of mutations are passengers
dNdScv	Martincorena et al 2017	Measuring recurrence of all mutations in a gene, estimating background using mutation covariates in the genome, synonymous mutations and trinucleotide context	Good sensitivity for commonly mutated genes	Poor sensitivity for infrequently mutated genes or driver genes where a large % of mutations are passengers
OncodriveFM	Gonzalez-Perez et al 2012	Measuring bias in mutation impact in a gene	Particularly good sensitivity for tumour suppressor genes which accumulate obviously high impact mutations (truncating)	Poorer sensitivity for infrequently mutated genes, genes that are never truncated (Oncogenes). Is dependent on accuracy of tools such as Polphen and SIFT to define impact which can be difficult
OncodriveCLUST	Tamborero et al 2013	Measure recurrence in a small portion of a gene defined by the primary sequence	Particularly good sensitivity for Oncogenes that commonly have mutational hotspots	Poorer sensitivity for tumour suppressor or other genes where mutations occur across the primary sequence
eDriver	Porta-Pardo et al 2014	Measure recurrence in a small portion of a gene defined by the primary sequence	Particularly good sensitivity for Oncogenes that commonly have mutational hotspots	Poorer sensitivity for tumour suppressor or other genes where mutations occur across the primary sequence
eDriver3D	Porta-Pardo et al 2015	Measure recurrence in a small portion of a gene defined by the tertiary sequence	Particularly good sensitivity for Oncogenes that commonly have mutational hotspots - will detect hotspots in tertiary rather than primary sequence	Poorer sensitivity for tumour suppressor or other genes where mutations occur across the primary sequence
GISTIC	Mermel et al 2011	Measures recurrence of CNAs and defines genomic regions recurrently deleted or amplified	Good sensitivity for common and consistently positioned CNAs	Can be difficult to define which genes in genomic regions driver a particular recurrent CNA and some are driven by non-selection-based mechanisms (e.g. fragile sites).
ActiveDriverWGS	Wadi et al 2017	Measures recurrence and mutation impact bias across the genome including non-coding regions. Uses mutations in adjacent genomic sequence to define background rates	Uses both recurrence and mutation impact in conjunction to increase sensitivity - enables detection in non-coding genome	In the non-coding genome synonymous mutations cannot be taken advantage of to estimate background rate (hence adjacent sequence is used). There are also fewer canonical non-coding drivers making it more difficult to assess the accuracy of this technique.

truncating mutations and were frequently called in OncodriveFM as well as other tools. This may be due to the different mutational dynamics, caused by reactive oxygen species from the mitochondrial electron transport chain, and the high number of mitochondrial genomes per cell which leads to significantly more heterogeneity. These factors prevent the tools used from calculating an accurate null model for mitochondrial genes however these may be worthy of functional investigation. For non-coding elements called by ActivedriverWGS filtering for expression or dN/dS was not possible and despite recent benchmarking<sup>16</sup> are not so well established. Hence, we took a more cautious approach with general significance cut offs of  $q < 0.001$ , and for previously identified elements in PCAWG  $q < 0.1$ . Q values were not recalculated for Driver elements only but  $q < 0.1$  for known elements was based on all elements. To calculate exome-wide mutational excess hypermutated cases (>500 exonic mutations) were removed and the global non-synonymous dN/dS ratios were applied to all dndscv annotated mutations excluding “synonymous” and “no SNV” annotations as described in Martincorena *et. al.*<sup>4</sup>.

## Detecting selection in CNAs

ASCAT raw CN values were used to detect frequently deleted or amplified regions of the genome using GISTIC2.0<sup>21</sup>. To determine which genes in these regions confer a selective advantage, CNAs from each gene within a GISTIC identified loci were correlated with TPM from matched RNA-seq in a sub-cohort of 116 samples and with mutations across all 551 samples. To call copy number in genes which spanned multiple copy number segments in ASCAT we considered the total number of full copies of the gene (*i.e.* the lowest total copy

number). Occasionally ASCAT is unable to confidently call the copy number in highly aberrant genomic regions. We found that the expression of genes in such regions matched what we would expect given the surrounding copy number and hence we used the mean of the two adjacent copy number fragments to call copy number for the gene in question. We found that amplification peak regions identified by GISTIC2.0 varied significantly in their precise location both in analysis of different sub-cohorts and when comparing to published GISTIC data from OACs<sup>52,55,60</sup>. A peak would often sit next to but not overlapping a well characterised oncogene or tumour suppressor. To account for this, we widened the amplification peak sizes upstream and downstream by twice the size of each peak to ensure we captured all possible drivers. Our expression analysis allows us to then remove false positives from this wider region and called drivers were still highly enriched for genes closer to the centre of GISTIC peak regions.

To detect genes in which amplification correlated with increased expression we compared expression of samples with a high copy number (CN) for that gene (above 10<sup>th</sup> percentile CN/Ploidy) with those which have a normal CN (median +/- 1) using the Wilcox rank-sum test and using the specific alternative hypothesis that high CN would lead to increased expression. Q-values were then generated based on Benjamini & Hochberg method, not considering genes without significant expression in amplified samples (at least 75% of the amplified samples with TPM > 0.1) and considering  $q < 0.001$  as significant. We also included an additional known driver gene only fdr reduction analysis as previously described for mutational drivers with  $q < 0.1$  considered as significant given the additional evidence for these genes in other cancer types. We also included *MYC* despite its  $q = 0.11$  for its expression correlation. This less significant correlation with expression is due to frequent non-amplification associated overexpression of *MYC* when compared to normal controls and

otherwise *MYC* is well evidence by a very close proximity to the peak centre (top 4 genes) and its high rate of amplification (19%). We took the same approach to detect genes in which homozygous deletion correlated with expression loss. Expression modulation was a highly specific marker for known CN driver genes and was not a widespread feature in most recurrently copy number variant genes. However, while expression modulation is a requirement for selection of CNA only drivers, it is not sufficient evidence alone and hence we grouped such genes into those which have been characterised as drivers previously in other cancer types (high confidence OAC CN drivers) and other genes (Candidate OAC CN drivers) which await functional validation. We used fragile site regions detected in Wala *et. al.*<sup>26</sup>. We also defined regions which may be recurrently heterozygous deleted, without any significant expression modulations, to allow loss of heterozygosity (LOH) of tumour suppressor gene mutations. To do this we analysed genes with at least 5 mutations in the matched RNA cohort for association between LOH (ASCAT minor allele = 0) and mutation using Fisher's exact test and generated q values using the Benjamini & Hochberg method. The analysis was repeated on known cancer genes only for reduced FDR and  $q < 0.05$  considered significant for both analyses. For those high confidence drivers, we chose to define amplification as CN/ploidy (referred to as ploidy-adjusted copy number) this produces superior correlation with expression. We chose a cut off for amplification at CN/ploidy = 2 as has been previously used, and as causes a highly significant increase in expression in our CN-driver genes.

## Pathways and relative distributions of genomic events

The relative distribution of driver events in each pathway was analysed using a Fisher's exact test in the case of pair-wise comparisons including wild type (WT) cases. In the case of multi-gene comparisons such as the cyclins we calculate the p value and odds ratio for each pair in the group by Fisher's exact test and combine p values using the Fisher method, Genes without comparable odds ratios to the rest of the genes in question were removed. Two sets of analyses were performed to assess mutual exclusivity and co-occurrence. Given we expected to observe particular relationships in certain pairs or sets of related genes, we first undertook a small number of hypothesis driven tests between frequented altered genes with well-known functional relationships (those tests indicated in Figure 17). Secondly, we undertook a hypothesis-free approach where all drivers with driver alterations in >5% of cases were test against each other and BH multiple hypothesis correction was applied (Figure 18). For both of these analyses we remove highly mutated cases (>500 exonic mutations, 41/551) as they bias distribution of genes towards co-occurrence. While the mutation rate per gene in the remaining samples is far lower than the mutation rates in drivers tested and hence is unlikely to have a strong effect on the co-occurrence analysis there may still be weak effects to bias towards co-occurrence due to the remaining variance in overall mutation rate. Approaches to account for the overall mutation rate in each sample when performing such analyses do exist. However, these may bias towards mutual exclusivity of drivers due to the non-linear relationship between driver mutation rate and overall mutation rate, *i.e.* the number of drivers does increase as fast as the number of overall mutations increases. We validated these relationships in independent TCGA cohorts of other GI cancers where we

could find cohorts with reasonable numbers of the genomic events in question (not possible for GATA4/6 for instance) using the cBioportal web interface tool<sup>95</sup>.

## Correlating genomics with the clinical phenotype

To find genomic markers for prognosis we undertook univariate Cox regression for those driver genes present in >5% of cases (16) along with Benjamini & Hochberg false discovery correction. We considered only these genes to reduce our false discover rate and because other genes were unlikely to impact on clinical practise given their low frequency in OAC. We validated *SMAD4*, in the TCGA gastroesophageal cohort which had a comparable frequency of these events, but notably is composed mainly of gastric cancers, and *GATA4* in the TCGA pancreatic cohort using the cBioportal web interface tool. We also validated these markers as independent predictors of survival both in respect of each other and stage using a multivariate Cox regression in our 551 case cohort. When assessing for genomic correlates with differentiation phenotypes we found only very few cases with well differentiated phenotypes (<5% cases) and hence for statistical analyses we collapse these cases with moderate differentiation to allow a binary Fisher's exact test to compare poorly differentiated with well-moderate differentiated phenotypes.

## Therapeutics

The cancer biomarker database was filtered for drugs linked to biomarkers found in OAC drivers. Ten OAC cell lines (SKGT4, OACP4C, OACM5.1, ESO26, ESO51, OE33, MFD, OE19, Flo-1 and JHesoAD) and 3 BO high grade dysplasia cell lines (CP-B, CP-C and CP-D) with WGS (Whole genome sequencing) data were used in proliferation assays to determine drug sensitivity to CDK4/6 inhibitors, Palbociclib (Biovision) and Ribociclib (Selleckchem). Cell lines were grown in their normal growth media as specified by the ATCC and ECACC excluding the BO cell lines which were grown in keratinocyte media (Thermofisher) supplemented by EGF and Bovine pituitary extract as supplied and with 5% FCS . Proliferation was measured using the Incucyte live cell analysis system (Incucyte ZOOM Essen biosciences). Each cell line was plated at a starting confluency of 10% and growth rate measured across 4-7 days depending on basal proliferation rate. For each cell-line drug combination concentrations of 16, 64, 250, 1000 and 4000 nanomolar were used each in 0.3% DMSO and compared to 0.3% DMSO only. Each condition was performed in at least triplicate. The time period of the exponential growth phase in the untreated (0.3% DMSO) condition was used to calculate GI50 and AUC. Accurate GI50s could not be calculated in cases where a cell line had >50% proliferation inhibition even with the highest drug concentration and hence AUC was used to compare cell line sensitivity. T47D had a highly similar GI50 for Palbociclib to that previously calculated in other studies (112 nM vs 127 nM). Primary organoid cultures were derived from OAC cases included in the OCCAMS/ICGC sequencing study. Detailed organoid culture and derivation method have been previously described<sup>96</sup>. Regarding the drug treatment, the seeding density for each line was optimised to ensure cell growth in the logarithmic growth phase. Cells were

seeded in complete medium for 24 hours then treated with compounds at a 5-point 4-fold serial dilutions for 6 days or 12 days. Cell viability was assessed using CellTiter-Glo (Promega) after drug incubation.

## Cell Passage

Cell lines were grown in flasks (Nunclon, thermofisher) containing each lines' respective media along with 10% FCS (fetal calf serum), for all lines excluding CP-B, CP-C and CP-D which were grown in 5% FCS, and 1% Penicillin and streptomycin solution (Sigma, 10,000 units Penicillin/5ml, 10 mg/ml Streptomycin). Between passages and experiments cells were kept at constant humidity, 37 °C and 5% CO<sub>2</sub>. All handling of cells requiring opening of their flasks was undertaken in a lamina flow hood to ensure sterility.

Cells were allowed to grow to 70-90% confluency before passage at which point they were washed in sterile PBS and incubated at 37 °C in sufficient Trypsin + EDTA solution (Sigma, diluted to x1 in sterile PBS) to cover the bottom of the flask, for 5-10 minutes until cells were no longer adherent to the flask. An equal volume to that of the Trypsin + EDTA solution of each line's respective media was then added to the flask and the solution was placed in a 15 ml tube and centrifuged for 5 minutes at 1000 rpm (600 g). The supernatant was discarded, and the cell pellet was suspended in its respective media (warmed at 37 °C) to a single cell suspension. 10-50% of that solution was seeded into a new flask of equal size. If cell counting was required, the Vi-Cell Coulter counter was used. This machine used Typhan blue staining along with image analysis software to consistently count cells.

## Protein Extraction

RIPA lysis buffer was used to extract protein lysates from each cell line. One protease inhibitor tablet (Roche) was diluted into 50 ml RIPA Buffer and the buffer was stored at -80 °C. T75 flasks were grown to 70-90% confluency and washed in PBS. 1ml Lysis buffer was added and cells were scraped from the bottom of the dish on ice. Cells + buffer were then removed from flasks and spun down at 14,000 rpm (bench top centrifuge) for 20 minutes to remove large pieces of cellular debris. Supernatant was collected and stored at -80 °C.

## SDS PAGE and Western Blotting

Protein lysates were thawed on ice, samples were mixed with 2x Lamellae buffer and incubated at 95 °C for 5 minutes in a heating block to denature the protein and break disulphide bridges. The resulting solution was loaded onto Biorad Mini-Protean pre-cast Gels and ran at 150V alongside PAGE-Ruler (Biorad) pre-stained protein ladder for 30-40 minutes until the lowest weight molecular marker had run close to the end of the gel, using a Biorad running tank. A PDVF transfer membrane was activated using immersion in methanol for 15 seconds, washed in dH<sub>2</sub>O and then transferred to transfer buffer. The membrane was kept moist at all times post activation. Protein was transferred onto the Membrane at 200V for 1 hour using a Biorad transfer Tank.

The membrane was then blocked to limit non-specific binding of antibody using 5% (w/v) Non Fat dry Milk in TBST (Tris-buffered saline + 0.1% Tween) at room temperature for

1hr. This block solution was also used to dilute primary antibodies. Anti-ARID1A mouse primary (1/1000). Primary antibody was incubated at the solution indicated either for 1hr at room temperature or overnight at 4 °C. Washes were then performed using incubation in TBST (0.1% tween) for 5 minutes (x3). Secondary antibodies, fused to the HRP (horse radish peroxidase) enzyme were diluted in block and incubated similarly for either 1 hr at room temperature or overnight at 4 °C. 3 washes in TBST for 5 minutes each were then repeated. Amersham ECL reagents (ECL prime) were allowed to reach room temperature, mixed (A+B), incubated for 5 minutes and then added to the membrane (approximately 50ul per 4 cm<sup>2</sup>) and excess was blotted off. Membranes were wrapped in Saran wrap, avoiding any air bubbles and then exposed in a Dark room to X ray film (Kodach, Sigma) with various lengths of exposure.

## DNA extraction

DNA was extracted using the Qiagen DNeasy blood and tissue extract kit.  $1-5 \times 10^6$  cells were centrifuged to a pellet and washed in PBS to remove excess media. Cell were resuspended in 200ul PBS and 20ul of Proteinase K solution, 200ul Lysis buffer AL was then added and the solution briefly vortexed. DNA was then precipitated via addition of 200ul of 100% ethanol and the solution is briefly vortexed. The solution was then added to the DNeasy mini spin column and spun through at 16,000 rpm for 1 minute. The column was then washed with buffer AW1 and then AW2 via identical spins. Elution was then performed by adding 100ul of elution buffer, incubation for 2 minutes on the column and similar

centrifugation as above to elute DNA. DNA concentration was then determined using the Nanodrop.

## PCR

For PCR of specific regions from genomic DNA primers were designed using primer blast so that product was between 300-1200 bps with 300 bps flanking the region of interest for further sequencing primer design. A Thermocycler was used with the program set as follows:

- Denaturation 95 degrees for 120 minutes
- Denaturation 95 degrees for 15s
- Annealing 55 degrees for 15s
- Extension 68 degrees for 45s
- Repeat to step 2 x30
- Final Extension 68 degrees for 120s

## Sanger sequencing

Source biosciences Sanger sequencing service was used. PCR products were ran on 1% agarose gels to ensure a clean single bands for each reaction. The reactions were also nanodropped alongside control reactions that had not be placed on the thermocycler. The

difference in DNA concentration was presumed to be the concentration of product. At least 5ul of 50 ng/ul of product was sent to Source bioscience where PCR clean up was performed to remove primers and any possible primer dimers. For plasmid DNA a solution of 100 ng/ul (5ul) was sent. Primers in both cases were sent at 3 nM. four colour capillary sanger sequencing was performed and traces with predicted sequences sent back to the lab.

## Plasmid Manipulation in *E.coli*

The DH5alpha strain of *E.coli* was used to grow and clone plasmids. These were grown in LB media or on LB Agar containing 100 ug/ml Ampicillin to which all our plasmids contained resistance markers. For transformation competent DH5alpha were bought from Invitrogen, thawed on ice for 30 minutes, mixed with either ligation mixture for cloning or 50ng of midi prepped plasmid DNA on ice for 20 minutes. For plasmid DNA this mixture was then added to pre-warmed LB-Agar plates and incubated o/n. For cloning the mixture was then heat shocked for 45 sec at 42 degrees then cooled on ice for 2 minutes before being added to pre-warmed plates.

## Qiagen Midipreps

For Midi preps Qiagen kits were used. 50ml of LB+Amp was inoculated from a transformation plate colony and incubate o/n for approximately 16 hrs. This *E.coli* broth was then

centrifuged at 4500 g to form a pellet. The pellet was then resuspended in 6ml pre-cooled buffer P1. 6ml Lysis buffer (P2) was added and 6ml neutralization buffer (P3 – pre-cooled) was then added causing precipitation. This mixture was then added to a Qiafilter cartridge and incubated for 10 minutes to allow the precipitate to rise to the top of the mixture. The lysate was then pushed through the filter to remove precipitate and added to a Hi-Speed Midi column pre-incubated in 4ml QBT buffer. The lysate moves through the column by gravity and the column is then washed with 20 ml QC buffer and eluted with 5ml QF buffer. The DNA from this elution is then precipitated with the addition of 3.5 ml 100% isopropanol, incubated for 5 minutes and added to the Qiaprecipitator. 2ml 70% ethanol is pushed through the precipitator and then the precipitator is dried via repeatedly pushing air through. 1 ml of Elution buffer is then added and pushed through the precipitator to elute plasmid DNA. The Nanodrop is then used to quantify the DNA yield which achieve between 50-100 ng/ul.

## **Designing and Cloning sgRNA expressing plasmid**

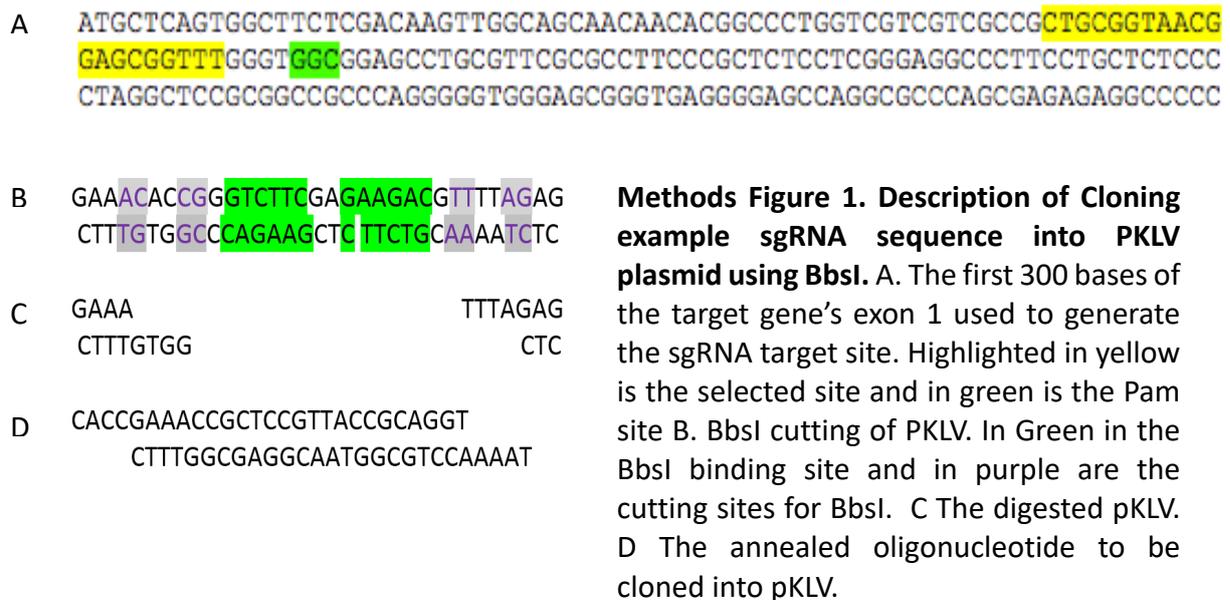
The sgRNA sequence targeting the beginning of the SMAD4 coding region was designed using the CRISPR guide design tool of the Zhang lab – (<http://crispr.mit.edu>). The first 300 bps of the ARID1A exon 1 were inputted and a suitable sequence 3' to a GCC PAM site was selected – “CTGCGGTAACGGAGCGGTTT” – (see methods figure 1).

The Zhang lab CRISPR 2 step cloning protocol was used (available at <https://www.addgene.org/crispr/zhang/>) to clone the sequence into the lenti-viral sgRNA expression plasmid pKLV-U6gRNA(BbsI)-PGKHygro2ABFP. The plasmid is double cut by BbsI, which cuts outside its binding sequence hence can produce different overhangs depending on the

binding sequence context, producing specific overhangs. Oligos were designed with our anti-ARID1A sequence and complementary overhangs for the BbsI cut pKLV plasmid to properly orientate itself relative to the promoter (see methods figure 1). Such oligos were ordered from sigma. They were phosphorylated and annealed as per the online protocol and then ligated into BbsI cut, gel extracted pKLV plasmid.

## Phenol-Chloroform DNA precipitation

To concentrate and clean plasmid DNA for transfections Phenol-chloroform DNA precipitations was used. 1 volume of DNA was mixed with 1 volume of TE-saturated Phenol-Chloroform-isoamyl (25:24:1) (Sigma) and vortexed for 1 minute. The mixture was then centrifuged at 13,200 rpm for 2 minutes. The upper layer from this mixture was then removed and mixed with 1/10 volume of 3M sodium acetate and 3 volumes of 100% ethanol. This mixture is then incubated at either -20 degrees for 16 hours or -80 degrees for 1 hr to precipitate DNA. The mixture is then centrifuged at 16,000 rpm for 20 minutes to pellet DNA. The supernatant is then removed and the pellet dried overnight at 4 degrees. The pellet is then resuspended in 20-50ul of DNase free water for transfections. Concentrations of 1-10 ug/ul are achieved and stored at -20 degrees.



## Transfection, infection and selection to perform CRISPR-Cas9

Second generation packing systems were used to make lenti-viral particles and infect Cas9 and sgRNA DNA into target OAC cell lines. HEK293T packaging cells were used which express a limited number of envelope proteins required for lenti-viral packaging. These cells were transfected with further plasmids that are required for lenti-viral packaging (GAG and POL genes) along with the transducing plasmid of interest. Transfections were performed in T25s using 26ul Lipofectamine 2000 5.7 ug of psPAX2 (Gag) 2.2 ug of pMD2.1 (Envelope) and 3.8 ug of the transduced plasmid. DNA and Lipofectamine were pipetted into two different vials of 416 ul of Opti-MEM serum free media and incubated as room temperature for 5 minutes. The Lipofactamine mixture was then added to the DNA mixture dropwise and incubated at room temperature to allow Lipid-DNA complexes to form for 30 minutes. These mixtures were held in polypropylene tubes to prevent lipid and DNA plastic adhesion. This mixture was then slowly added to 2.6 ml of DMEM +10% FCS (antibiotic free) media in which HEK293Ts

were 60% confluent. This media was removed and replaced by antibiotic containing media after 6 hours. The HEK293s were incubated for 24 hours at which point the first batch of Virus soup was harvested from the media and was replaced by fresh media. After a further 24hrs a second batch of viral soup can be collected. This viral soup was filtered through 0.45 um pore sized filters to remove any contaminating HEK293 cells that could allow further viral production. The soup was then added to a variety of OAC cancer cell lines and HaCaT keratinocytes in a 6 well plate format. 1 ml of appropriate media with 8 mg/ml polybrene was added to each well and then 200 ul of viral soup was added and incubated with target cells for 24 hours. Target cells were grown with extra wells for no virus controls and at infection were 50% confluent. After 24 hours the cells were treated with anti-biotic concentrations (either Hygromycin, Blastocidin or Puromycin from sigma) depending on their sensitivity determined using dose response kill curves. The concentrations of drug aimed to kill all control cells in 3-5 days and were used (1-2ug/ml). Once no virus controls were dead, remaining cells in virus treated wells remained in antibiotic for two further passages and then antibiotic was removed. Some cell lines were very easily transducible with no apparent death in virus treated conditions (such as OACP4C) and some left only small antibiotic resistant clones (OE33) which took days/weeks to grow out. This process was first performed with pCWCas9 (addgene plasmid #50661), lentiviral plasmid encoding a doxycycline inducible humanized *S. pyogenes* Cas9, using puromycin resistance then once Cas9 expression had been confirmed the process was repeated with both cloned pKLV (described in above in "Cloning") and Empty vector pKLV using hygromycin. Once hygromycin resistant cells were isolated from this the cells were treated with 1 ug/ml Doxycycline for 5-7 days while they grew up to induce Cas9 expression and so genomic DNA cleavage in *SMAD4*. Protein and DNA were then isolated from these lines to confirm *SMAD4* mutations and protein loss. These

lines were then mixtures of different *SMAD4* mutations as different mutations occurred in every cell. To isolate clones single cells were sorted into 96 well plates and clones grown up with the same mutation(s) in every cell. Both EV and cloned pKLV lines were brought through this entire process and EV lines used as controls in functional assays.

# Results Chapter 1: Detecting novel genetics drivers for OAC

## OAC driver detection using published tools and 551 cases

To detect positively selected events in Oesophageal adenocarcinoma (OAC) we accumulated a cohort of 551 OACs with either whole genome (379) or whole exome sequencing (172) and with a number of whole genome matched RNA sequenced cases (116). The clinical characteristics of this cohort are shown in Table 1. As is expected of this disease, our cohort is male dominated and generally late stage<sup>97</sup>, however a greater proportion of our cases followed a curative pathway than would be expected due to the greater availability of samples in those cases that underwent surgery.

In these 551 OACs we called a total of 11,813,333 single nucleotide variants (SNVs) and small insertions or deletions (Indels), with a median of 6.4 such mutations / Mb (figure 1), and 286,965 copy number aberrations (CNAs). We also identified 134,697 structural variants (SVs) in WGS cases (355/case). This is broadly in line with what has been previously observed, for example whole genome sequencing of 22 WGS OACs in Nones *et. al.* identified 8 mutations/Mb and 263 SVs per case<sup>53,54,81</sup>.

Mutations or copy number aberrations under selection were detected using a variety of methodologies which detect sets of mutations observed more frequently than we would expect by neutral chance, indicating that at least some are driving clonal expansions, and hence pushing these mutations into detectable variant allele frequencies (VAFs). Figure 2 describes four classes of methodology to detect recurrence of different groups of genomic

**Table 1. Clinical Characteristics of OACs used in study**

Item		Full cohort (N=551)	ICGC cohort with full clinical anotation used for Prognostication (n=379)
Age	Years (median, IQR)	67.3 (59.3-75.0)	70.0 (59.2-74.4)
Sex	Female	81/551	55/380
	Male	470/551	325/380
Barrett's associated	Positive	252/469	161/324
	Negative	217/469	163/324
Siewert type	Esophageal	127/417	56/300
	GEJ (unspec.)**	46/117	NA
	GEJ1	101/300	101/300
	GEJ2	115/300	115/300
	GEJ3	28/300	28/300
Histopatholigcal grading	Well	22/482	10/310
	Moderate	185/482	113/310
	Poor	275/482	187/310
Pre-treatment Tumour Stage	T0	5/483	5/311
	T1	86/483	56/311
	T2	68/483	39/311
	T3	289/483	184/311
	T4	35/483	27/311
Pretreatment nodal involvement	Positive	156/475	199/303
	Negative	319/475	104/303
Pretreatment distant metastases	Positive	27/322	10/150
	Negative	195/322	140/140
Treatment pathway***	Endoscopic Resection	8/358	8/358
	Surgical Resection only	68/358	68/358
	Chemoradiotherapy + Surgery	244/358	244/358
	Definitive chemo-radiotherapy	6/358	6/358
	Palliative treatment	23/358	23/358
	Palliative support care	9/358	9/358
Overall Survival****	Weeks (median, IQR)	101 (54-161)	103 (55-166)

\* Denominators smaller that total cohorts indicate missing data

\*\* In publically available studies tumour location is reported as Esophageal vs GOJ only, the siewert type is not specified

\*\*\*Not reported in publically available studies

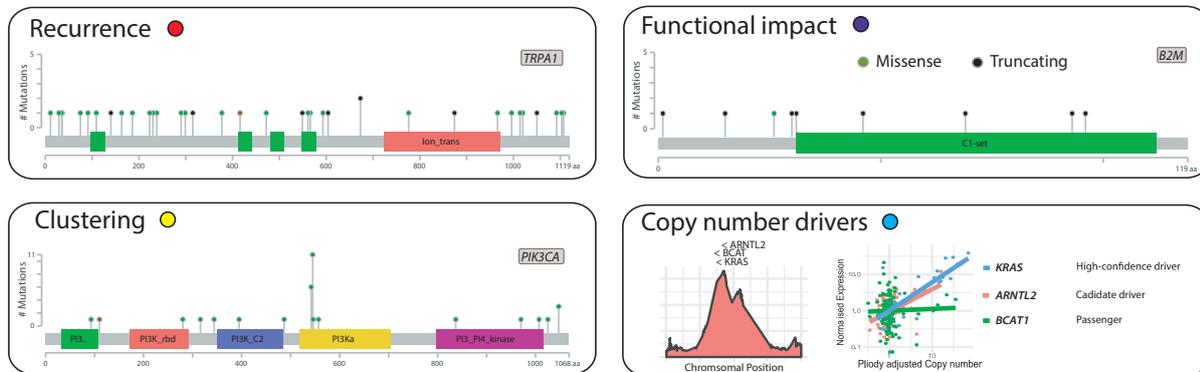
\*\*\*\*Not reported in Dulak et al 2013



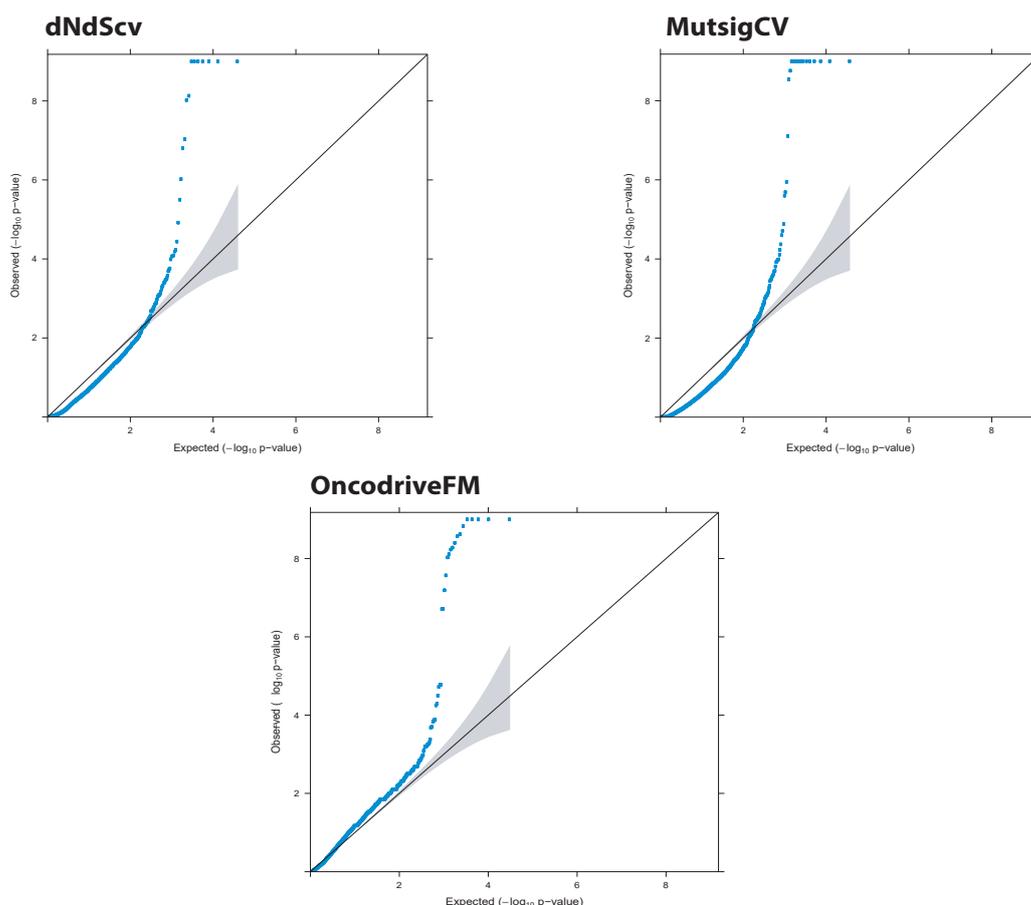
**Figure 1.** Distribution of small scale mutations (SNVs and Indels) across the 551 OAC cohort. Red line indicates the median mutations per case (6.4)

abberations; “Recurrence” methods measure mutation rate in a whole gene and test against expected mutation rates (dNdScv<sup>4</sup>, ActivedriverWGS<sup>85</sup>, MutsigCV<sup>15</sup>), “Functional impact” methods essentially test the mutation rate of high functional impact mutations compared to expected rate calculated using the number of low impact mutations (OncodriveFM<sup>87</sup>, ActivedriverWGS), “Clustering” methods measure mutational recurrence in only a small portion of a gene (OncodriveClust<sup>90</sup>, eDriver<sup>84</sup> and eDriver3D<sup>86</sup>) and our “Copy number driver” method measures recurrence of copy number amplifications or deletions to detection selectively aberrant regions of the genome (using GISTIC<sup>21</sup>) and then identifies driver genes within these regions using expression and copy number correlations (In house analyses, see methods page 41). We tested 8 different publicly available tools or in-house methods to detect mutations under positive selection as described and 6 of these could produce uninflated Q-Q plots (Figure 3) along with recovery of a reasonable number (>30%) of known cancer genes. Two methods were rejected due to a high rate of infiltration of likely false positive genes in our hands (Hotspot<sup>98</sup> and ActiveDriver<sup>91</sup>). We also undertook filtering to remove any possible false positives from our driver list produced by these QC’ed tools which included filtering out genes that were not expressed (<50% of cases with <0.1 TPM) or had no evidence of enrichment in mutation rate (odds ratio < 1.5) in any specific mutation type using dNdScv, our best performing driver detection tool (measured by sensitivity and specificity for known cancer genes). None of those genes filtered were known drivers. This multi-tool approach has become the gold-standard in driver gene detection studies as has been exemplified by several driver detection efforts by large consortia such as the TCGA<sup>99</sup>, ICGC<sup>13</sup> and others<sup>100</sup>.

These complementary methods produced highly significant agreement in calling OAC driver genes, particularly within the same feature-type (Figure 4A) and on average more than



**Figure 2.** Methodologies for detecting positive selection in cancer genomes.

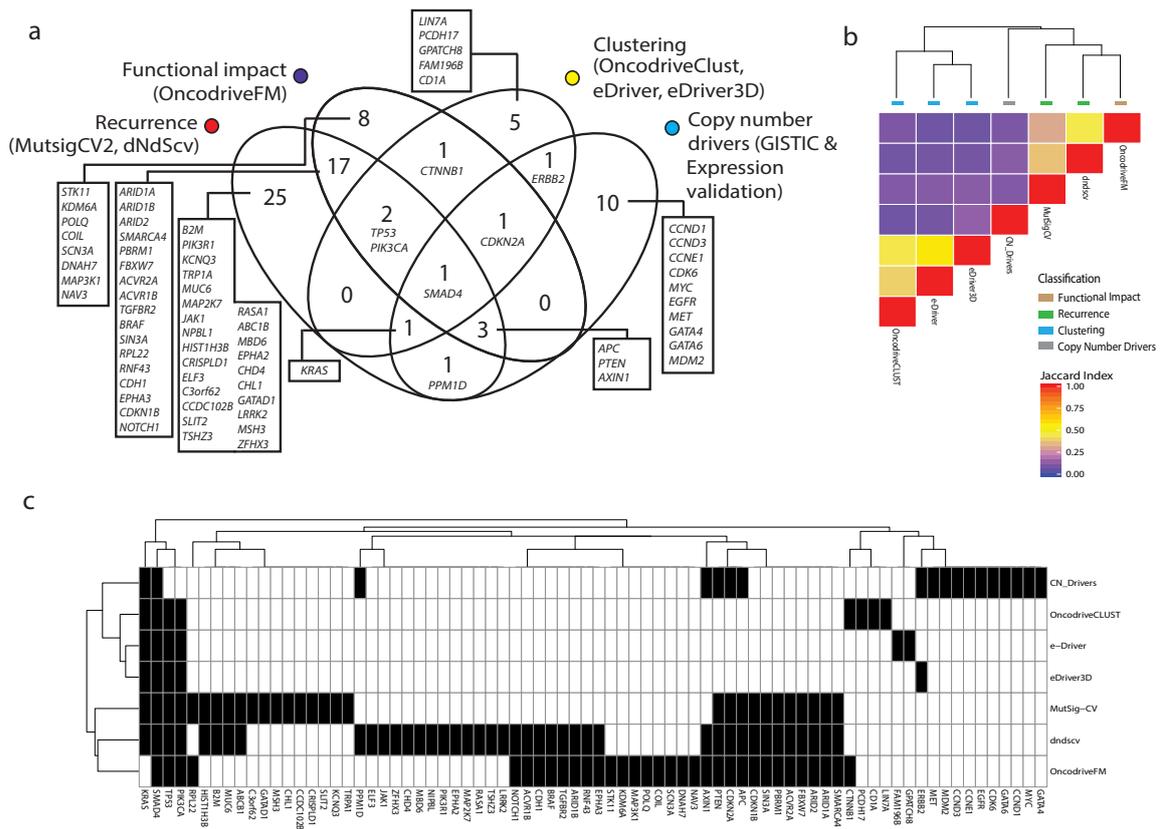


**Figure 3.** Quality Control for published driver gene detection tools in our hands using Q-Q plots. Q-Q plots were not appropriate for cluster based analysis due to high thresholds set for cluster discovery causing many fewer genes to be assessed which were more often significant than would otherwise be expected. These tools (OncodriveClust, eDriver and eDriver3D) passed the second QC step with >30% known cancer genes.

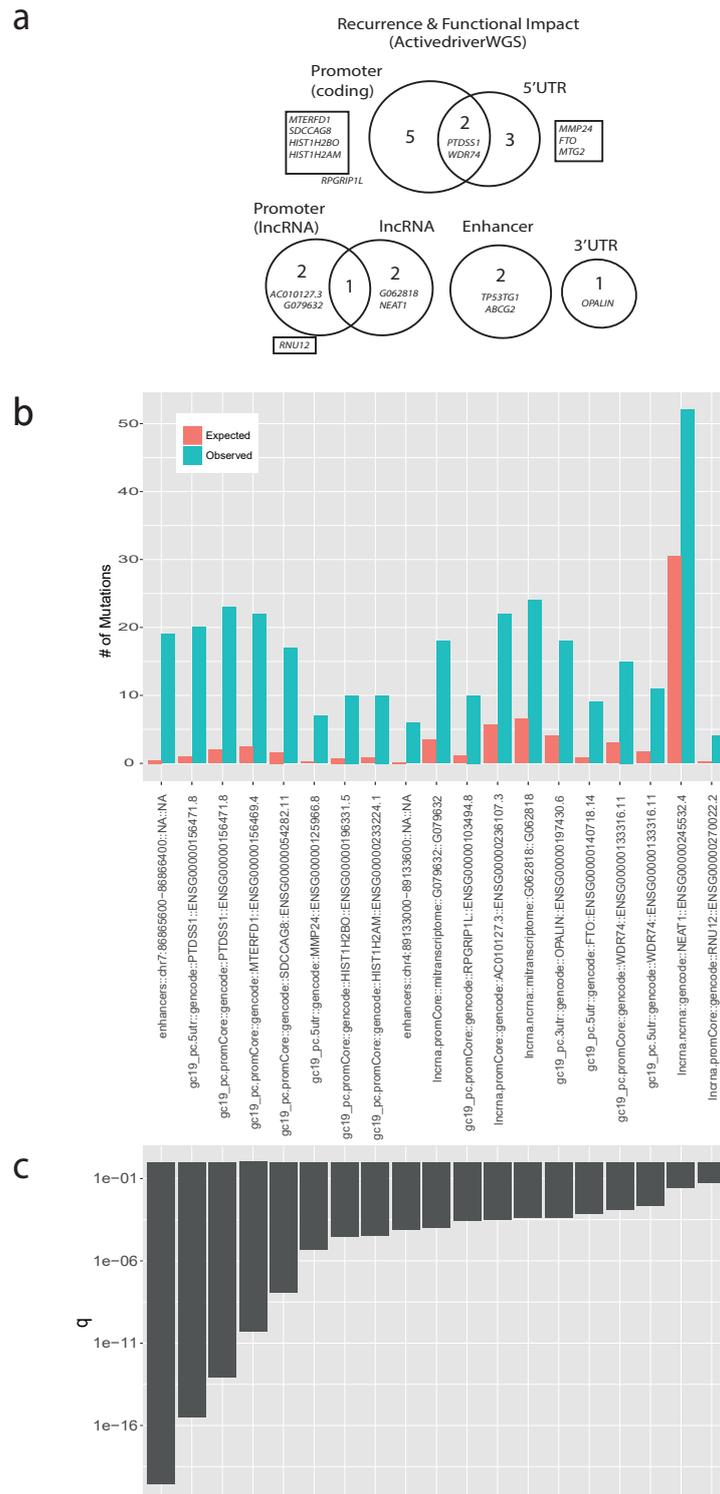
half of the genes identified by one feature were also identified by other features (Figure 4B). Hierarchical clustering indicated the driver tools clustered as expected given the feature used, with strong relationships between functional impact and recurrence methods, as well as surprisingly large overlap between copy number and mutational drivers (Figure 4C). In total seventy-six putative OAC driver genes were discovered, 86% of which have not been detected in OAC previously and 69% are known drivers in independent pan-cancer analyses giving us confidence in our methods. To detect driver elements in the non-coding genome we used ActiveDriverWGS a recently benchmarked<sup>16</sup> tool that uses recurrence and functional impact to detect selection across the genome (Figure 5). We discovered 21 putative non-coding driver elements using this method which each contained driver mutation in 5% or less of cases, however in total they still contribute several hundred non-coding driver events in this cohort, consistent with the PCAWG analysis<sup>19</sup>. We have recovered several known non-coding driver elements from the pan-cancer PCAWG analysis including an enhancer on chr7 linked to *TP53TG1*, a gene required for TP53 action, the only non-coding driver found in OAC in PCAWG<sup>19</sup> and the promoter/5'UTR regions of *PTDSS1* and *WRD74* which are novel in OAC but were found in other cancer types. We also identified completely novel non-coding cancer driver elements including in the 5'UTR of *MMP24* and promoters of two related histones (*HIST1H2BO* and *HIST1H2AM*).

## Selection in copy number aberrations

OAC is notable among cancer types for harbouring a high degree of chromosomal instability<sup>81</sup>. Using GISTIC we identified 149 recurrently deleted or amplified loci across the



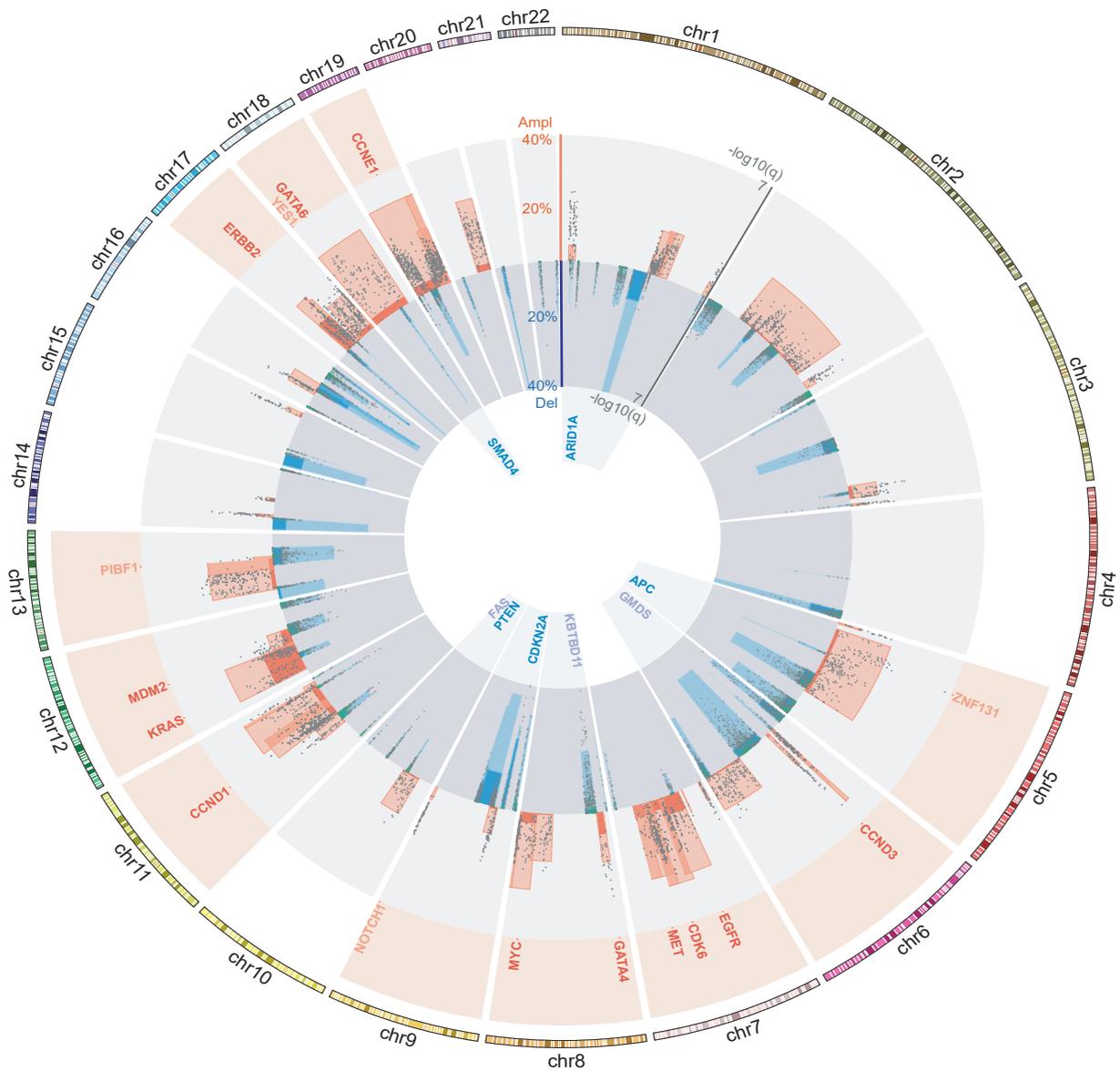
**Figure 4.** Coding Drivers identified in OAC. **a.** A Venn diagram indicating drivers identified using different Driver detection methodologies and their overlap. **b.** Hierarchical clustering between tools based on genes identified. **c.** Genes identified by each tool.



**Figure 5. Frequency and significance of OAC non-coding drivers from ActiveDriverWGS using 379 WGS ICGC OACs. a.** Non-coding driver elements identified in OAC organised by element type. **b.** The observed and expected mutation counts found on each element in ActiveDriverWGS. **c.** The fdr for each element in ActiveDriverWGS

genome (Figure 6). To determine which genes within these loci confer a selective advantage when they undergo CNAs we used a subset of 116 cases with matched RNA-seq to detect genes within these loci in which homozygous deletion or amplification causes a significant under or over-expression respectively, a prerequisite for selection of CNA-only drivers. The majority of genes in these regions showed no significant CNA-associated expression change after false discovery rate correction (74%), although work in larger cohorts suggests we may be underpowered to detect small expression changes. We observed highly significant expression changes in 17 known cancer genes within GISTIC peaks such as *ERBB2*, *KRAS* and *SMAD4* which we designate high-confidence OAC drivers (see Figure 10, page 69 for full list). We also found five tumour suppressor genes where copy number loss was not necessarily associated with expression modulation but tightly associated with presence of mutations leading to LOH; *ARID1A*, *CDKN2A*, *SMAD4*, *APC* and *CDH11*. *CDH11* was not identified by our driver gene detection methods but this would suggest it may be a promising candidate for further validation. To determine whether copy number changes in genes not previously associated with cancer may contribute to oncogenesis we searched for genes with similar expression-CN profile as most of our high-confidence drivers (see methods). We found 140 such cases which we designated “candidate copy number (CN) drivers”. Not all candidate drivers are likely to be true CN-drivers. However, several candidate drivers such as *ZNF131*, *YES1* and *PIBF1* are not accompanied by other drivers in their GISTIC peak and contain extrachromosomal-like events (referred to below), hence are promising candidates for further study.

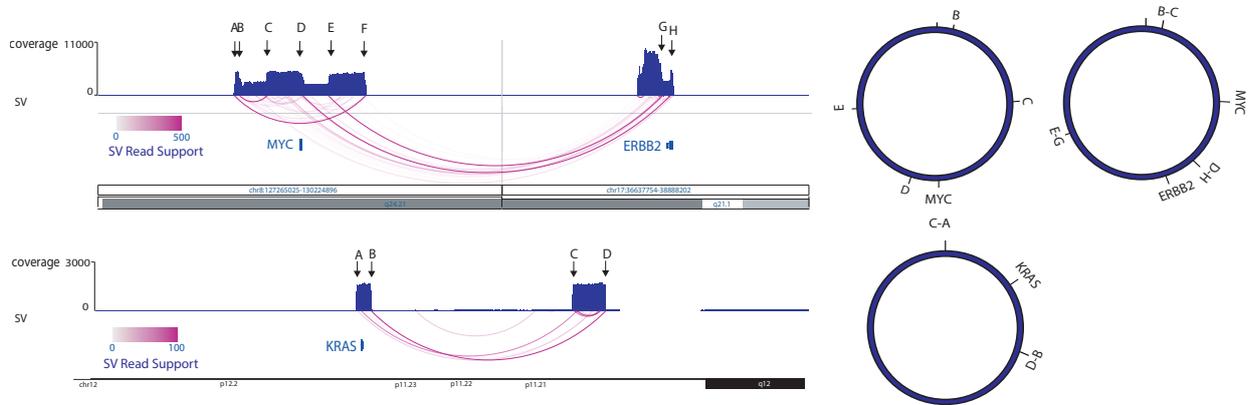
In a subset of GISTIC loci, we observed extremely high copy number amplification, commonly greater than 100 copies, and we hypothesized, based on previous finding in OAC<sup>81</sup> and other cancer types<sup>101</sup> that these were likely to be extrachromosomal events, such as



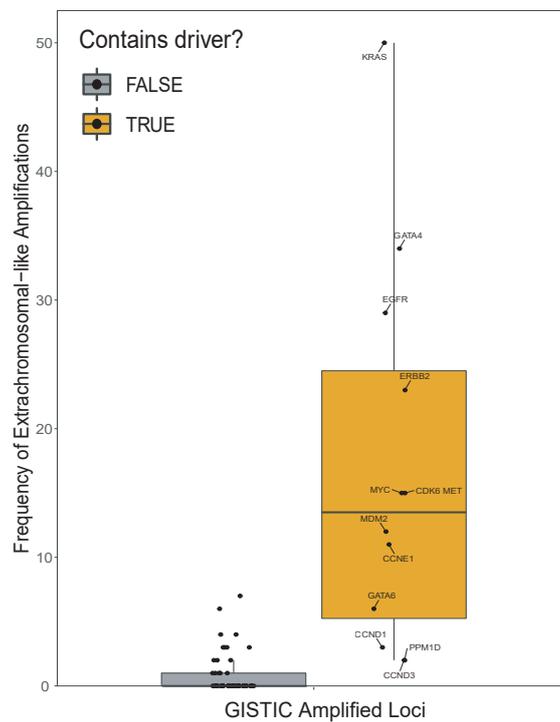
**Figure 6. Copy number variation under positive selection. a.** Recurrent copy number changes across the genome identified by GISTIC in 551 OACs. Frequency of different CNV types are indicated (dark blue = Homozygous deletion, light blue = heterozygous deletion, dark red = extrachromosomal-like amplification, light red = amplification) as well as the position of CNV high confidence driver genes and candidate driver genes. The q value for expression correlation with amplification and homozygous deletion is shown for each gene within each amplification (wilcox test, one sided, expression compared above and below 90th percentile of ploidy-adjusted CN) and deletion peak (wilcox test, one sided, expression compared between homozygous deleted and all other cases) respectively and occasions of significant association between LOH and mutation are indicated in green (fisher's exact test, one sided). Benjamini & Hochberg false discovery correction was applied in each of these cases.

double minutes, which can allow rapid amplification of relatively small pieces of circular DNA. This process would leave specific hallmarks on the genome. Because amplification of the double minute chromosome does not require additional structural variants to cause amplification once in a circular form, this causes the copy number steps surrounding the circular region to be highly precipitous, rather than step-like as in other kinds of amplification. It would also cause the existing structural variants already contained within the circular DNA to be copied precisely, very many times giving them very high read support. Hence, we searched for these features within these ultra-high amplification events and consistent with previous data in OAC<sup>81</sup> we found them to be common, accounting for the large majority of such ultra-high amplifications (examples shown in figure 7). In the first example circularisation and amplification occurred around MYC and also incorporated ERBB2 from an entirely different chromosome and in the second an inversion has been followed by circularisation and amplification of KRAS. Hence, we defined these ultra-high amplifications (CN-adjusted ploidy >10) as “extrachromosomal-like” amplifications. Interestingly we found that the recurrently amplified loci with such extrachromosomal-like events were highly correlated with presence of CN-drivers (Ploidy adjusted Copy number >10, Wilcox test,  $P < 10^{-6}$ ) and such events were almost completely absent in GISTIC loci without a known CN driver (Figure 8). This high specificity for Driver containing regions suggests such events may be generally very rare but highly selected for and could possibly be used to detect copy number drivers in the future.

We use copy number adjusted ploidy to define amplifications as it produces superior correlation with expression data than absolute CN alone. Ploidy of our samples varies from 1.4-6.2 (median 2.8), and hence ploidy adjusted copy number of >10 cut off translates into



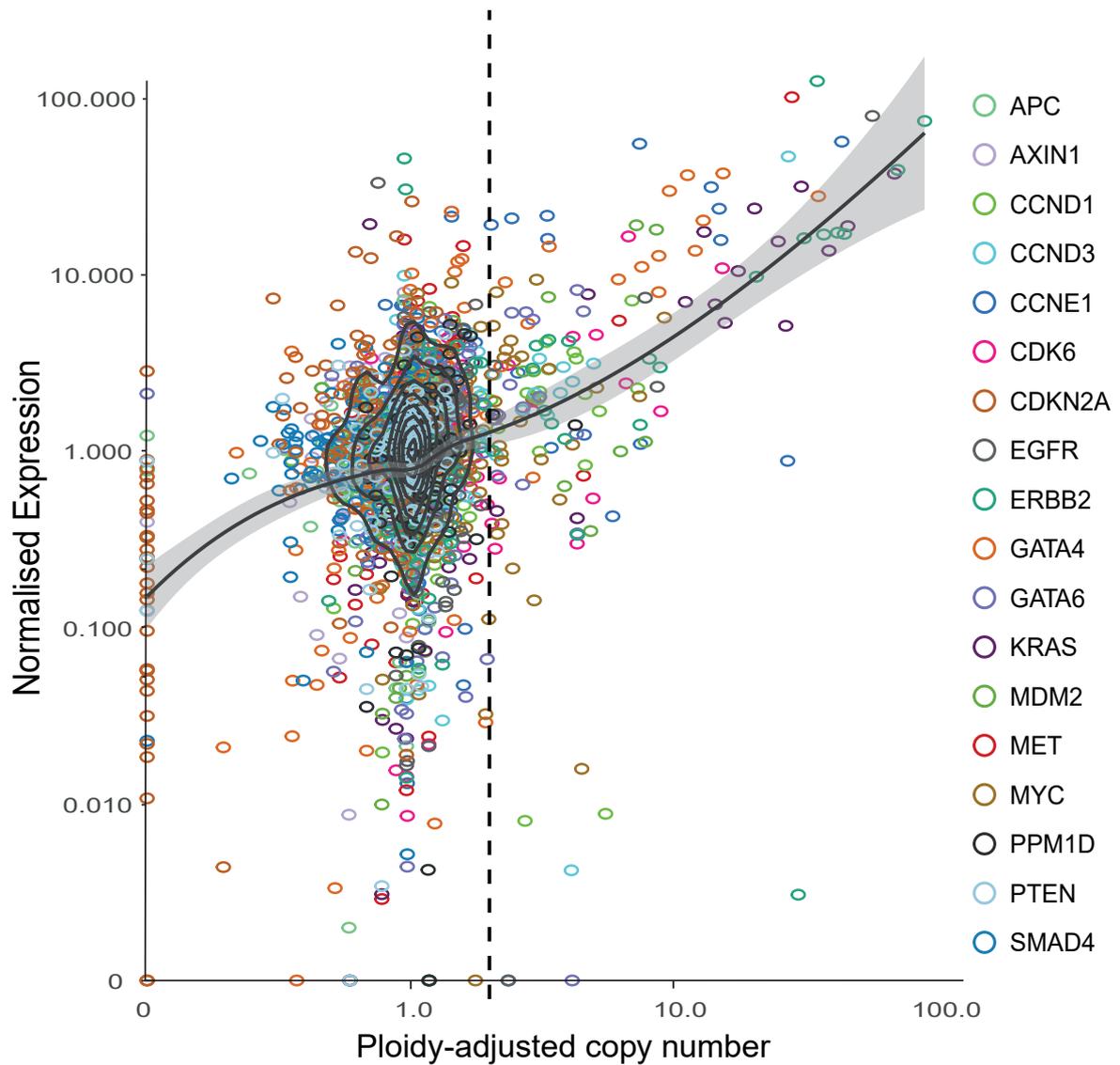
**Figure 7. Examples of Extrachromosomal-like amplifications.** Extra-chromosomal events suggested by very high read support SVs at the boundaries of highly amplified regions produced from a single copy number step. In the first example two populations of extrachromosomal DNA are apparent, one amplifying only MYC and the second also incorporating ERBB2 from a different chromosome. In the second example an inversion has occurred before circularization and amplification around KRAS



**Figure 8.** Frequency of Extrachromosomal-like events (CN-adjusted Ploidy >10) in GISTIC amplification loci using 551 OACs and showing the presence of any high confidence driver genes. Boxplots indicate median and interquartile range.

>14-62 absolute copies (on average 28 copies). For some cases we may have been unable to identify drivers in loci simply because the aberrations do not occur in the smaller RNA-seq matched cohort.

We found extrachromosomal-like amplifications had an extreme and highly penetrant effects on expression while moderate amplification (ploidy adjusted copy number > 2) and homozygous deletion had highly significant (Wilcox test,  $P < 10^{-4}$  and  $P < 10^{-3}$  respectively) but less dramatic effects on expression with a lower penetrance (Figure 9). This lack of penetrance was associated with low cellularity (Fisher's exact test, expression cut off = 2.5 normalised FPKM,  $P < 0.01$ ) however many samples with moderate amplification but without overexpression were of good cellularity hence this also likely reflects that genetic mechanisms other than gene-dosage modulate expression in a rearranged genome. We also detected several cases of over expression or complete expression loss without associated CN changes which may reflect non-genetic mechanisms for driver dysregulation. For example, one case overexpressed *ERBB2* at 28-fold median expression however had entirely diploid CN in and surrounding *ERBB2* and a second case contained almost complete loss of *SMAD4* expression (0.008-fold median expression) despite possessing 5 copies of *SMAD4*.

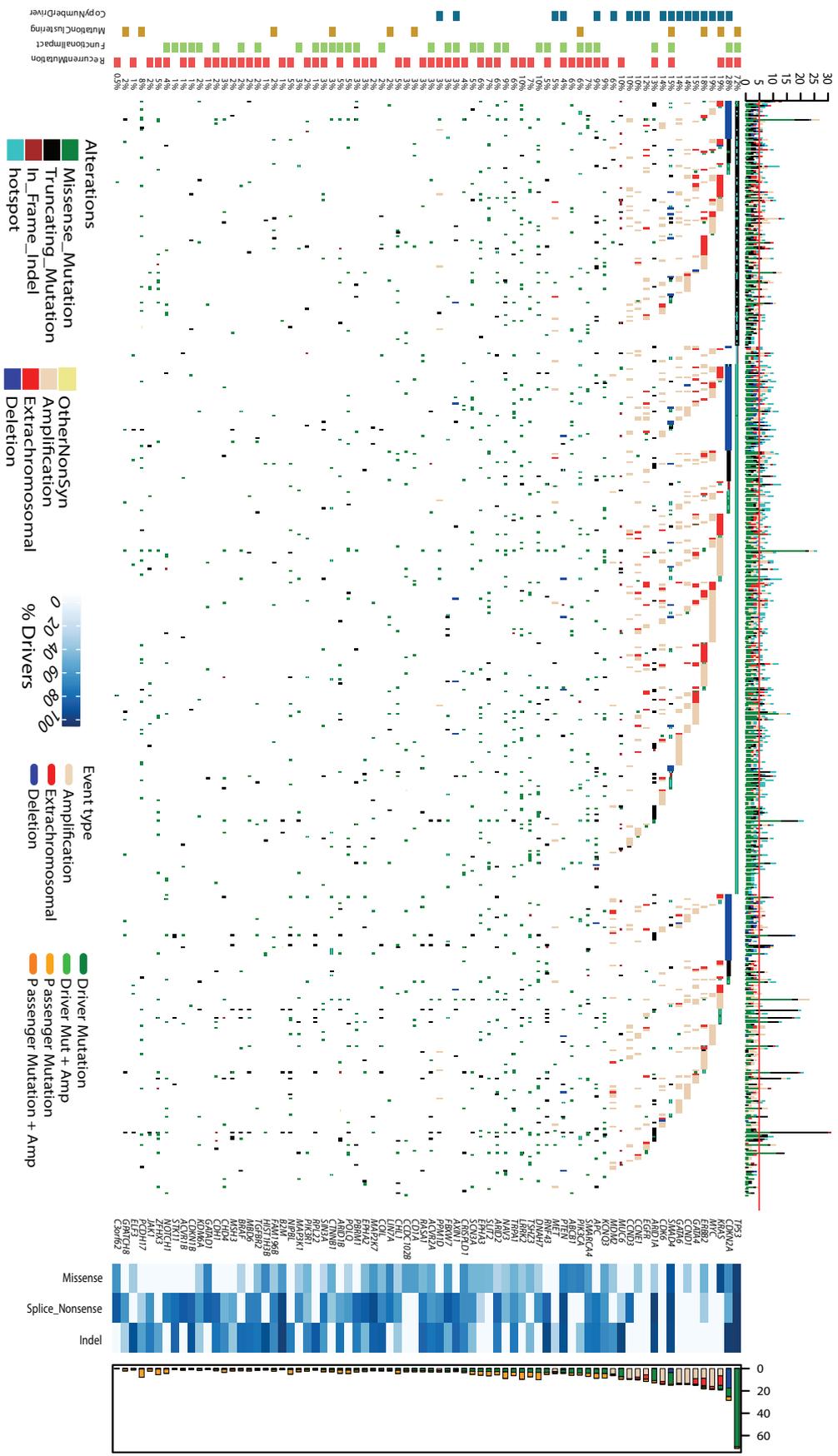


**Figure 9. Relationship between copy number and expression in CN driver genes in RNA matched sub-cohort (n=116).** A 2D kernel density estimation and a loess regression curve with 95% CIs (grey) are also shown to describe the data. The dashed line indicates Ploidy adjusted CN = 2, the cut off for defining amplification.

# Results chapter 2: Characterising OAC Drivers

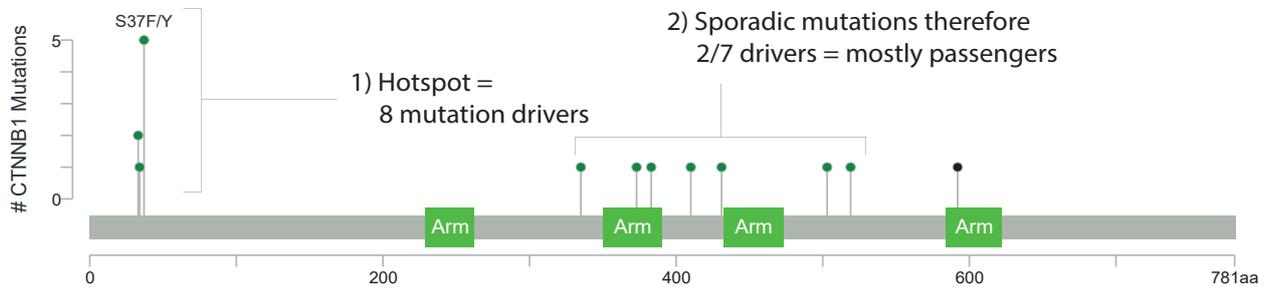
## Landscape of driver Events in OAC

The overall landscape of driver gene mutations and copy number alterations per case is depicted in Figure 10. These comprise both oncogenes and tumour suppressor genes which have been activated or repressed via different mechanisms. Occasionally different types of events are selected for in the same gene, such as *KRAS* and *ERBB2* which both harbour activating mutations and amplifications in 19% and 18% of cases respectively. Passenger mutations occur by chance in most driver genes. To quantify this we have used the observed:expected mutation ratios (calculated by dNdScv) to estimate the percentage of driver mutations in each gene and in different mutation classes. For many genes, only specific mutation classes appear to be under selection. Many tumour suppressor genes; *ARID2*, *RNF43*, *ARID1B* for example, are only under selection for truncating mutations; *i.e.* splice site, nonsense and frameshift Indel mutations, but not missense mutations which are not under selection. However, oncogenes, like *ERBB2*, only contain missense drivers which form clusters to activate gene function in a specific manner. Where a mutation class is <100% driver mutations, mutational clustering can help us define the driver vs passenger status of a mutation (Figure 11). Clusters of mutations occurring in OAC or mutations on amino acids which are mutation hotspots in other cancer types<sup>98</sup> are indicated in Figure 10. Novel OAC drivers of particular interest include *B2M*, a core component of the MHC (major

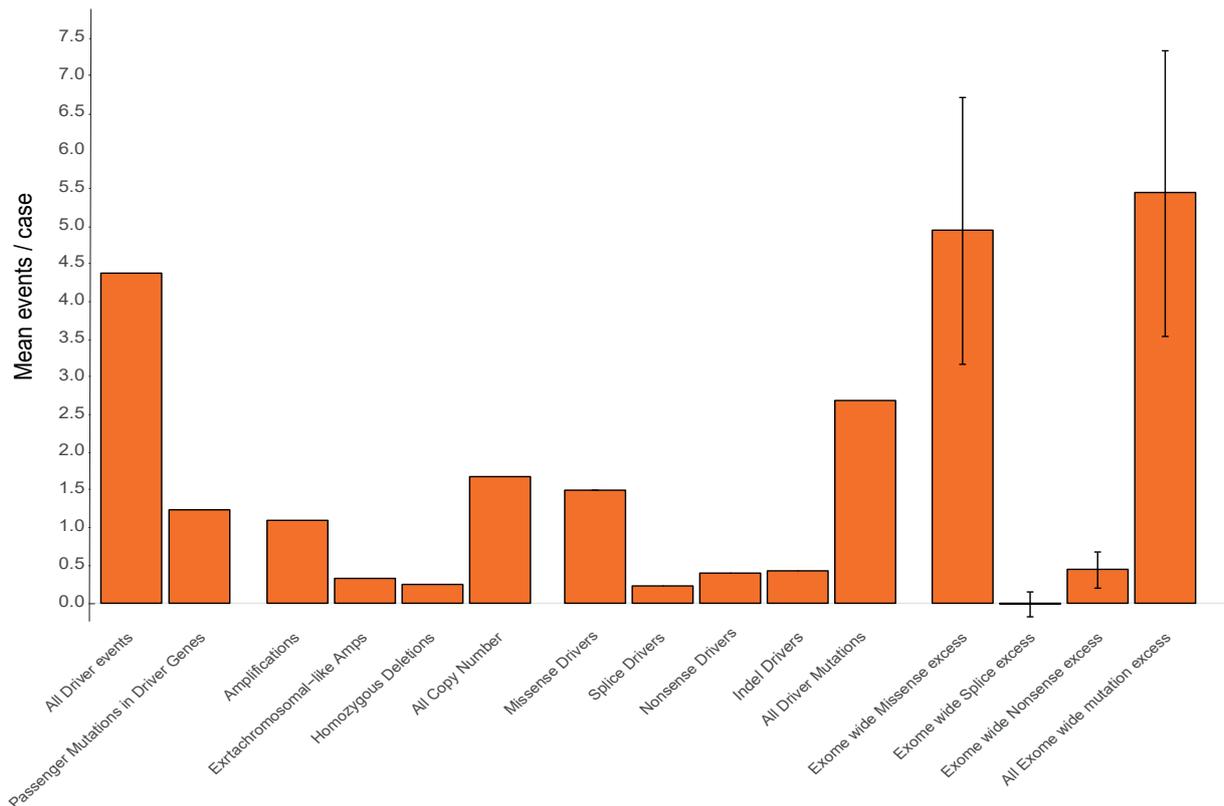


**Figure 10. The driver gene landscape of OAC.** Driver mutations or CNVs are shown for each patient. Amplification is defined as >2 Copy number adjusted ploidy (2 x ploidy of that case) and extrachromosomal amplification as >10 Copy number adjusted ploidy (10 x ploidy for that case). Driver associated features for each driver gene are displayed to the left. On the right the percentages of different mutation and copy number changes are displayed, differentiating between driver and passenger mutations using dNdScv, and the % of predicted drivers by mutation type is shown. Above the plot are the number of driver mutations per sample with an indication of the mean (red line = 5).

CTNNB1 mutations: Missense observed:expected = 3:1 therefore 2/3s missense mutations likely drivers = 10/15 total missense.



**Figure 11.** A scheme demonstrating how to use mutational clustering along with dn/ds ratios to estimate the probability of a particular mutation being a driver. In this case, the dn/ds ratio suggests 2/3 of missense mutations are drivers hence 10/15 mutations. 8 missense mutations lie in a mutational cluster, in this case of known significance in the N-terminal of B-Catenin, making it likely that these are drivers and hence most (2/7) other mutations are likely to be passengers. Similarly, mutations on amino acids known to be drivers in other cancer types (see Supplementary table 3, eg KRAS G12 mutations) can be considered likely drivers.



**Figure 12. A detailed breakdown of mutation and copy number types per case and the exome wide dn/ds excess for different mutation types in 551 OACs.** Exome wide excess indel rates cannot be calculated as they have no synonymous mutation equivalent. However a null model based on indel rates in genes presumed to be under no selection is used in the per gene dn/ds method. Error bars indicate 95% confidence intervals for exome wide dn/ds mutation excess assessment and bars indicate the mean.

histocompatibility complex) class I complex and resistance marker for Immunotherapy<sup>102</sup>, MUC6 a secreted glycoprotein involved in gastric acid resistance and *ABCB1* a channel pump protein which is associated with multiple instances of drug resistance<sup>103</sup>. We note that several of these drivers have been previously associated with gastric and colorectal cancer<sup>104,105</sup> (Table 2) .

The identification of driver events provides a rich information about the molecular history of each OAC tumour. We detect a median of five events in driver genes per tumour (IQR = 3-7, Mean = 5.6) and only a very small fraction of cases have no such events detected (6 cases, 1%). When we remove the predicted percentage of passenger mutations using dN/dS ratios we find a mean of 4.4 true driver events per case which derive more commonly from mutations than CNAs (Figure 12). dNdScv, one of the driver gene detection methods used, also analyses the genome-wide excess of non-synonymous mutations based on expected mutation rates to assess the total number of driver mutations across the exome which is calculated at 5.4 (95% CIs: 3.5-7.3), in comparison to 2.7 driver mutations which we calculate in our gene-centric analysis after passenger removal. This suggests low frequency driver genes may be prevalent in the OAC mutational landscape (see discussion). Further analysis suggests these missing mutations are mostly missense mutations and our gene-centric analysis captures almost all predicted splice and nonsense drivers (Figure 12). Some of our methods use enrichment of nonsense and splice mutations as a marker of driver genes and hence have a higher sensitivity for these mutations.

To determine whether distinct subgroups of OAC, driven by specific sets of drivers, might exist we performed hierarchical clustering on a binary matrix of tumours and the presence or not of events in all 76 drivers. We did not find a very distinct structure in the resulting clustering, which seemed to be mostly determined by the number of drivers and

Table 2. Q values for every tool across all mutational drivers. Driver Novelty in OAC, pan-cancer and in gastric and colorectal cancers.

gene_name	Mutation Frequency (%)	edriver	edriver3d	dndscv	OncodriveFM	OncodriveClust	Mutsig	Tools_identifying	Novel_Drivers?*	Pan-cancer drivers**	Gastric_Drivers?***	Colorectal_Drivers?****
TP53	72.23	0	0	0	0	0	0	6	FALSE	TRUE	TRUE	TRUE
SMAD4	11.25	0	0	0	0	0	0	6	FALSE	TRUE	TRUE	TRUE
PIK3CA	5.99	0	0	0	0.0348	0	0.0255	6	FALSE	TRUE	TRUE	TRUE
KRAS	3.81	0.00528	0	0	0.459	0	0	5	FALSE	TRUE	TRUE	TRUE
ARID1A	12.7	1	1	0	0	1	0	3	FALSE	TRUE	TRUE	TRUE
CDKN2A	11.25	1	1	0	0	1	0	3	FALSE	TRUE	FALSE	TRUE****
APC	8.35	1	1	0	0	1	0	3	TRUE	TRUE	TRUE	TRUE
SMARCA4	7.44	1	1	0.000535	0.0000297	1	0.00275	3	FALSE	TRUE	FALSE	FALSE
ARID2	5.81	1	1	0	0.0135	1	0.0906	3	FALSE	TRUE	FALSE	FALSE
PBRM1	4.36	1	1	0.000629	0.00167	1	0.0737	3	TRUE	TRUE	FALSE	FALSE
PTEN	3.45	1	1	0	0.002	1	0	3	TRUE	TRUE	FALSE	TRUE
ACVR2A	3.09	1	1	0.0476	0	1	0.0188	3	TRUE	TRUE	FALSE	TRUE
FBXW7	3.27	1	1	0.000323	0	1	0.00672	3	FALSE	TRUE	FALSE	FALSE
SIN3A	2.9	1	1	0.0354	0.0198	1	0.0555	3	TRUE	TRUE	FALSE	TRUE
CDKN1B	1.27	1	1	0.00557	0.0372	1	0.00275	3	TRUE	TRUE	FALSE	FALSE
MUC6	9.8	1	1	0	1	1	0	2	TRUE	TRUE	TRUE	FALSE
EPHA3	6.53	1	1	0.0837	0.037	1	0.292	2	TRUE	TRUE	FALSE	FALSE
RNF43	5.26	1	1	0	0	1	0.451	2	TRUE	TRUE	TRUE	TRUE
NOTCH1	4.54	1	1	0.0584	0.0745	1	1	2	TRUE	TRUE	FALSE	FALSE
ARID1B	4.54	1	1	0.00495	0.0377	1	0.163	2	TRUE	TRUE	FALSE	FALSE
CTNNB1	2.9	1	1	1	0.0325	0	0.209	2	TRUE	TRUE	TRUE	TRUE
BRAF	2	1	1	0.0375	0.0182	1	1	2	TRUE	TRUE	FALSE	TRUE
CDH1	2	1	1	0.0155	0.0387	1	1	2	TRUE	TRUE	TRUE	FALSE
ACVR1B	1.63	1	1	0.0265	0.029	1	0.598	2	TRUE	TRUE	FALSE	FALSE
AXIN1	1.63	1	1	0.00133	0.0815	1	0.494	2	TRUE	TRUE	FALSE	FALSE
B2M	1.27	1	1	0	1	1	0.0274	2	TRUE	TRUE	FALSE	FALSE
TGFB2	1.63	1	1	0.0697	0.00571	1	1	2	TRUE	TRUE	FALSE	FALSE
HIST1H3B	1.27	1	1	0.00517	1	1	0.0274	2	TRUE	TRUE	FALSE	FALSE
RPL22	1.27	1	1	1	0.0372	1	0.0007	2	TRUE	TRUE	FALSE	FALSE
DNAH7	10.71	1	1	1	0.0149	1	1	1	TRUE	FALSE	FALSE	FALSE
LRK2	10.89	1	1	0.059	0.336	1	1	1	TRUE	TRUE	FALSE	FALSE
NAV3	9.8	1	1	1	0.0164	1	1	1	TRUE	TRUE	FALSE	FALSE
KCNQ3	9.44	1	1	1	1	1	0	1	FALSE	FALSE	FALSE	FALSE
PCDH17	8.17	1	1	1	1	0.000181	1	1	TRUE	TRUE	FALSE	FALSE
SLIT2	7.8	1	1	1	1	1	0.00214	1	TRUE	FALSE	FALSE	FALSE
TRPA1	7.44	1	1	1	1	1	0.00124	1	TRUE	FALSE	FALSE	FALSE
TSHZ3	7.44	1	1	0.00387	0.378	1	1	1	TRUE	TRUE	FALSE	FALSE
ABCB1	6.35	1	1	1	1	1	0.0123	1	TRUE	FALSE	FALSE	FALSE
POLQ	5.08	1	1	1	0.037	1	1	1	TRUE	TRUE	FALSE	FALSE
ZFXH3	5.44	1	1	0.0787	1	1	1	1	TRUE	TRUE	FALSE	FALSE
CHL1	6.17	1	1	1	1	1	0.0352	1	TRUE	FALSE	FALSE	FALSE
NIPBL	5.44	1	1	0.0354	1	1	0.2	1	TRUE	TRUE	FALSE	FALSE
SCN3A	5.26	1	1	1	0.0135	1	1	1	TRUE	FALSE	FALSE	FALSE
CRISPLD1	3.99	1	1	1	1	1	0.00248	1	TRUE	FALSE	FALSE	FALSE
CCDC102B	3.99	1	1	1	1	1	0.0729	1	TRUE	FALSE	FALSE	FALSE
CHD4	3.63	1	1	0.0646	1	1	1	1	TRUE	TRUE	FALSE	TRUE
ERBB2	3.09	1	1	0.0000263	1	1	0.694	1	FALSE	TRUE	TRUE	FALSE
RASA1	2.9	1	1	1	0.00114	1	1	1	TRUE	TRUE	TRUE	FALSE
CD1A	3.27	1	1	1	1	0	1	1	TRUE	FALSE	FALSE	FALSE
MSH3	2.9	1	1	1	1	1	0.0468	1	TRUE	TRUE	FALSE	FALSE
MAP3K1	3.09	1	1	1	0.0542	1	1	1	TRUE	TRUE	FALSE	FALSE
EPHA2	2.9	1	1	0.00991	1	1	1	1	TRUE	TRUE	FALSE	FALSE
JAK1	2.54	1	1	0.0418	1	1	1	1	TRUE	TRUE	FALSE	FALSE
COIL	2.36	1	1	1	0.0815	1	0.132	1	TRUE	TRUE	FALSE	FALSE
PIK3R1	2.36	1	1	0.0787	0.135	1	1	1	TRUE	TRUE	FALSE	FALSE
FAM196B	2.36	0.03	1	1	1	1	1	1	TRUE	FALSE	FALSE	FALSE
LIN7A	2.36	1	1	1	1	0	0.118	1	TRUE	FALSE	FALSE	FALSE
GPATCH8	2.18	0.0491	1	1	1	1	1	1	TRUE	FALSE	FALSE	FALSE
MBD6	2.18	1	1	0.014	1	1	0.694	1	TRUE	TRUE	FALSE	FALSE
KDM6A	2	1	1	1	0.0325	1	1	1	TRUE	TRUE	FALSE	FALSE
MAP2K7	2	1	1	0.000149	1	1	0.117	1	TRUE	TRUE	FALSE	FALSE
PPM1D	1.45	1	1	0.0924	1	1	0.817	1	TRUE	TRUE	FALSE	FALSE
ELF3	1.27	1	1	0.0319	0.865	1	1	1	TRUE	TRUE	FALSE	TRUE
GATAD1	0.91	1	1	1	1	1	0	1	TRUE	FALSE	FALSE	FALSE
C3orf62	0.36	1	1	1	1	1	0	1	TRUE	FALSE	FALSE	FALSE
STK11	0.91	1	1	1	0.0202	1	0.684	1	TRUE	TRUE	FALSE	FALSE

\* Known drivers from Dulak et al 2013 Nature genetics, Secrier et al 2016 Nature genetics, TCGA 2017 Nature and Lin et al 2017 Gut

\*\* Defined as those called in Kandoth et al 2013 Nature and Martincorena et al 2017 Cell and tier 1 Cancer gene consensus genes (excluding blood cancer associated translocation drivers).

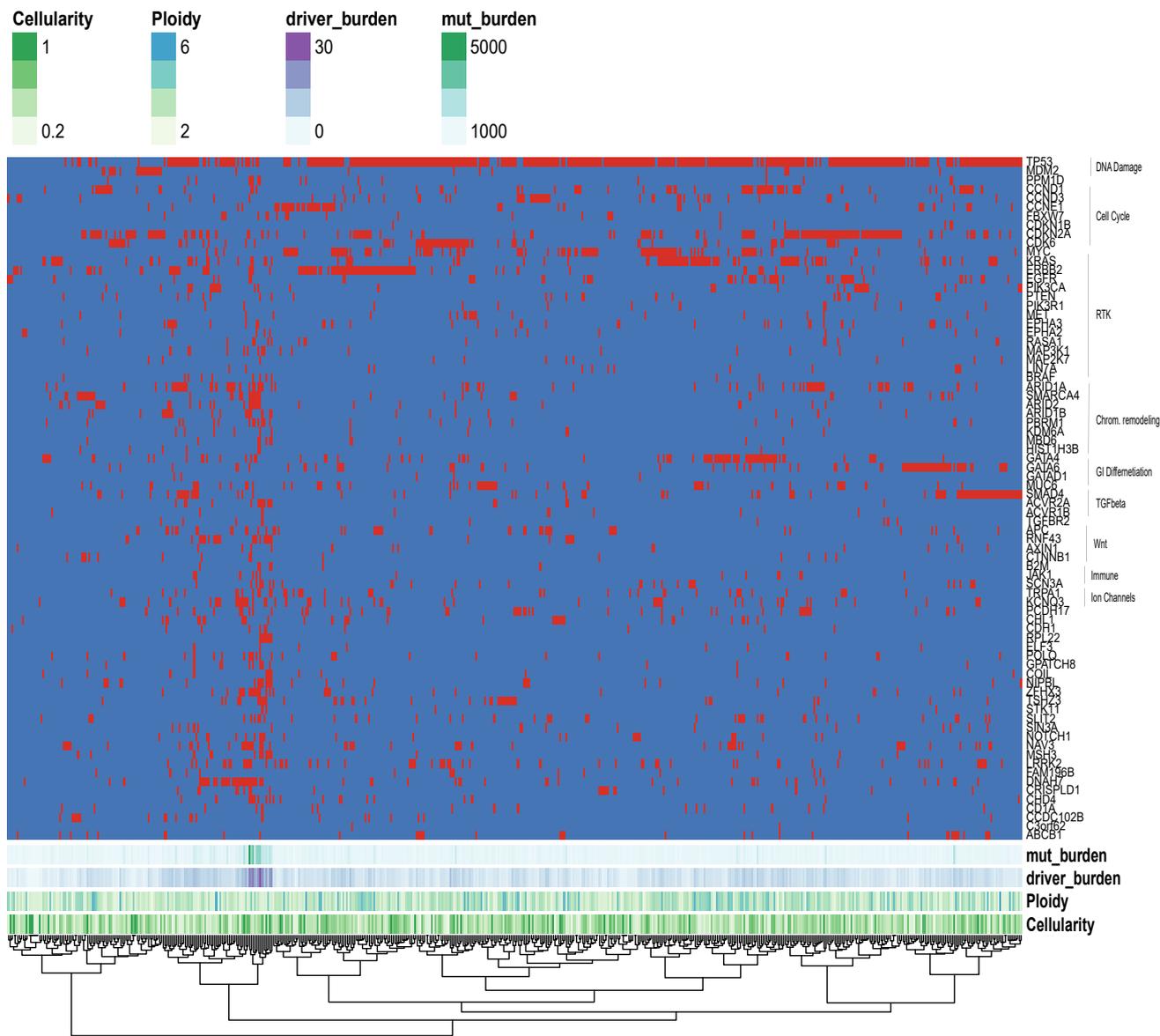
\*\*\* Gastric cancer drivers defined from TCGA 2014 Nature

\*\*\*\* Colorectal cancer drivers defined from Grasso et al 2018 Cancer Discovery

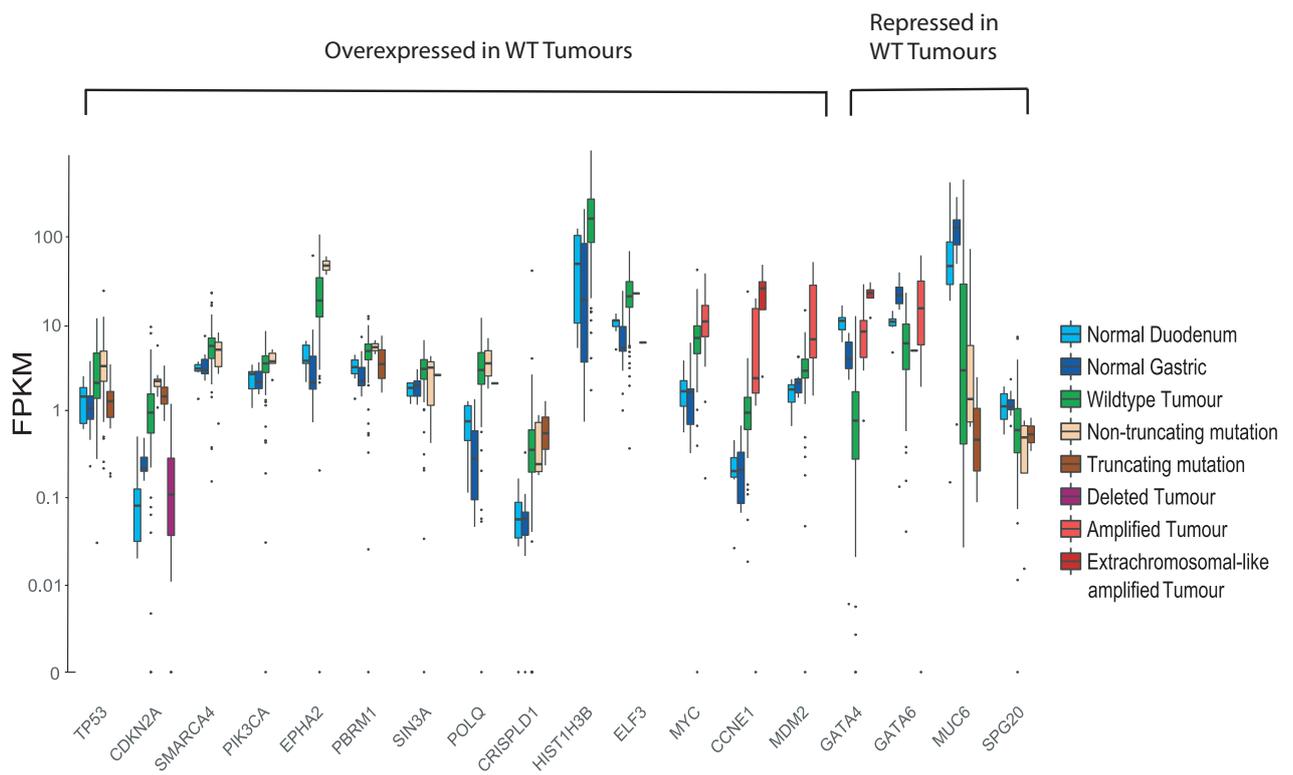
\*\*\*\*\* CDKN2A a known Gastric cancer driver (eg Huang et al 2015 Int J Clin Exp Med) but was not called in TCGA 2014. This may be due to expression filtering which can remove highly deleted/truncated tumour suppressor genes when these alterations lead to expression loss.

singular presence or not of more frequent drivers, however we noted that *TP53* mutant cases had significantly more CN drivers when inspecting the *TP53* mutant dominated cluster (Wilcoxon test,  $p = 0.0032$ , Figure 13).

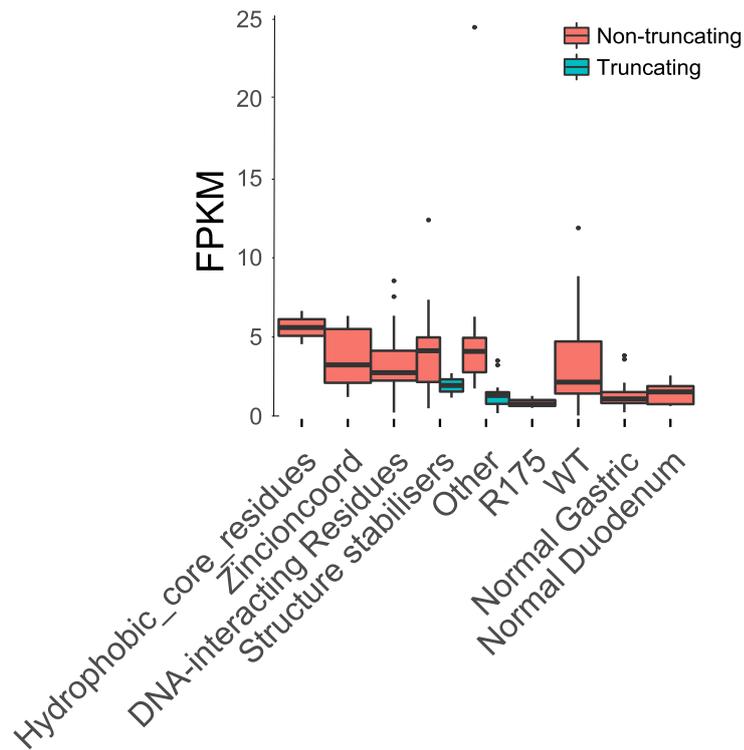
To better understand the functional impact of driver mutations we analysed expression of driver genes with different mutation types and compared their expression to normal tissue RNA, which was sequenced alongside our tumour samples (Figure 14). Since surrounding squamous epithelium is a fundamentally different tissue, from which OAC does not directly arise, we have used duodenum and gastric cardia samples as gastrointestinal phenotype controls, likely to be similar to the, as yet unconfirmed, tissue of origin in OAC. A large number of driver genes have upregulated expression in comparison to normal controls, for example *TP53* has upregulated RNA expression in WT tumour tissue and in cases with missense (see non-truncating, Figure 14) mutations but RNA expression is lost upon gene truncation. In depth analysis of different *TP53* mutation types reveals significant heterogeneity within non-truncating mutations, for example R175H mutations correlate with low RNA expression (Figure 15). Normal tissue expression of *CDKN2A* suggests that *CDKN2A* is generally upregulated in OAC, likely due to genotoxic or other cancer-associated stresses<sup>106</sup> and returns to physiologically normal levels when deleted. High expression in *CDKN2A* mutated cases suggests this upregulation is sustained and the cancer cell escapes the effects of such expression by preventing *CDKN2A* function using mutations (either missense mutations, which are often clustered (figure 16), preventing protein function, or nonsense mutations preventing protein nuclear export and translation). Heterogeneous expression in WT *CDKN2A* cases suggest a different mechanism of inhibition such as methylation in some cases and *CDKN2A* independent mechanisms in cases with high *CDKN2A* expression, for example amplification and overexpression of *CDK6* may allow proliferation even in the



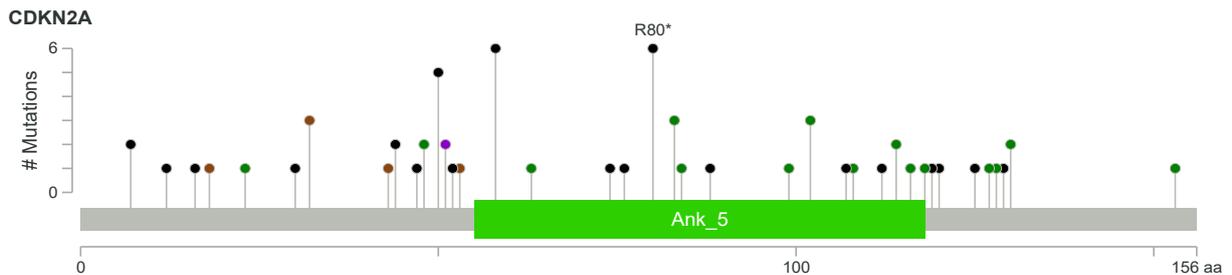
**Figure 13.** Hierarchical Clustering of samples based on presence of driver variants with genes ordered by pathway membership.



**Figure 14. Expression changes in OAC driver genes in comparison to normal intestinal tissues.** Genes with expression changes of note are shown. WT = Wild type.



**Figure 15. TP53 expression in different TP53 mutation types in comparison to TP53 WT tumours, normal duodenum and gastric cardia tissues in 116 WGS OAC cases with matched RNAseq data.** Boxplots represent the median and interquartile range.



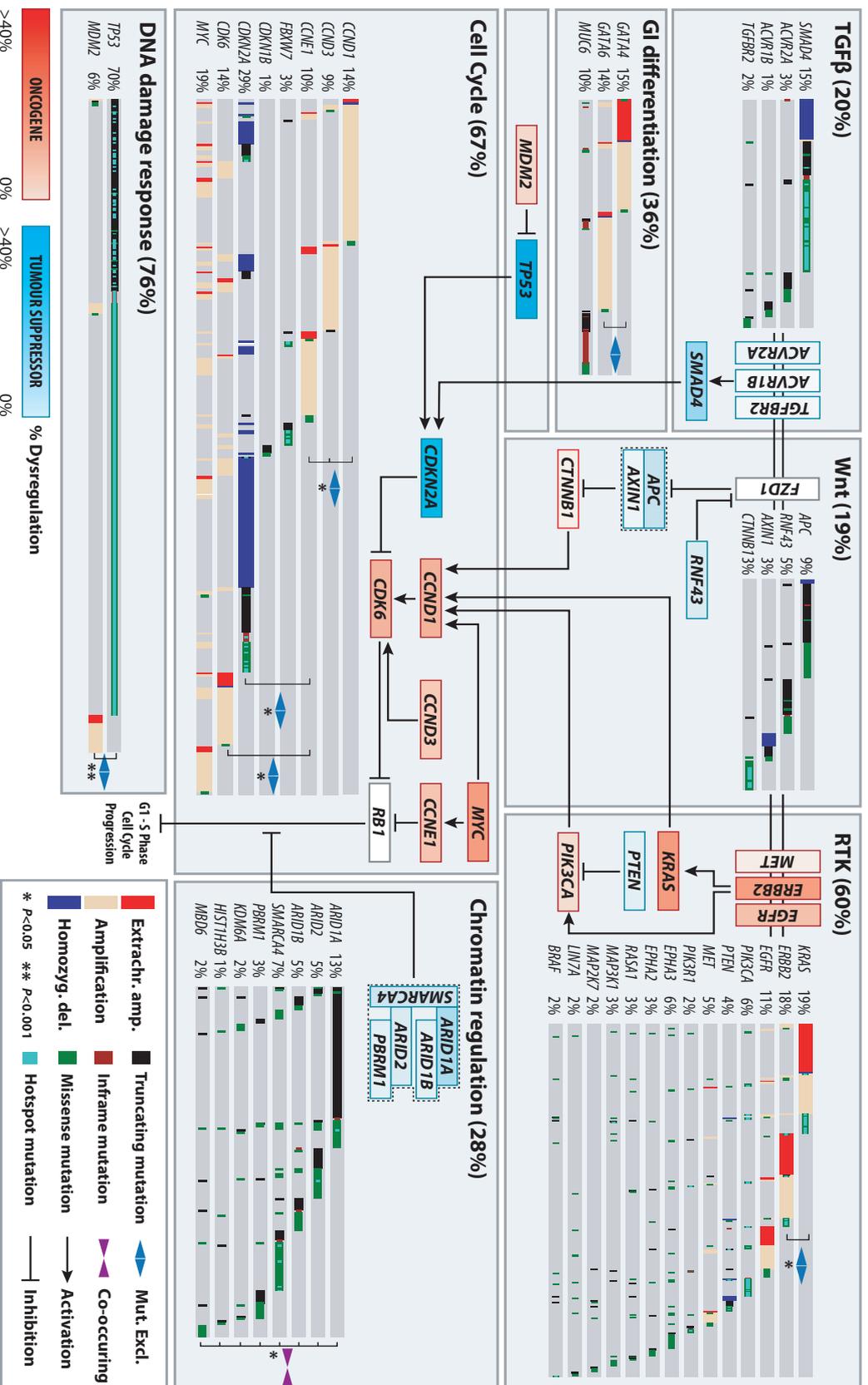
**Figure 16.** Lollipop plot showing mutations in CDKN2A. Note Missense mutations (green) are commonly in hotspots - 12/21. Brown = inframe indels, black = truncating mutations (Nonsense or frameshift indel).

presence of *CDKN2A* upregulation. Overexpression of other genes in wild type tumours, such as *SIN3A*, may confer a selective advantage due to their oncogenic properties, in this case cooperating with *MYC*, which is also overexpressed in OACs (Figure 15). A smaller number of driver genes are downregulated in OAC tissue and 3/4 of these (*GATA4*, *GATA6* and *MUC6*) are involved in the differentiated phenotype of gastrointestinal tissues and may be lost with tumour de-differentiation. Driving alterations in these genes have been observed in other GI cancers however their oncogenic mechanism is poorly understood.

## Dysregulation of specific pathways and processes in OAC

It is known that selection preferentially dysregulates certain functionally related groups of genes and biological pathways in cancer<sup>107</sup>. This phenomenon is highly evident in OAC, as shown in Figure 17 which depicts the functional relationships between OAC drivers. This provides greater functional homogeneity to the landscape of driver events.

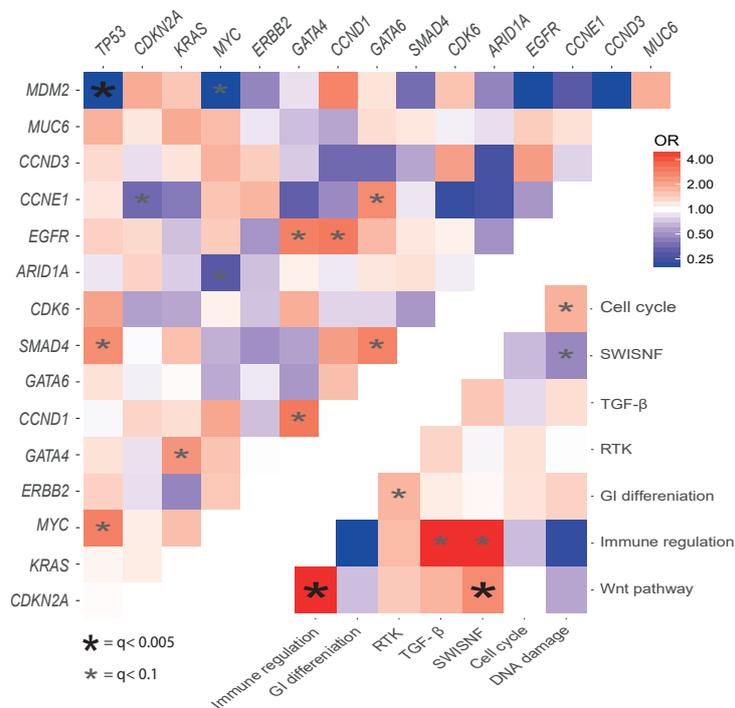
While *TP53* is the dominant driver in OAC, 28% of cases remain *TP53* wildtype. *MDM2* is a E3 ubiquitin ligase that targets *TP53* for degradation. Its selective amplification and overexpression is mutually exclusive with *TP53* mutation suggesting it can functionally substitute the effect of *TP53* mutation via its degradation. Similar mutually exclusive relationships are observed between; *KRAS* and *ERBB2*, *GATA4* and *GATA6* and Cyclin genes (*CCNE1*, *CCND1* and *CCND3*). Activation of the Wnt pathway occurs in 19% of cases either by mutation of phospho-residues at the N terminus of  $\beta$ -catenin, which prevent degradation, or loss of Wnt destruction complex components like *APC*. Many different chromatin modifying



**Figure 17. Biological pathways undergoing selective dysregulation in OAC. a.** Biological Pathways dysregulated by driver gene mutation and/or CNVs in 551 cases. WT cases for a pathway are not shown. Mutual exclusivities and/or associations between genes in a pathway are annotated. GAT44/GAT76 amplifications have a mutually exclusive relationship (ie GAT44 amplification is more common in GAT76 WT cases) although this does not reach statistical significance (fisher's exact test, two sided,  $p=0.07$  OR=0.52).

genes, often belonging to the SWI/SNF complex, are also selectively mutated (28% of cases). In contrast SWI/SNF genes are co-mutated significantly more often than we would expect by chance (Fisher's exact test,  $P < 0.01$  see methods page 43), suggesting an increased advantage to further mutations once one has been acquired. We also assessed mutual exclusivity and co-occurrence in genes in different pathways and between pathways themselves (Figure 18). Of particular note are co-occurring relationships between *TP53* and *MYC*, *GATA6* and *SMAD4*, Wnt and Immune pathways as well as mutually exclusive relationships between *ARID1A* and *MYC*, gastrointestinal (GI) differentiation and RTK pathways and SWI-SNF and DNA-Damage response pathways. Wnt dysregulation has been previously linked to immune escape and interestingly was also associated with hyper-mutated cases ( $> 50,000$  SNVs or Indels, fisher's exact test,  $p = 0.021$ , OR= 2.4). We were able to confirm some of these relationships in independent cohorts in different cancer types (Table 3) suggesting some of these may be pan-cancer phenomenon. As shown in Figure 17, all of these pathways interact to stimulate the G1 to S phase transition of the cell cycle via promoting phosphorylation of Rb, although many of these pathways have multiple oncogenic or tumour suppressive functions.

A number of other driver genes have highly related functional roles including core transcriptional components (*TAF1* and *POLQ*), drivers of immune escape (*JAK1* and *B2M*), cell adhesion receptors (*CDH1*, *CHDL* and *PCDH17*), core ribosome components (*ELF3* and *RPL22*), core RNA processing components (*GPATCH8* and *COIL*), ion channels (*KCNQ3* and *TRPA1*) and Ephrin type-A receptors (*EPHA2* and *EPHA3*).



**Figure 18. Pairwise assessment of mutual exclusivity and association in OAC driver genes and pathways.**

**Table 3. Validation of mutual exclusivity and cooccurrence of genes and pathways in independent cohorts.**

P values and ORs are calculated from Fisher's exact test (two tailed)

Cohort	Gene A	Gene B	Neither	A Not B	B Not A	Both	Log OR	OR	p-Value	Tendency
MSKCC pan-cancer cohort (n= 10,945)	B2M*	CTNNB1*	10443	130	358	14	1.145	13.964	<0.001	Co-occurrence
MSKCC pan-cancer cohort (n= 10,945)	TP53	SMAD4	6097	4286	204	358	0.915	8.2224	<0.001	Co-occurrence
MSKCC pan-cancer cohort (n= 10,945)	TP53	MYC	6130	4293	171	351	1.075	11.885	<0.001	Co-occurrence
MSKCC pan-cancer cohort (n= 10,945)	TP53	MDM2	5913	4570	388	74	-1.399	0.0399	<0.001	Mutual exclusivity
MSKCC pan-cancer cohort (n= 10,945)	ERBB2	KRAS	8478	661	1741	65	-0.736	0.1837	<0.001	Mutual exclusivity
MSKCC pan-cancer cohort (n= 10,945)	EGFR	KRAS	8336	803	1754	52	-1.178	0.0664	<0.001	Mutual exclusivity
Pancreatic adenocarcinoma**	SMAD4	GATA6	373	129	16	14	0.928	8.4723	0.013	Co-occurrence
Pancreatic adenocarcinoma**	GATA4	EGFR	522	8	1	1	>3	>1000	0.034	Co-occurrence

\* Two genes from Immune and Wnt pathways which co-occur in our data

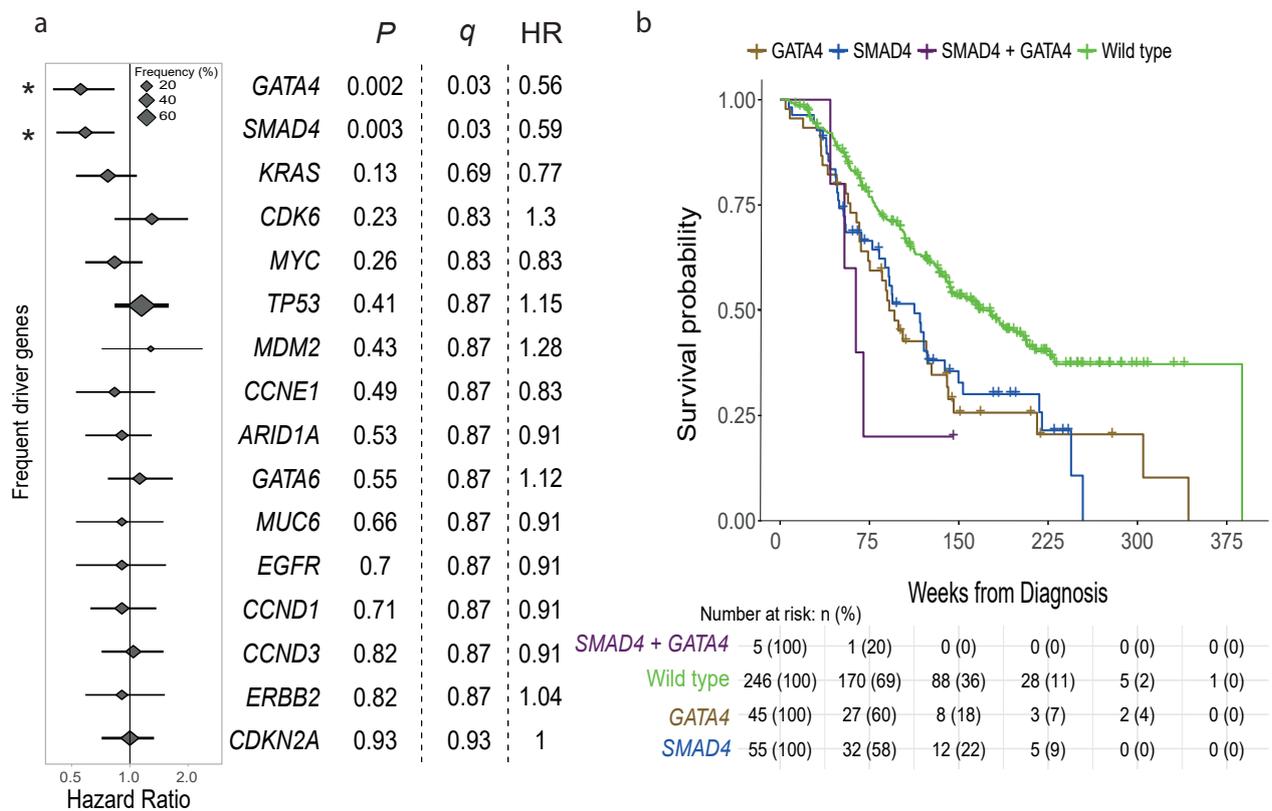
\*\*Pancreatic adenocarcinoma (TCGA + QCMG study 2016) used to assess GATA factors as not in MSKCC gene panel (n = 433)

# Results chapter 3: Using OAC drivers as clinical biomarkers

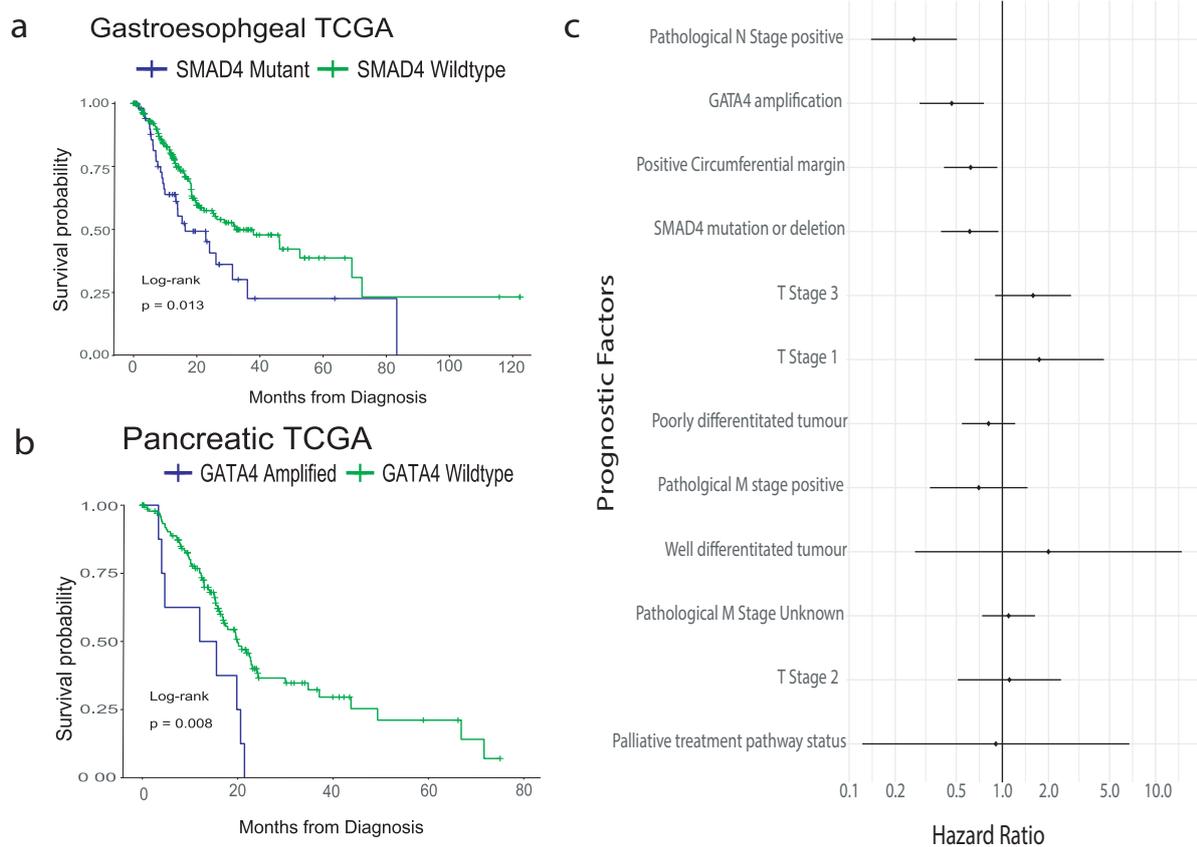
## Clinical significance of driver variants

Events undergoing selection during cancer evolution influence tumour biology and thus impact tumour aggressiveness, response to treatment and patient prognosis as well as other clinical parameters. Clinical-genomic correlations can provide useful biomarkers but also give insights into the biology of these events.

Univariate Cox regression was performed for events in each driver gene with driver events occurring in greater than 5% of OACs (*i.e.* after removal of predicted passengers, 16 genes) to detect prognostic biomarkers (Figure 19). Events in two genes conferred significantly poorer prognosis after multiple hypothesis correction, *GATA4* amplification (HR : 0.54 , 95% CI : 0.38 – 0.78, *P* value = 0.0008) and *SMAD4* mutation or homozygous deletion (HR : 0.60 , 95% CI : 0.42 – 0.84, *P* value = 0.003). Both genes remained significant in multivariate Cox regression including pathological TNM staging, resection margin, curative vs palliative treatment intent and differentiation status (*GATA4* = HR adjusted : 0.47, 95% CIs adjusted : 0.29 - 0.76, *P* value = 0.002 and *SMAD4* = HR adjusted : 0.61, 95% CI adjusted : 0.40 – 0.94, *P* value = 0.026, Figure 20) and were among the most predictive of the clinical variants with only N stage and positive circumferential margin also remaining significant. 31% of OACs contain either *SMAD4* mutation or homozygous deletion or *GATA4* amplification and cases with both genes altered had a poorer prognosis. We validated the poor prognostic



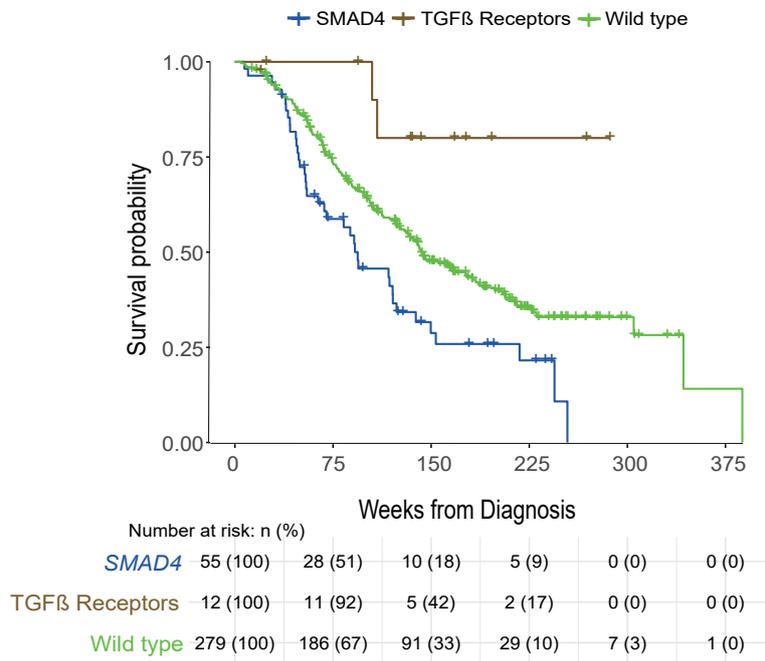
**Figure 19. Discovery of SMAD4 and GATA4 as prognostic indicators in OAC.** a. Hazard ratios and 95% confidence intervals for Cox regression analysis across all driver genes with at least a 5% frequency of driver alterations \* =  $q < 0.05$  after BH adjustment. b. Kaplan-Meier curves for OACs with different status of significant prognostic indicators (GATA4 and SMAD4).



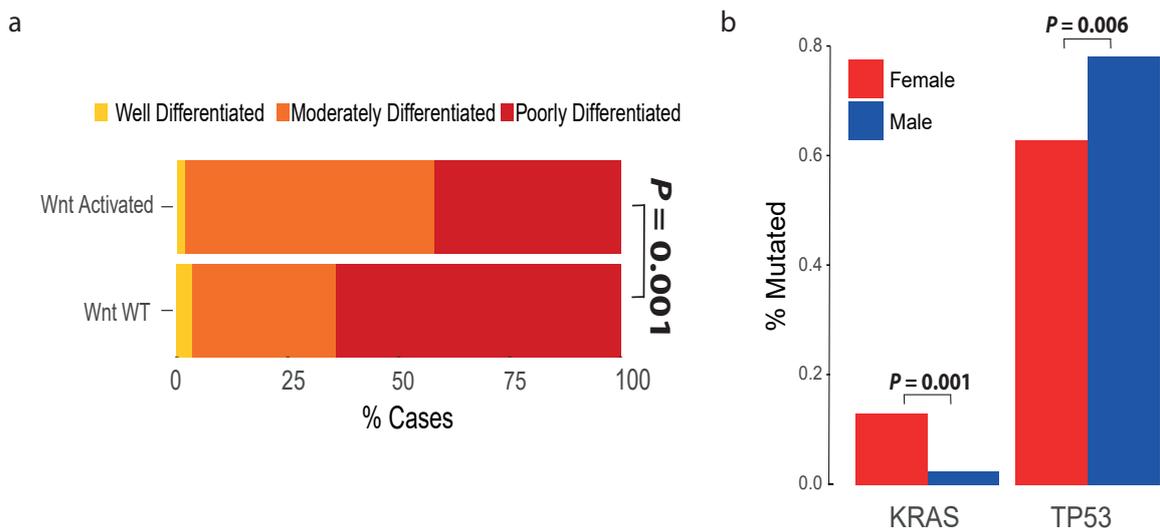
**Figure 20. Validation of GATA4 and SMAD4 prognostic markers.** a. Validation of SMAD4 mutations in a Gastroesophageal cohort. b. Validation of GATA4 amplifications in a pancreatic cohort (no OAC cohorts with enough GATA4 amplifications were found). c. Adjusted Hazard ratios and 95% CIs for clinical and molecular prognostic factors in 379 WGS OACs in a multivariate Cox regression. Both molecular prognostic factors remain significant in the multivariate analysis and are among the most predictive biomarkers when compared to clinical predictors of prognosis - particularly GATA4 amplifications.

impact of *SMAD4* events in an independent TCGA gastroesophageal cohort (HR = 0.58, 95% CI = 0.37 – 0.90, *P* value = 0.014) (Figure 20) and we also found *GATA4* amplifications were prognostic in a cohort of TCGA pancreatic cancers (HR = 0.38 95% CI: 0.18 – 0.80, *P* value = 0.011) (Figure 20), the only available cohort containing a feasible number of *GATA4* amplifications. The prognostic impact of *GATA4* has been suggested in previously published independent OAC cohort<sup>60</sup> although it did not reach statistical significance after FDR correction and *SMAD4* expression loss has been previously linked to poor prognosis in OAC<sup>108</sup>. We also noted stark survival differences between cases with *SMAD4* events and cases in which TGFβ receptors were mutated (HR = 5.6, 95% CI : 1.7 – 18.2, *P* value = 0.005) in keeping with the biology of the TGFβ pathway where non-SMAD TGFβ signalling is known to be oncogenic<sup>109</sup> (Figure 21).

In additional to survival analyses we also assessed driver gene events for correlation with various other clinical factors including differentiation status, sex, age and treatment response. We generally did not find a strong correlation between OAC genomics and most clinical factors. However, we found Wnt pathway mutations had a strong association with well differentiated tumours (*p*=0.001, OR = 2.9, Fisher's test, see methods, Figure 22) and we also noted interesting differences between female (*n*=81) and male (*n*=470) cases. Female cases were enriched for *KRAS* mutation (*p* = 0.001, Fisher's exact test) and *TP53* wildtype status (*p* = 0.006, Fisher's exact test) (Figure 22). This is of particular interest given the male predominance of OAC.



**Figure 21. Kaplan-Meier curves for different alterations in the TGF-β pathway; SMAD4 mutations or deletions and mutations in TGF-β pathway receptor drivers in OAC (TGFBR2, ACVR2A and ACVR1B).**

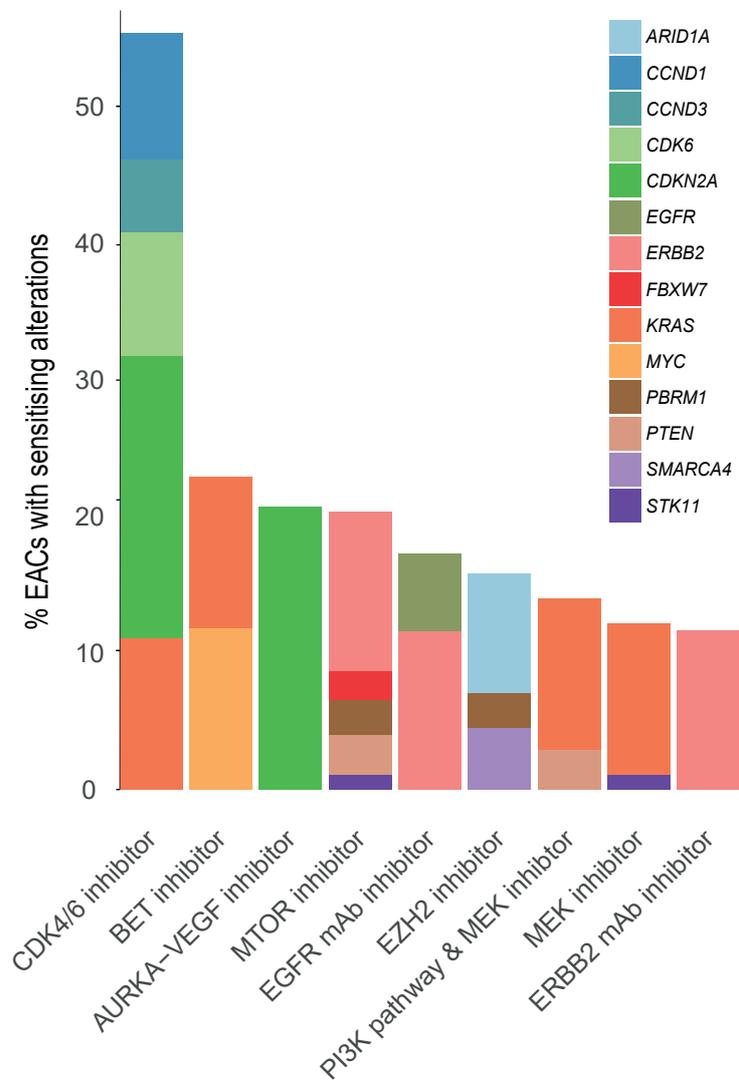


**Figure 22. Correlations between Drivers and Clinical Factors** a. Differentiation bias in tumours containing events in Wnt pathway driver genes. b. Relative frequency of KRAS mutations and TP53 mutations driver gene events in females vs males (Fisher's exact test).

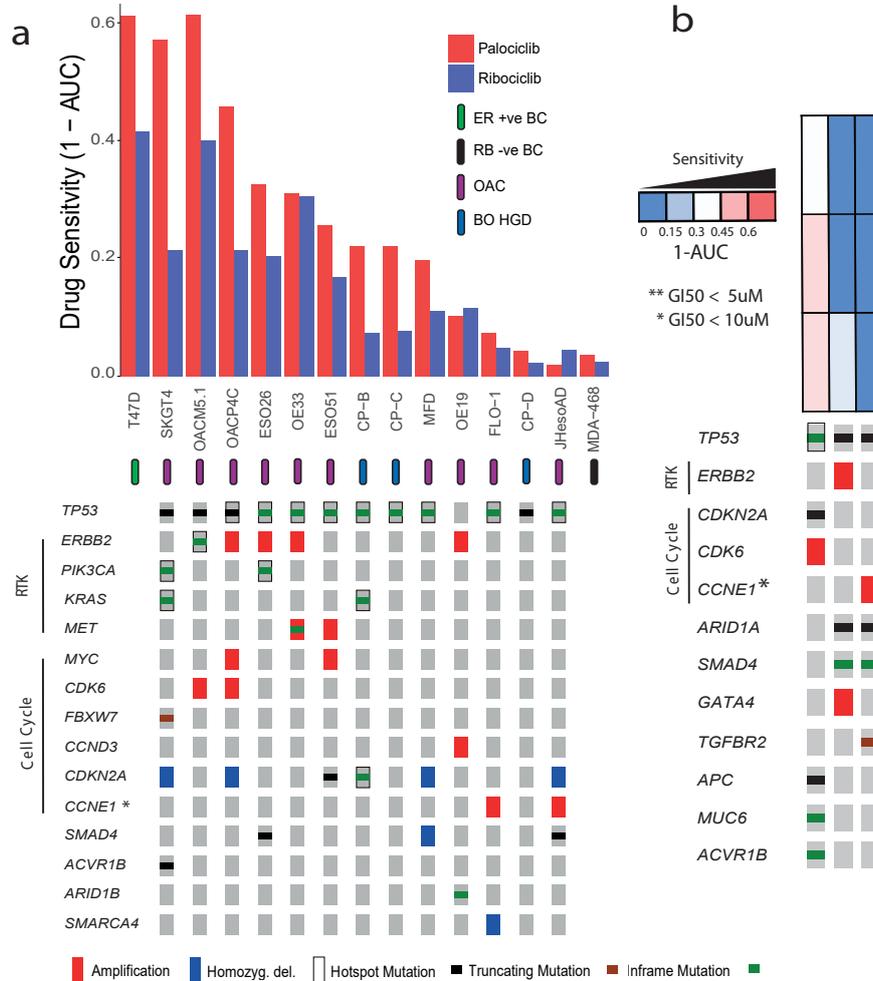
## Targeted therapeutics using OAC driver events

The biological distinctions between normal and cancer cells provided by driver events can be used to derive clinical strategies for selective cancer cell killing. To investigate whether the driver events in particular genes and/or pathways might sensitise OAC cells to certain targeted therapeutic agents we used the Cancer Biomarkers database<sup>110</sup>. We calculated the percentage of our cases which contain OAC-driver biomarkers of response to each drug class in the database (Figure 23). Aside from *TP53*, which has been problematic to target clinically so far, we found a number of drugs with predicted sensitivity in >10% of OACs including EZH2 inhibitors for SWI/SNF mutant cancers (23%, and 28% including other SWI/SNF OAC drivers), and BET inhibitors which target *KRAS* activated and *MYC* amplified cases (25%). However, by far the most significantly effective drug was predicted to be CDK4/6 inhibitors where in >50% of cases harboured sensitivity causing events in the receptor tyrosine kinase (RTK) and core cell cycle pathways (*e.g.* in *CCND1*, *CCND3* and *KRAS*).

To verify that these driver events would also sensitise OAC tumours to such inhibitors we used a panel of thirteen OAC or Barrett's HGD cell lines, which share similar genomic changes and driver events, which have undergone whole genome sequencing<sup>111</sup> and assessed them for presence of OAC driver events (Figure 24). The mutational landscape of these lines was broadly representative of OAC tumours. We found that the presence of cell cycle and or RTK activating driver events was highly correlated with response to two FDA approved CDK4/6 inhibitors, Ribociclib and Palbociclib and several cell lines were sensitive below maximum tolerated blood concentrations in humans<sup>112</sup> (Figure 24, Table 4, Figure 25). Such OAC cell lines had comparable sensitivity to T47D which is derived from an ER +ve



**Figure 23. Drug classes for which sensitivity is indicated by OAC driver genes with data from the Cancer Biomarkers database**



**Figure 24. Drug responses to CDK4/6 inhibitors in OAC in vitro models.**

a. Area under the curve (AUC) of sensitivity is shown in a panel of 13 OAC and BO high grade dysplasia cell lines with associated WGS and their corresponding driver events, based on primary tumour analysis. Also AUC is shown for two control lines T47D, an ER +ve breast cancer line (+ve control) and MDA-MB-468 a Rb negative breast cancer (-ve control). \*CCNE1 is a known marker of resistance to CDK4/6 inhibitors due to its regulation of Rb downstream of CDK4/6 hence bypassing the need for CDK4/6 activity (see figure 4). b. Response of organoid cultures to three FDA approved CDK4/6 inhibitors and corresponding driver events.

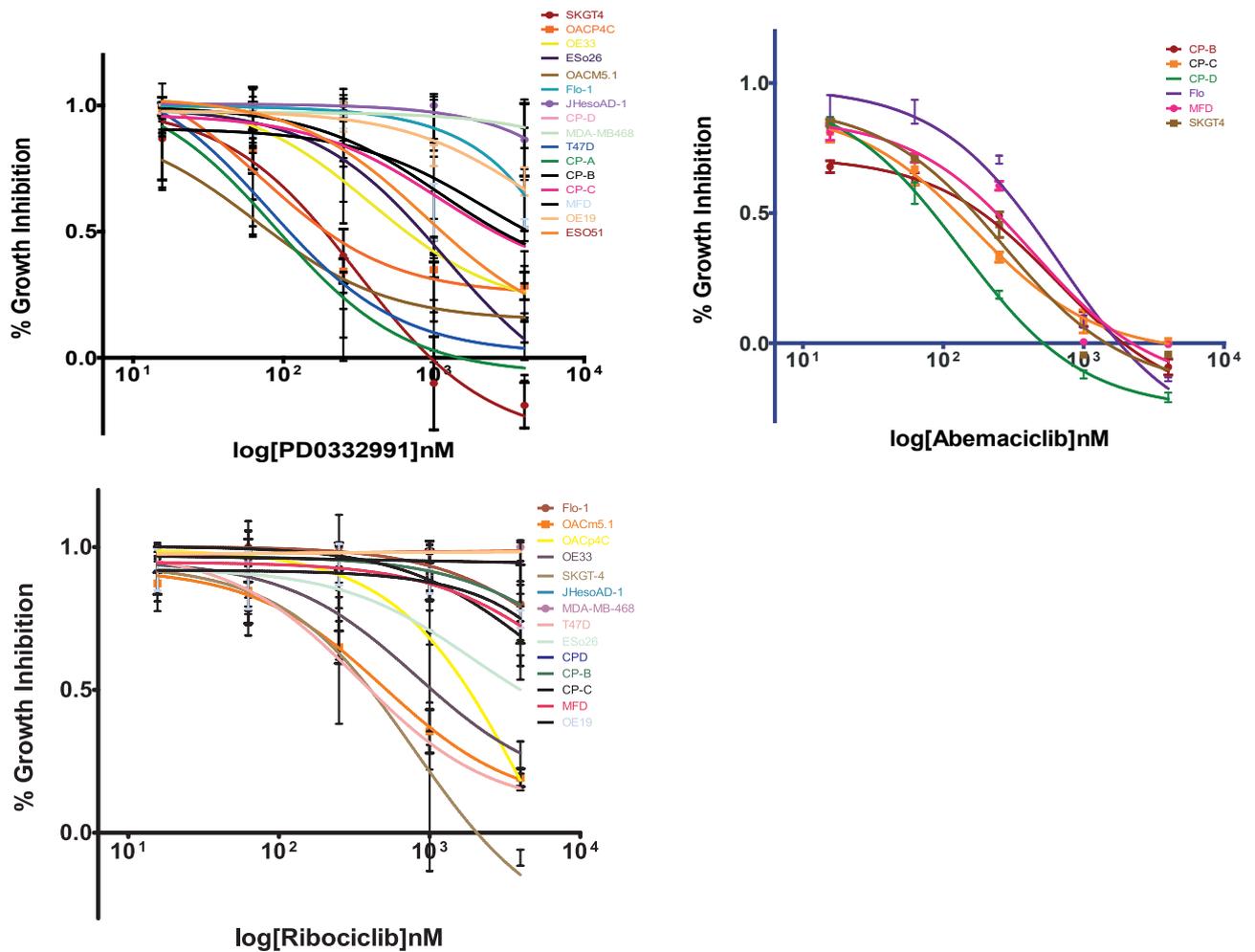
**Table 4. GI50s for Palbociclib and Ribociclib across all cell lines.** GI50s were considered calculable whe at least one drug concentration provided 50% growth inhibition.

Cell line	GI50 (nM)	
	Palbociclib	Ribociclib
T47D (+ve Cntrl)	112	400
OACM5.1	82	497
SKGT4	217	458
OACp4C	149	1870
ESO26	823	6114
OE33	547	1166
ESO51	1132	2567
CP-B	>4000	>4000
CP-C	2518	>4000
OE19	>4000	>4000
MFD	2773	>4000
Flo-1	5574	>8000
CP-D	>8000	>8000
JHesoAD	5668	>8000
MDA-468 (-ve Cntrl)	>8000	>8000

breast cancer where CDK4/6 inhibitors have been FDA approved. We noted three cell lines without sensitising events which were highly resistant, with little drug effect even at 4000 nanomolar concentrations, similar to a known Rb mutant resistant line breast cancer cell line (MDA-MB-468). Two of these three cell lines harbour amplification of CCNE1 which is known to drive resistance to CDK4/6 inhibitors by bypassing CDK4/6 and causing Rb phosphorylation via CDK2 activation<sup>113</sup>. To verify these effects in a more representative model of OAC we treated three whole genome sequenced OAC organoid cultures<sup>96</sup> with Palbociclib and Ribociclib as well as a more recently approved CDK4/6 inhibitor, Abemaciclib. As was observed in cell lines, cell cycle and RTK driver events were present only in the more sensitive organoids and *CCNE1* activation in the most resistant (Figure 24). We found Abemaciclib to be significantly more potent in comparison to both other CDK4/6 inhibitors, both in organoids and cell lines (Figure 25). We note that the maximum tolerated blood doses of Abemaciclib achieved in the clinic were also higher than the other CDK4/6 inhibitors, within the range of sensitivity achieved in several cell lines and organoids cultures.

## Investigation of other targeted therapeutic options for OAC

Although CDK4/6 inhibitors were the most promising drugs given the genetics of OAC, several other drugs were predicted to be effective in reasonable proportions of tumours and could potentially be more potent. CDK4/6is have generally been most effective used in combination with Oestrogen receptor antagonists in breast cancer<sup>114</sup>, although they have also been approved as a monotherapy<sup>115</sup> and hence we sought a possible combination therapy for CDK4/6i in OAC. In a series of small studies, designed and supervised



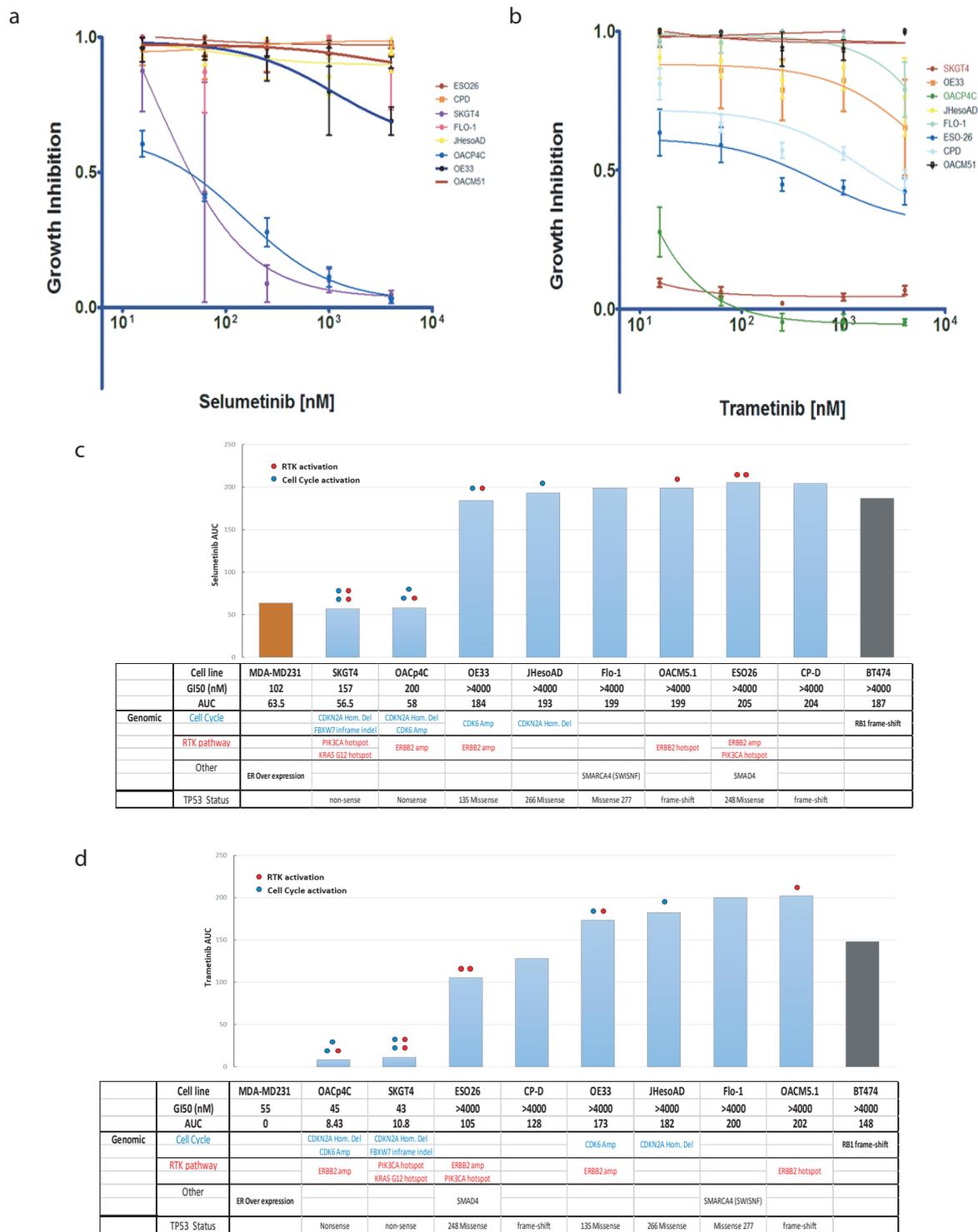
**Figure 25. Growth inhibition responses of 15 OAC, high grade dysplasia BO and control cell lines to CDK4/6 inhibitors Palbociclib, Ribociclib.** A subset of cell lines also recieved treatment with Abemaciclib which shows efficacy in cell lines as well as organiods (fig 6C). Data pionts indicate the mean and error bars the standard error of the mean (SEM) accross technical replicates.

by the PhD candidate and with experiments mostly undertaken by pre-doctoral students in the lab, we investigated the efficacy of some of the other most promising therapeutics *in vitro*.

## MEK inhibitors

The activation of the oestrogen receptor promotes *CCND1* expression and therefore CDK4/6 activity to drive breast cancer cell proliferation<sup>116</sup>. This may explain the high efficacy of CDK4/6is in combination with oestrogen receptor antagonists in breast cancers where CDK4/6 is highly activated. To investigate a similar combinatorial approach in OAC we chose MEK inhibitors which have been FDA approved for various cancer types to inhibit the receptor tyrosine kinase pathway, most effectively for *BRAF* mutant melanoma<sup>117</sup>. The receptor tyrosine kinase pathway, similar to the oestrogen receptor, promotes cell proliferation in part via upregulation of *CCND1* and activation of CDK4/6 and hence may be a similarly effective combination in OAC as with oestrogen receptor antagonists in breast cancer.

To test this hypothesis, we first investigated the effects of two different FDA approved MEK inhibitors (Trametinib and Selumetinib) as a monotherapy on OAC cell lines (Figure 26). We found that two cell lines (OACP4C and SKGT4) were highly sensitive to both MEK inhibitors while all other lines were highly resistant. These two lines contained the highest number of driver events in the cell cycle and receptor tyrosine kinase pathways as compared to other lines, however many other RTK or core cell cycle activated lines were insensitive. Most of these MEK resistant lines (all but OACM5.1) were only moderately



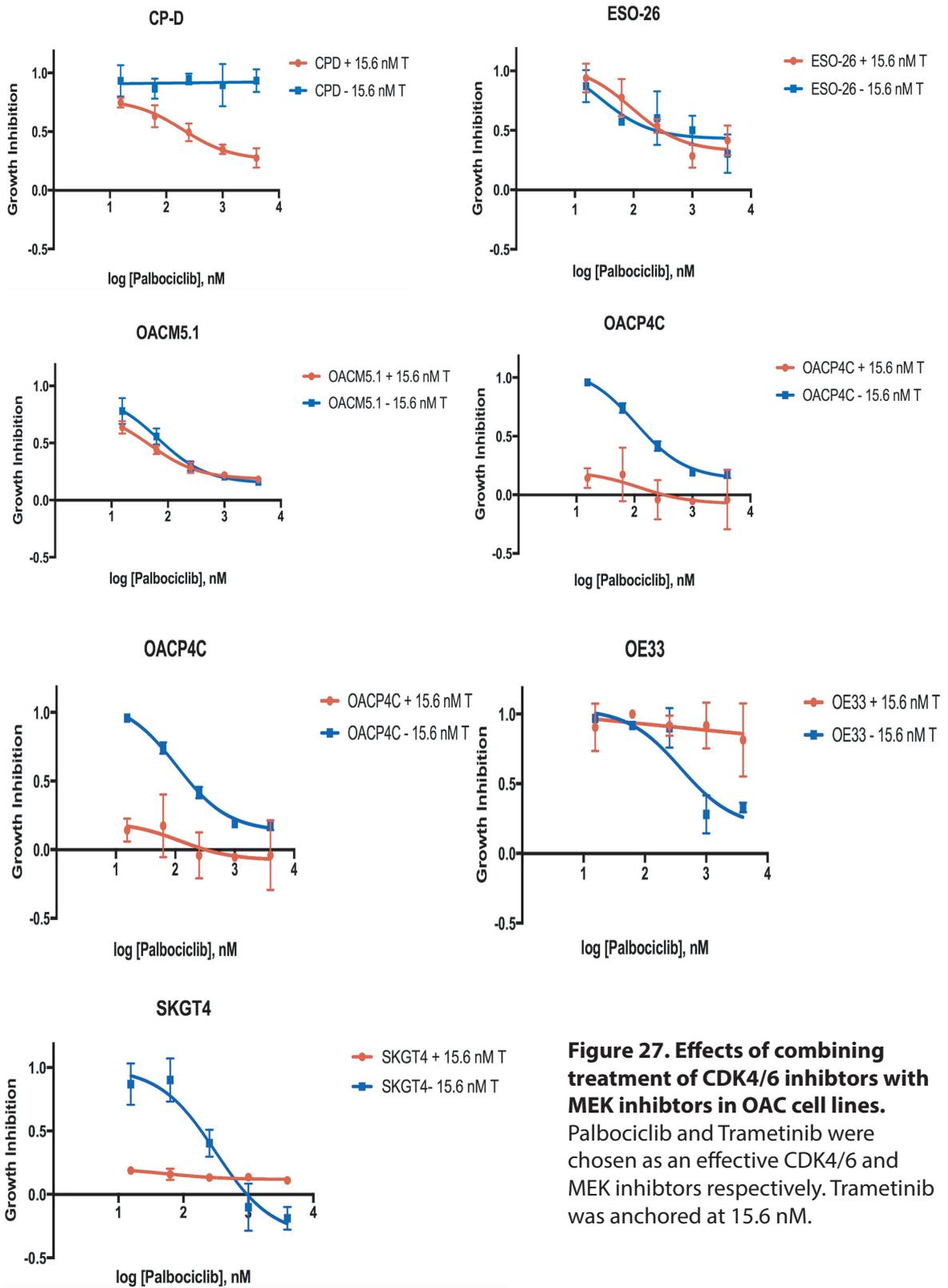
**Figure 26. Efficacy of two MEK inhibitors, Selumetinib and Trametinib, in OAC in vitro OAC models.** a. Growth inhibition across the range of Selumetinib concentrations in various OAC cell lines. b. Growth inhibition across the range of Trametinib concentrations in various OAC cell lines. c. Relationship between genomic drivers and selumetinib response. d. Relationship between genomic drivers and Trametinib response.

sensitive to or completely resistant to CDK4/6 inhibitors. We therefore hoped that by combining the two drugs we might observe synergy and cause a much larger number of cell lines to be highly sensitive in comparison to either drug alone. We chose to anchor our most effective MEK inhibitor, trametinib at a low concentration (15.6 nM) and apply a range of Palcobiciclib, a CDK4/6i, concentrations. However, we did not observe synergy between the two drugs and the CDK4/6i sensitivity of those MEK inhibitors resistant cell lines was not altered by addition of MEK inhibitors (Figure 27).

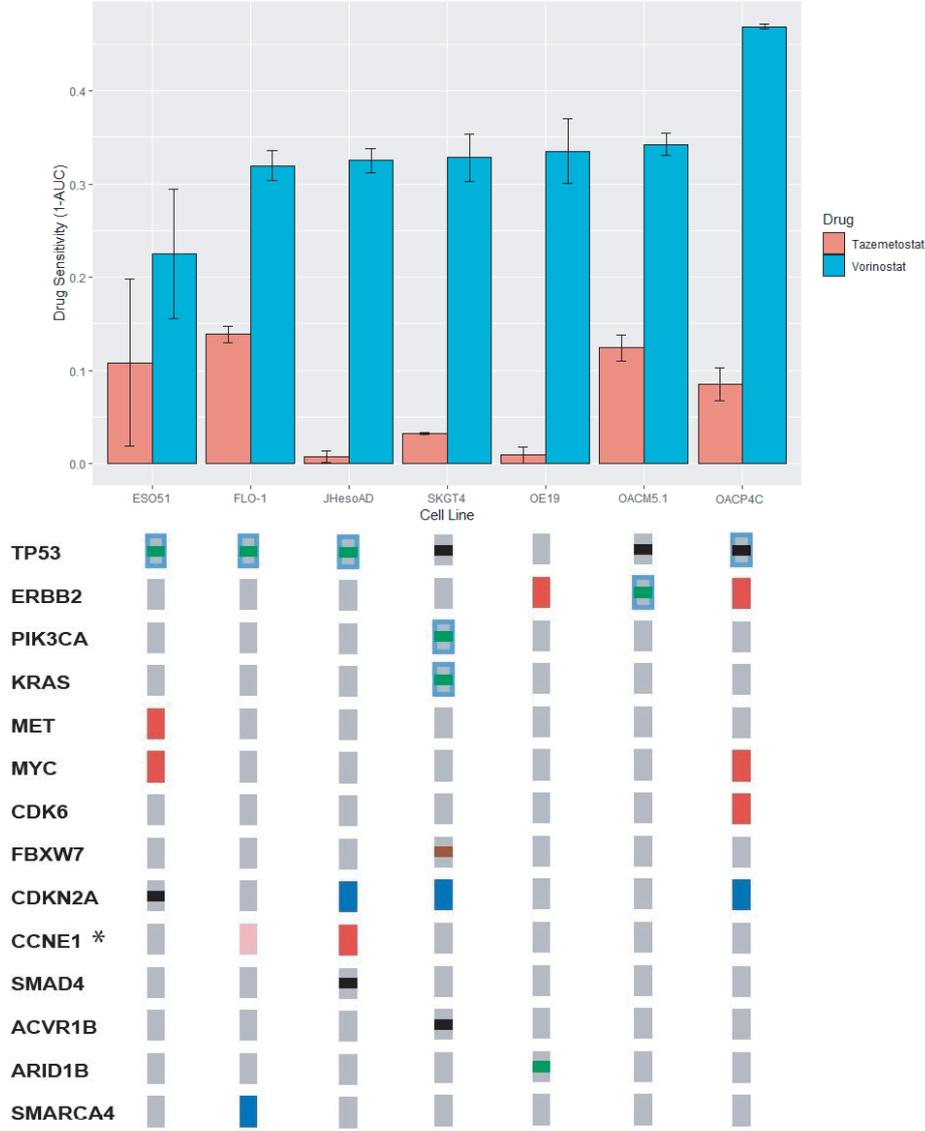
### **EZH2/HDAC inhibitors**

One of the most commonly dysregulated pathways we discovered was the SWI/SNF, chromatin remodelling complex in 28% of OACs. Using the cancer biomarker database we noted that several of these SWI/SNF mutants had been observed sensitive to EZH2 inhibitors in other cancer types. We also noted a recent publication which showed ARID1A mutants were sensitive to treatment by HDAC inhibitors<sup>118</sup>. Both HDAC and EZH2 are proteins involved in chromatin regulation, interacting with the SWI/SNF complex. Hence, we hypothesised in these cases that all SWI/SNF mutants may confer sensitivity.

We profiled sensitivity of seven, whole genome sequenced, OAC cell lines to an EZH2 inhibitor (EPZ-6438) and HDAC inhibitor (SAHA) (Figure 28). OAC cell lines were all resistant to the EZH2 inhibitor at the concentrations applied, however some sensitivity to the HDAC inhibitor was observed. Despite this we did not observe particularly high sensitivity in the SWI/SNF mutated (in SMAR4) line Flo-1. Unfortunately, only a single endogenously SWI/SNF mutant OAC line is known and many other factors may determine the drug response of a



**Figure 27. Effects of combining treatment of CDK4/6 inhibitors with MEK inhibitors in OAC cell lines.** Palbociclib and Trametinib were chosen as an effective CDK4/6 and MEK inhibitors respectively. Trametinib was anchored at 15.6 nM.

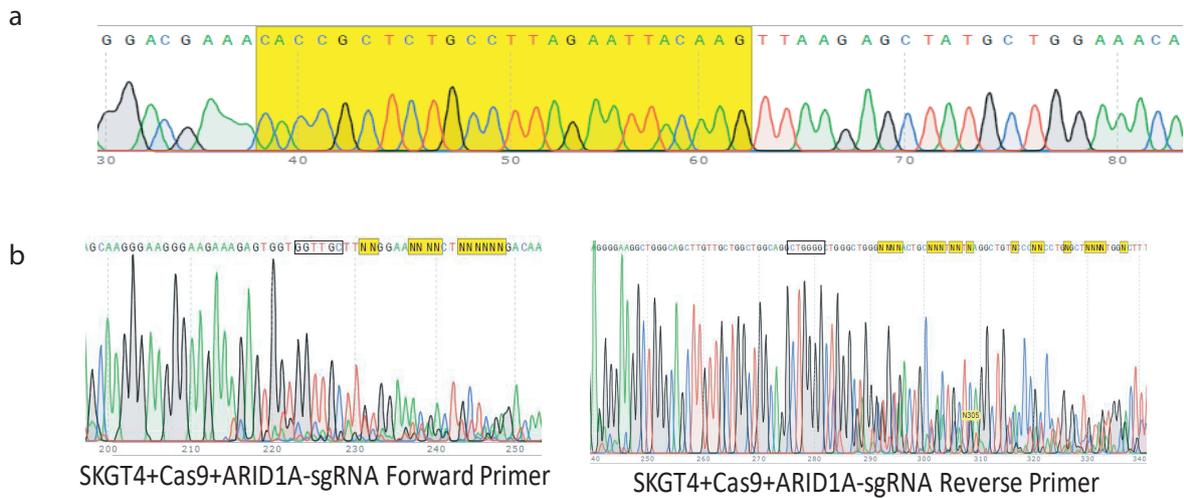


**Figure 28. Drug sensitivity to EZH2 (Vorinostat) and HDAC (Tazemetostat) inhibitors across a panel of OAC cell lines.**

specific cell line, making it difficult in this case to rule out SWI/SNF mutant as possible biomarkers for EZH2i response. To overcome this problem, we engineered *ARID1A* deletions in a SWI/SNF wildtype cell lines, to compare the effect of SWI/SNF mutation on EZH2 sensitivity in the same genetic background. To do this *ARID1A* knock out sgRNAs were designed and cloned into a lenti-viral vector (Figure 29A). This vector was then transduced into a tet-inducible Cas9 expressing OAC cell line, SKGT4 and Cas9 expression was induced for 7 days. Sanger sequencing then identified large deletions in the *ARID1A* gene (Figure 29B), and we observed a loss of *ARID1A* expression in western blots (Figure 29C). Unfortunately, the *ARID1A* mutant OAC cell lines were no more sensitive to EZH2 inhibitors than non-targeting control transduced control lines (Figure 30) confirming that this is not a promising further avenue of investigation for OAC therapeutics.

## **BET inhibitors**

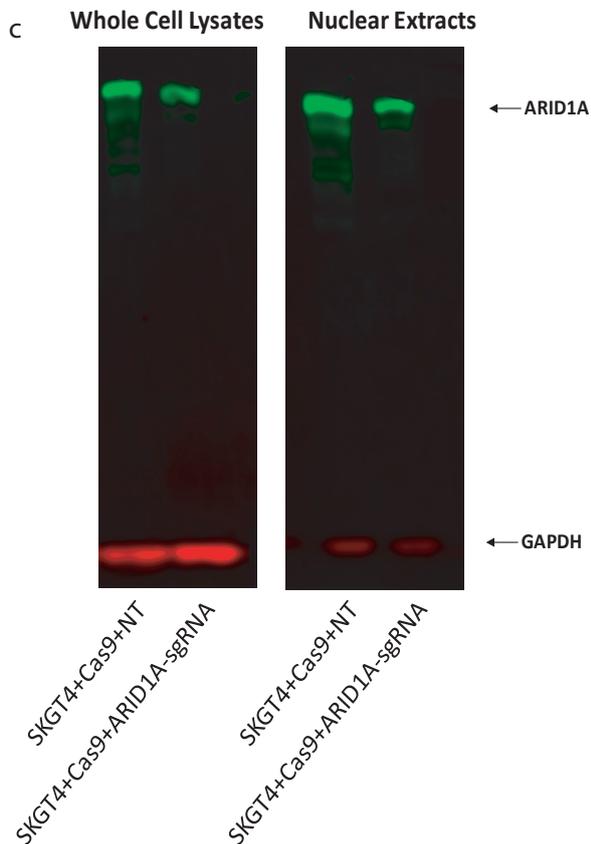
BET (Bromodomain and Extra-Terminal domain) proteins promote the expression of *MYC* and inhibitors of these proteins have been found to be effective against *MYC* amplified cancers. The cancer biomarker database also highlighted studies showing *KRAS* activated cancers were also often sensitive to BET inhibitors. This may be because activation of RTK pathway also leads to *MYC* upregulation. We applied BET inhibitors to seven whole genome sequenced OAC cell lines and found a correlation between sensitivity and *MYC* amplification and *ERBB2* activation, although the *KRAS* mutant line was not particularly sensitive (Figure 31). These results are suggestive of efficacy of BET inhibitors in *MYC* amplified OAC (almost 1/5 of cases) and possibly in other *MYC* over expressed cases which are common in *MYC*



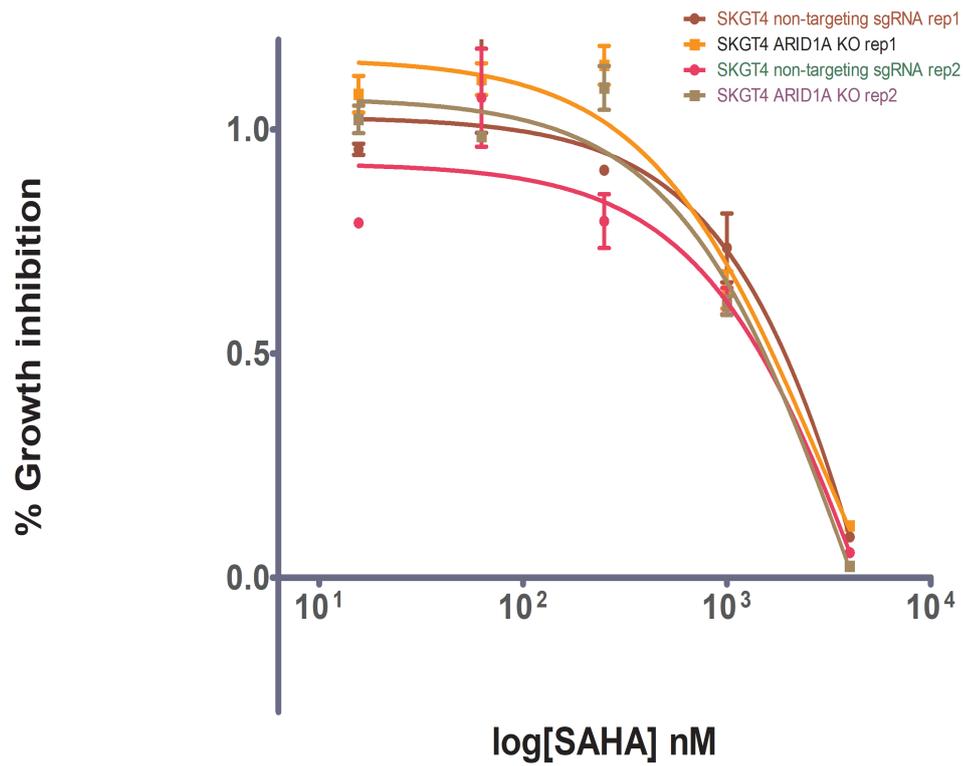
```

TCTCTCACCTGACTGGCCAGTCCTGCCTGAAGAGCCAGTCCTCAATCTCTTCTATTTGGAGTTGGAA
GGGGCTAGAAATAGACCTTATTTTGTATTTTAGGTTTTGTATATTTTCTACTTAAGCAAGGGAAGGAA
GAAAGAGTGGTGGTTGCTTTTGGAAACAACCTCARRAAGACAATTTGTTAAGGTGATCCCATGTTTTCTT
GGAGTCTGTGTCACCAAGCATCTGGTTGTAGCCATCTTGGCATCTGTGGGCTTTATGTCCTGAGTGCA
GAGTATTAACCTCCCCTCTGCTTGTCTCTGCCTTAGAATTACAAGGGCCCAATGGATGGCACATATGGCC
CTCCTGCCAAGCGGCACGAAGGGGAGATGTACAGCGTGCCATACAGCACTGGGCAGGGGCAGCCTCAGCA
GCAGCAGTTGCCCCAGCCAGCCCCAGCCTGCCAGCCAGCAACAAGCTGCCAGCCTTCCCCTCAGCAA
GATGTATAACAACAGTATGGCAATGCCTATCCTGCCACTGCCACAGCTGCTACTGAGCGCCGACCAGCAG
GGGGCCCCAGAACCAATTTCCATTCCAGTTTGGCCGAGACCGTGTCTCTGCACCCCTGGCCACCAATGC

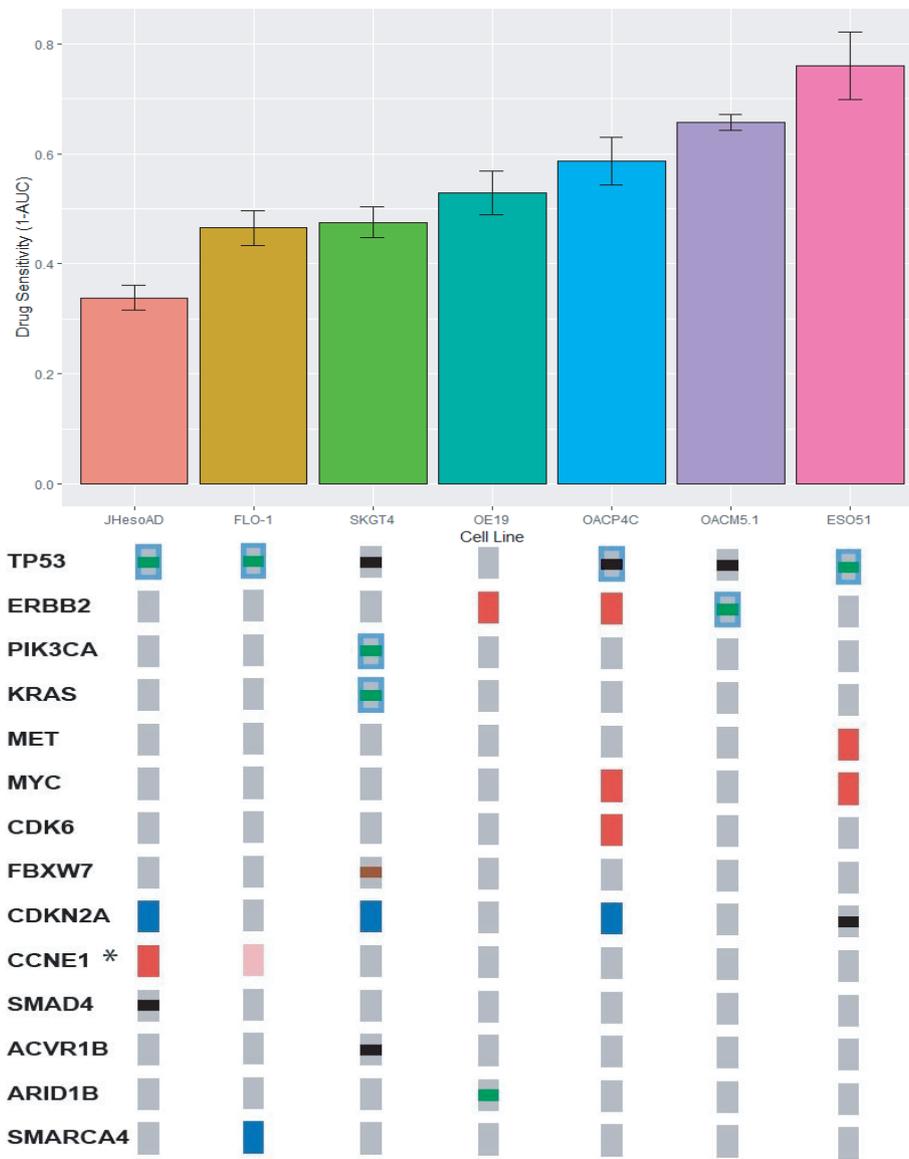
```



**Figure 29. CRISPR-mediated genome engineering of ARID1A deletions into SKGT4.** **a.** Sanger sequencing confirming successful cloning in backbone, sgRNA vector. **b.** Sanger sequencing showing large deletions around the targeted PAM site indicated in red. Regions around the PAM site are highlighted in the ARID1A sequence and the sanger sequencing traces. **c.** Western blots in whole cell and nuclear extracts for ARID1A showing loss of expression in cell lines with ARID1A deletions as compared to controls with non-targeting (NT) sgRNA.



**Figure 30. Sensitivity of SKGT4, an OAC cell line, to SAHA, a HDAC inhibitor with and without CRISPR-mediated ARID1A Knock out.**



**Figure 31. Drug sensitivity to BET inhibitor (JQ-1) across a panel of OAC cell lines.**

wildtype OACs (Figure 14). It is also suggestive that *ERBB2* activated cases may also be sensitive but a larger panel of OAC model would be required to confirm these findings.

# Discussion

## Identification and characterisation of OAC drivers

We present here a detailed catalogue of putative coding and non-coding genomic events that have been selected for during the evolution of oesophageal adenocarcinoma. These events have been characterised in terms of their relative impact, related functions, mutual exclusivity and co-occurrence and expression in comparison to normal tissues, producing insights into OAC biology.

## Limitations to driver gene detection in this study

While clinical annotation and matched RNA data is a strength of this study, in some cases we may have been unable to assess selected variants for survival associations or expression changes which were detected in the full 551 cohort, due to lack of representation in clinically annotated or RNA matched sub cohorts. Despite rigorous analyses to detect selected events, assessment of the global excess of mutations by dNdScv suggests we are unable to detect all events selected in OAC, similar to many other cancer types. All driver gene detection methods which we have used ultimately rely on driver mutation re-occurrence within some specific genomic region or mutation class. Many of these undetected driver mutations are hence likely to be spread across a large number of genes<sup>14</sup>

whereby each is mutated at low frequency across OAC patients. This tendency for low frequency OAC drivers may be responsible for the low yield of MutsigCV in previous cohorts and may suggest that C-type cancers such as OAC, are not less 'mutation-driven' than M-type cancers but rather that their mutational drivers are spread across a larger number of genes. This is in fact apparent from the identified list of drivers where only 4/66 are predicted to contain driver variants in >5% of cases (*i.e.* after the predicted percentage of passengers has been removed in each gene, *TP53*, *ARID1A*, *SMAD4* and *CDKN2A*), in significant contrast to other cancer types<sup>4</sup>. Many of these low frequency drivers are well evidenced however with highly significant q-values and identification by multiple driver detection tools (Table 2). The identification of extreme very low frequency mutations, perhaps unique in 1000s of tumours, will require substantially different detection techniques to those which are currently in wide spread use and such methods are in development<sup>119</sup> although they require validation. Undoubtedly many copy number drivers are also left undiscovered and validation of candidates identified here is an important avenue of future work.

In addition to false negative drivers it is likely that, despite our best endeavours, some identified drivers in this study may also be false positives. This is inevitable given an FDR cut off of 0.1 was used, hence there is a 1 in 10 probability that drivers identified just below this significance threshold are false positives. This is also amplified by our use of many different driver detection tools each with FDR thresholds of 0.1. However, it should be noted that only a small minority of drivers have an FDR close to this threshold. In addition to these false positives there may also be instances where, although a gene truly is mutated more commonly than we would expect given our model, our model is in fact incomplete and does not account for an as yet unknown mutational process or informatic artefact which would explain these additional mutations. This is particularly likely to occur when considering non-

coding elements in which the background mutation rate becomes more difficult to model due to difficulties already discussed (see Introduction, page 20). Specifically, the PCWAG consortium has noted several artefacts that make specific non-coding elements suspicious. For instance, some non-coding elements can be subject to mapping artefacts, where due to high homology, reads derived from a different genomic region map repeatedly to a specific, incorrect site in the genome and small differences in sequence between the two sites are detected as SNVs. This is particularly common for elements frequently repeated in the genome. This is clearly identified in some instances where precisely the same mutations apparently arise in both cancer and in normal genomes. Occasionally higher than expected mutation rates can also be found over large genomic regions, rather than in specific non-coding elements, suggesting our background mutation rate model is failing in these regions.

### Comparing detected drivers to those known in the literature

Previous reports on OAC drivers have had a limited yield per case. The first such study<sup>53</sup> used methods that, despite being well regarded at the time, were subsequently discredited<sup>15</sup>. Since then, several reports, including our own, using MutSigCV<sup>54,55,57</sup> on medium and large cohort sizes, have detected only a small number of mutational driver genes (7, 5 and 15 in each study, respectively). By using both a large cohort and more comprehensive methodologies, we markedly increased this figure to 66 mutational driver genes (excluding copy number drivers). This includes all 12 previously detected OAC drivers that are known drivers in other cancer types and 3/10 previously detected OAC-unique drivers (*KCNQ3*, *EPHA3* and *CCDC102B*). All those genes identified in other OAC studies but

not ours are mutated at a low frequency hence this discrepancy may be due to sampling, however it may also represent real biological differences between different OAC cohorts. Such differences are apparent in the *ERBB2* mutation rate which is significantly higher in the TCGA<sup>55</sup> (13%) as compared to this ICGC cohort (3%) which also contains a lower rate of *ERBB2* amplification (15%) in comparison to the TCGA (28%). It may also be due to different cut offs used to remove unexpressed genes, a standard step to remove false positives, which is implemented in different ways and was not implemented in Secrier *et. al.*<sup>54</sup> for instance. Table 2 indicates all identified genes and which are novel in OAC.

Detection of driver CNAs has previously relied on GISTIC to detect recurrently mutated regions<sup>52,54,55,60</sup> but no analyses have been performed to determine which genes in these large regions are true drivers. Many of the genes annotated by such papers are unlikely to be CNA-only drivers from this analysis due to; their lack of expression modulation with CNAs (*e.g. YEATS4* and *MCL1*), the role of recurrent heterozygous losses to drive LOH in some mutational drivers (*ARID1A* and *CDH11*), which has not previously been noted in GISTIC loci to our knowledge, or their association with fragile sites (*PDE4D*, *WWOX*, *FHIT*). Conversely, we have been able to identify novel OAC copy number drivers (*e.g. CCND3*, *AXIN1*, *PPM1D* and *APC*). There still remains a very large number of recurrently amplified or deleted regions of the genome identified by GISTIC that lack an obvious driver event (133/149 without high confidence drivers). As has been described many of these recurrently amplified regions contain relatively low-level copy number gains in comparison to those which contain high confidence drivers, and some seem to be mostly driven by copy number gains too small to be even considered as “amplifications” by our definition (>2x Ploidy). It seems likely that at least many of these may be caused by random mutational processes, yet to be understood, rather than being under selection. Similarly, deletion peaks are also generally dominated by

heterozygous deletions excluding fragile site regions and there are only a few regions which contain clear deletion driver genes (*CDKN2A* and *SMAD4* for instance). The functional effect of these heterozygous deletions is also unknown. Many heterozygous deletions do correlate with loss of expression however and a prominent recent example of heterozygous losses at the HLA locus appears to have been selected for in parallel across different clones within various tumours<sup>120</sup>. However, given the high frequency of such heterozygous deletion dominated GISTIC loci it seems likely that many are also the product of random mutational processes rather than selection. It will be important to better understand the mutational processes that underlie rates of amplification and deletion across the genome so as to better reveal low level or infrequent CNA drivers. Much progress has been made on this in the past decade with regard to mutational drivers<sup>4</sup> but most of the field still relies on GISTIC2.0, released in 2011 for detection of CNA drivers.

## Opportunities for further investigations

A number of discoveries made in this work require further investigation. Functional characterisation of many of the driver genes described is needed to understand why they are advantageous to OAC tumours and how they modify OAC biology. Particularly interesting are the GI specific genes *GATA4*, *GATA6* and *MUC6* of which *GATA4* modulates prognosis and all have expression loss during the transition from normal to tumour tissue. Biological pathways and processes that are selectively dysregulated deserve particular attention in this regard as do the gene pairs or groups with mutually exclusive or co-occurring relationships such as *MYC* and *TP53* or SWI/SNF factors, suggestive of particular functional relationships. We have

also noted known cancer-associated mutational hotspots, likely to be true driver mutations, some of which lie outside the 66 mutational driver genes identified with a significantly high mutation rate than expected. These mutations lie in genes; *ATM*, *SF3B1*, *WT1*, *ERCC2*, *PIK3R1*, *ERBB3*, *GNAS* and *IL7R*. Such genes, despite being mutated at very a low frequency can inform us about OAC biology, and may also warrant functional investigation, particularly if involved in wide OAC-associated pathways such as *ATM* (DNA damage) or *ERBB3* and *PIK3R1* (Receptor tyrosine kinase).

We noted a three-way association among hypermutation, Wnt activation and loss of immune-signaling genes such as *B2M*. Microsatellite-instability-driven hypermutation has been associated with higher immune activity<sup>74,121</sup>. However, Wnt dysregulation and mutation of immune-pathway genes such as *B2M*<sup>102</sup> have been linked to immunological escape<sup>122</sup>, thus suggesting that this may be an acquired mechanism to prevent immune surveillance caused by hypermutation, requiring further investigation.

We found ultra-high amplifications to be prevalent in OAC, with a high specificity for driver gene containing recurrently amplified regions of the genome. We also found that many recurrently amplified regions without known drivers were dominated by low-level copy number gains with many regions never reaching 2x ploidy, our definition of amplification, in any one case. These observations suggest integration of the absolute levels of copy number amplification is of use, in addition to recurrence of events, in defining regions of the genome amplified by selection-based mechanisms, rather than random processes. We also found that these high-level amplification also had features of double minute chromosomes, as has previously been described in OAC<sup>81</sup>. It is possible these provide a distinct mechanism for selective amplification given the unlinking of such DNA from the mitotic apparatus. This could allow random segregation of double minutes between cells at cell division, producing a large

amount of variability in copy number in cell of a tumour. This could allow selection to act, continually increasing the copy number of oncogenic double minutes to significantly greater levels than would otherwise be likely to occur by chance. This effect could be studied *in vitro* by single cell cloning a cell line containing such a double minute event and then measuring the variability in double minute copy number across the resulting population. An additional question of interest would be how these double minutes are replicated in division given they are unlikely to contain replication origins.

In summary this thesis provides a detailed compendium of mutations and copy number alterations undergoing selection in OAC which have functional impact on tumour behaviour. This comprehensive study provides us with useful insights into the nature of OAC tumours.

## Using novel OAC drivers and clinical biomarkers

### Genotype-clinical phenotype correlations

Poor prognostic implications of *SMAD4* mutation have been noted in several other cancer types from the GI tract where *SMAD4* mutations are common<sup>123,124</sup>. This has been linked to what is known as the “TGF- $\beta$  switch”, a phenomenon noted in mouse models<sup>125</sup>, whereby early stage tumours are growth inhibited by TGF- $\beta$  pathway activation however in later stage tumours TGF- $\beta$  pathway activation causes growth and a more aggressive phenotype. This dual nature of the pathway is most commonly explained by opposing oncogenic roles of canonical and non-canonical TGF- $\beta$  signalling. Canonical TGF- $\beta$  signalling, also known as the

SMAD pathway in which *SMAD4* and several other tumour suppressors sit, promotes *CDKN2A* and inhibits *MYC* expression leading to cytostasis. In contrast Non-canonical signalling via non-SMAD proteins can promote invasion, EMT and resistance to apoptosis<sup>126,127</sup>. A tumour promoting role for TGF- $\beta$  signalling has also been noted in the stroma<sup>128-130</sup>. *SMAD4* mutation occurs late in the progression of CRC and OAC and correlates with the invasive transition<sup>62,131</sup>, hence it is thought that *SMAD4* mutations, and possibly other mechanisms to dampen the effects of the SMAD pathway, cause the TGF- $\beta$  switch by inhibiting the tumour suppressive nature of the pathway and allowing the oncogenic components to continue. Consistent with this, many tumours secrete TGF- $\beta$  ligands in an autocrine fashion. This also explains why TGF- $\beta$  receptor mutations lead to a more promising prognosis as this mechanism of SMAD-pathway inhibition also inhibits the non-canonical, oncogenic components of the pathway.

Much less is known about the role in tumour biology of *GATA4* amplification, the second poor prognostic indicator we found in our study. Both *GATA4* and *GATA6*, another amplified driver gene in our study, are two of three zinc finger transcription factors involved in the early embryogenesis of the gut, the other being *GATA5*, and they are particularly expressed in the upper gastrointestinal tract as well as in Barrett's Oesophagus<sup>132</sup>. In fact, *GATA4* is thought to be the first transcription factor to specify endoderm in embryogenesis. As well as promoting differentiation these GATA factor also has a role in tissue regeneration and mucin secretion. Oncogenic pathways stimulated by specifically *GATA4* not *GATA6* must promote this aggressive phenotype, given *GATA6* amplification showed no correlation with prognosis (Figure 19).

It is surprising, given the presumably important role of genetic drivers in carcinogenesis, and hence tumour aggressiveness, and the large sample size we have

accrued, that we did not detect significant correlations with survival with other genetic drivers. Several factors may have diluted our statistical power in this regard. Firstly there is heterogeneity in treatment and clinical course for our cohort (Table 1) which will modulate patient survival and confound the effects of innate tumour aggressiveness. Secondly many of the genetic drivers in OAC are found at a low frequency and hence although our cohort size is large (379 with high quality clinical data) some of the genes assessed were still found altered in less than 20 cases. Thirdly to maximise our sensitivity for clinically useful biomarkers we chose not to assess many of the very low frequency drivers (those with driving alterations in <5% of cases) because by assessing too many genes we would increase our false discovery rate. In these low frequency genes will would have a poor power to detect prognostic biomarkers and even if these were detected they would be of limited clinical use because of their low frequency of occurrence in OAC. Lastly, recent reports have noted that *NOTCH1* mutations, although under strong selection during the evolution of oesophageal squamous cell carcinoma (OSCC), are actually even more frequently abundant in histologically normal oesophageal squamous mucosa of apparently healthy individuals, particularly in older generations such that develop OSCC<sup>133</sup>. The mechanism of this selection is well understood<sup>134</sup> where *NOTCH1* mutation allows for a differentiation imbalance in stem cells towards renewal, but does not directly promote proliferation or other hallmarks of cancer. Therefore, selection for *NOTCH1* mutation does occur in the clonal ancestors of OSCCs but only in normal oesophageal mucosa and this does not necessarily indicate it has a role in driving OSCCs. In fact, the rate of *NOTCH1* mutation in OSCCs is actually significantly lower than in normal oesophageal mucosa, suggesting it could even have a preventative effect. Alternative explanations to this phenomenon are also plausible. It is thus possible that highly frequent genes under selection in OAC have a role in non-neoplastic precursors, but not OAC, and

hence do not modulate prognosis. *CDKN2A* is commonly deleted or mutations in approximately 1/3 of OACs making it the second most common driver in OAC, however has the least indication of any correlation with prognosis of all the genes which we have assessed (Figure 19, HR = 1.00,  $P = 0.93$ ). It is also known to occur very commonly in non-dysplastic Barrett's oesophagus and it is also very commonly found in cases that never-progress to cancer, even after long follow up<sup>58</sup>. Formal assessments of *CDKN2A* mutation selection in Barret's in comparison to OACs have not been yet been made.

It is also of note that we were unable to find many correlations between clinical factors, for example with chemotherapy treatment, and genetic drivers. Clinical heterogeneity as has been discussed makes these comparisons difficult and the lack of correlation with chemotherapy treatment, may be due to the very low rate of chemotherapy response in our cancer type and is consistent with our previous reports<sup>56</sup>.

## Targeted therapeutics for OAC

While we observed significant genotype specific sensitivity in our OAC in vitro models to CDK4/6 inhibitors, it has been observed in some cancer types that *TP53* mutations correlate with resistance to these agents<sup>135</sup>. Almost all of our cell line models are *TP53* mutant apart from one line (OE19) which was relatively resistant despite activating events in receptor tyrosine kinase and core cell cycle pathways, hence, due to the available models, we cannot properly assess the effect of *TP53* mutation on drug sensitivity. Our data suggests that most cell lines without activating events in RTK or cell cycle pathways are resistant, perhaps due to *TP53* mutations, but only with these activating events in these pathways is this basal

resistance overcome to achieve sensitivity. However, the numbers of available models makes this difficult to assess.

In vivo work with CDK4/6i has been previously published using orthotopic Xenograft models with three OAC cell lines<sup>136</sup>, which showed efficacy of these drugs in vivo, however they did not use a large enough panel of models or genotyping to assess the genomic biomarkers we discover in this work. This in vivo evidence allows us to consider moving these inhibitors straight into human clinical trials and we are incorporating these inhibitors into an adaptive, multi-arm clinical trial, Oelixir, which will whole genome sequence OAC cases across the UK and match them to appropriate targeted therapeutics.

## **Future directions**

While OAC is a poor prognosis cancer type, significant heterogeneity of survival outcome makes triaging patients in treatment groups an important part of clinic practice which could be improved using better prognostication. Prospective clinical work to verify and implement SMAD4 and GATA4 biomarkers in this study would be the next stage to move these markers into routine practice. Whole genome or whole exome sequencing may be impractical for use in the clinic, however targeted NGS panels to detect mutations and copy number alterations have been implemented to detect genomic biomarkers in a cost effective and sensitive manner for some cancer types. In OAC development of a customised panel is likely to be required on the basis of this analysis.

A number of targeted therapeutics may provide clinic benefit to OAC cases based on their individual genomic profile. In particular CDK4/6 inhibitors deserve considerable attention as an option for OAC treatment as they are, by a significant margin, the treatment to which the most OACs harbour sensitivity-causing driver events, excluding *TP53* as an unlikely therapeutic biomarker. The in vitro validation of these biomarkers for CDK4/6 inhibitors in OAC is also persuasive of possible clinical benefit using a targeted approach. The next phase of the OCCAMS/ICGC project, the Oelixir project, is currently in the planning stage. In Oelixir, oesophageal adenocarcinomas diagnosed across the UK will be recruited at diagnosis and whole genome sequenced, irrespective of pathological cellularity, to 70X as part of the ICGC-ARGO project. The cases will then be matched to various clinical trials based on the genomics of their specific tumour. Among other compounds, we will be including CDK4/6 inhibitors in one of these arms and we are in discussions with partners in the pharmaceutical industry to enable this. We will also be analysing circulating tumour DNA and variety of other aspects of each tumour to build a databank for research purposes along with their WGS and clinical information. We will also use the Oelixir trail as a platform to prospectively assess our prognostic biomarkers (*SMAD4* mutation or deletion and *GATA4* amplification) for validation of their clinical utility.

# Bibliography

1. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
3. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13**, 795–806 (2012).
4. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
5. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
6. Tarabichi, M. *et al.* Neutral tumor evolution? *Nat. Genet.* **50**, 1630–1633 (2018).
7. McDonald, T. O., Chakrabarti, S. & Michor, F. Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution. *Nat. Genet.* **50**, 1620–1623 (2018).
8. Balaparya, A. & De, S. Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data. *Nat. Genet.* **50**, 1626–1628 (2018).
9. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
10. Turajlic, S. *et al.* Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell* **173**, 595–610.e11 (2018).
11. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *bioRxiv* 161562 (2018).  
doi:10.1101/161562

12. Mitchell, T. J. *et al.* Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. *Cell* **173**, 611–623.e17 (2018).
13. Sabarinathan, R. *et al.* The whole-genome panorama of cancer drivers. *bioRxiv* 190330 (2017). doi:10.1101/190330
14. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
15. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
16. Shuai, S., Gallinger, S., Stein, L. D., Group, P. D. and F. I. & Net, I. P.-C. A. of W. G. DriverPower: Combined burden and functional impact tests for cancer driver discovery. *bioRxiv* 215244 (2017). doi:10.1101/215244
17. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
18. Porta-Pardo, E. *et al.* Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nat. Methods* **14**, 782–788 (2017).
19. Rheinbay, E. *et al.* Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. *bioRxiv* 237313 (2017). doi:10.1101/237313
20. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018).
21. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).

22. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
23. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10 11 (2008).
24. Raine, K. M. *et al.* ascatNgs: Identifying Somatic Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr. Protoc. Bioinforma.* **56**, 15.9.1-15.9.17 (2016).
25. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
26. Wala, J. A. *et al.* Selective and mechanistic sources of recurrent rearrangements across the cancer genome. *bioRxiv* 187609 (2017). doi:10.1101/187609
27. Vaughan, T. L. & Fitzgerald, R. C. Precision prevention of oesophageal adenocarcinoma. *Nat. Rev. Gastroenterol. Hepatol.* **12**, 243–248 (2015).
28. Bhat, S. *et al.* Risk of Malignant Progression in Barrett’s Esophagus Patients: Results from a Large Population-Based Study. *JNCI J. Natl. Cancer Inst.* **103**, 1049–1057 (2011).
29. Wong, A. & Fitzgerald, R. C. Epidemiologic risk factors for Barrett’s esophagus and associated adenocarcinoma. *Clin. Gastroenterol. Hepatol.* **3**, 1–10 (2005).
30. Haggitt, R. C. Barrett’s esophagus, dysplasia, and adenocarcinoma. *Hum. Pathol.* **25**, 982–93 (1994).
31. Jiang, M. *et al.* Transitional basal cells at the squamous–columnar junction generate Barrett’s oesophagus. *Nature* **550**, 529–533 (2017).
32. Nicholson, A. M. *et al.* Barrett’s metaplasia glands are clonal, contain multiple stem cells and share a common squamous progenitor. doi:10.1136/gutjnl-2011-301174
33. Shaheen, N. J. *et al.* Durability of Radiofrequency Ablation in Barrett’s Esophagus With

- Dysplasia. *Gastroenterology* **141**, 460–468 (2011).
34. Shaheen, N. J. *et al.* Radiofrequency Ablation in Barrett's Esophagus with Dysplasia. *N. Engl. J. Med.* **360**, 2277–2288 (2009).
  35. Zhang, Y. Epidemiology of esophageal cancer. *World J. Gastroenterol.* **19**, 5598 (2013).
  36. Hayeck, T. J., Kong, C. Y., Spechler, S. J., Gazelle, G. S. & Hur, C. Original article: The prevalence of Barrett's esophagus in the US: estimates from a simulation model confirmed by SEER data. *Dis. Esophagus* **23**, 451–457 (2010).
  37. Pereira, A. D. & Chaves, P. Low risk of adenocarcinoma and high-grade dysplasia in patients with non-dysplastic Barrett's esophagus: Results from a cohort from a country with low esophageal adenocarcinoma incidence. *United Eur. Gastroenterol. J.* **4**, 343–352 (2016).
  38. Offman, J. *et al.* Barrett's oESophagus trial 3 (BEST3): study protocol for a randomised controlled trial comparing the Cytosponge-TFF3 test with usual care to facilitate the diagnosis of oesophageal pre-cancer in primary care patients with chronic acid reflux. *BMC Cancer* **18**, 784 (2018).
  39. Ross-Innes, C. S. *et al.* Risk stratification of Barrett's oesophagus using a non-endoscopic sampling method coupled with a biomarker panel: a cohort study. *Lancet Gastroenterol. Hepatol.* **2**, 23–31 (2017).
  40. Reid, B. J., Haggitt, R. C., Rubin, C. E. & Rabinovitch, P. S. Barrett's esophagus: Correlation between flow cytometry and histology in detection of patients at risk for adenocarcinoma. *Gastroenterology* **93**, 1–11 (1987).
  41. Blount, P. L. *et al.* 17p allelic deletions and p53 protein overexpression in Barrett's adenocarcinoma. *Cancer Res.* **51**, 5482–6 (1991).
  42. Blount, P. L. *et al.* Clonal ordering of 17p and 5q allelic losses in Barrett dysplasia and

- adenocarcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 3221–5 (1993).
43. Al-Kasspoles, M., Moore, J. H., Orringer, M. B. & Beer, D. G. Amplification and over-expression of the EGFR anderbB-2 genes in human esophageal adenocarcinomas. *Int. J. Cancer* **54**, 213–219 (1993).
  44. Neshat, K. *et al.* p53 mutations in Barrett's adenocarcinoma and high-grade dysplasia. *Gastroenterology* **106**, 1589–95 (1994).
  45. Krishnadath, K. K. *et al.* Accumulation of genetic abnormalities during neoplastic progression in Barrett's esophagus. *Cancer Res.* **55**, 1971–6 (1995).
  46. Barrett, M. T. *et al.* Allelic loss of 9p21 and mutation of the CDKN2/p16 gene develop as early lesions during neoplastic progression in Barrett's esophagus. *Oncogene* **13**, 1867–73 (1996).
  47. Barrett, M. T., Schutte, M., Kern, S. E. & Reid, B. J. Allelic loss and mutational analysis of the DPC4 gene in esophageal adenocarcinoma. *Cancer Res.* **56**, 4351–3 (1996).
  48. Phillips, W. A. *et al.* Mutation analysis of PIK3CA and PIK3CB in esophageal cancer and Barrett's esophagus. *Int. J. Cancer* **118**, 2644–2646 (2006).
  49. Trautmann, B. *et al.* K-ras point mutations are rare events in premalignant forms of Barrett's oesophagus. *Eur. J. Gastroenterol. Hepatol.* **8**, 799–804 (1996).
  50. González, M. V *et al.* Mutation analysis of the p53, APC, and p16 genes in the Barrett's oesophagus, dysplasia, and adenocarcinoma. *J. Clin. Pathol.* **50**, 212–7 (1997).
  51. Agrawal, N. *et al.* Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discov.* **2**, 899–905 (2012).
  52. Dulak, A. M. *et al.* Gastrointestinal Adenocarcinomas of the Esophagus, Stomach, and Colon Exhibit Distinct Patterns of Genome Instability and Oncogenesis. *Cancer Res.* **72**, 4383–4393 (2012).

53. Dulak, A. M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478–486 (2013).
54. Secrier, M. *et al.* Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.* **48**, 1131–1141 (2016).
55. Network, T. C. G. A. R. Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175 (2017).
56. Noorani, A. *et al.* A comparative analysis of whole genome sequencing of esophageal adenocarcinoma pre- and post-chemotherapy. *Genome Res.* **27**, 902–912 (2017).
57. Lin, D.-C. *et al.* Identification of distinct mutational patterns and new driver genes in oesophageal squamous cell carcinomas and adenocarcinomas. *Gut* **67**, 1769–1779 (2018).
58. Li, X. *et al.* Temporal and Spatial Evolution of Somatic Chromosomal Alterations: A Case-Cohort Study of Barrett’s Esophagus. *Cancer Prev. Res.* **7**, 114–127 (2014).
59. Stachler, M. D. *et al.* Paired exome analysis of Barrett’s esophagus and adenocarcinoma. *Nat. Genet.* **47**, 1047–1055 (2015).
60. Frankel, A. *et al.* Genome-wide analysis of esophageal adenocarcinoma yields specific copy number aberrations that correlate with prognosis. *Genes, Chromosom. Cancer* **53**, 324–338 (2014).
61. Vousden, K. H. & Prives, C. Blinded by the Light: The Growing Complexity of p53. *Cell* **137**, 413–431 (2009).
62. Weaver, J. M. J. *et al.* Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat. Genet.* **46**, 837–843 (2014).

63. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
64. Martinez, P. *et al.* Dynamic clonal equilibrium and predetermined cancer risk in Barrett's oesophagus. *Nat. Commun.* **7**, 12158 (2016).
65. Martinez, P. *et al.* Evolution of Barrett's esophagus through space and time at single-crypt and whole-biopsy levels. *Nat. Commun.* **9**, 794 (2018).
66. Lordick, F. *et al.* Oesophageal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up † on behalf of the ESMO. (2016).  
doi:10.1093/annonc/mdw329
67. Nguyen, P. L. *et al.* Breast Cancer Subtype Approximated by Estrogen Receptor, Progesterone Receptor, and HER-2 Is Associated With Local and Distant Recurrence After Breast-Conserving Therapy. *J. Clin. Oncol.* **26**, 2373–2378 (2008).
68. Harvey, J. M., Clark, G. M., Osborne, C. K. & Allred, D. C. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *J. Clin. Oncol.* **17**, 1474–81 (1999).
69. Rice, T. W., Patil, D. T. & Blackstone, E. H. 8th edition AJCC/UICC staging of cancers of the esophagus and esophagogastric junction: application to clinical practice. *Ann. Cardiothorac. Surg.* **6**, 119–130 (2017).
70. McCormick Matthews, L. H. *et al.* Systematic review and meta-analysis of immunohistochemical prognostic biomarkers in resected oesophageal adenocarcinoma. *Br. J. Cancer* **113**, 107–18 (2015).
71. Blayney, J. K. *et al.* Glucose transporter 1 expression as a marker of prognosis in oesophageal adenocarcinoma. *Oncotarget* **9**, 18518–18528 (2018).
72. Elimova, E. *et al.* 18-fluorodeoxy-glucose positron emission computed tomography as

- predictive of response after chemoradiation in oesophageal cancer patients. *Eur. J. Cancer* **51**, 2545–52 (2015).
73. Bang, Y.-J. *et al.* 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. *Lancet* **376**, 687–697 (2010).
74. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
75. Muro, K. *et al.* Pembrolizumab for patients with PD-L1-positive advanced gastric cancer (KEYNOTE-012): a multicentre, open-label, phase 1b trial. *Lancet. Oncol.* **17**, 717–726 (2016).
76. Fuchs, C. S. *et al.* Adjuvant Chemoradiotherapy With Epirubicin, Cisplatin, and Fluorouracil Compared With Adjuvant Chemoradiotherapy With Fluorouracil and Leucovorin After Curative Resection of Gastric Cancer: Results From CALGB 80101 (Alliance). *J. Clin. Oncol.* **35**, 3671–3677 (2017).
77. Janjigian, Y. Y. *et al.* Genetic Predictors of Response to Systemic Therapy in Esophagogastric Cancer. *Cancer Discov.* **8**, 49–58 (2018).
78. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
79. Lee, A. Y. *et al.* Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biol.* **19**, 188 (2018).
80. Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).

81. Nones, K. *et al.* Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nat. Commun.* **5**, 5224 (2014).
82. Ohkohchi, N. *et al.* Differential expression profiles of sense and antisense transcripts between HCV-associated hepatocellular carcinoma and corresponding non-cancerous liver tissue. *Int. J. Oncol.* **40**, 1813–20 (2012).
83. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
84. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**, 3109–3114 (2014).
85. Wadi, L. *et al.* Candidate cancer driver mutations in super-enhancers and long-range chromatin interaction networks. *bioRxiv* 236802 (2017). doi:10.1101/236802
86. Porta-Pardo, E., Hrabe, T. & Godzik, A. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res.* **43**, D968–D973 (2015).
87. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* **40**, e169 (2012).
88. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7.20.1-7.20.41 (2013).
89. Ng, P. C. & Henikoff, S. Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80 (2006).
90. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
91. Reimand, J., Wagih, O. & Bader, G. D. The mutational landscape of phosphorylation

- signaling in cancer. *Sci. Rep.* **3**, 2651 (2013).
92. Whelan, R. *et al.* Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature* **512**, 185–189 (2014).
  93. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
  94. Northcott, P. A. *et al.* The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).
  95. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
  96. Li, X. *et al.* Organoid cultures recapitulate esophageal adenocarcinoma heterogeneity providing a model for clonality studies and precision therapeutics. *Nat. Commun.* **9**, 2983 (2018).
  97. Ferlay, J. *et al.* Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *Eur. J. Cancer* **103**, 356–387 (2018).
  98. Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol* **34**, 155–163 (2016).
  99. Matthew Bailey, A. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations Article Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385 (2018).
  100. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
  101. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).
  102. Zaretsky, J. M. *et al.* Mutations Associated with Acquired Resistance to PD-1 Blockade

- in Melanoma. *N Engl J Med* **375**, 819–829 (2016).
103. Chen, Z. *et al.* Mammalian drug efflux transporters of the ATP binding cassette (ABC) family in multidrug resistance: A review of the past decade. *Cancer Lett.* **370**, 153–164 (2016).
  104. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
  105. Giannakis, M. *et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep.* **17**, 1206 (2016).
  106. Pei, X.-H. & Xiong, Y. Biochemical and cellular mechanisms of mammalian CDK inhibitors: a few unresolved issues. *Oncogene* **24**, 2787–2795 (2005).
  107. Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**, 106–114 (2015).
  108. Singhi, A. D. *et al.* Smad4 Loss in Esophageal Adenocarcinoma Is Associated With an Increased Propensity for Disease Recurrence and Poor Survival. *Am. J. Surg. Pathol.* **39**, 487–495 (2015).
  109. Levy, L. & Hill, C. S. Alterations in components of the TGF- $\beta$  superfamily signaling pathways in human cancer. *Cytokine Growth Factor Rev.* **17**, 41–58 (2006).
  110. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
  111. Contino, G. *et al.* Whole-genome sequencing of nine esophageal adenocarcinoma cell lines. *F1000Res* **5**, 1336 (2016).
  112. Liston, D. R. & Davis, M. Clinically Relevant Concentrations of Anticancer Drugs: A Guide for Nonclinical Studies. *Clin. Cancer Res.* **23**, 3489–3498 (2017).

113. Herrera-Abreu, M. T. *et al.* Early Adaptation and Acquired Resistance to CDK4/6 Inhibition in Estrogen Receptor–Positive Breast Cancer. *Cancer Res.* **76**, 2301–2313 (2016).
114. O’Leary, B., Finn, R. S. & Turner, N. C. Treating cancer with selective CDK4/6 inhibitors. *Nat Rev Clin Oncol* **13**, 417–430 (2016).
115. Goetz, M. P. *et al.* MONARCH 3: Abemaciclib As Initial Therapy for Advanced Breast Cancer. *J. Clin. Oncol.* **35**, 3638–3646 (2017).
116. Cicatiello, L. *et al.* Estrogens and Progesterone Promote Persistent CCND1 Gene Activation during G1 by Inducing Transcriptional Derepression via c-Jun/c-Fos/Estrogen Receptor (Progesterone Receptor) Complex Assembly to a Distal Regulatory Element and Recruitment of Cyclin D1 to Its Own Gene Promoter. *Mol. Cell. Biol.* **24**, 7260–7274 (2004).
117. Pascale, F. *et al.* Encorafenib plus binimetinib versus vemurafenib or encorafenib in patients with BRAF-mutant melanoma (COLUMBUS): a multicentre, open-label, randomised phase 3 trial. *Artic. Lancet Oncol* **19**, 603–618 (2018).
118. Fukumoto, T. *et al.* Repurposing Pan-HDAC Inhibitors for ARID1A-Mutated Ovarian Cancer. *Cell Rep.* **22**, 3393–3400 (2018).
119. D’Antonio, M. & Ciccarelli, F. D. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol.* **14**, R52 (2013).
120. McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171**, 1259–1271.e11 (2017).
121. Llosa, N. J. *et al.* The Vigorous Immune Microenvironment of Microsatellite Instable Colon Cancer Is Balanced by Multiple Counter-Inhibitory Checkpoints. *Cancer Discov.* **5**, 43–51 (2015).

122. Grasso, C. S. *et al.* Genetic Mechanisms of Immune Evasion in Colorectal Cancer. *Cancer Discov.* **8**, 730–749 (2018).
123. Sarshekeh, A. M. *et al.* Association of SMAD4 mutation with patient demographics, tumor characteristics, and clinical outcomes in colorectal cancer. *PLoS One* **12**, e0173345 (2017).
124. Shugang, X. *et al.* Prognostic Value of SMAD4 in Pancreatic Cancer: A Meta-Analysis. *Transl. Oncol.* **9**, 1–7 (2016).
125. Tang, B. *et al.* TGF- $\beta$  switches from tumor suppressor to prometastatic factor in a model of breast cancer progression. *J. Clin. Invest.* **112**, 1116–1124 (2003).
126. Zhang, Y. E. Non-Smad pathways in TGF- $\beta$  signaling. *Cell Res.* **19**, 128–139 (2009).
127. David, C. J. *et al.* TGF- $\beta$  Tumor Suppression through a Lethal EMT. *Cell* **164**, 1015–30 (2016).
128. Battegay, E. J., Raines, E. W., Seifert, R. A., Bowen-Pope, D. F. & Ross, R. TGF-beta induces bimodal proliferation of connective tissue cells via complex control of an autocrine PDGF loop. *Cell* **63**, 515–24 (1990).
129. Hasegawa, Y. *et al.* Transforming growth factor-beta1 level correlates with angiogenesis, tumor progression, and prognosis in patients with nonsmall cell lung carcinoma. *Cancer* **91**, 964–71 (2001).
130. Yang, L., Pang, Y. & Moses, H. L. TGF-beta and immune cells: an important regulatory axis in the tumor microenvironment and progression. *Trends Immunol.* **31**, 220–7 (2010).
131. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–67 (1990).
132. Haveri, H. *et al.* Transcription factors GATA-4 and GATA-6 in normal and neoplastic

- human gastrointestinal mucosa. (2008). doi:10.1186/1471-230X-8-9
133. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* (80-. ). eaau3879 (2018). doi:10.1126/SCIENCE.AAU3879
  134. Alcolea, M. P. *et al.* Differentiation imbalance in single oesophageal progenitor cells causes clonal immortalization and field change. *Nat. Cell Biol.* **16**, 612–619 (2014).
  135. Shapiro, G. *et al.* Efficacy and Safety of Abemaciclib, an Inhibitor of CDK4 and CDK6, for Patients with Breast Cancer, Non-Small Cell Lung Cancer, and Other Solid Tumors. (2016). doi:10.1158/2159-8290.CD-16-0095
  136. Kosovec, J. E. *et al.* CDK4/6 dual inhibitor abemaciclib demonstrates compelling preclinical activity against esophageal adenocarcinoma: a novel therapeutic option for a deadly disease. *Oncotarget* **8**, 100421–100432 (2017).

# The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic

Alexander M. Frankell<sup>1</sup> , SriGanesh Jammula<sup>2</sup>, Xiaodun Li<sup>1</sup>, Gianmarco Contino<sup>1</sup>, Sarah Killcoyne<sup>1,3</sup>, Sujath Abbas<sup>1</sup>, Juliane Perner<sup>2</sup>, Lawrence Bower<sup>2</sup>, Ginny Devonshire<sup>2</sup> , Emma Ococks<sup>1</sup>, Nicola Grehan<sup>1</sup>, James Mok<sup>1</sup>, Maria O'Donovan<sup>4</sup>, Shona MacRae<sup>1</sup>, Matthew D. Eldridge<sup>2</sup>, Simon Tavaré<sup>2</sup>, the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium<sup>5</sup> and Rebecca C. Fitzgerald<sup>1\*</sup> 

**Esophageal adenocarcinoma (EAC) is a poor-prognosis cancer type with rapidly rising incidence. Understanding of the genetic events driving EAC development is limited, and there are few molecular biomarkers for prognostication or therapeutics. Using a cohort of 551 genomically characterized EACs with matched RNA sequencing data, we discovered 77 EAC driver genes and 21 noncoding driver elements. We identified a mean of 4.4 driver events per tumor, which were derived more commonly from mutations than copy number alterations, and compared the prevalence of these mutations to the exome-wide mutational excess calculated using non-synonymous to synonymous mutation ratios (*dN/dS*). We observed mutual exclusivity or co-occurrence of events within and between several dysregulated EAC pathways, a result suggestive of strong functional relationships. Indicators of poor prognosis (*SMAD4* and *GATA4*) were verified in independent cohorts with significant predictive value. Over 50% of EACs contained sensitizing events for CDK4 and CDK6 inhibitors, which were highly correlated with clinically relevant sensitivity in a panel of EAC cell lines and organoids.**

Esophageal cancer is the eighth most common form of cancer worldwide and the sixth most common cause of cancer-related death<sup>1</sup>. Esophageal adenocarcinoma (EAC) is the predominant subtype in the West, and its incidence has been rapidly rising<sup>2</sup>. EAC is a highly aggressive neoplasm that usually presents at a late stage and is generally resistant to chemotherapy, thus leading to an overall 5-year survival of <15% (refs. 1,3). In comparison to other cancer types, it is characterized by very high mutation rates<sup>4</sup> but also, paradoxically, by a paucity of recurrently mutated genes. EAC displays marked chromosomal instability and thus may be classified as a C-type neoplasm, which may be driven mainly by structural variation rather than mutations<sup>5,6</sup>. Currently, the understanding of precisely which genetic events drive the development of EAC is limited, and consequently there are few available molecular biomarkers for prognosis or targeted therapeutics.

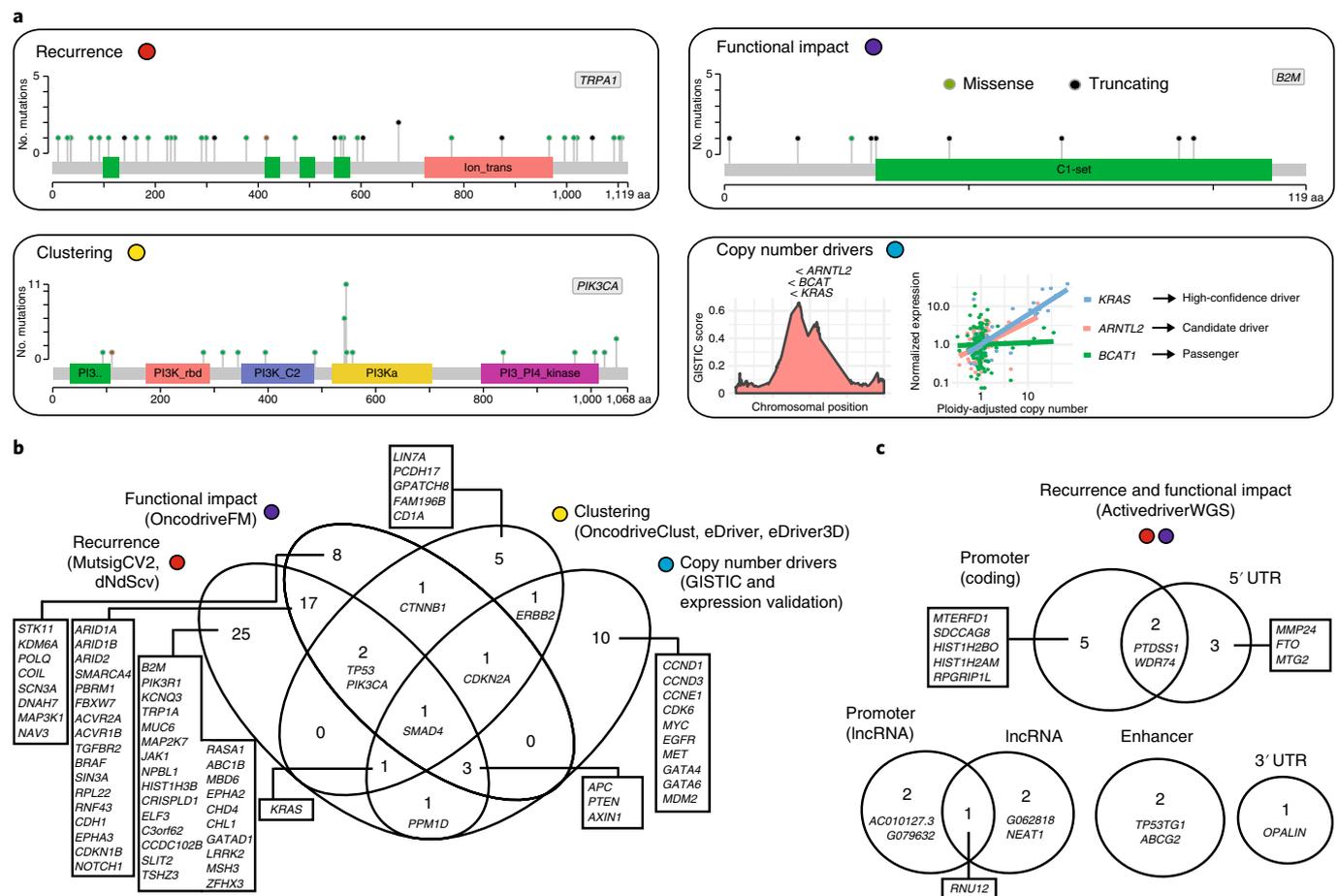
Methods to differentiate driver mutations from passenger mutations use features associated with known drivers to detect regions of the genome in which mutations are enriched in these features<sup>7</sup>. The simplest of these features is the tendency of a mutation to co-occur with other mutations in the same gene at a high frequency, as detected by MutSigCV<sup>8</sup>. MutSigCV has identified 12 known cancer genes as EAC drivers (*TP53*, *CDKN2A*, *SMAD4*, *ARID1A*, *ERBB2*, *KRAS*, *PIK3CA*, *SMARCA4*, *CTNNB1*, *ARID2*, *PBRM1* and *FBXW7*)<sup>6,9,10</sup>. The Pancancer Analysis of Whole Genomes (PCA-WG) International Cancer Genome Consortium (ICGC) analysis has also identified a significantly mutated enhancer associated with

*TP53TGI* (ref. 11). However, these analyses leave most EAC cases with only one known driver mutation, usually *TP53*. Equivalent analyses in other cancer types have identified three or four drivers per case<sup>12,13</sup>. Similarly, detection of copy number driver events in EAC has relied on identifying regions of the genome recurrently deleted or amplified by using GISTIC (genomic identification of significant targets in cancer)<sup>9,14–17</sup>. GISTIC often identifies relatively large regions of the genome, and there is little indication of which specific gene copy number aberrations (CNAs) actually confer a selective advantage. There are also several non-selection-based mechanisms that can cause recurrent CNAs, such as genomic fragile sites, which have not been well differentiated from selection-based CNAs<sup>18</sup>. Epigenetic events such as methylation may also be important sources of driver events in EAC but are much more difficult to formally assess for selection.

To address these issues, by using our esophageal ICGC project, we accumulated a cohort of 551 genomically characterized EACs with high-quality clinical annotation and associated whole-genome sequencing (WGS) and RNA sequencing (RNA-seq) data on cases with sufficient material. We augmented our ICGC WGS cohort with publicly available whole-exome sequencing<sup>19</sup> and WGS<sup>20</sup> data and applied several complementary driver-detection methods to produce a comprehensive assessment of mutations and CNAs under selection in EAC. We used these events to define functional cell processes that have been selectively dysregulated in EAC and identified new, verifiable and clinically relevant biomarkers for

<sup>1</sup>MRC cancer unit, Hutchison/MRC research Centre, University of Cambridge, Cambridge, UK. <sup>2</sup>Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. <sup>3</sup>European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. <sup>4</sup>Department of Histopathology, Cambridge University Hospital NHS Trust, Cambridge, UK. <sup>5</sup>A list of members and affiliations appears at the end of the paper.

\*e-mail: [rcf29@mrc-cu.cam.ac.uk](mailto:rcf29@mrc-cu.cam.ac.uk)



**Fig. 1 | Detection of EAC driver genes. a**, Types of driver-associated features used to detect positive selection in mutations and copy number events with examples of genes containing such features. **b**, Coding driver genes identified and their driver-associated features. **c**, Noncoding driver elements detected and their element types. UTR, untranslated region.

prognostication. Finally, we used this compendium of EAC driver events to provide an evidence base for targeted therapeutics, which we tested in vitro.

**Results**

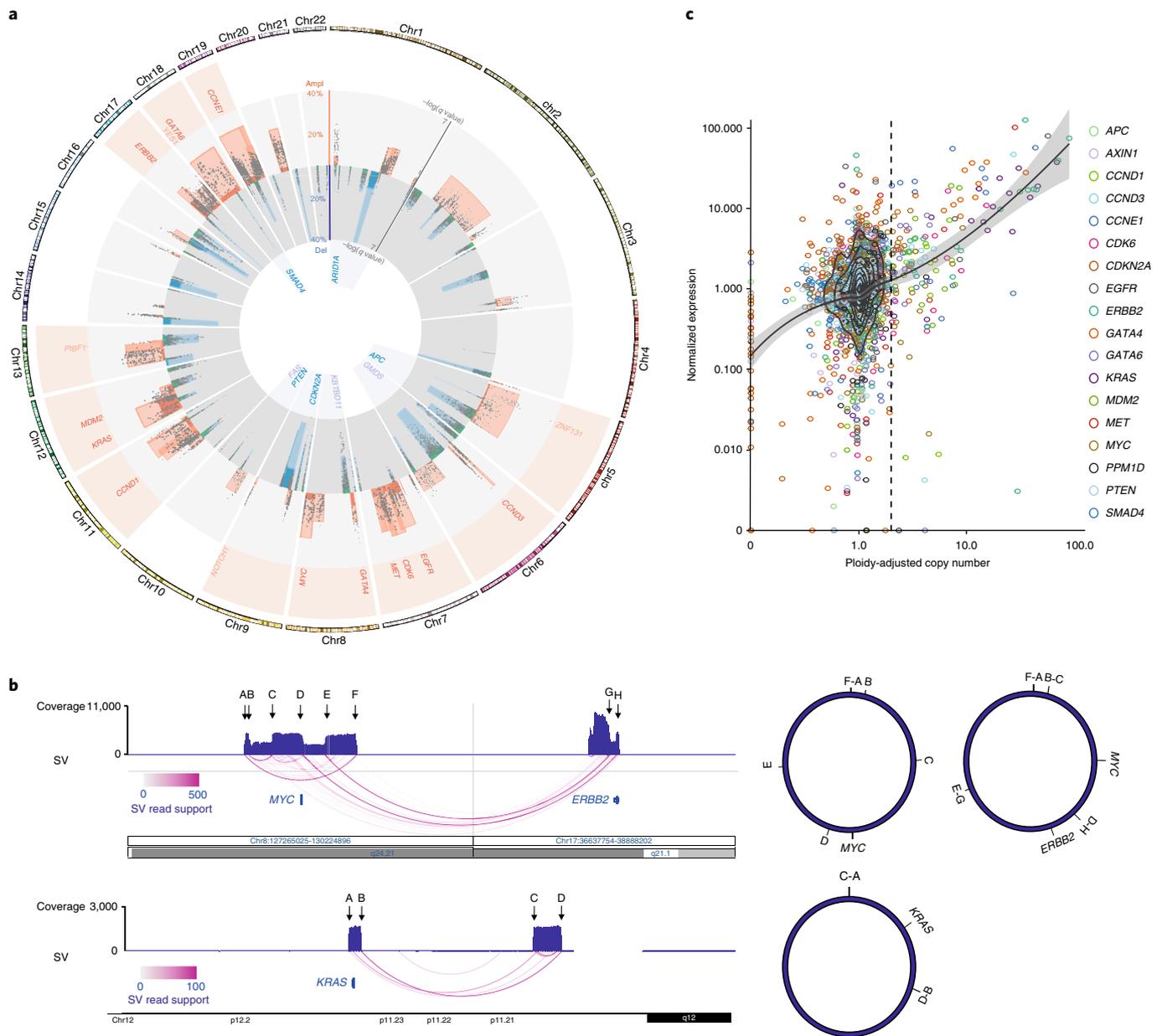
**Compendium of EAC driver events and their functional impact.**

In 551 EACs, we identified a total of 11,813,333 single-nucleotide variants (SNVs) and small insertions or deletions (indels), with a median of 6.4 such mutations per megabase (Supplementary Fig. 1), and 286,965 CNAs. We also identified 134,697 structural variants in WGS cases. We use several complementary driver-detection tools to detect driver-associated features in mutations and CNAs (Fig. 1a). Each tool underwent quality control to ensure the reliability of the results (Methods). These features included highly recurrent mutations within a gene (dNdScv<sup>21</sup>, ActiveDriverWGS<sup>22</sup> and MutSigCV2 (ref. <sup>8</sup>)), high-functional-impact mutations within a gene (OncodriveFM<sup>23</sup> and ActiveDriverWGS<sup>22</sup>), mutation clustering (OncodriveClust<sup>24</sup>, eDriver<sup>25</sup> and eDriver3D<sup>26</sup>) and recurrent amplification or deletion of genes (GISTIC<sup>14</sup>) undergoing concurrent over- or underexpression<sup>7</sup> (Methods and Fig. 1a).

These complementary methods produced highly significant agreement in calling EAC driver genes, particularly within the same feature type (Supplementary Fig. 2); on average, more than half the genes identified by one feature were also identified by other features (Fig. 1b). In total, 76 EAC driver genes were discovered, 71% of which had not previously been detected in EAC<sup>9,10,15–17,19</sup> and 69% of which are known drivers in pancancer analyses<sup>21,27,28</sup>. To detect driver

elements in the noncoding genome, we used ActiveDriverWGS<sup>22</sup>, a recently benchmarked method<sup>29</sup> using both functional impact and recurrence to determine driver status (Fig. 1c and Supplementary Fig. 3). We discovered 21 noncoding driver elements by using this method. We recovered several known noncoding driver elements from the pancancer PCAWG analysis<sup>11</sup>, including the enhancer on chromosome 7, which is linked to *TP53TG1* and has been identified in EAC; the promoter and 5' untranslated regions of *PTDSS1* and *WDR74*. We also identified new noncoding cancer driver elements, including in the 5' untranslated region of *MMP24* and promoters of two related histone-encoding genes (*HIST1H2BO* and *HIST1H2AM*).

EAC is notable among cancer types for its high degree of chromosomal instability<sup>20</sup>. Using GISTIC, we identified 149 recurrently deleted or amplified loci across the genome (Fig. 2a and Supplementary Tables 1 and 2). To determine which genes within these loci confer a selective advantage when they undergo CNAs, we used a subset of 116 cases with matched RNA-seq to detect genes in which homozygous deletion or amplification caused significant under- or overexpression, respectively (Supplementary Note and Supplementary Tables 3–6). Most genes in these regions showed no significant copy-number-associated expression change (74%), although work in larger cohorts suggests that we might have lacked the power to detect small expression changes<sup>30</sup>. We observed highly significant expression changes in 17 known cancer genes within GISTIC loci, such as *ERBB2*, *KRAS* and *SMAD4*, which we designated high-confidence EAC drivers (Methods). We also found five



**Fig. 2 | Copy number variation under positive selection. a**, Recurrent copy number changes across the genome, identified by GISTIC in 551 EACs. Frequencies of different CNV types are indicated (dark blue, homozygous deletion; light blue, heterozygous deletion; dark red, extrachromosomal-like amplification; light red, amplification) as well as the positions of CNV high-confidence driver genes and candidate driver genes. The  $q$  value for expression correlation with amplification and homozygous deletion is shown for each gene within each amplification (one-sided Wilcoxon rank-sum test, expression compared above and below the ninetieth percentile of ploidy-adjusted copy number) and deletion peak (one-sided Wilcoxon rank-sum test, with expression compared between homozygous deleted and all other cases), respectively, and occasions of significant association between LOH and mutation are indicated in green (one-sided Fisher's exact test). Benjamini-Hochberg false-discovery correction was applied in each case. **b**, Examples of extrachromosomal-like amplifications suggested by very high read-support structural variants at the boundaries of highly amplified regions produced from a single copy number step. In the first example, two populations of extrachromosomal DNA are apparent, one amplifying only *MYC* and the second also incorporating *ERBB2* from a different chromosome. In the second example, an inversion has occurred before circularization and amplification around *KRAS*. **c**, Relationship between copy number and expression in copy number driver genes in an RNA-matched subcohort ( $n = 116$ ). A two-dimensional kernel density estimation and a LOESS regression curve with 95% CIs (gray) are shown to describe the data. Chr, chromosome.

tumor-suppressor genes for which copy number loss was not necessarily associated with expression modulation but was tightly associated with the presence of mutations leading to loss of heterozygosity (LOH), for example, *ARID1A* and *CDH11*.

In a subset of GISTIC loci, we observed extremely high copy number amplification, commonly  $>100$  copies, and these events

were highly enriched in recurrently amplified regions containing driver genes rather than those that seemed to contain only passengers (ploidy-adjusted copy number  $>10$ , two-sided Wilcoxon rank-sum test,  $P = 4.97 \times 10^{-8}$ ) (Supplementary Fig. 4). We used ploidy-adjusted copy number to define amplifications, because it produces superior correlation with expression data than absolute

copy number alone. The ploidy of our samples varied from 1.4 to 6.2 (median 2.8), and hence a ploidy-adjusted copy number cutoff of  $>10$  translated into  $>14$  to 62 absolute copies (on average 28 copies). To discern a mechanism for these ultrahigh amplifications, we assessed structural variants associated with these events. For many of these events, the extreme amplification was produced largely from a single copy number step whose edges were linked by structural variants with ultrahigh read support. Two examples are in Fig. 2b, and further randomly selected examples are in Supplementary Fig. 5. In the first example, circularization and amplification initially occurred around *MYC* but subsequently incorporated *ERBB2* from an entirely different chromosome, and in the second, an inversion was followed by circularization and amplification of *KRAS*. Such a pattern of extrachromosomal amplification via double minutes has been noted in EAC<sup>20</sup> and other neoplasms<sup>31</sup>, and hence we refer to this amplification class with ultrahigh amplification (ploidy-adjusted copy number  $>10$ ) as extrachromosomal-like amplifications.

We found that extrachromosomal-like amplifications had extreme and highly penetrant effects on expression, whereas moderate amplification (ploidy-adjusted copy number  $>2$  but  $<10$ ) and homozygous deletion had highly significant (two-sided Wilcoxon rank-sum test,  $P=9.62 \times 10^{-16}$  and  $P=7.64 \times 10^{-11}$ , respectively) but less marked effects on expression with a lower penetrance (Fig. 2c). This lack of penetrance was associated with low cellularity, as calculated by ASCAT (allele-specific copy number analysis of tumors) (two-sided Wilcoxon rank-sum test, overexpression cutoff of  $2.5 \times$  normalized expression,  $P=0.011$ ) in nonextrachromosomal-like amplified cases, but also probably reflects that specific genetic rearrangements, not just gene dosage, can modulate expression. We also detected several cases of overexpression or complete expression loss without associated copy number changes, results reflecting nongenetic mechanisms for driver dysregulation. One case overexpressed *ERBB2* at 28-fold median expression but had entirely diploid copy number in and surrounding *ERBB2*, and a second case lost *SMAD4* expression (0.008-fold median expression) despite having five copies of *SMAD4*.

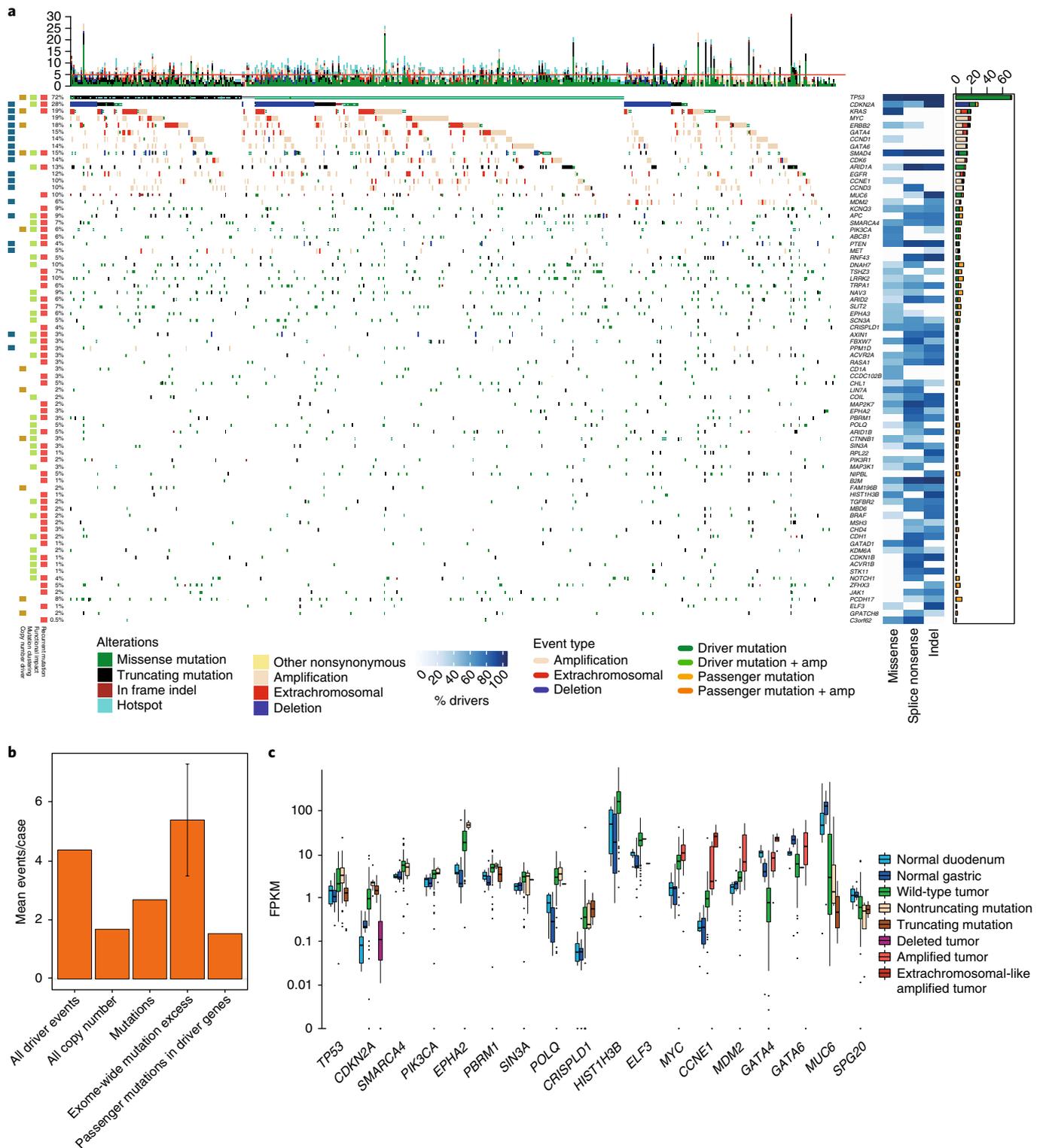
**Landscape of driver events in EAC.** The overall landscape of driver-gene mutations and CNAs per case is depicted in Fig. 3a. These genes comprise both oncogenes and tumor-suppressor genes activated or repressed via different mechanisms. Passenger mutations occur by chance in most driver genes. For quantification, we used the observed/expected mutation ratios (calculated by dNdScv) to estimate the percentage of driver mutations in each gene and in different mutation classes. For many drivers, only specific mutation classes seemed to be under selection. Many tumor-suppressor genes (*ARID2*, *RNF43* and *ARID1B*, for example) are under selection for only truncating mutations, that is, splice-site, nonsense and frameshift indel mutations, but not missense mutations, which are passengers. However, oncogenes such as *ERBB2* contain only missense drivers that form clusters that activate gene function in a specific manner. When a mutation class is  $<100\%$  driver mutations, mutational clustering can help to define the driver versus passenger status of a mutation (Supplementary Fig. 6). Mutational hotspots in EAC or other cancer types<sup>32</sup> (Supplementary Table 7 and Supplementary Data) are indicated in Fig. 3a. Novel EAC drivers of particular interest include *B2M*, which encodes a core component of the MHC class I complex and is a marker of acquired resistance to immunotherapy<sup>33</sup>; *MUC6*, which encodes a secreted glycoprotein involved in gastric acid resistance; and *ABCB1*, which encodes a channel pump protein associated with multiple instances of drug resistance<sup>34</sup>. Notably, several of these drivers are associated with gastric and colorectal cancer<sup>13,35</sup> (Supplementary Table 8).

The identification of driver events provides rich information about the molecular history of each EAC tumor. We detected a

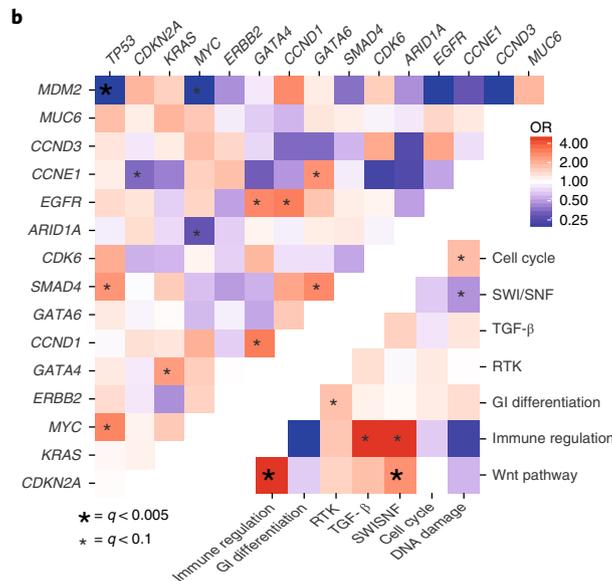
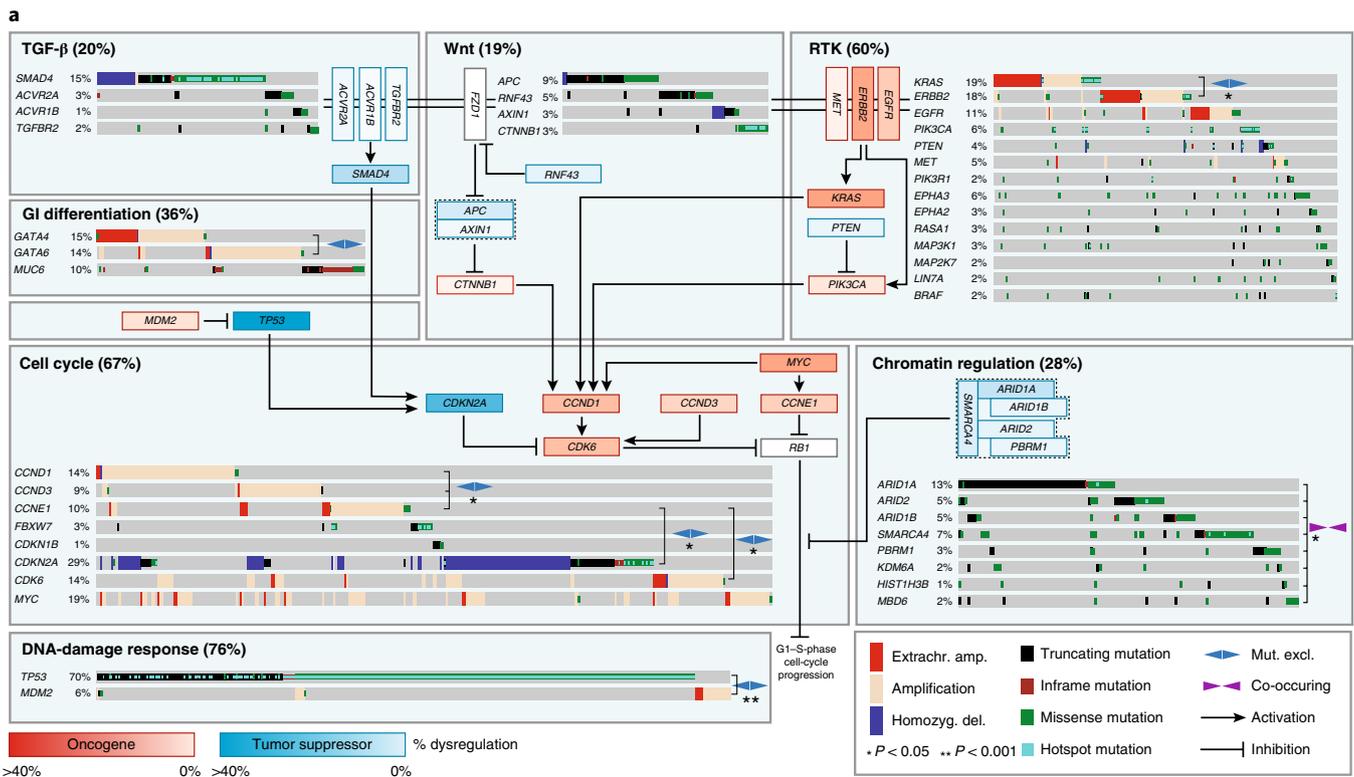
median of five events in driver genes per tumor (interquartile range of 3–7; mean, 5.6), and only a very small fraction of cases had no such events detected (six cases, 1%). When we removed the predicted percentage of passenger mutations by using the observed/expected mutation ratios calculated by dNdScv, one of the driver-gene-detection methods used, we found a mean of 4.4 true driver events per case. These driver events were derived more commonly from mutations than copy number events (Fig. 3b and Supplementary Table 9). Using hierarchical clustering of drivers, we noted that *TP53*-mutant cases had significantly more copy number drivers (two-sided Wilcoxon rank-sum test,  $P=0.0032$ , Supplementary Figs. 7 and 8). dNdScv also analyzes the genome-wide excess of nonsynonymous mutations on the basis of nonsynonymous/synonymous mutation ratios ( $dN/dS$ ) to assess the mean number of exonic driver mutations per case, which was calculated at 5.4 (95% confidence interval (CI) 3.5–7.3) in comparison to a mean excess of 2.7 driver mutations in specific EAC driver genes, thus suggesting that additional low-frequency driver genes are yet to be discovered in EAC.

To better understand the functional impact of driver mutations, we analyzed the expression of driver genes with different mutation types and compared their expression to normal tissue RNA (Fig. 3c and Supplementary Fig. 10). Because the surrounding squamous epithelium is a fundamentally different tissue from which EAC does not directly arise, we used duodenum and gastric cardia samples as gastrointestinal phenotype controls, which also have a columnar phenotype similar to EAC and Barrett's. Many driver genes had higher expression than that in normal controls; for example, *TP53* had upregulated RNA expression in wild-type tumor tissue and in cases with nontruncating mutations, but RNA expression was lost after gene truncation. In-depth analysis of different *TP53* mutation types revealed substantial heterogeneity within nontruncating mutations (Supplementary Fig. 9). The normal tissue expression of *CDKN2A* suggested that *CDKN2A* is generally activated in EAC, probably because of genotoxic or other cancer-associated cellular stresses<sup>36</sup>, and returns to physiologically normal levels when deleted. Heterogeneous expression in wild-type *CDKN2A* cases suggested a different mechanism of inhibition, perhaps methylation, in some cases. Overexpression of some oncogenes occurs without genomic aberrations, such as *MYC*, which was overexpressed in *MYC*-wild-type EACs relative to normal tissues (Fig. 3c). Fewer driver genes were downregulated in EACs without genomic aberrations. Three-quarters of these genes (*GATA4*, *GATA6* and *MUC6*) are involved in the differentiated phenotype of gastrointestinal tissues and may be lost with tumor dedifferentiation.

**Dysregulation of specific pathways and processes in EAC.** Selection preferentially dysregulates certain functionally related groups of genes and biological pathways in cancer<sup>37</sup>. This phenomenon is highly evident in EAC, as shown in Fig. 4, which depicts the functional relationships between EAC drivers (Supplementary Note). Whereas *TP53* is the dominant driver in EAC, 28% of cases remain *TP53* wild type. MDM2 is an E3 ubiquitin ligase that targets *TP53* for degradation. Its selective amplification and overexpression is mutually exclusive with *TP53* mutation, thus suggesting that its degradation can functionally substitute for the effect of *TP53* mutation. Similar mutually exclusive relationships were observed among *KRAS* and *ERBB2*, *GATA4* and *GATA6*, and cyclin genes (*CCNE1*, *CCND1* and *CCND3*). Activation of the Wnt pathway occurred in 19% of cases, either by mutation of phosphorylated residues at the N terminus of  $\beta$ -catenin, preventing degradation, or loss of Wnt destruction-complex components such as APC. Many different chromatin-modifying genes, often belonging to the SWI-SNF complex, were also selectively mutated (28% of cases). In contrast to genes involved in other pathways, SWI-SNF genes were mutated significantly more often than expected by chance (two-sided Fisher's exact test,  $q < 0.05$  for each gene; Methods), thus



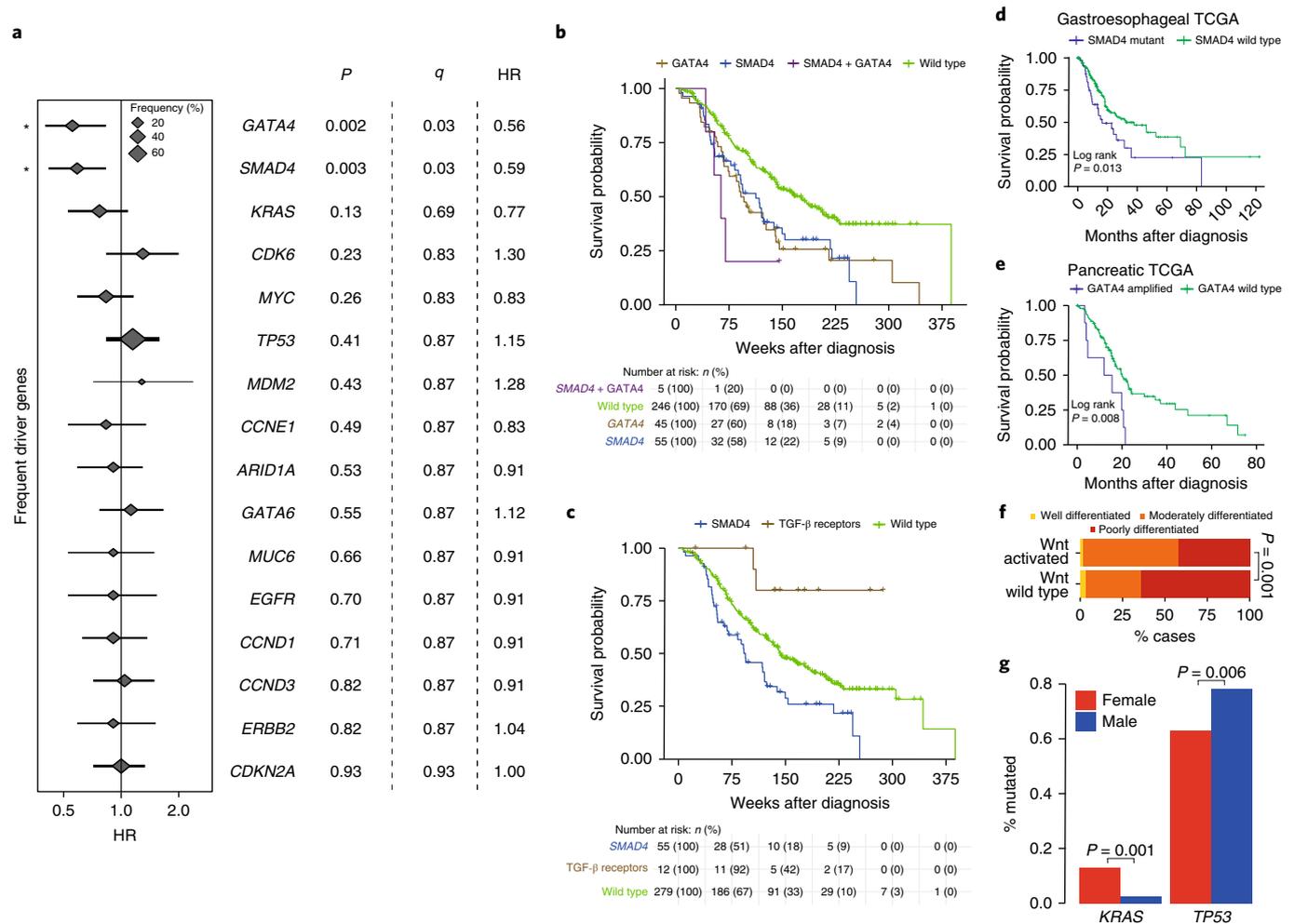
**Fig. 3 | The driver-gene landscape of EAC. a**, Driver mutations or CNVs are shown for each subject of 551 EACs. Amplification (amp) is defined as copy-number-adjusted ploidy >2 (2× ploidy of that case) and extrachromosomal amplification as >10 copy-number-adjusted ploidy (10× ploidy for that case). Driver-associated features for each driver gene are at left. At right, the percentages of different mutation and copy number changes are shown, differentiated between driver and passenger mutations, and the percentage of predicted drivers by mutation type is shown. Passenger-mutation rates were determined by using observed-to-expected mutation rates, as calculated by dNdScv. Above the plot, the number of driver mutations per sample is shown, with an indication of the mean (red line = 5). **b**, Mean driver events per case in 551 EACs and comparison to exome-wide excess of mutations generated by dNdScv. **c**, Expression changes in EAC driver genes in comparison to normal intestinal tissues in RNA-matched samples ( $n = 116$ ). FPKM, fragments per kilobase of transcript per million mapped reads. Only genes with notable expression changes are shown.



**Fig. 4 | Biological pathways undergoing selective dysregulation in EAC. a**, Biological pathways dysregulated by driver-gene mutation and/or CNVs in 551 cases. Wild-type cases for a pathway are not shown. Inter- and intrapathway interactions are described, and mutual exclusivities and/or associations between genes in a pathway are annotated. *GATA4* and *GATA6* amplifications had a mutually exclusive (mut. excl.) relationship, although it did not reach statistical significance (two-sided Fisher’s exact test,  $P=0.07$ ,  $OR=0.52$ ). Extrachr. amp., extrachromosomal amplification; homozyg. del., homozygous deletion. **b**, Pairwise assessment of mutual exclusivity and association in EAC driver genes and pathways. Two-sided Fisher’s exact tests were used, and hypermutated cases (>500 exonic mutations) were removed to avoid bias toward co-occurrence, hence  $n=510$ .

suggesting that these mutations are synergistic. We also assessed mutual exclusivity and co-occurrence in genes in different pathways and between pathways themselves (Fig. 4b). Of particular note were co-occurring relationships between *TP53* and *MYC*, *GATA6* and *SMAD4*, and Wnt and immune pathways, as well as mutually exclusive relationships between *ARID1A* and *MYC*, gastrointestinal differentiation and receptor tyrosine kinase (RTK) pathways,

and SWI–SNF and DNA-damage-response pathways. We confirmed some of these relationships in independent cohorts in different cancer types (Supplementary Table 10), thus suggesting that some may represent pancancer phenomena. Wnt dysregulation was associated with hypermutated cases (>500 exonic SNVs or indels, two-sided Fisher’s exact test,  $P=2.98 \times 10^{-5}$ , odds ratio (OR)=9.3), as was mutation in immune-pathway genes (*B2M* and *JAK1*, >500



**Fig. 5 | Clinical importance of driver events in 379 clinically annotated EACs. a**, HRs and 95% CIs for Cox regression analysis across all driver genes with at least a 5% frequency of driver alterations. \* $q < 0.05$  after Benjamini-Hochberg adjustment. **b**, Kaplan-Meier curves for EACs with different status of significant prognostic indicators (*GATA4* and *SMAD4*). **c**, Kaplan-Meier curves showing verification of *GATA4* prognostic value in gastrointestinal cancers in a pancreatic TCGA cohort. **d**, Kaplan-Meier curves showing verification of *SMAD4* prognostic value in gastroesophageal cancers in a gastroesophageal TCGA cohort. **e**, Kaplan-Meier curves for different alterations in the TGF- $\beta$  pathway. **f**, Differentiation bias in tumors containing events in Wnt-pathway driver genes. **g**, Relative frequency of *KRAS*-mutation and *TP53*-mutation driver-gene events in females versus males (two-sided Fisher's exact test).

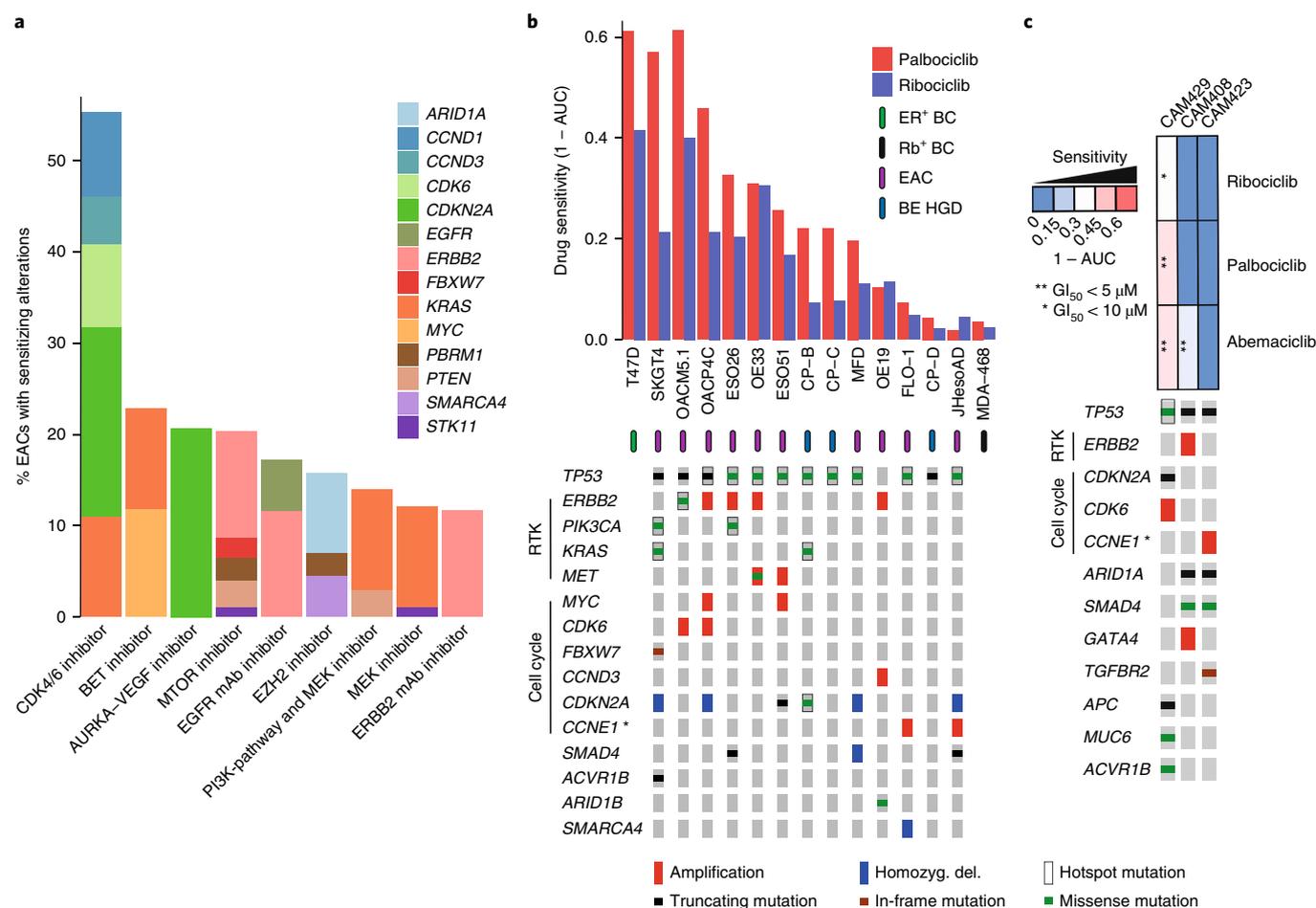
exonic SNVs or indels, two-sided Fisher's exact test,  $P = 6.27 \times 10^{-6}$ , OR = 35.7).

**EAC driver events are correlated with clinical phenotype.** Events undergoing selection during cancer evolution influence tumor biology and thus affect tumor aggressiveness, response to treatment and patient prognosis, and other clinical parameters.

To detect prognostic biomarkers, we performed univariate Cox regression for events in each driver gene with driver events occurring in >5% of EACs after passenger removal (Fig. 5a). Events in two genes were associated with significantly poorer prognosis after multiple-hypothesis correction: *GATA4* amplification (hazard ratio (HR) = 0.54, 95% CI = 0.38–0.78,  $P = 0.0008$ ) and *SMAD4* mutation or homozygous deletion (HR = 0.60, 95% CI = 0.42–0.84,  $P = 0.003$ ), which were present in 31% of EACs (Fig. 5b). Both genes remained significant in multivariate Cox regression, including pathological tumor node metastasis staging, resection margin, curative versus palliative treatment intent and differentiation status (*GATA4*, HR adjusted = 0.47, 95% CI adjusted = 0.29–0.76,  $P = 0.002$ ; *SMAD4*, HR adjusted = 0.61, 95% CI adjusted = 0.40–0.94,  $P = 0.026$ ) (Fig. 5b and Supplementary Fig. 11). We validated the poor-prognosis-associated effects of *SMAD4* events in an independent The Cancer

Genome Atlas (TCGA) gastroesophageal cohort (HR = 0.58, 95% CI = 0.37–0.90,  $P = 0.014$ ) (Fig. 5c), and we also found that *GATA4* amplifications were prognostic in a cohort of TCGA pancreatic cancers (HR = 0.38, 95% CI = 0.18–0.80,  $P = 0.011$ ) (Fig. 5d), the only available cohort containing a feasible number of *GATA4* amplifications. The prognostic effect of *GATA4* has been suggested in a previously published independent EAC cohort<sup>16</sup>, although it did not reach statistical significance after false discovery rate (FDR) correction in that study, and *SMAD4* expression loss has been linked to poor prognosis in EAC<sup>38</sup>. We also noted stark survival differences between cases with *SMAD4* events and cases in which TGF- $\beta$  receptors were mutated (Fig. 5e, HR = 5.6, 95% CI = 1.7–18.2,  $P = 0.005$ ), in keeping with the biology of the TGF- $\beta$  pathway, in which non-*SMAD* TGF- $\beta$  signaling is oncogenic<sup>39</sup>.

In addition to survival analyses, we also assessed driver-gene events for correlation with various other clinical factors, including differentiation status, sex, age and treatment response. We found that Wnt-pathway mutations had a strong association with well-differentiated tumors ( $P = 0.001$ , OR = 2.9, two-sided Fisher's exact test, Methods and Fig. 5f). Female cases ( $n = 81$ ) were enriched in *KRAS* mutation ( $P = 0.001$ , two-sided Fisher's exact test) and *TP53* wild-type status ( $P = 0.006$ , two-sided Fisher's exact test) (Fig. 5g).



**Fig. 6 | CDK4/CDK6 inhibitors in EAC.** **a**, Drug classes for which sensitivity is indicated by EAC driver genes with data from the Cancer Biomarkers database<sup>36</sup>. mAb, monoclonal antibody. **b**, AUC of sensitivity, shown for a panel of 13 EAC and Barrett's esophagus high-grade dysplasia (BE HGD) cell lines with associated WGS and their corresponding driver events, on the basis of primary tumor analysis. AUC is also shown for two control lines: T47D, an estrogen receptor (ER)-positive breast cancer (BC) line (positive control), and MDA-MB-468, a retinoblastoma-negative breast cancer (negative control). \**CCNE1* is a known marker of resistance to CDK4/CDK6 inhibitors, owing to its regulation of retinoblastoma downstream of CDK4/CDK6, thus bypassing the need for CDK4/CDK6 activity (Fig. 4). **c**, Response of organoid cultures to three FDA-approved CDK4/CDK6 inhibitors and corresponding driver events.

This finding is of particular interest, given the male predominance of EAC<sup>3</sup>.

**Targeted therapeutics based on EAC driver events.** To investigate whether driver events, particularly genes and/or pathways, might sensitize EAC cells to certain targeted therapeutic agents, we used the Cancer Biomarkers database<sup>40</sup>. We calculated the percentage of our cases that contained EAC-driver biomarkers of response to each drug class in the database (Fig. 6a and full data in Supplementary Table 11). Aside from *TP53*, which has been problematic to target clinically to date, we found several drugs with predicted sensitivity in >10% of EACs, including EZH2 inhibitors for some SWI-SNF-mutant cancers (23%, and 28% including all SWI-SNF EAC drivers), and BET inhibitors, which target *KRAS*-activated and *MYC*-amplified cases (23%). However, by far the most significantly effective class of drug was predicted to be inhibitors of CDK4 and CDK6 (CDK4/CDK6): >50% of cases had sensitivity-causing events in the RTK and core cell-cycle pathways (for example, in *CCND1*, *CCND3* and *KRAS*).

To verify that these driver events would also sensitize EAC tumors to such inhibitors, we used a panel of 13 EAC or Barrett's high-grade dysplasia cell lines that had undergone WGS<sup>41</sup> and assessed them for the presence of EAC driver events (Fig. 6b).

The mutational landscape of these lines was broadly representative of EAC tumors. We found that the presence of cell-cycle and/or RTK-activating driver events was highly correlated with the response to two US Food and Drug Administration (FDA)-approved CDK4/CDK6 inhibitors, ribociclib and palbociclib, and several cell lines were sensitive below maximum tolerated blood concentrations in humans<sup>42</sup> (Fig. 6b, Supplementary Table 12 and Supplementary Fig. 12). Such EAC cell lines had comparable sensitivity to T47D, which is derived from an estrogen-receptor-positive breast cancer in which CDK4/CDK6 inhibitors have been FDA approved. We noted three cell lines that were highly resistant, with little drug effect even at a concentration of 4,000 nM, similarly to a known retinoblastoma-mutant resistant breast cancer cell line (MDA-MB-468). Two of these three cell lines have amplification of *CCNE1*, which is known to drive resistance to CDK4/CDK6 inhibitors by bypassing CDK4/CDK6 and causing retinoblastoma phosphorylation via CDK2 activation<sup>43</sup>. To verify these effects in a more representative model of EAC, we treated three whole-genome-sequenced EAC organoid cultures<sup>44</sup> with palbociclib and ribociclib, as well as a more recently approved CDK4/CDK6 inhibitor, abemaciclib. As observed in cell lines, cell-cycle and RTK driver events were present in only the more sensitive organoids, and *CCNE1* activation was present in only the most resistant organoid (Fig. 6c).

## Discussion

We present a detailed catalog of coding and noncoding genomic events that have been selected for during the evolution of EAC. These events were characterized in terms of their relative impact, related functions, mutual exclusivity and co-occurrence and expression in comparison to those in normal tissues. We used this set of biologically important gene alterations to identify prognostic biomarkers and actionable genomic events for personalized medicine.

Although the matched RNA-sequencing data are a strength of this study, we may not have been able to assess some uncommon variants for expression changes if these variants, detected in the full 551-patient cohort, were not well represented in the RNA-matched subcohort of 116 cases. Despite rigorous analyses to detect selected events, assessment of the global excess of mutations by dNdScv suggested that we could not detect all mutations selected in EAC, as in many other cancer types<sup>21</sup>. All driver-gene-detection methods that we used rely on driver-mutation recurrence in a genomic region to some degree. Many of these undetected driver mutations are hence probably spread across many genes, such that each is mutated at very low frequency across individuals with EAC. This tendency for low-frequency EAC drivers may be responsible for the low yield of MutSigCV in previous cohorts and may suggest that C-type cancers such as EAC are not less 'mutation driven' than M-type cancers but instead that their mutational drivers may be spread across a larger number of genes<sup>5</sup>. Copy number driver-gene identification is even more challenging because of the large size and lower frequency of these events, and hence many more EAC copy number drivers may remain to be discovered, some of which may have been identified as candidates here.

Although some previous reports have attempted to detect EAC drivers, they have had a limited yield per case. The first such study<sup>19</sup> used methods that, despite being well regarded at the time, were subsequently discredited<sup>8</sup>. Since then, several reports, including our own, using MutSigCV<sup>9,10,17</sup> on medium and large cohort sizes, have detected only a small number of mutational driver genes (7, 5 and 15 in each study, respectively). By using both a large cohort and more comprehensive methodologies, we markedly increased this figure to 66 mutational driver genes (excluding copy number drivers). Detection of driver CNAs has previously relied on GISTIC to detect regions with recurrent CNAs<sup>9,14–17</sup>, but no analyses have been performed to determine which genes in these large regions are true drivers. Many of the genes annotated by such papers are unlikely to be copy number drivers, owing to their lack of expression modulation with CNAs (for example, *YEATS4* and *MCL1*), the role of recurrent heterozygous losses in driving LOH in some mutational drivers (*ARID1A* and *CDH11*) or their association with fragile sites (*PDE4D*, *WWOX* and *FHIT*). In contrast, we identified new EAC copy number drivers (for example, *CCND3*, *AXIN1*, *PPM1D* and *APC*).

We noted a three-way association among hypermutation, Wnt activation and loss of immune-signaling genes such as *B2M*. Microsatellite-instability-driven hypermutation has been associated with higher immune activity<sup>45,46</sup>. However, Wnt dysregulation and mutation of immune-pathway genes such as *B2M*<sup>33</sup> have been linked to immunological escape<sup>47</sup>, thus suggesting that this may be an acquired mechanism to prevent immune surveillance caused by hypermutation.

Many of the driver genes that we described will require further functional characterisation to understand why they are advantageous to EAC tumors and how they modify EAC biology. Biological pathways and processes that are selectively dysregulated deserve particular attention in this regard, as do the gene pairs or groups with mutually exclusive or co-occurring relationships, such as *MYC* and *TP53* or SWI-SNF factors, which are suggestive of particular functional relationships. Prospective clinical work to verify and implement *SMAD4* and *GATA4* biomarkers in this study would be

worthwhile. Although EAC is a poor-prognosis cancer type, substantial heterogeneity in survival outcomes makes triaging patients in treatment groups an important part of clinical practice that could be improved through better prognostication. Several targeted therapeutics may provide clinical benefit for EAC cases on the basis of individual genomic profiles. In particular, CDK4/CDK6 inhibitors deserve considerable attention as an option for EAC treatment because they are, by a large margin, the treatment for which the most EACs have sensitivity-causing driver events, excluding TP53 as an unlikely therapeutic biomarker at the current time. Previous work has noted the activity of the CDK4/CDK6 inhibitor palbociclib in a small number of EAC cell lines<sup>48</sup>, but biomarkers were not investigated. The extensive in vitro validation of identified biomarkers for CDK4/CDK6 inhibitors in EAC across 16 cell lines and organoids suggests possible clinical benefit through use of a targeted approach.

In summary, this work provides a detailed compendium of mutations and copy number alterations undergoing selection in EAC that have clinically relevant effects on tumor behavior. This comprehensive study provides insights into the nature of EAC tumors and should pave the way for evidence-based clinical trials in this poor-prognosis disease.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41588-018-0331-5>.

Received: 27 April 2018; Accepted: 10 December 2018;

Published online: 4 February 2019

## References

1. Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
2. Coleman, H. G., Xie, S. H. & Lagergren, J. The epidemiology of esophageal adenocarcinoma. *Gastroenterology* **154**, 390–405 (2018).
3. Smyth, E. C. et al. Oesophageal cancer. *Nat. Rev. Dis. Primers* **3**, 17048 (2017).
4. Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O. & Stein, L.D. Pan-cancer analysis of whole genomes. Preprint at <https://www.biorxiv.org/content/early/2017/07/12/162784> (2017).
5. Ciriello, G. et al. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
6. Secrier, M. et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.* **48**, 1131–1141 (2016).
7. Tamborero, D. et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
8. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
9. Cancer Genome Atlas Research Network. et al. Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175 (2017).
10. Lin, D. C. et al. Identification of distinct mutational patterns and new driver genes in oesophageal squamous cell carcinomas and adenocarcinomas. *Gut* **67**, 1769–1779 (2017).
11. Rheinbay, E. et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. Preprint at <https://www.biorxiv.org/content/early/2017/12/23/237313> (2017).
12. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).
13. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
14. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
15. Dulak, A. M. et al. Gastrointestinal adenocarcinomas of the esophagus, stomach, and colon exhibit distinct patterns of genome instability and oncogenesis. *Cancer Res.* **72**, 4383–4393 (2012).
16. Frankel, A. et al. Genome-wide analysis of esophageal adenocarcinoma yields specific copy number aberrations that correlate with prognosis. *Genes Chromosom. Cancer* **53**, 324–338 (2014).

17. Secrier, M. & Fitzgerald, R. C. Signatures of mutational processes and associated risk factors in esophageal squamous cell carcinoma: a geographically independent stratification strategy? *Gastroenterology* **150**, 1080–1083 (2016).
18. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
19. Dulak, A. M. et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478–486 (2013).
20. Nones, K. et al. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nat. Commun.* **5**, 5224 (2014).
21. Martincorena I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21.
22. Wadi, L. et al. Candidate cancer driver mutations in super-enhancers and long-range chromatin interaction networks. Preprint at <https://www.biorxiv.org/content/early/2017/12/19/236802> (2017).
23. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).
24. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
25. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**, 3109–3114 (2014).
26. Porta-Pardo, E., Hrade, T. & Godzik, A. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res.* **43**, D968–D973 (2015).
27. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
28. Kandath, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
29. Shuai, S., Gallinger, S. & Stein, L.D. DriverPower: combined burden and functional impact tests for cancer driver discovery. Preprint at <https://www.biorxiv.org/content/early/2017/11/06/215244> (2017).
30. Taylor, A. M. et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).
31. Turner, K. M. et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).
32. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
33. Zaretsky, J. M. et al. Mutations associated with acquired resistance to PD-1 blockade in melanoma. *N. Engl. J. Med.* **375**, 819–829 (2016).
34. Chen, Z. et al. Mammalian drug efflux transporters of the ATP binding cassette (ABC) family in multidrug resistance: a review of the past decade. *Cancer Lett.* **370**, 153–164 (2016).
35. Giannakis, M. et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Rep.* **17**, 1206 (2016).
36. Pei, X. H. & Xiong, Y. Biochemical and cellular mechanisms of mammalian CDK inhibitors: a few unresolved issues. *Oncogene* **24**, 2787–2795 (2005).
37. Leiserson, M. D. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
38. Singhi, A. D. et al. Smad4 loss in esophageal adenocarcinoma is associated with an increased propensity for disease recurrence and poor survival. *Am. J. Surg. Pathol.* **39**, 487–495 (2015).
39. Levy, L. & Hill, C. S. Alterations in components of the TGF-beta superfamily signaling pathways in human cancer. *Cytokine Growth Factor Rev.* **17**, 41–58 (2006).
40. Tamborero, D. et al. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. Preprint at <https://www.biorxiv.org/content/early/2017/06/21/140475> (2017).
41. Contino, G. et al. Whole-genome sequencing of nine esophageal adenocarcinoma cell lines. *F1000Res.* **5**, 1336 (2016).
42. Liston, D. R. & Davis, M. Clinically relevant concentrations of anticancer drugs: a guide for nonclinical studies. *Clin. Cancer Res.* **23**, 3489–3498 (2017).
43. Herrera-Abreu, M. T. et al. Early adaptation and acquired resistance to CDK4/6 inhibition in estrogen receptor-positive breast cancer. *Cancer Res.* **76**, 2301–2313 (2016).
44. Li, X. et al. Organoid cultures recapitulate esophageal adenocarcinoma heterogeneity providing a model for clonality studies and precision therapeutics. *Nat. Commun.* **9**, 2983 (2018).
45. Llosa, N. J. et al. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discov.* **5**, 43–51 (2015).
46. Le, D. T. et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
47. Grasso, C. S. et al. Genetic mechanisms of immune evasion in colorectal cancer. *Cancer Discov.* **8**, 730–749 (2018).
48. Ismail, A. et al. Early G1 cyclin-dependent kinases as prognostic markers and potential therapeutic targets in esophageal adenocarcinoma. *Clin. Cancer Res.* **17**, 4513–4522 (2011).

### Acknowledgements

We thank A. J. Bass and N. Waddell for providing data in Dulak et al.<sup>19</sup> and Nones et al.<sup>20</sup>, respectively, which were also included in our previous publication<sup>18</sup>. Inclusion of these data allowed for augmentation of our ICGC cohort and greater sensitivity for the detection of EAC driver variants. OCCAMS was funded by a Programme Grant from Cancer Research UK (RG66287), and the laboratory of R.C.F. is funded by a Core Programme Grant from the Medical Research Council. We thank the Human Research Tissue Bank, which is supported by the UK National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre, from Addenbrooke's Hospital. Additional infrastructure support was provided from the Cancer Research UK-funded Experimental Cancer Medicine Centre.

### Author contributions

R.C.F. and A.M.F. conceived the overall study. A.M.F. and S.J. analyzed the genomic data and performed statistical analyses. R.C.F., A.M.F. and X.L. designed the experiments. A.M.F., X.L. and J.M. performed the experiments. G.C. contributed to the structural variant analysis and data visualization. S.K. helped compile the clinical data and aided in statistical analyses. J.P. and S.A. produced and performed quality control on the RNA-seq data. E.O. aided in WGS of EAC cell lines. S.M. and N.G. coordinated the clinical centers and were responsible for sample collection. M.D.E. benchmarked our mutation-calling pipelines. M.O. led the pathological sample quality control for sequencing. L.B. and G.D. constructed and managed the sequencing alignment and variant-calling pipelines. R.C.F. and S.T. supervised the research. R.C.F. and S.T. obtained funding. A.M.F. and R.C.F. wrote the manuscript. All authors approved the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41588-018-0331-5>.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Correspondence and requests for materials** should be addressed to R.C.F.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

## the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium

Rebecca C. Fitzgerald<sup>1</sup>, Ayesha Noorani<sup>1</sup>, Paul A. W. Edwards<sup>1,2</sup>, Nicola Grehan<sup>1</sup>, Barbara Nutzinger<sup>1</sup>, Caitriona Hughes<sup>1</sup>, Elwira Fidziukiewicz<sup>1</sup>, Shona MacRae<sup>1</sup>, Alex Northrop<sup>1</sup>, Gianmarco Contino<sup>1</sup>, Xiaodun Li<sup>1</sup>, Rachel de la Rue<sup>1</sup>, Annalise Katz-Summercorn<sup>1</sup>, Sujath Abbas<sup>1</sup>, Daniel Loureda<sup>1</sup>, Maria O'Donovan<sup>1,4</sup>, Ahmad Miremadi<sup>1,4</sup>, Shalini Malhotra<sup>1,4</sup>, Monika Tripathi<sup>1,4</sup>, Simon Tavaré<sup>2</sup>, Andy G. Lynch<sup>2</sup>, Matthew Eldridge<sup>2</sup>, Maria Secrier<sup>6</sup>, Ginny Devonshire<sup>2</sup>, Juliane Perner<sup>2</sup>, SriGanesh Jammula<sup>2</sup>, Jim Davies<sup>7</sup>, Charles Crichton<sup>7</sup>, Nick Carroll<sup>8</sup>, Peter Safranek<sup>8</sup>, Andrew Hindmarsh<sup>8</sup>, Vijayendran Sujendran<sup>8</sup>, Stephen J. Hayes<sup>9,10</sup>, Yeng Ang<sup>9,11,12</sup>, Andrew Sharrocks<sup>12</sup>, Shaun R. Preston<sup>13</sup>, Sarah Oakes<sup>13</sup>, Izhar Bagwan<sup>13</sup>, Vicki Save<sup>14</sup>, Richard J. E. Skipworth<sup>14</sup>, Ted R. Hupp<sup>14</sup>, J. Robert O'Neill<sup>14,15</sup>, Olga Tucker<sup>16,17</sup>, Andrew Beggs<sup>16,18</sup>, Philippe Tanriere<sup>16</sup>, Sonia Puig<sup>16</sup>, Timothy J. Underwood<sup>19,20</sup>, Robert C. Walker<sup>19,20</sup>, Ben L. Grace<sup>19</sup>, Hugh Barr<sup>21</sup>, Neil Shepherd<sup>21</sup>, Oliver Old<sup>21</sup>, Jesper Lagergren<sup>22,23</sup>, James Gossage<sup>22,24</sup>, Andrew Davies<sup>22,24</sup>, Fujun Chang<sup>22,24</sup>, Janine Zylstra<sup>22,24</sup>, Ula Mahadeva<sup>22</sup>, Vicky Goh<sup>24</sup>, Francesca D. Ciccarelli<sup>24</sup>, Grant Sanders<sup>25</sup>, Richard Berrisford<sup>25</sup>, Catherine Harden<sup>25</sup>, Mike Lewis<sup>26</sup>, Ed Cheong<sup>26</sup>, Bhaskar Kumar<sup>26</sup>, Simon L. Parsons<sup>27</sup>, Irshad Soomro<sup>27</sup>, Philip Kaye<sup>27</sup>, John Saunders<sup>27</sup>, Laurence Lovat<sup>28</sup>, Rehan Haidry<sup>28</sup>, Laszlo Igali<sup>29</sup>, Michael Scott<sup>30</sup>, Sharmila Sothi<sup>31</sup>, Sari Suortamo<sup>31</sup>, Suzy Lishman<sup>32</sup>, George B. Hanna<sup>33</sup>, Krishna Moorthy<sup>33</sup>, Christopher J. Peters<sup>33</sup>, Anna Grabowska<sup>34</sup>, Richard Turkington<sup>35</sup>, Damian McManus<sup>35</sup>, Helen Coleman<sup>35</sup>, David Khoo<sup>36</sup> and Will Fickling<sup>36</sup>

<sup>6</sup>Department of Genetics, Evolution and Environment, UCL Genetics Institute, University College London, London, UK. <sup>7</sup>Department of Computer Science, University of Oxford, Oxford, UK. <sup>8</sup>Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. <sup>9</sup>Salford Royal NHS Foundation Trust, Salford, UK. <sup>10</sup>Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK. <sup>11</sup>Wigan and Leigh NHS Foundation Trust, Wigan, Manchester, UK. <sup>12</sup>GI Science Centre, University of Manchester, Manchester, UK. <sup>13</sup>Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK. <sup>14</sup>Edinburgh Royal Infirmary, Edinburgh, UK. <sup>15</sup>Edinburgh University, Edinburgh, UK. <sup>16</sup>University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. <sup>17</sup>Heart of England NHS Foundation Trust, Birmingham, UK. <sup>18</sup>Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK. <sup>19</sup>University Hospital Southampton NHS Foundation Trust, Southampton, UK. <sup>20</sup>Cancer Sciences Division, University of Southampton, Southampton, UK. <sup>21</sup>Gloucester Royal Hospital, Gloucester, UK. <sup>22</sup>Guy's and St Thomas's NHS Foundation Trust, London, UK. <sup>23</sup>Karolinska Institutet, Stockholm, Sweden. <sup>24</sup>King's College London, London, UK. <sup>25</sup>Plymouth Hospitals NHS Trust, Plymouth, UK. <sup>26</sup>Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, UK. <sup>27</sup>Nottingham University Hospitals NHS Trust, Nottingham, UK. <sup>28</sup>University College London, London, UK. <sup>29</sup>Norfolk and Waveney Cellular Pathology Network, Norwich, UK. <sup>30</sup>Wythenshawe Hospital, Manchester, UK. <sup>31</sup>University Hospitals Coventry and Warwickshire NHS, Trust, Coventry, UK. <sup>32</sup>Peterborough Hospitals NHS Trust, Peterborough City Hospital, Peterborough, UK. <sup>33</sup>Department of Surgery and Cancer, Imperial College London, London, UK. <sup>34</sup>Queen's Medical Centre, University of Nottingham, Nottingham, UK. <sup>35</sup>Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, UK. <sup>36</sup>Queen's Hospital, Romford, UK.

## Methods

**Cohort, sequencing and calling of genomic events.** 379 cases (69%) of our EAC cohort were derived from the EAC WGS ICGC study, for which samples are collected through the UK-wide Oesophageal Cancer Classification and Molecular Stratification (OCCAMS) consortium. The procedures for obtaining the samples, quality-control processes, extractions and WGS were as previously described<sup>17</sup>. Strict pathology consensus review was observed for these samples, with a 70% cellularity requirement before inclusion. Comprehensive clinical information was available for the ICGC–OCCAMS cases (Supplementary Table 13). In addition, previously published samples were included in the analysis from Dulak et al.<sup>19</sup> (149 whole-exome sequencing samples; 27%) and Nones et al.<sup>20</sup> (22 WGS samples; 4%), for a total of 551 genome-characterized EACs. RNA-seq data were available from our ICGC WGS samples (116 of 379 samples). BAM files for all samples (including those from Dulak et al.<sup>19</sup> and Nones et al.<sup>20</sup>) were run through our alignment (BWA-MEM), mutation (Strelka), copy number (ASCAT) and structural-variant (Manta) calling pipelines, as described<sup>17</sup>. Our methods were benchmarked against various other available methods and have among the best sensitivity and specificity for variant calling (ICGC benchmarking exercise<sup>49,50</sup>). Cell lines were subjected to WGS at 30× coverage with 150-bp paired-end reads on an Illumina HiSeq4000 instrument. Copy number calling was performed by Freec as described<sup>41</sup>. Mutations were called by GATK as described<sup>41</sup> and filtered for germline variants in the 1000 Genomes Project, and any known oncogenic hotspots<sup>23</sup> were recovered. Amplifications were defined as genes with 2× the median copy number of the host chromosome or greater.

Total RNA was extracted with an All Prep DNA/RNA kit from Qiagen, and the quality was checked on an Agilent 2100 Bioanalyzer with an RNA 6000 nano kit (Agilent). A Qubit High Sensitivity RNA assay kit from Thermo Fisher was used for quantification. Libraries were prepared from 250 ng RNA, with a TruSeq Stranded Total RNA Library Prep Gold (Ribo-zero) kit, and ribosomal RNA (nuclear, cytoplasmic and mitochondrial rRNA) was depleted with biotinylated probes that selectively bind rRNA molecules, forming probe–rRNA hybrids. These hybrids were pulled down with magnetic beads, and rRNA-depleted total RNA was reverse transcribed. The libraries were prepared according to Illumina's protocol<sup>21</sup>. Paired-end 75-bp sequencing on a HiSeq4000 instrument generated the paired-end reads. For normal expression controls, we chose gastric cardia tissue, from which some hypothesize Barrett's esophagus may arise, and duodenum with intestinal histology, including goblet cells, which mimics the histology of Barrett's esophagus. We did not use Barrett's esophagus tissue itself as a normal control, given the heterogeneous and plentiful phenotypic and genomic changes that it undergoes early in its pathogenesis.

**Analyzing EAC mutations for selection.** To detect positively selected mutations in our EAC cohort, a multitool approach across various selection-related 'features' (recurrence, functional impact and clustering) was implemented to provide a comprehensive analysis. This procedure is broadly similar to those of several previous approaches<sup>7,11</sup>. dNdScv<sup>21</sup>, MutSigCV<sup>8</sup>, e-Driver<sup>25</sup>, ActivedriverWGS<sup>22</sup> and e-Driver3D<sup>26</sup> were run with the default parameters. To run OncodriverFM<sup>23</sup>, Polyphen<sup>32</sup> and SIFT<sup>33</sup> were used to score the functional impact of each missense nonsynonymous mutation (from 0, indicating nonimpactful, to 1, indicating highly impactful); synonymous mutations were given a score of 0 impact, and truncating mutations (nonsense and frameshift mutations) were given a score of 1. Any gene that had fewer than seven mutations and was unlikely to contain detectable drivers with this method was not considered to decrease the FDR. OncodriverClust was run with a minimum cluster distance of 3, a minimum number of mutations for a gene to be considered of 7 and a stringent probability cutoff to find cluster seeds of  $P = 1 \times 10^{-13}$  to prevent infiltration of large numbers of probable false-positive genes. For all tool outputs, we undertook quality control including quantile–quantile plots to ensure that no tool produced inflated  $q$  values, and each tool produced at least 30% known cancer genes. Two tools were removed from the analysis, owing to the failure of both of these parameters in quality control in our hands (Activedriver<sup>24</sup> and Hotspot<sup>33</sup>). For three of the quality-control-approved tools (dNdScv, OncodriverFM and MutSigCV) for which it was possible, we also undertook an additional FDR-reducing analysis by recalculating  $q$  values on the basis of analysis of known cancer genes only<sup>21,27,28</sup>, as previously implemented<sup>21,55</sup>. Significance cutoffs were set at  $q < 0.1$  for coding genes. Tool outputs were then put through various filters to remove any further possible false-positive genes. Specifically, genes for which  $< 50\%$  of EAC cases had no expression (transcripts per kilobase million (TPM)  $< 0.1$ ) in our matched RNA-seq cohort were removed and, with dNdScv, genes with no or only a small mutation excess (observed/expected ratio  $> 1.5:1$ ) of any single mutation type were also removed. We also removed two mitochondrial genes (*MT-MD2* and *MT-MD4*) that were highly enriched in truncating mutations and were frequently called in OncodriverFM as well as other tools, possibly because of the different mutational dynamics caused by reactive oxygen species from the mitochondrial electron-transport chain and the high number of mitochondrial genomes per cell, which enables significantly more heterogeneity. These factors prevent the tools used from calculating an accurate null model for these genes, but they may be worthy of functional investigation. ActivedriverWGS calculates an expected background mutation rate on the basis of mutation rates of local, adjacent sequence for each tested element

while correcting for the differential mutation rates within each trinucleotide context; it thus tests observed mutation rates against this predicted background for each element. ActivedriverWGS also detects elements with mutations enriched in binding-site regions (high impact). For noncoding elements called by ActivedriverWGS, filtering for expression or  $dN/dS$  was not possible, and despite recent benchmarking<sup>29</sup>, such methods are not well established. Hence, we took a more cautious approach with general significance cutoffs of  $q < 0.001$  and  $q < 0.1$  for previously identified elements in other cancer types<sup>11</sup>.  $q$  values were not recalculated for previously identified elements alone, as with coding genes, but the  $q < 0.1$  cutoff was calculated on the basis of  $P$  values for all assessed elements. To calculate exome-wide mutational excess, we removed hypermutated cases ( $> 500$  exonic mutations) and applied the global nonsynonymous  $dN/dS$  ratios to all dNdScv-annotated mutations, excluding 'synonymous' and 'no SNV' annotations, as described in Martincorena et al.<sup>21</sup>.

**Detecting selection in copy number values.** ASCAT raw copy number values (CNVs) were used to detect frequently deleted or amplified regions of the genome with GISTIC2.0 (ref. 14). To determine which genes in these regions confer a selective advantage, we examined the correlation of CNVs from each gene within GISTIC-identified loci with TPM from matched RNA-seq in a subcohort of 116 samples and with mutations across all 551 samples. To call copy numbers in genes that spanned multiple copy number segments in ASCAT, we considered the total number of full copies of the gene (the lowest total copy number). Occasionally ASCAT is unable to confidently call the copy number in highly aberrant genomic regions. We found that the expression of genes in such regions matched well with what we would expect given the surrounding copy number, and hence we used the mean of the two adjacent copy number fragments to call copy number for the gene in question. We found that amplification peak regions identified by GISTIC2.0 varied significantly in precise location both in analysis of different subcohorts and in comparison to published GISTIC data from EACs<sup>9,15,16</sup>. A peak would often sit next to, but not overlap, a well-characterized oncogene or tumor suppressor. To account for this tendency, we widened the amplification peak sizes upstream and downstream by twice the size of each peak to ensure that we captured all possible drivers. Our expression analysis allowed us to then remove false positives from this wider region, and called drivers were still highly enriched in genes closer to the centers of GISTIC peak regions.

To detect genes for which amplification was correlated with increased expression, we compared the expression of samples with a high copy number for that gene (above the tenth-percentile copy number/ploidy) with those that had a normal copy number (median  $\pm 1$ ), by using the Wilcoxon rank-sum test with the specific alternative hypothesis that a high copy number would lead to increased expression.  $q$  values were then generated with the Benjamini–Hochberg method, not considering genes without significant expression in amplified samples (at least 75% amplified samples with TPM  $> 0.1$ ) and considering  $q < 0.001$  as significant. We also included an additional known driver-gene-only FDR-reduction analysis, as we previously described for mutational drivers, with  $q < 0.1$  considered significant, given the additional evidence of these genes in other cancer types. We also included *MYC* despite its  $P = 0.11$  for expression correlation resulting from frequent nonamplification-associated overexpression of *MYC* compared with the expression in normal controls. Otherwise, *MYC* was well evidenced for inclusion as an EAC driver by a proximity to the peak center (top four genes) and its high rate of amplification (19%). We used the same approach to detect genes for which homozygous deletion was correlated with expression loss, comparing cases with copy number = 0 to all others. Large expression modulation was a highly specific marker for known copy number driver genes and was not a widespread feature in most recurrently CNV genes. Whereas expression modulation is a requirement for selection of CNV-only drivers, it is not sufficient evidence alone, and hence we grouped such genes into those previously characterized as drivers in other cancer types (high-confidence EAC copy number drivers) and other genes (candidate EAC copy number drivers), which await functional validation. We used fragile-site regions detected in Wala et al.<sup>56</sup>. We also defined regions that might be recurrently heterozygously deleted, without any significant expression modulations, to allow for LOH of tumor-suppressor-gene mutations. To do so, we analyzed genes with at least five mutations for association between LOH (ASCAT minor allele = 0) and mutation with Fisher's exact test and generated  $q$  values with the Benjamini–Hochberg method. The analysis was repeated on known cancer genes only for decreased FDR, and  $q < 0.1$  was considered significant for both analyses. For those high-confidence drivers, we chose to define amplification as total copy number/ploidy (referred to as ploidy-adjusted copy number) because this procedure produces superior correlation with expression. We chose a cutoff for amplification at ploidy-adjusted copy number = 2, as has been previously used, thus resulting in a highly significant increase in expression in our copy number driver genes when amplified.

**Pathways and relative distributions of genomic events.** The relative distribution of driver events in each pathway was analyzed with Fisher's exact test in the case of pairwise comparisons including wild-type cases. In the case of multigene comparisons, such as those for cyclins, we calculated the  $P$  value and OR for each gene compared to all other genes in the group with a two-sided Fisher's exact test

with Benjamini–Hochberg correction, and combined the resulting  $q$  values with the Fisher method; genes without  $OR > 2$  for co-occurrence and  $< 0.5$  for mutual exclusivity were removed. For this analysis, we also removed highly mutated cases ( $> 500$  exonic mutations, 41 of 551), because they bias the distribution of mutations toward co-occurrence. To ensure that a nonrandom distribution of mutations across samples did not affect the strong co-occurrence of SWI–SNF genes (all genes  $q < 0.05$  before  $q$  values were combined), we repeated the analysis, randomly iterating 30,000 times over all the other eight driver–gene combinations (excluding SWI–SNF genes) and found that only 0.01% (4 of 30,000) of random combinations had all genes  $q < 0.05$ , as found in SWI–SNF genes. We then performed these analyses across all pairs of driver genes with two-sided Fisher's exact tests and Benjamini–Hochberg multiple-hypothesis correction ( $q$  values  $< 0.1$  are shown in Fig. 4b). We validated these relationships in independent TCGA cohorts of other gastrointestinal cancers in which we found cohorts with reasonable numbers of the genomic events in question (this procedure was not possible for *GATA4/GATA6*, for instance) with the cBioportal web interface tool<sup>37</sup>.

**Correlation of genomics with clinical phenotype.** To find genomic markers for prognosis, we performed univariate Cox regression for those driver genes present in  $> 5\%$  of cases ( $n = 16$ ) along with Benjamini–Hochberg false-discovery correction. We considered only these genes to reduce our FDR, because other genes were unlikely to affect clinical practice, given their low frequency in EAC. We validated *SMAD4* in the TCGA gastroesophageal cohort, which has a comparable frequency of these events but notably is composed mainly of gastric cancers, and *GATA4* in the TCGA pancreatic cohort with the cBioportal web interface tool. We also validated these markers as independent predictors of survival with respect to each other and to stage with a multivariate Cox regression in our 379 clinically annotated ICGC cohort. When assessing genomic correlates with differentiation phenotypes, we found only very few cases with well differentiated phenotypes ( $< 5\%$  of cases), and hence for statistical analyses, we collapsed these cases with moderate differentiation to allow a binary Fisher's exact test to compare poorly differentiated and well-differentiated or moderately differentiated phenotypes.

**Therapeutics.** The cancer-biomarker database was filtered for drugs linked to biomarkers found in EAC drivers, and Supplementary Table 8 was constructed with the cohort frequencies of EAC biomarkers. Ten EAC cell lines (SKGT4, OACP4C, OACM5.1, ESO26, ESO51, OE33, MFD, OE19, Flo-1 and JHesoAD) and three Barrett's esophagus high-grade dysplasia cell lines (CP-B, CP-C and CP-D) with WGS data<sup>41</sup> were used in proliferation assays to determine drug sensitivity to CDK4/CDK6 inhibitors, palbociclib (Biovision) and ribociclib (Selleckchem). Cell lines were grown in their normal growth media. Proliferation was measured with an Incucyte live-cell analysis system (Incucyte ZOOM Essen Biosciences). Each cell line was plated at a starting confluence of 10%, and the growth rate was measured over 4–7 d, depending on the basal proliferation rate (until 90% confluent in DMSO control). For each cell line–drug combination, concentrations of 16, 64, 250, 1,000 and 4,000 nM in 0.3% dimethylsulfoxide (DMSO) were used and compared to 0.3% DMSO only. Each condition was performed in at least triplicate (technical replicates) and for 12 of 12 randomly chosen cell lines, the drug combinations were successfully replicated with biological replicates (independent experiments). The time period of treatment to growth cessation in the control (0.3% DMSO) condition was used to calculate half-maximal growth inhibition ( $GI_{50}$ ) and area under the curve (AUC). Accurate  $GI_{50}$  values could not be calculated in cases in which a cell line had  $> 50\%$  proliferation inhibition even

with the highest drug concentration, and hence AUC was used to compare cell-line sensitivity. T47D had a highly similar  $GI_{50}$  for palbociclib to that previously calculated in other studies (112 nM versus 127 nM)<sup>38</sup>. Primary organoid cultures were derived from EAC cases included in the OCCAMS–ICGC sequencing study. Detailed organoid culture and derivation methods have been described<sup>44</sup>. Regarding the drug treatment, the seeding density for each organoid line was optimized to ensure cell growth in the logarithmic growth phase. Cells were seeded in complete medium for 24 h and then treated with compounds at five-point four-fold serial dilutions for 6 or 12 d. Cell viability was assessed with CellTiter-Glo (Promega) after drug incubation.

**Ethics.** The study was registered (UKCRNID 8880) and approved by the Institutional Ethics Committees (REC 07/H0305/52 and 10/H0305/1), and all subjects gave individual informed consent.

**Reporting Summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

## Code availability

Code associated with the analysis is available upon request.

## Data availability

The WGS and RNA expression data can be found at the European Genome-phenome Archive under accession numbers EGAD00001004417 and EGAD00001004423, respectively.

## References

- Ding, J. et al. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat. Commun.* **6**, 8554 (2015).
- Lee, A. Y. et al. Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biol.* **19**, 188 (2018).
- Nagai, K. et al. Differential expression profiles of sense and antisense transcripts between HCV-associated hepatocellular carcinoma and corresponding non-cancerous liver tissue. *Int. J. Oncol.* **40**, 1813–1820 (2012).
- Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **79**(7), 20 (2013).
- Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80 (2006).
- Reimand, J., Wagih, O. & Bader, G. D. The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* **3**, 2651 (2013).
- Northcott, P. A. et al. The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).
- Wala, J.A. et al. Selective and mechanistic sources of recurrent rearrangements across the cancer genome. Preprint at <https://www.biorxiv.org/content/early/2017/09/14/187609> (2017).
- Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
- Finn, R. S. et al. PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. *Breast Cancer Res.* **11**, R77 (2009).