# The landscape of selection in 551 Esophageal Adenocarcinomas defines genomic biomarkers for the clinic

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

[1] MRC cancer unit, Hutchison/MRC research centre, University of Cambridge, Cambridge, UK

[2] Cancer Research UK Cambridge institute, University of Cambridge, Cambridge, UK.

[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

[4] Department of Histopathology, Cambridge University Hospital NHS Trust, Cambridge, UK

†A full list of contributors from the OCCAMS Consortium is available at the end of the manuscript

*Corresponding author: rcf29@mrc-cu.cam.ac.uk

## Abstract:

Esophageal Adenocarcinoma (EAC) is a poor prognosis cancer type with rapidly rising incidence. Our understanding of genetic events which drive EAC development is limited and there are few molecular biomarkers for prognostication or therapeutics. Using cohort of 551 genomically characterised EACs with matched RNA-seq, we discover 77 EAC driver genes and 21 non-coding driver elements. We identify a mean of 4.4 driver events per tumour, derived more commonly from mutations rather than copy number alterations, and compare these mutations to the exome-wide mutational excess using dN/dS calculations. We observe mutual exclusivity or co-occurrence of events within and between a number of dysregulated EAC pathways suggestive of important functional relationships. Poor prognostic indicators (*SMAD4*, *GATA4*) are verified in independent cohorts with significant predictive

value. Over 50% of EACs contain sensitising events for CDK4/6 inhibitors which are highly correlated with clinically relevant sensitivity in a panel EAC cell lines and organoids.

## Introduction

Esophageal cancer is the eighth most common form of cancer world-wide and the sixth most common cause of cancer related death[1]. Esophageal Adenocarcinoma (EAC) is the predominant subtype in the west and incidence has been rapidly rising[2]. EAC is a highly aggressive neoplasm, usually presenting at a late stage and is generally resistant to chemotherapy, leading to five-year survival rates below 15%[3]. It is characterised by very high mutation rates in comparison to other cancer types[4] but also, paradoxically, by a paucity of recurrently mutated genes. EAC displays dramatic chromosomal instability and thus may be classified as a C-type neoplasm, which may be driven mainly by structural variation rather than mutations[5,6]. Currently our understanding of precisely which genetic events drive the development of EAC is limited and consequentially there is a paucity of molecular biomarkers for prognosis or targeted therapeutics available.

Methods to differentiate driver mutations from passenger mutations use features associated with known drivers to detect regions of the genomein which mutations are enriched for these features[8]. 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[9]. MutsigCV has identified ten known cancer genes as EAC drivers (*TP53*, *CDKN2A*, *SMAD4*, *ARID1A*, *ERBB2*, *KRAS*, *PIK3CA*, *SMARCA4*, *CTNNB1* and *FBXW7*) [6,10,11]. The PCAWG ICGC analysis also identified a significantly mutated enhancer associated with *TP53TG1*[12]. 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[13,14]. Similarly, detection of copy number driver events in EAC has relied on identifying regions of the genome recurrently deleted or amplified, as detected by

GISTIC[10,15-18]. However, GISTIC often identifies relatively large regions of the genome, with little indication of which specific gene-copy number aberrations (CNAs) may actually confer a selective advantage. There are also several non-selection based mechanisms which can cause recurrent CNAs, such as genomic fragile sites which have not been well differentiated from selection-based CNAs[19]. . Epigenetic events, for example methylation, may also be important sources of driver events in EAC but are much more difficult to assess formally for selection.

. To address these issues, we accumulated a cohort of 551 genomically characterised EACs using our esophageal ICGC project, which have high quality clinical annotation, associated whole genome sequencing (WGS) and RNA-seq on cases with sufficient material. We augmented our ICGC WGS cohort with publicly available whole exome[20] and whole genome sequencing[21] data and applied a number of complementary driver detection methodsto produce a comprehensive assessment of mutations and CNAs under selection in EAC. We use these events to define functional cell processes that have been selectively dysregulated in EAC and identify novel, verifiable and clinically relevant biomarkers for prognostication. Finally, we have used this compendium of EAC driver variants to provide an evidence base for targeted therapeutics, which we have tested *in vitro*.

# Results

**A 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 / Mb (supplementary figure 1), and 286,965 copy number aberrations (CNAs). We also identified 134,697 structural variants (SVs) 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 reliability of results (see methods). These features include highly recurrent mutations within a gene (dNdScv[22],

ActivedriverWGS[23], MutsigCV2[9]), high functional impact mutations within a gene (OncodriveFM[24], ActivedriverWGS[23]), mutation clustering (OncodriveClust[25], eDriver[26] and eDriver3D[27]) and recurrent amplification or deletion of genes (GISTIC[15]) undergoing concurrent over or under-expression (see methods) (Fig 1A)[8].

These complementary methods produced highly significant agreement in calling EAC driver genes, particularly within the same feature-type (supplementary figure 2) and on average more than half of the genes identified by one feature were also identified by other features (Fig 1B). In total seventy six EAC driver genes were discovered, 86% of which have not been detected in EAC previously[10,11,16-18,20] and 69% are known drivers in pan-cancer analyses [22,28,29]. To detect driver elements in the non-coding genome we used ActiveDriverWGS[23] a recently benchmarked[30] method using both functional impact and recurrence to determine driver status (Fig 1C, supplementary figure 3). We discovered 21 non-coding driver elements using this method. We have recovered several known non-coding driver elements from the pan-cancer PCAWG analysis[12] including the enhancer on chr7 linked to *TP53TG1*previously identified in EAC and the promoter/5'UTR regions of *PTDSS1* and *WRD74* were found in other cancer types. We also identified non-coding cancer driver elements novel to cancer including in the 5'UTR of *MMP24* and promoters of two related histones (*HIST1H2BO* and *HIST1H2AM*).

EAC is notable among cancer types for harbouring a high degree of chromosomal instability[21]. Using GISTIC we identified 149 recurrently deleted or amplified loci across the genome (Fig 2A, Supplementary tables 1 and 2). To determine which genes within these loci confer a selective advantage when they undergo CNAs we use a subset of 116 cases with matched RNA-seq to detect genes in which homozygous deletion or amplification causes a significant under or over-expression respectively (supplementary note and Supplementary tables 3-6). The majority of genes in these regions showed no significant CN associated expression change (74%), although work in larger cohorts suggests we may be underpowered to detect small expression changes[31]. 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 EAC drivers (see methods). 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, for example *ARID1A* and *CDH11*. .

In a subset of GISTIC loci, we observed extremely high copy number amplification, commonly greater than 100 copies, and these events were highly enriched in recurrently amplified regions containing driver genes rather than those which appear to contain only passengers (ploidy adjusted copy number >10, two sided Wilcox test, $P$ =4.97Ex10$^{-8}$) (supplementary figure 4). We use ploidy adjusted copy number to define amplifications as it produces superior correlation with expression data than absolute CN alone. Ploidy of our samples varies from 2-6 (3.5 on average) and hence ploidy adjusted copy number of >10 cut off translates into >20-60 absolute copies (on average 35 copies). To discern a mechanism for these ultra-high amplifications we assessed structural variants (SVs) associated with these events. For many of these events the extreme amplification was produced largely from a single copy number step the edges of which were supported by structural variants with ultra-high read support. Two examples are shown in Fig 2B and further randomly selected examples in supplementary figure 5. In the first example circularisation and amplification initially occurred around *MYC* but subsequently incorporated *ERBB2* from an entirely different chromosome and in the second an inversion has been followed by circularisation and amplification of *KRAS*. Such a pattern of extrachromosomal amplification via double minutes has been previously noted in EAC[21] and other neoplasms[32], and hence we refer to this amplification class with ultra-high amplification (ploidy adjusted copy number >10) as 'extrachromosomal-like'.

We found extrachromosomal-like amplifications had an extreme and highly penetrant effects on expression while moderate amplification (ploidy adjusted copy number > 2 but < 10) and homozygous deletion had highly significant (Wilcox test, two sided, $P$ = 9.62Ex10$^{-16}$ and $P$ = 7.64Ex10$^{-}$

[11] respectively) but less dramatic effects on expression with a lower penetrance (Fig 2C). This lack of penetrance was associated with low cellularity as calculated by ASCAT (Wilcox test, two sided,, overexpression cut off = 2.5x median FPKM, *P* = 0.011) in non-extrachromosomal-like amplified cases but also likely reflects that genetic mechanisms other than gene-dosage can modulate expression. We also detected several cases of overexpression or complete expression loss without associated CN changes reflecting non-genetic mechanisms for driver dysregulation. One case overexpressed *ERBB2* at 28-fold median expression however had entirely diploid CN in and surrounding *ERBB2* and a second case lost *SMAD4* expression (0.008-fold median expression) despite possessing five copies of SMAD4.

**Landscape of driver events in EAC**

The overall landscape of driver gene mutations and copy number alterations per case is depicted in Fig 3A. These comprise both oncogenes and tumour suppressor genes activated or repressed via different mechanisms. . 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 drivers, 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; ie splice site, nonsense and frameshift Indel mutations, but not missense mutations which are passengers. 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 (supplementary figure 6). Mutational hotspots in EAC or other cancer types[33] (supplementary table 7, Supplementary data) are indicated in Fig 3A. Novel EAC drivers of particular interest include *B2M*, a core component of the MHC class I complex and a marker of acquired resistance to Immunotherapy[34], *MUC6* a secreted

148 glycoprotein involved in gastric acid resistance and *ABCB1* a channel pump protein which is
149 associated with multiple instances of drug resistance[35]. We note that several of these drivers have
150 been previously associated with gastric and colorectal cancer (supplementary table 8)[14,36].

151 The identification of driver events provides a rich information about the molecular history of
152 each EAC tumour. We detect a median of five events in driver genes per tumour (IQR = 3-7, Mean =
153 5.6) and only a very small fraction of cases has no such events detected (6 cases, 1%). When we
154 remove the predicted percentage of passenger mutations using dN/dS ratios we find a mean of 4.4
155 true driver events per case which derive more commonly from mutations than CN events (Fig 3B,
156 Supplementary table 9). Using hierarchal clustering of drivers we noted that *TP53* mutant cases had
157 significantly more CN drivers (Wilcox test, two sided, $P$ = 0.0032, supplementary figures 7 and 8).
158 dNdScv, one of the driver gene detection methods used, also analyses the genome-wide excess of
159 non-synonymous mutations based on expected mutation rates to assess the mean of driver
160 mutations across the exome which is calculated at 5.4 (95% CIs: 3.5-7.3) in comparison to 2.7 in EAC
161 driver genes after passenger removal.

162 To better understand the functional impact of driver mutations we analysed expression of
163 driver genes with different mutation types and compared their expression to normal tissue RNA, (Fig
164 3C, supplementary figure 10). Since surrounding squamous epithelium is a fundamentally different
165 tissue, from which EAC does not directly arise, we have used duodenum and gastric cardia samples
166 as gastrointestinal phenotype controls, likely to be similar to the, as yet unconfirmed, tissue of origin
167 in EAC. A large number of driver genes have upregulated expression in comparison to normal
168 controls, for example *TP53* has upregulated RNA expression in WT tumour tissue and in cases with
169 non-truncating mutations but RNA expression is lost upon gene truncation. In depth analysis of
170 different *TP53* mutation types reveals significant heterogeneity within non-truncating mutations
171 (supplementary figure 9). Normal tissue expression of *CDKN2A* suggests that *CDKN2A* is generally
172 activated in EAC, likely due to genotoxic or other cancer-associated cellular stresses[37] and returns to

physiologically normal levels when deleted. Heterogeneous expression in WT *CDKN2A* cases suggest a different mechanism of inhibition, perhaps methylation, in some cases. Overexpression of some oncogenes occurs without genomic aberrations, such as *MYC*, which is overexpressed in MYC-wildtype EACs relative to normal tissues (Fig 3C). A smaller number of driver genes are downregulated in EACs without genomic aberrations. 3/4 of these genes (*GATA4*, *GATA6* and *MUC6*) are involved in the differentiated phenotype of gastrointestinal tissues and may be lost with tumour de-differentiation.

**Dysregulation of specific pathways and processes in EAC**

It is known that selection preferentially dysregulates certain functionally related groups of genes and biological pathways in cancer[40]. This phenomenon is highly evident in EAC, as shown in Fig 4 which depicts the functional relationships between EAC drivers (supplementary note). While *TP53* is the dominant driver in EAC, 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 β-catenin, which prevent degradation, or loss of Wnt destruction complex components like APC. Many different chromatin modifying genes, often belonging to the SWI/SNF complex, are also selectively mutated (28% of cases). In contrast to other pathways, SWI/SNF genes are co-mutated significantly more often than we would expect by chance (fisher's exact test, two sided, $q<0.05$ for each gene see methods), 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 (Fig 4B). Of particular note are co-occurring relationships between *TP53* and

198 *MYC*, *GATA6* and *SMAD4*, Wnt and Immune pathways as well as mutually exclusive relationships
199 between *ARID1A* and *MYC*, gastrointestinal (GI) differentiation and RTK pathways and SWI-SNF and
200 DNA-Damage response pathways. We were able to confirm some of these relationships in
201 independent cohorts in different cancer types (supplementary table 10) suggesting some of these
202 may be pan-cancer phenomenon. Wnt dysregulation was associated with hyper-mutated cases (>
203 500 exonic SNVs or Indels, fisher's exact test, two sided, $P = 2.98Ex10^{-5}$, OR= 9.3) as was mutation in
204 immune pathway genes (B2M and JAK1, > 500 exonic SNVs or Indels, fisher's exact test, two sided, *P*
205 $= 6.27Ex10^{-6}$, OR = 35.7).

206

207 **EAC driver events correlate with the clinical phenotype**

208 Events undergoing selection during cancer evolution influence tumour biology and thus impact
209 tumour aggressiveness, response to treatment and patient prognosis as well as other clinical
210 parameters.

211 Univariate Cox regression was performed for events in each driver gene with driver events
212 occurring in greater than 5% of EACs after passenger removal to detect prognostic biomarkers (Fig
213 5A). Events in two genes, present in 31% of EACs (Fig 5B), conferred significantly poorer prognosis
214 after multiple hypothesis correction, *GATA4* amplification (HR : 0.54 , 95% CI : 0.38 – 0.78, *P* value =
215 0.0008) and *SMAD4* mutation or homozygous deletion (HR : 0.60 , 95% CI : 0.42 – 0.84, *P* value =
216 0.003). Both genes remained significant in multivariate Cox regression including pathological TNM
217 staging, resection margin, curative vs palliative treatment intent and differentiation status (*GATA4* =
218 HR adjusted : 0.47, 95% CIs adjusted : 0.29 - 0.76, *P* value = 0.002 and *SMAD4* = HR adjusted : 0.61,
219 95% CI adjusted : 0.40 – 0.94, *P* value = 0.026). (Fig 5B). We validated the poor prognostic impact of
220 *SMAD4* events in an independent TCGA gastroesophageal cohort (HR = 0.58, 95% CI = 0.37 – 0.90, *P*
221 value =0.014) (Fig 5C) and we also found *GATA4* amplifications were prognostic in a cohort of TCGA
222 pancreatic cancers (HR = 0.38 95% CI: 0.18 – 0.80, *P* value = 0.011) (Fig 5D), the only available cohort

containing a feasible number of *GATA4* amplifications. The prognostic impact of *GATA4* has been suggested in previously published independent EAC cohort[17] although it did not reach statistical significance after fdr correction and SMAD4 expression loss has been previously linked to poor prognosis in EAC[41]. We also noted stark survival differences between cases with *SMAD4* events and cases in which TGFβ receptors were mutated (Fig 5E, 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[42].

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 found Wnt pathway mutations had a strong association with well differentiated tumours (*P* = 0.001, OR = 2.9, fisher's test, two sided, see methods, Fig 5F).. Female cases (n=81) were enriched for *KRAS* mutation (*P* = 0.001, fisher's exact test, two sided) and *TP53* wildtype status (*P* = 0.006, fisher's exact test, two sided) (Fig 5G). This is of particular interest given the male predominance of EAC[3].

**Targeted therapeutics using EAC driver events.**

To investigate whether the driver events in particular genes and/or pathways might sensitise EAC cells to certain targeted therapeutic agents we used the Cancer Biomarkers database[43]. We calculated the percentage of our cases which contain EAC-driver biomarkers of response to each drug class in the database (Fig 6A, and full data supplementary table 11). Aside from TP53, which has been problematic to target clinically so far, we found a number of drugs with predicted sensitivity in >10% of EACs including EZH2 inhibitors for SWI/SNF mutant cancers (23%, and 28% including other SWI/SNF EAC 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 >50% of cases harboured sensitivity

causing events in the receptor tyrosine kinase (RTK) and core cell cycle pathways (eg in *CCND1*, *CCND3* and *KRAS*).

To verify that these driver events would also sensitise EAC tumours to such inhibitors we used a panel of thirteen EAC or Barrett's HGD cell lineswhich have undergone whole genome sequencing[46] and assessed them for presence of EAC driver events (Fig 6B). The mutational landscape of these lines was broadly representative of EAC 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 (Fig 6B, supplementary table 12, supplementary figure 11)[47]. Such EAC cell lines had comparable sensitivity to T47D which is derived from an ER +ve breast cancer where CDK4/6 inhibitors have been FDA approved. We noted three cell lines 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[48]. To verify these effects in a more representative model of EAC we treated three whole genome sequenced EAC organoid cultures[49] 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 (Fig 6C, supplementary figure 10).

## Discussion

We present here a detailed catalogue of coding and non-coding genomic events that have been selected for during the evolution of esophageal 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. We have used this set of biologically important gene alterations to identify prognostic biomarkers and actionable genomic events for personalised medicine.

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 expression changes which were detected in the full 551 cohort, due to lack of representation RNA matched sub cohort. Despite rigorous analyses to detect selected events, assessment of the global excess of mutations by dNdScv suggests we are unable to detect all mutations selected in EAC, similar to many other cancer types[22]. All driver gene detection methods which we have used rely on driver mutation re-occurrence in a gene to some degree. Many of these undetected driver mutations are hence likely to be spread across a large number of genes whereby each is mutated at very low frequency across EAC patients. This tendency for low frequency EAC drivers may be responsible for the low yield of MutsigCV in previous cohorts and may suggests that C-type cancers such as EAC, are not less 'mutation-driven' than M-type cancers but rather that their mutational drivers are spread across a larger number of genes[5]. CN driver gene identification is even more challenging due to the large size and lower frequency of these events and hence is it also possible there are significantly more EAC CN drivers yet to discovered, possibly already identified as candidates here..

While a number of previous reports have attempted to detect EAC drivers, they have had a limited yield per case. The first such study[20] used methods which, despite being well regarded at the time, were subsequently discredited[9]. Since then a number of reports, including our own, on medium and large cohort sizes using MutsigCV[10,11,18] were only able to detect a small number of mutational driver genes (7, 5 and 15 in each study). By using both a large cohort and more comprehensive methodologies we have significantly increased this figure to 66 mutational driver genes (excluding CN drivers). Detection of driver CNAs has previously relied on GISTIC to detect

296 recurrently CN aberrant regions[10,15-18] but no analyses have been performed to evidence which
297 genes in these large regions are true drivers. Many of the genes annotated by such papers are
298 unlikely to be CN drivers due to their lack of expression modulation with CNAs (eg *YEATS4* and
299 *MCL1*), the role of recurrent heterozygous losses to drive LOH in some mutational drivers (*ARID1A*
300 and *CDH11*) or their association with fragile sites (*PDE4D*, *WWOX*, *FHIT*). Conversely, we have been
301 able to identify novel EAC copy number drivers (eg *CCND3*, *AXIN1*, *PPM1D* and *APC*).

302 We have noted a three-way association between hyper-mutation, Wnt activation and loss of
303 Immune signalling genes such as *B2M*. MSI-driven hyper-mutation has been previously associated
304 with higher immune activity[52,53]. However, Wnt dysregulation and mutation of immune pathway
305 genes such as *B2M*[34] have been previously linked to immune escape[54] suggesting this may be an
306 acquired mechanism to prevent immune surveillance caused by hyper-mutation.

307 Functional characterisation of many of the driver genes described is needed to understand
308 why they are advantageous to EAC tumours and how they modify EAC biology.. Biological pathways
309 and processes that are selectively dysregulated deserve particular attention in this regard as do the
310 gene pairs or groups with mutually exclusive or co-occurring relationships such as *MYC* and *TP53* or
311 SWI/SNF factors, suggestive of particular functional relationships. Prospective clinical work to verify
312 and implement *SMAD4* and *GATA4* biomarkers in this study would be worthwhile. While EAC is a
313 poor prognosis cancer type, significant heterogeneity of survival outcome makes triaging patients in
314 treatment groups an important part of clinic practice which could be improved using better
315 prognostication. A number of targeted therapeutics may provide clinic benefit to EAC cases based on
316 their individual genomic profile. In particular CDK4/6 inhibitors deserve considerable attention as an
317 option for EAC treatment as they are, by a significant margin, the treatment to which the most EACs
318 harbour sensitivity-causing driver events, excluding TP53 as an unlikely therapeutic biomarker.
319 Previously work has noted activity of the CDK4/6 inhibitor Palbociclib in a small number of EAC cell
320 lines[57], however biomarkers were not investigated. The extensive *in vitro* validation of identified

biomarkers for CDK4/6 inhibitors in EAC across sixteen cell lines and organoids is persuasive of possible clinical benefit using a targeted approach.

In summary this work provides a detailed compendium of mutations and copy number alterations undergoing selection in EAC which have clinically relevant impact on tumour behaviour. This comprehensive study provides us with useful insights into the nature of EAC tumours and should pave the way for evidence based clinical trials in this poor prognosis disease.

## References


1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer.* Mar 1 2015;136(5):E359-386.
2. Coleman HG, Xie SH, Lagergren J. The Epidemiology of Esophageal Adenocarcinoma. *Gastroenterology.* Jan 2018;154(2):390-405.
3. Smyth EC, Lagergren J, Fitzgerald RC, et al. Oesophageal cancer. *Nat Rev Dis Primers.* Jul 27 2017;3:17048.
4. Campbell PJ, Getz G, Stuart JM, Korbel JO, Stein LD. Pan-cancer analysis of whole genomes. *bioRxiv.* 2017.
5. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet.* Oct 2013;45(10):1127-1133.
6. Secrier M, Li X, de Silva N, et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat Genet.* Oct 2016;48(10):1131-1141.
7. Stratton MR, Futreal PA. Cancer: understanding the target. *Nature.* Jul 1 2004;430(6995):30.
8. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep.* Oct 2 2013;3:2650.
9. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* Jul 11 2013;499(7457):214-218.
10. Integrated genomic characterization of oesophageal carcinoma. *Nature.* Jan 12 2017;541(7636):169-175.
11. Lin DC, Dinh HQ, Xie JJ, et al. Identification of distinct mutational patterns and new driver genes in oesophageal squamous cell carcinomas and adenocarcinomas. *Gut.* Aug 31 2017.
12. Rheinbay E, Nielsen MM, Abascal F, et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. *bioRxiv.* 2017.
13. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature.* Mar 20 2014;507(7492):315-322.
14. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature.* Sep 11 2014;513(7517):202-209.
15. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 2011;12(4):R41.

16. Dulak AM, Schumacher SE, van Lieshout J, et al. Gastrointestinal adenocarcinomas of the esophagus, stomach, and colon exhibit distinct patterns of genome instability and oncogenesis. *Cancer Res.* Sep 1 2012;72(17):4383-4393.
17. Frankel A, Armour N, Nancarrow D, et al. Genome-wide analysis of esophageal adenocarcinoma yields specific copy number aberrations that correlate with prognosis. *Genes Chromosomes Cancer.* Apr 2014;53(4):324-338.
18. Secrier M, Fitzgerald RC. Signatures of Mutational Processes and Associated Risk Factors in Esophageal Squamous Cell Carcinoma: A Geographically Independent Stratification Strategy? *Gastroenterology.* May 2016;150(5):1080-1083.
19. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet.* Oct 2013;45(10):1134-1140.
20. Dulak AM, Stojanov P, Peng S, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet.* May 2013;45(5):478-486.
21. Nones K, Waddell N, Wayte N, et al. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nat Commun.* Oct 29 2014;5:5224.
22. Martincorena I, Raine KM, Gerstung M, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell.* Nov 16 2017;171(5):1029-1041 e1021.
23. Wadi L, Uuskula-Reimand L, Isaev K, et al. Candidate cancer driver mutations in super-enhancers and long-range chromatin interaction networks. *bioRxiv.* 2017.
24. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* Nov 2012;40(21):e169.
25. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics.* Sep 15 2013;29(18):2238-2244.
26. Porta-Pardo E, Godzik A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics.* Nov 1 2014;30(21):3109-3114.
27. Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res.* Jan 2015;43(Database issue):D968-973.
28. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer.* Mar 2004;4(3):177-183.
29. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. *Nature.* Oct 17 2013;502(7471):333-339.
30. Shuai S, Gallinger S, Stein LD. DriverPower: Combined burden and functional impact tests for cancer driver discovery. *bioRxiv.* 2017.
31. Taylor AM, Shih J, Ha G, et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer cell.* Apr 9 2018;33(4):676-689 e673.
32. Turner KM, Deshpande V, Beyter D, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature.* Mar 2 2017;543(7643):122-125.
33. Chang MT, Asthana S, Gao SP, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol.* Feb 2016;34(2):155-163.
34. Zaretsky JM, Garcia-Diaz A, Shin DS, et al. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *The New England journal of medicine.* Sep 1 2016;375(9):819-829.
35. Chen Z, Shi T, Zhang L, et al. Mammalian drug efflux transporters of the ATP binding cassette (ABC) family in multidrug resistance: A review of the past decade. *Cancer Lett.* Jan 1 2016;370(1):153-164.
36. Giannakis M, Mu XJ, Shukla SA, et al. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell reports.* Oct 18 2016;17(4):1206.

37. Pei XH, Xiong Y. Biochemical and cellular mechanisms of mammalian CDK inhibitors: a few unresolved issues. *Oncogene.* Apr 18 2005;24(17):2787-2795.
38. Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* Jul 18 2012;487(7407):330-337.
39. Waddell N, Pajic M, Patch AM, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature.* Feb 26 2015;518(7540):495-501.
40. Leiserson MD, Vandin F, Wu HT, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet.* Feb 2015;47(2):106-114.
41. Singhi AD, Foxwell TJ, Nason K, et al. Smad4 loss in esophageal adenocarcinoma is associated with an increased propensity for disease recurrence and poor survival. *Am J Surg Pathol.* Apr 2015;39(4):487-495.
42. Levy L, Hill CS. Alterations in components of the TGF-beta superfamily signaling pathways in human cancer. *Cytokine Growth Factor Rev.* Feb-Apr 2006;17(1-2):41-58.
43. Tamborero D, Rubio-Perez C, Deu-Pons J, et al. Cancer Genome Interpreter Annotates The Biological And Clinical Relevance Of Tumor Alterations. *bioRxiv.* 2017.
44. Weaver JMJ, Ross-Innes CS, Shannon N, et al. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nature genetics.* Aug 2014;46(8):837-843.
45. Ross-Innes CS, Becq J, Warren A, et al. Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nature genetics.* Sep 2015;47(9):1038-1046.
46. Contino G, Eldridge MD, Secrier M, et al. Whole-genome sequencing of nine esophageal adenocarcinoma cell lines. *F1000Res.* 2016;5:1336.
47. Liston DR, Davis M. Clinically Relevant Concentrations of Anticancer Drugs: A Guide for Nonclinical Studies. *Clin Cancer Res.* Jul 15 2017;23(14):3489-3498.
48. Herrera-Abreu MT, Palafox M, Asghar U, et al. Early Adaptation and Acquired Resistance to CDK4/6 Inhibition in Estrogen Receptor-Positive Breast Cancer. *Cancer Res.* Apr 15 2016;76(8):2301-2313.


**Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium:**


Rebecca C. Fitzgerald[1], Ayesha Noorani[1], Paul A.W. Edwards[1,2], Nicola Grehan[1], Barbara Nutzinger[1], Caitriona Hughes[1], Elwira Fidziukiewicz[1], Jan Bornschein[1], Shona MacRae[1], Jason Crawte[1], Alex Northrop[1], Gianmarco Contino[1], Xiaodun Li[1], Rachel de la Rue[1], Maria O'Donovan[1,4], Ahmad Miremadi[1,4], Shalini Malhotra[1,4], Monika Tripathi[1,4], Simon Tavaré[2], Andy G. Lynch[2], Matthew Eldridge[2], Maria Secrier[5], Lawrence Bower[2], Ginny Devonshire[2], Juliane Perner[2], SriGanesh Jammula[2], Jim Davies[6], Charles Crichton[6], Nick Carroll[7], Peter Safranek[7], Andrew Hindmarsh[7], Vijayendran Sujendran[7], Stephen J. Hayes[8,15], Yeng Ang[8,9,30], Shaun R. Preston[10], Sarah Oakes[10], Izhar Bagwan[10], Vicki Save[11], Richard J.E. Skipworth[11], Ted R. Hupp[11], J. Robert O'Neill[11,24], Olga Tucker[12,34], Andrew Beggs[12,29], Philippe Taniere[12], Sonia Puig[12],Timothy J. Underwood[13,14], Fergus Noble[13], Jack Owsley[13], Hugh Barr[16], Neil Shepherd[16], Oliver Old[16], Jesper Lagergren[17,26], James Gossage[17,25], Andrew Davies[17,25,], Fuju Chang[17,25], Janine Zylstra[17,25], Ula Mahadeva[17], Vicky Goh[25], Francesca D. Ciccarelli[25], Grant Sanders[18], Richard Berrisford[18], Catherine Harden[18], Mike Lewis[19], Ed Cheong[19], Bhaskar Kumar[19],Simon L Parsons[20], Irshad Soomro[20], Philip Kaye[20], John Saunders[20], Laurence Lovat[21], Rehan Haidry[21], Laszlo Igali[22], Michael Scott[23], Sharmila Sothi[27], Sari Suortamo[27], Suzy Lishman[28], George B. Hanna[32], Krishna Moorthy[32], Christopher J. Peters[32], Anna Grabowska[33], Richard Turkington[35.]

[5] UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, WC1E 6BT, London, UK

[6]Department of Computer Science, University of Oxford, UK, OX1 3QD
[7]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK, CB2 0QQ
[8]Salford Royal NHS Foundation Trust, Salford, UK, M6 8HD
[9]Wigan and Leigh NHS Foundation Trust, Wigan, Manchester, UK, WN1 2NN
[10]Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK, GU2 7XX
[11]Edinburgh Royal Infirmary, Edinburgh, UK, EH16 4SA
[12]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK, B15 2GW
[13]University Hospital Southampton NHS Foundation Trust, Southampton, UK, SO16 6YD
[14]Cancer Sciences Division, University of Southampton, Southampton, UK, SO17 1BJ
[15]Faculty of Medical and Human Sciences, University of Manchester, UK, M13 9PL
[16]Gloucester Royal Hospital, Gloucester, UK, GL1 3NN
[17] Guy's and St Thomas's NHS Foundation Trust, London, UK, SE1 7EH
[18]Plymouth Hospitals NHS Trust, Plymouth, UK, PL6 8DH
[19]Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, UK, NR4 7UY
[20]Nottingham University Hospitals NHS Trust, Nottingham, UK, NG7 2UH
[21]University College London, London, UK, WC1E 6BT
[22]Norfolk and Waveney Cellular Pathology Network, Norwich, UK, NR4 7UY
[23]Wythenshawe Hospital, Manchester, UK, M23 9LT
[24]Edinburgh University, Edinburgh, UK, EH8 9YL
[25]King's College London, London, UK, WC2R 2LS
[26]Karolinska Institutet, Stockholm, Sweden, SE-171 77
[27]University Hospitals Coventry and Warwickshire NHS, Trust, Coventry, UK, CV2 2DX
[28]Peterborough Hospitals NHS Trust, Peterborough City Hospital, Peterborough, UK, PE3 9GZ
[29]Institute of Cancer and Genomic sciences, University of Birmingham, B15 2TT
[30]GI science centre, University of Manchester, UK, M13 9PL.
[31]Queen's Medical Centre, University of Nottingham, Nottingham, UK, NG7 2UH
[32]Department of Surgery and Cancer, Imperial College London, UK, W2 1NY
[33]Queen's Medical Centre, University of Nottingham, Nottingham, UK
[34]Heart of England NHS Foundation Trust, Birmingham, UK, B9 5SS.
[35]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Northern Ireland, UK, BT7 1NN.

## Author contributions:

RCF and AMF conceived the overall study. AMF and SJ analysed the genomic data and performed statistical analyses. RCF, AMF and XL designed the experiments. AMF, XL and JM performed the experiments. GC contributed to the Structural variant analysis and data visualisation. SK helped compile the clinical data and aided statistical analyses. JP and SA produced and QC'ed the RNA-seq

data. EO aided the whole genome sequencing of EAC cell lines. SM and NG coordinated the clinical centres and were responsible for sample collections. ME benchmarked our mutation calling pipelines. MO led the pathological sample QC for sequencing. LB and GD constructed and managed the sequencing alignment and variant calling pipelines. RCF and ST supervised the research. RCF and ST obtained funding. AMF and RCF wrote the manuscript. All authors approved the manuscript.

**The authors declare no competing interests.**

## Data availability

**The WGS and RNA expression data can be found at the European Genome-phenome Archive (EGA) under accessions EGAD00001004417 and EGAD00001004423 respectively.**

## Code availability

**Code associated with the analysis is available upon request.**

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

**OCCAMS was funded by a programme grant from Cancer Research UK (RG66287) and the Fitzgerald laboratory is funded by a Core Programme Grant from the Medical Research Council. We thank the Human Research Tissue Bank, which is supported by the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre, from**

**Addenbrooke's Hospital. Additional infrastructure support was provided from the CRUK funded Experimental Cancer Medicine Centre.**

## Acknowledgements

We would like to thank Bass AJ and Waddell N for providing data in Dulak et al 2013 and Nones et al 2014 respectively, also included in our previous publication Secrier et al 2016. Inclusion of this data allowed augmentation of our ICGC cohort and a greater sensitivity for the detection of EAC driver variants.

## Figure Legends:

**Figure 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**. Non-coding driver elements detected and their element types.

**Figure 2. Copy number variation under positive selection. a.** Recurrent copy number changes across the genome identified by GISTIC. 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 and deletion peaks respectively and occasions of significant association between LOH and mutation are indicated in green. Purple deletion

peaks indicate fragile sites. **b.** Examples of Extrachromosomal-like amplifications 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*. **c.** Relationship between copy number and expression in CN driver genes.

**Figure 3. The driver gene landscape of EAC. a.** 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). **b.** Assessment of driver event types per case and comparison to exome-wide excess of mutations generated by dNdScv. **c.** Expression changes in EAC driver genes in comparison to normal intestinal tissues. Genes with expression changes of note are shown**.**

**Figure 4. Biological pathways undergoing selective dysregulation in EAC. a.** Biological Pathways dysregulated by driver gene mutation and/or CNVs. WT cases for a pathway are not shown. Inter and intra-pathway interactions are described and mutual exclusivities and/or associations between genes in a pathway are annotated. *GATA4*/*GATA6* amplifications have a mutually exclusive relationship although this does not reach statistical significance (fisher's exact test, two sided, *P* = 0.07 OR = 0.52). **b.** Pairwise assessment of mutual exclusivity and association in EAC driver genes and pathways.

**Figure 5. Clinical significance of driver events in EAC. a.** Hazard rations 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 EACs with different status of significant prognostic indicators (*GATA4* and *SMAD4*). **c.** Kaplan-Meier curves for different alterations in the TGF-β pathway. **d.** Kaplan-Meier curves showing verification *GATA4* prognostic value in GI cancers using a pancreatic TCGA cohort**. e.** Kaplan-Meier curves showing verification *SMAD4* prognostic value in Gastroesophageal cancers using a gastroesophageal TCGA cohort. **f.** Differentiation bias in tumours containing events in Wnt pathway driver genes**. g.** Relative frequency of *KRAS* mutations and *TP53* mutations driver gene events in females vs males (fishers exact test).

**Figure 6. CDK4/6 inhibitors in EAC. a.** Drug classes for which sensitivity is indicated by EAC driver genes with data from the Cancer Biomarkers database[36]**. b.** Area under the curve (AUC) of sensitivity is shown in a panel of 13 EAC and Be 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). c. Response of organoid cultures to three FDA approved CDK4/6 inhibitors and corresponding driver events.

# Methods

### Cohort, sequencing and calling of genomic events

379 cases (69%) of our EAC cohort were derived from the esophageal adenocarcinoma WGS ICGC study, for which samples are collected through the UK wide OCCAMS (Oesophageal Cancer

591 Classification and Molecular Stratification) consortium. The procedures for obtaining the samples,
592 quality control processes, extractions and whole genome sequencing are as previously described[18].
593 Strict pathology consensus review was observed for these samples with a 70% cellularity
594 requirement before inclusion. Comprehensive clinical information was available for the ICGC-
595 OCCAMS cases (Supplementary Table 13). In addition, previously published samples were included in
596 the analysis from Dulak et al 2013[20] – 139 WES and 10 WGS (total 27%) and Nones et al 2014[21] with
597 22 WGS samples (4%) to total 551 genome characterised EACs. RNA-seq data was available from our
598 ICGC WGS samples (116/379). BAM files for all samples (include those from Dulak et al 2013 and
599 Nones et al 2014) were run through our alignment (BWA-MEM), mutation (Strelka), copy number
600 (ASCAT) and structural variant (Manta) calling pipelines, as previously described[18]. Our methods
601 were benchmarked against various other available methods and have among the best sensitivity and
602 specificity for variant calling (ICGC benchmarking exercise[58]). Mutation and copy number calling on
603 cell lines was performed as previously described[46]. Amplifications were defined as genes with 2x the
604 average copy number of the host chromosome or greater.

605 Total RNA was extracted using All Prep DNA/RNA kit from Qiagen and the quality was checked on
606 Agilent 2100 Bioanalyzer using RNA 6000 nano kit (Agilent). Qubit High sensitivity RNA assay kit from
607 thermo fisher was used for quantification. Libraries were prepared from 250ng RNA, using TruSeq
608 Stranded Total RNA Library Prep Gold (Ribo-zero) kit and ribosomal RNA (nuclear, cytoplasmic and
609 mitochondrial rRNA) was depleted, whereby biotinylated probes selectively bind to ribosomal RNA
610 molecules forming probe-rRNA hybrids. These hybrids were pulled down using magnetic beads and
611 rRNA depleted total RNA was reverse transcribed. The libraries were prepared according to Illumina
612 protocol[59]. Paired end 75bp sequencing on HiSeq4000 generated the paired end reads. For normal
613 expression controls we chose gastric cardia tissue, from which some hypothesise Barrett's may arise,
614 and duodenum which contains intestinal histology, including goblet cells, which mimics that of
615 Barrett's. We did not use Barrett's tissue itself as a normal control given the heterogeneous and
616 plentiful phenotypic and genomic changes which it undergoes early in its pathogenesis.

**Analysing EAC mutations for selection**

To detect positively selected mutations in our EAC cohort, a multi-tool approach across various selection related 'features' (recurrence, functional impact, clustering) was implemented in order to provide a comprehensive analysis. This is broadly similar to several previous approaches[8,12]. dNdScv[22], MutsigCV[9], e-Driver[26], ActivedriverWGS and e-Driver3D[27] were run using the default parameters. To run OncodriverFM[24], Polyphen[60] and SIFT[61] were used to score the functional impact of each missense non-synonymous mutation (from 0, non-impactful to 1 highly impactful), synonymous mutation were given a score of 0 impact and truncating mutations (Nonsense and frameshift mutations) were given a score of 1. Any gene with less than 7 mutations, unlikely to contain detectable drivers using this method, was not considered to decrease the false discovery rate. OncodriveClust 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 $P = Ex10^{-13}$ 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[62] and Hotspot[33]). For three of the QC-approved tools (dNdScv, OncodriveFM, MutsigCV) where this was possible we also undertook an additional fdr reducing analysis by re-calculating q values based on analysis of known cancer genes only[22,28,29] as has been previously implemented[22,63]. 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 EAC cases had no expression (TPM<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 truncating

mutations and were frequently called in OncodriveFM as well as other tools. This is may be due to the different mutational dynamics, caused by ROS 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 however they may be worthy of functional investigation. ActiveDriverWGS calculates an expected background mutation rate based on 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 non-coding elements called by ActivedriverWGS filtering for expression or dN/dS was not possible and despite recent benchmarking[30] such methods are not so well established. Hence we took a more cautious approach with general significance cut offs of $q < 0.001$ and $q < 0.1$ for previously identified elements in other cancer types[12]. Q values were not recalculated for previously identified elements alone but the $q < 0.1$ cut off was calculated based on p values for all assessed 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[22].

**Detecting selection in CNVs**

ASCAT raw CN values were used to detected frequently deleted or amplified regions of the genome using GISTIC2.0[15]. To determine which genes in these regions confer a selective advantage, CNVs 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 (ie the lowest total copy number). Occasionally ASCAT is unable to confidently call

667 the copy number in highly aberrant genomic regions. We found that the expression of genes in such
668 regions matched well what we would expect given the surrounding copy number and hence we used
669 the mean of the two adjacent copy number fragments to call copy number in the gene in question.
670 We found amplification peak regions identified by GISTIC2.0 varied significantly in precise location
671 both in analysis of different sub-cohorts and when comparing to published GISTIC data from
672 EACs[10,16,17]. A peak would often sit next to but not overlapping a well characterised oncogene or
673 tumour suppressor. To account for this, we widened the amplification peak sizes upstream and
674 downstream by twice the size of each peak to ensure we captured all possible drivers. Our
675 expression analysis allows us to then remove false positives from this wider region and called drivers
676 were still highly enriched for genes closer to the centre of GISTIC peak regions.

677 To detect genes in which amplification correlated with increased expression we compared
678 expression of samples with a high CN for that gene (above 10$^{th}$ percentile CN/ploidy) with those
679 which have a normal CN (median +/- 1) using the Wilcox rank-sum test and using the specific
680 alternative hypothesis that high CN would lead to increased expression. Q-values were then
681 generated based on Benjamini & Hochberg method, not considering genes without significant
682 expression in amplified samples (at least 75% amplified samples with TPM > 0.1) and considering
683 q<0.001 as significant. We also included an additional known driver gene only fdr reduction analysis
684 as previously described for mutational drivers with q<0.1 considered as significant given the
685 additional evidence for these genes in other cancer types. We also included MYC despite its p= 0.11
686 for expression correlation. This is due to frequent non-amplification associated overexpression of
687 *MYC* when compared to normal controls and otherwise *MYC* is well evidence by a very close
688 proximity to the peak centre (top 4 genes) and its high rate of amplification (19%). We took the
689 same approach to detect genes in which homozygous deletion correlated with expression loss,
690 comparing cases with CN = 0 to all others. Large expression modulation was a highly specific marker
691 for known CN driver genes and was not a widespread feature in most recurrently copy number
692 variant genes. However, while expression modulation is a requirement for selection of CNV only

693 drivers, it is not sufficient evidence alone and hence we grouped such genes into those which have
694 been characterised as drivers previously in other cancer types (high confidence EAC CN drivers) and
695 other genes (Candidate EAC CN drivers) which await functional validation. We used fragile site
696 regions detected in Wala et al 2017[64]. We also defined regions which may be recurrently
697 heterozygous deleted, without any significant expression modulations, to allow LOH of tumour
698 suppressor gene mutations. To do this we analysed genes with at least 5 mutations for association
699 between LOH (ASCAT minor allele = 0) and mutation using fisher's exact test and generated q values
700 using the Benjamini & Hochberg method. The analysis was repeated on known cancer genes only for
701 reduced fdr and $q < 0.1$ considered significant for both analyses. For those high confidence drivers
702 we chose to define amplification as CN/ploidy (referred to as loidy adjusted copy number) because
703 this produces superior correlation with expression. We chose a cut off for amplification at CN/ploidy
704 = 2 as has been previously used, and causes a highly significant increase in expression in our CN-
705 driver genes when amplified.

706

707 **Pathways and relative distributions of genomic events**

708 The relative distribution of driver events in each pathway was analysed using a fisher's exact test in
709 the case of pair-wise comparisons including WT cases. In the case of multi-gene comparisons such as
710 the cyclins we calculate the p value and odds ratio for gene in the group using a two sided fisher's
711 exact test, correct by Benjamini & Hochberg, and combine resulting q values using the Fisher
712 method, genes without odds ratios > 2 for cooccurrence and < 0.5 for mutual exclusivity were
713 removed. For this analysis we also remove highly mutated cases (>500 exonic mutations, 41/551) as
714 they bias distribution of genes towards co-occurrence. To ensure that a non-random distribution of
715 mutations across samples was still not affecting the strong co-occurrence of SWI/SNF genes (all
716 genes $q < 0.05$ before combining q values) we repeated the analysis randomly iterating 30,000 times
717 over other gene eight combinations (excluding SWI/SNF genes) and found only 0.01% (4/30,000) of

718 random combinations had all genes q < 0.05 as found in SWI/SNF genes. We then performed this
719 analyses across all pairs of driver genes using two sided Fisher's exact tests and Benjamini &
720 Hochberg multiple hypothesis correction, q values < 0.1 are shown in Fig4B. We validated these
721 relationships in independent TGCA cohorts of other GI cancers where we could find cohorts with
722 reasonable numbers of the genomic events in question (not possible for GATA4/6 for instance) using
723 the cBioportal web interface tool[65].

724

725

726 **Correlating genomics with the clinical phenotype**

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

739

740 **Therapeutics**

741 The cancer biomarker database was filtered for drugs linked to biomarkers found in EAC drivers and
742 supplementary table 8 constructed using the cohort frequencies of EAC biomarkers. 10 EAC cell lines
743 (SKGT4, OACP4C, OACM5.1, ESO26, ESO51, OE33, MFD, OE19, Flo-1 and JHesoAD) and 3 BE high
744 grade dysplasia cell lines (CP-B, CP-C and CP-D) with WGS data[46] were used in proliferation assays to
745 determine drug sensitivity to CDK4/6 inhibitors, Palbociclib (Biovision) and Ribociclib (Selleckchem).
746 Cell lines were grown in their normal growth media. Proliferation was measured using the Incucyte
747 live cell analysis system (Incucyte ZOOM Essen biosciences). Each cell line was plated at a starting
748 confluency of 10% and growth rate measured across 4-7 days depending on basal proliferation rate
749 (until 90% confluent). For each cell-line drug combination concentrations of 16, 64, 250, 1000 and
750 4000 nanomolar were used each in 0.3% DMSO and compared to 0.3% DMSO only. Each condition
751 was performed in at least triplicate (technical replicates) and 12/12 randomly chosen cell line – drug
752 combinations were successfully replicated with biological replicates (independent experiments). The
753 time period of treatment to growth cessation in the control (0.3% DMSO) condition was used to
754 calculate GI50 and AUC. Accurate GI50s could not be calculated in cases where a cell line had >50%
755 proliferation inhibition even with the highest drug concentration and hence AUC was used to
756 compare cell line sensitivity. T47D had a highly similar GI50 for Palbociclib to that previously
757 calculated in other studies (112 nM vs 127 nM)[66]. Primary organoid cultures were derived from EAC
758 cases included in the OCCAMS/ICGC sequencing study. Detailed organoid culture and derivation
759 method have been previously described[49]. Regarding the drug treatment, the seeding density for
760 each organoid line was optimised to ensure cell growth in the logarithmic growth phase. Cells were
761 seeded in complete medium for 24 hours then treated with compounds at a 5-point 4-fold serial
762 dilutions for 6 days or 12 days. Cell viability was assessed using CellTiter-Glo (Promega) after drug
763 incubation. The Life sciences reporting summary is also provided in supplementary matierals
764 detailing exact software and biological materials used and efforts made to ensure reproducibility
765 results.

766

# References


49. Li X, Francies HE, Secrier M, et al. Organoid cultures recapitulate esophageal adenocarcinoma heterogeneity providing a model for clonality studies and precision therapeutics. *Nature communications.* Jul 30 2018;9(1):2983.
50. Patnaik A, Rosen LS, Tolaney SM, 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. *Cancer discovery.* Jul 2016;6(7):740-753.
51. D'Antonio M, Ciccarelli FD. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol.* May 29 2013;14(5):R52.
52. Llosa NJ, Cruise M, Tam A, et al. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer discovery.* Jan 2015;5(1):43-51.
53. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *The New England journal of medicine.* Jun 25 2015;372(26):2509-2520.
54. Grasso CS, Giannakis M, Wells DK, et al. Genetic Mechanisms of Immune Evasion in Colorectal Cancer. *Cancer discovery.* Jun 2018;8(6):730-749.
55. Lin L, Bass AJ, Lockwood WW, et al. Activation of GATA binding protein 6 (GATA6) sustains oncogenic lineage-survival in esophageal adenocarcinoma. *Proceedings of the National Academy of Sciences of the United States of America.* Mar 13 2012;109(11):4251-4256.
56. Zehir A, Benayed R, Shah RH, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med.* Jun 2017;23(6):703-713.
57. Ismail A, Bandla S, Reveiller M, et al. Early G(1) cyclin-dependent kinases as prognostic markers and potential therapeutic targets in esophageal adenocarcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Jul 1 2011;17(13):4513-4522.

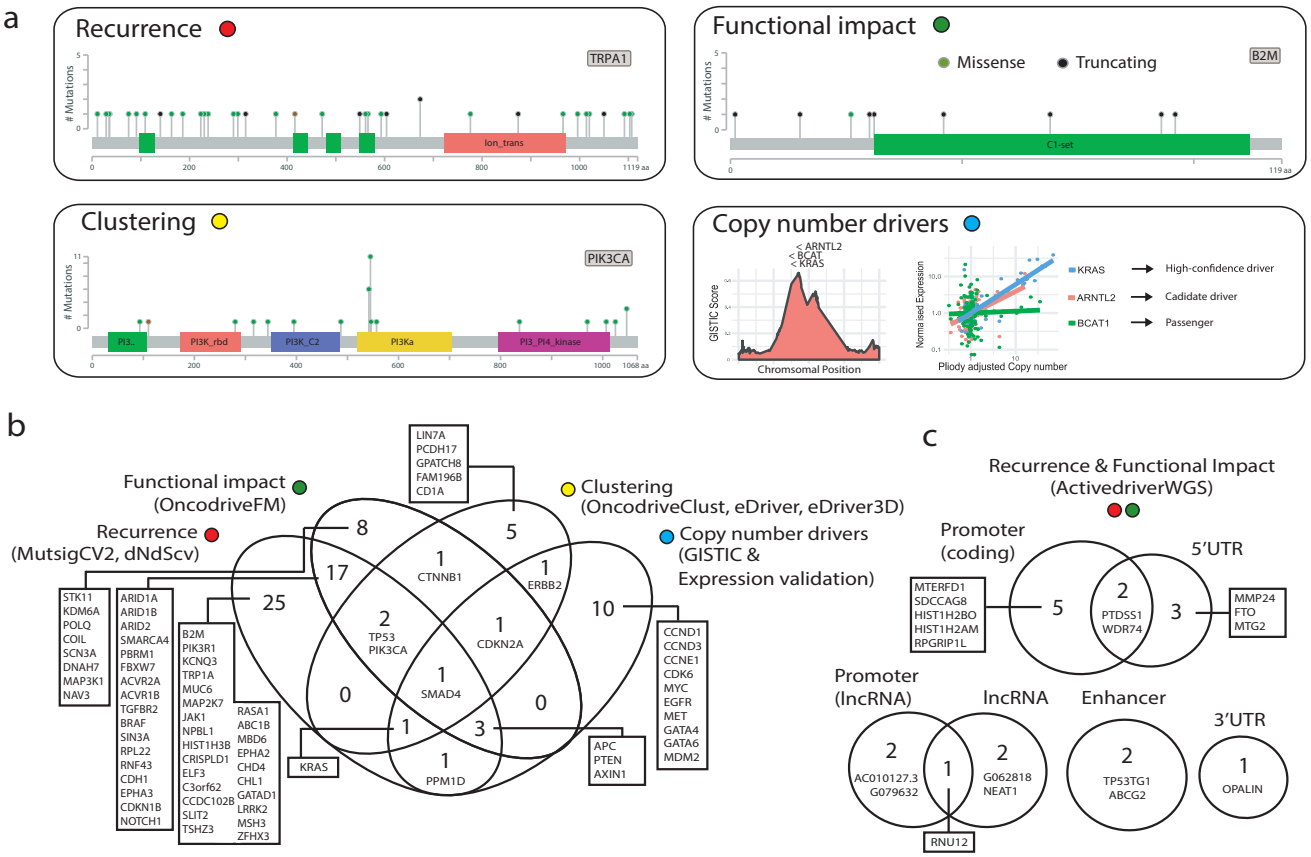



**Figure 1 Detection of EAC driver genes** . **a**. Types of driver-associated features used to detect positive selection in mutations and copy number events with examples of genes containing such features **b**. Coding driver genes identified and their driver-associated features. c. Non-coding driver elements detected and thier element types.

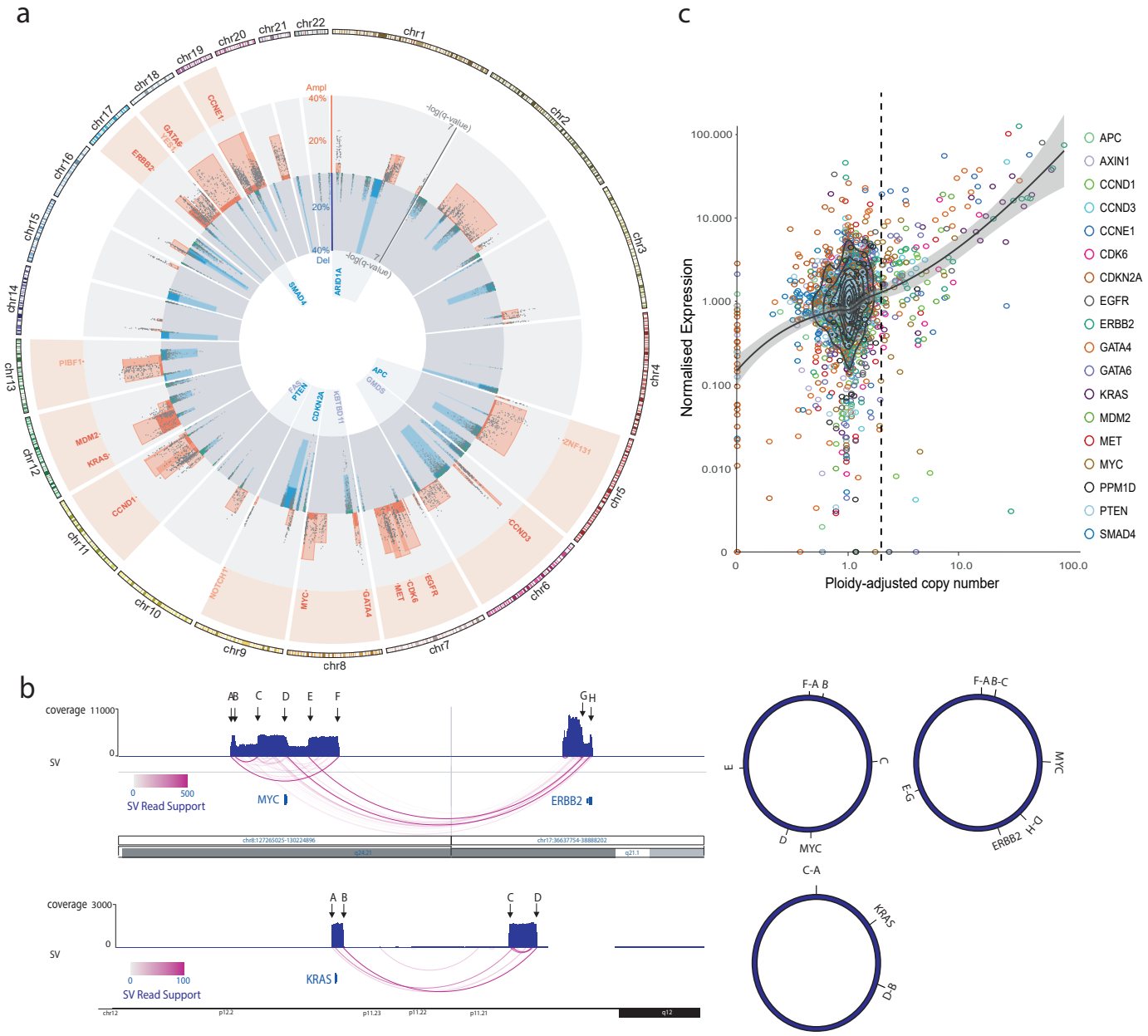



**Figure 2. Copy number variation under positive selection. a.** Recurrent copy number changes across the genome identified by GISTIC in 551 EACs. Frequency of different CNV types are indicated (dark blue = Homozygous deletion, light blue = heterozygous deletion, dark red = extrachromosomal-like amplification, 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 pliody-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. Purple deletion peaks indicate fragile sites. **b.** Examples of Extrachromosomal-like amplifications 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. **c.** Relationship between copy number and expression in CN driver genes in RNA matched sub-cohort (n=116). A 2D kernel density estimation and a leoss regression curve with 95% CIs (grey) are shown to describe the data.

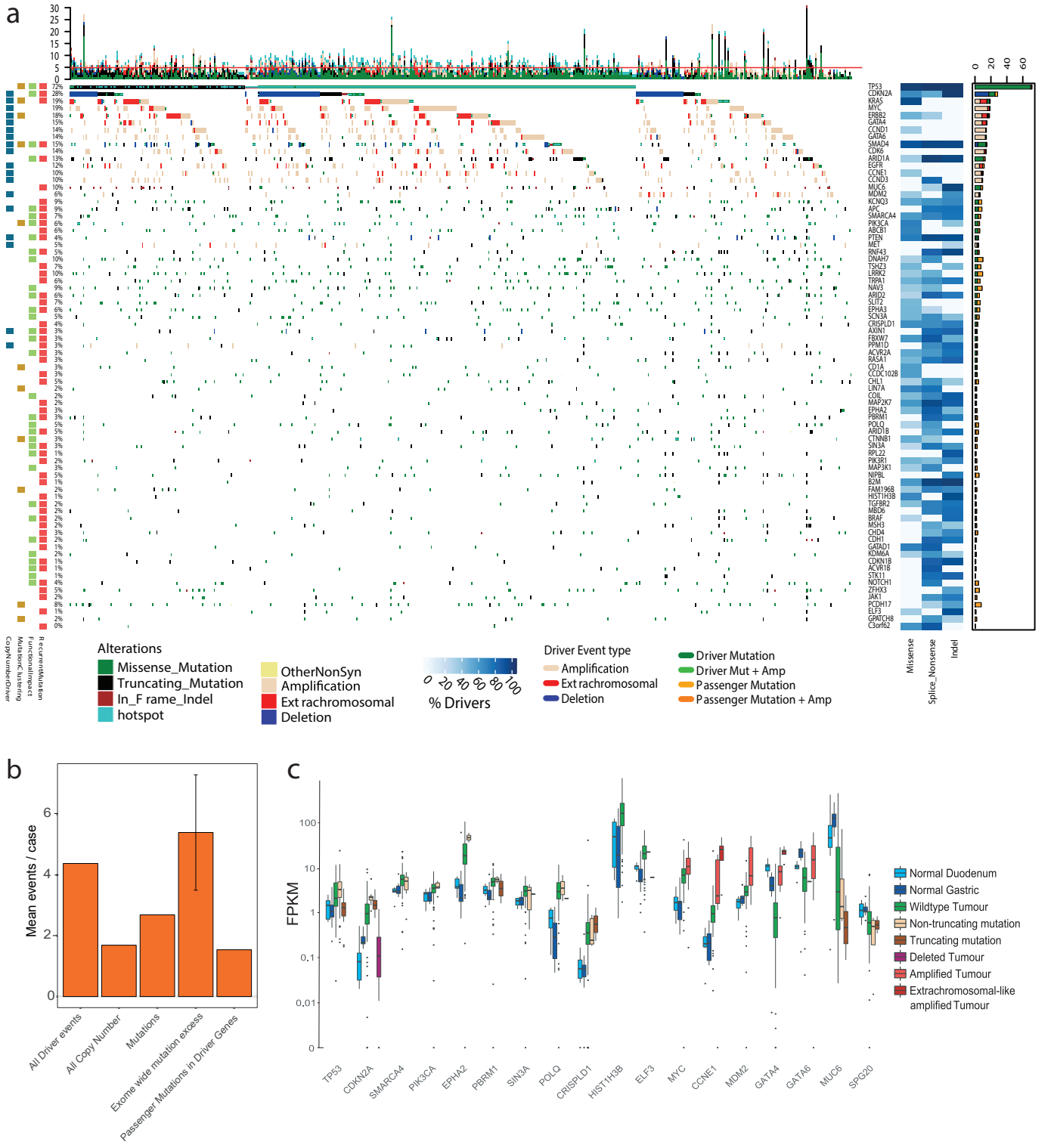
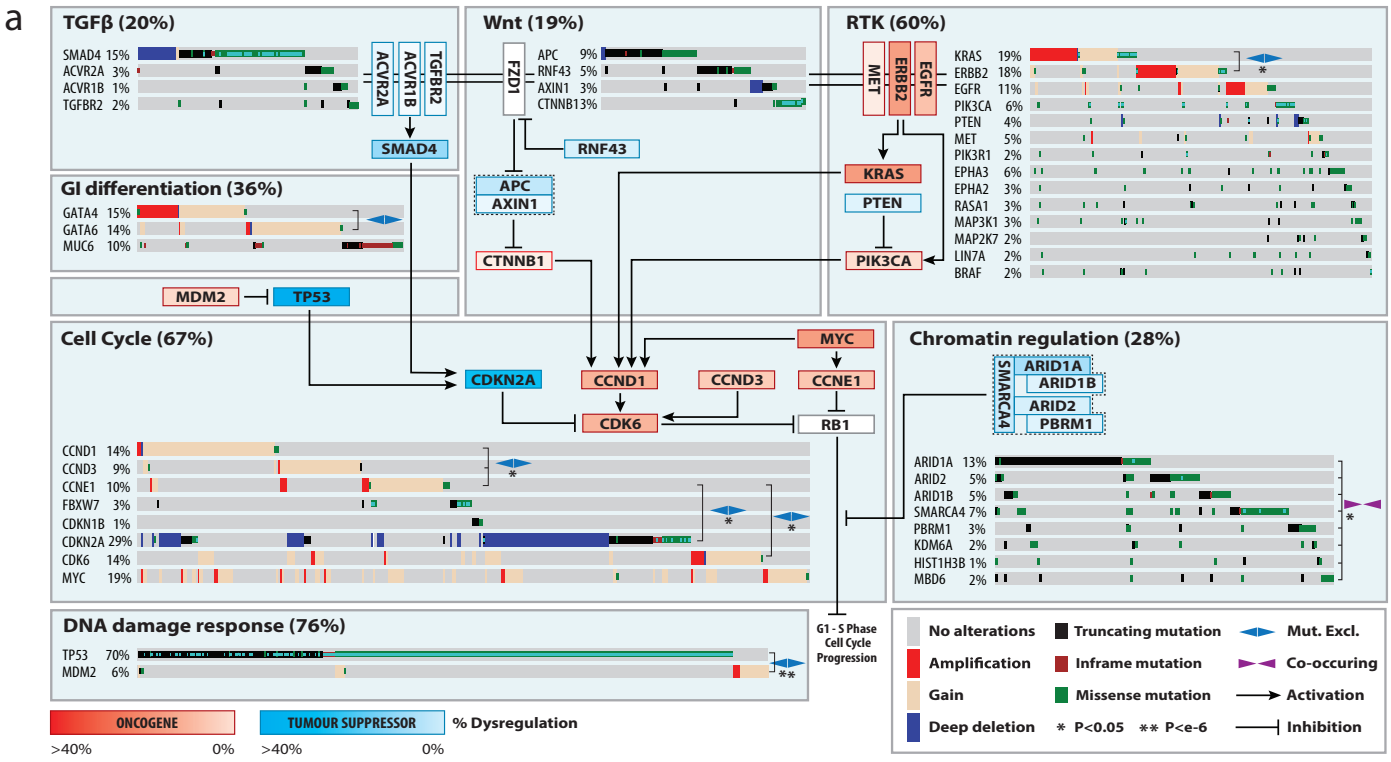



**Figure 3. The driver gene landscape of EAC. a**. Driver mutations or CNVs are shown for each patient of 551 cases. Amplifcation is defined as >2 copy number adjusted ploidy (2 x ploidy of that case) and extrachromosomal amplifcation as >10 copy number adjusted ploidy (10 x ploidy for that case). Driver 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 median (red line = 5). **b.** Mean driver event types per case in 551 EACs and comparison to exome-wide excess of mutations generated by dNdScv. Error bar indicates 95% confidence intervals. **c.** Expression changes in EAC driver genes in comparison to normal intestinal tissues in RNA matched samples (n=116). Boxplots indicate the median and interquartile range. Only genes with expression changes of note are shown.



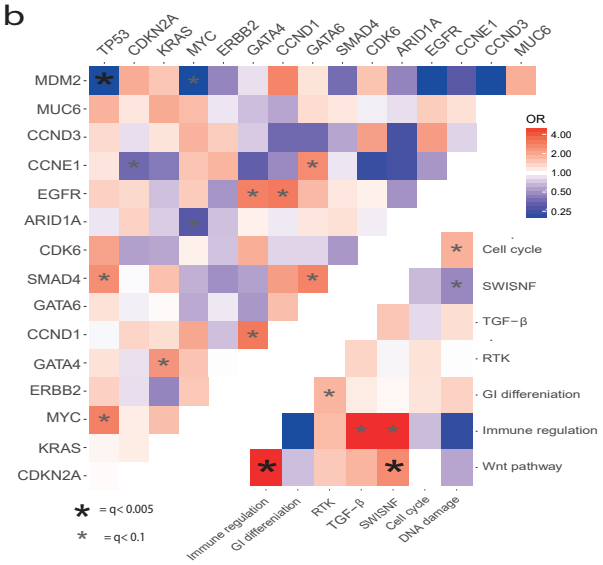




**Figure 4. Biological pathways undergoing selective dysregulation in EAC. a.** Biological Pathways dysregulated by driver gene mutation and/or CNVs in 551 cases. WT cases for a pathway are not shown. Mutual exclusivities and/or associations between genes in a pathway are annotated. *GATA4*/*GATA6* amplifications have a mutually exclusive relationship (ie *GATA4* ampification is more common in *GATA6* WT cases) although this does not reach statistical significance (fisher's exact test, two sided, p=0.07 OR =0.52). **b.** Pairwise assessment of mutual exclusivity and association in EAC driver genes and pathways. Two sided fishers exact tests were used and hypermutated cases (>500 exonic mutations) were removed to aviod bais towards co-occurance, hence n= 510.

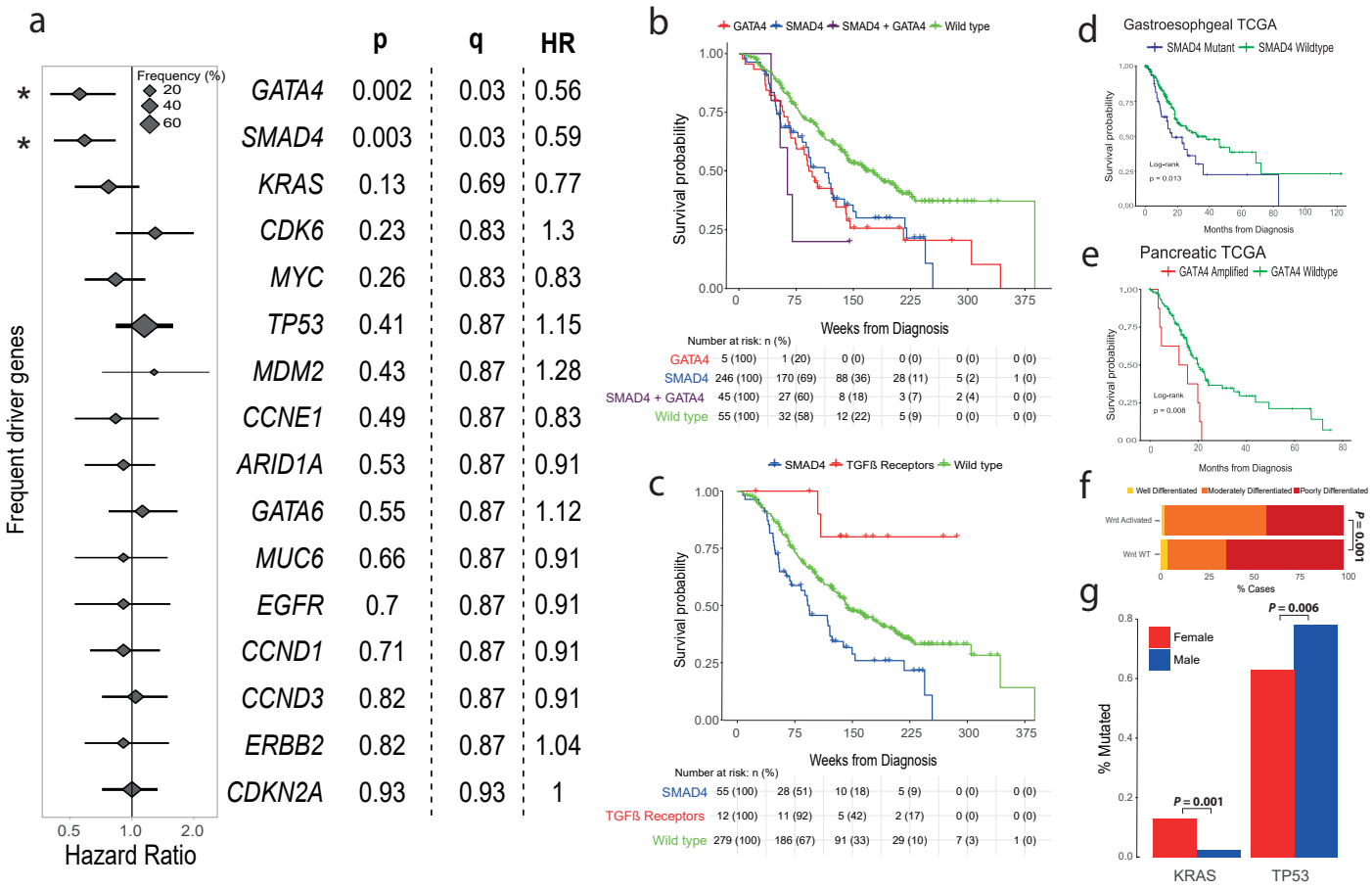



**Figure 5. Clinical significance of driver events in 379 clinical annotated ICGC WGS'd EACs. a.** Hazard ratios and 95% confidence intervals for Cox regression analysis across all drivers genes with at least a 5% frequency of driver alterations. P values are generated from the wald test and q values generated using BH correction. **b.** Kaplan-Meier curves for EACs with different status of significant prognostic indicators (*GATA4* and *SMAD4*). **c.** Kaplan-Meier curves for different alterations in the TGF-ß pathway. **d.** Kaplan-Meier curves showing verification *GATA4* prognostic value in GI cancers using a pancreatic TCGA cohort. **e.** Kaplan-Meier curves showing verification *SMAD4* prognostic value in gastroesophageal cancers using a gastroesophageal TCGA cohort. **f.** Differentiation bias in tumours containing events in Wnt pathway driver genes. **g.** Relative frequency of *KRAS* mutations and *TP53* mutations driver gene events in females vs males (fisher's exact test, two sided).

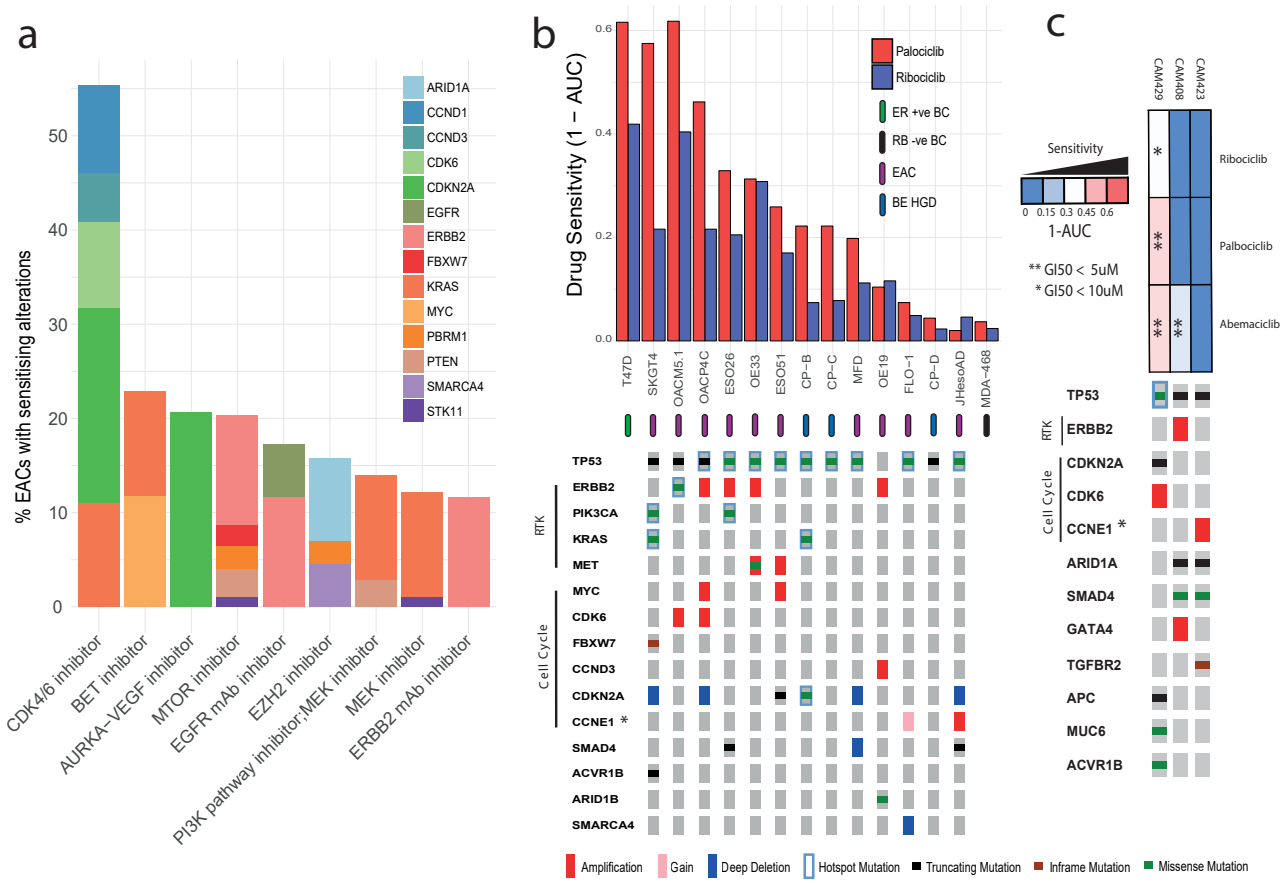



**Figure 6. CDK4/6 inhibitors in EAC. a.** Drug classes for which sensitivity is indicated by EAC driver genes with frequencies shwon from 551 EACs using data from the Cancer Biomarkers database. **b.** Area under the curve (AUC) of sensitivity is shown in a panel of 13 EAC and BE high grade dysplasia 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). **c.** Response of organoid cultures to three FDA approved CDK4/6 inhibitors