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

Frankell AM[1], Jammula S[2], Contino G[1], Killcoyne S[1,3],Abbas S[1], Perner J[2], Bower L[2], Devonshire G[2], Grehan N[1], Mok J[1], O'Donovan M[4], MacRae S[1], Tavare S[2], Fitzgerald RC[1] and the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium[5]

[1] MRC cancer unit, Hutchison/MRC research centre, University of Cambridge, Cambridge, UK
[2] CRUK 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
[5] A full list of contributors from the OCCAMS Consortium is available at the end of the manuscript

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. We have accumulated a cohort of 551 genomically characterised EACs (73% WGS and 27% WES) with clinical annotation and matched RNA-seq. Using a variety of driver gene detection methods we discover 65 EAC drivers (66% novel) and describe mutation and CNV types with specific functional impact. We identify a mean of 3.7 driver events per case derived almost equally from copy number events and mutations. We compare driver mutation rates to the exome-wide mutational excess calculated using Non-synonymous vs Synonymous mutation rates (dNdS). We see mutual exclusivity or co-occurrence of events within and between a number of EAC pathways (GATA factors, Core Cell cycle genes, TP53 regulators and the SWI/SNF complex) suggestive of important functional relationships. These driver variants correlate with tumour differentiation, sex and prognosis. 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.

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, including the UK and the US. The incidence of EAC in such countries has been rapidly rising, with a seven-fold increase in incidence over the last 35 years in the US[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, there is a paucity of recurrently mutated genes. EACs also display 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 highly limited and consequentially there is a paucity of molecular biomarkers for prognosis or targeted therapeutics available in the clinic.

Driver events undergoing positive selection during cancer evolution are a small proportion of total number of genetic events that occur in each tumour[7]. Methods to differentiate driver mutations from passenger mutations use features associated with known driver events to detect regions of the genome, often genes, in 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 been applied on several occasions to EAC cohorts[6,10,11] and has identified ten known cancer genes as high confidence EAC drivers (TP53, CDKN2A, SMAD4, ARID1A, ERBB2, KRAS, PIK3CA, SMARCA4, CTNNB1 and FBXW7).

However these analyses leave most EAC cases with only one known driver mutation, usually TP53, due to the low frequency at which other drivers occur. Equivalent analyses in other cancer types have identified three or four drivers per case[12,13]. 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[10,14-17]. However, GISTIC identifies relatively large regions of the genome, containing hundreds of genes, 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 fragile sites where a low density of DNA replication origins causes frequent structural events at a particular loci. These have not been differentiated properly from selection based recurrent CNAs[18].

Without proper annotation of the genomic variants which drive the biology of EAC tumours we are left with a very large number of events, most of which are likely to be inconsequential, making it extremely difficult to detect statistical associations between genomic variants and various biological and clinical parameters. To address these issues, we have 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 have augmented our ICGC WGS cohort with publically available whole exome[19] and whole genome sequencing[20] data. We have applied a number of complementary driver detection tools to this cohort, using a range of driver associated features combined with analyses of RNA expression to produce a comprehensive assessment and characterisation of mutations and CNAs under selection in EAC. We then use these events to define functional cell processes that have been selectively dysregulated in EAC and identify novel, clinically relevant biomarkers for prognostication, which we have verified in independent cohorts. 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 effects

In 551 EACs 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 (supplementary figure 1), and 286,965 copy number aberrations (CNAs). We also identified 134,697 structural variants (SVs) in WGS cases. Mutations or copy number variants under selection were detected using specific driver associated-mutation features (Fig 1A). We use several complementary driver detection tools to detect each feature, and each tool underwent quality control to ensure reliability of results (see methods). These features include highly recurrent mutations within a gene (dNdScv[21] and MutsigCV2[9]), high functional impact mutations within a gene (OncodriveFM[22]), mutation clustering (OncodriveClust[23], eDriver[24] and eDriver3D[25]) and recurrent amplification or deletion of genes (GISTIC[14]) 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 sixty five EAC driver genes were discovered, 64% of which have not been detected in EAC previously[10,11,15-17,19]. Of the sixty five gene identified, 82% are known drivers in pan-cancer analyses giving confidence in our methods[21,26,27].

EAC is notable among cancer types for harbouring a high degree of chromosomal instability[20]. Using GISTIC we identified 126 recurrently deleted or amplified loci across the genome (Fig 2A). To determine which genes within these loci confer a selective advantage when they undergo CNAs we use a subset of 119 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 CNAs. The majority of genes in these regions showed no

CN associated expression change (74%). 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. 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. 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 74 such cases which we designated "candidate copy number (CN) drivers" (supplementary tables 1 and 2). Several GISTIC loci contained only one candidate driver such as ZNF131, PRKCI and MYBL2 which 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 these loci were highly correlated with presence of CN-drivers (Wilcox test, $p<Ex10^{-6}$) (Supplementary Figure 3). To discern a mechanism for these ultra-high amplifications we assessed structural variants (SVs) associated with these events and the copy number steps surrounding them. 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 Fig2B. In the first example an inversion has been followed by circularisation and amplification KRAS and in the second circularisation and amplification initially occurred around MYC but subsequently incorporated ERBB2 from an entirely different chromosome. A pattern of extrachromosomal amplification via double minutes has been previously noted in EAC[20], and hence we refer to this amplification class with ultra-high amplification (Ploidy adjusted Copy number >10) as 'extrachromosomal-like'. Several deletion loci co-align with fragile sites (Fig 2A). Most deletion loci were dominated by heterozygous deletions while a small subset had a far higher percentage of homozygous deletions including CDKN2A and several associated with fragile site loci (Fig. 2A). 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, $p<Ex10^{-4}$ and $p<Ex10^{-3}$ respectively) but less dramatic effects on expression with a lower penetrance (Fig 2C). This lack of penetrance was associated with low cellularity (fisher's exact test, expression cut off = 2.5 normalised FPKM, $p<0.01$) in amplified cases but also likely reflects that genetic mechanisms other than gene-dosage can 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.

Landscape of driver Events in EAC

The overall landscape of driver gene mutations and copy number alterations per case is depicted in Figure 3A. These comprise both oncogenes and tumour suppressor genes activated or repressed via different mechanisms. Occasionally different types of events are selected for in the same gene, such as KRAS which harbours both activating mutations and amplifications in 19% of cases. 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; 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 4). Clusters of mutations occurring in EAC or mutations on amino acids which are mutation hotspots in other cancer types[28] (supplementary table 3) are indicated in figure 3A. Novel EAC drivers of particular interest include B2M, a core component of the MHC class I complex and resistance marker for Immunotherapy[29], MUC6 a secreted glycoprotein involved in gastric acid resistance and ABCB1 a channel pump protein which is associated with multiple instances of drug resistance[30]. Lollipop plots showing primary sequence distribution of mutations in these genes are provided (supplementary data).

The identification of driver events provides a rich information about the molecular history of each EAC tumour. We detect a median of four events in driver genes per tumour (IQR = 3-6, Mean = 5.1) and only a very small fraction of cases have no such events detected (11 cases, 2%). When we remove the predicted percentage of passenger mutations using dnds ratios we find a mean of 3.7 true driver events per case which derive quite evenly from both copy number events and mutations (Fig 3B). 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 4.8 (95% CIs: 3.7-5.9) in comparison to 2.1 driver mutations which we calculate in our gene-centric analysis after passenger removal. This suggests low frequency driver genes may be prevalent in the EAC 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 (Supplementary Figure 5). 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 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 3C). Since surrounding squamous

177 epithelium is a fundamentally different tissue, from which EAC does not directly arise, we have used
178 duodenum and gastric cardia samples as gastrointestinal phenotype controls, likely to be similar to
179 the, as yet unconfirmed, tissue of origin in EAC. A large number of driver genes have upregulated
180 expression in comparison to normal controls, for example TP53 has upregulated RNA expression in
181 WT tumour tissue and in cases with missense (non-truncating) mutations but RNA expression is lost
182 upon gene truncation. In depth analysis of different TP53 mutation types reveals significant
183 heterogeneity within non-truncating mutations, for example R175H mutations correlate with low
184 RNA expression (supplementary figure 6). Normal tissue expression of CDKN2A suggests that
185 CDKN2A is generally activated in EAC and returns to physiologically normal levels when deleted.
186 Heterogeneous expression in WT CDKN2A cases suggest a different mechanism of inhibition such as
187 methylation in some cases. Overexpression of other genes in wild type tumours, such as SIN3A, may
188 be a confer selective advantage due to their oncogenic properties, in this case cooperating with
189 MYC, which is also overexpressed in EACs (Fig 3C). A smaller number of driver genes are
190 downregulated in EAC tissue- 3/4 of these (GATA4, GATA6 and MUC6) are involved in the
191 differentiated phenotype of gastrointestinal tissues and may be lost with tumour de-differentiation.
192 Driving alterations in these genes have been observed in other GI cancers[13,31,32] however their
193 oncogenic mechanism is unknown. In most genes we did not observe expression loss at the RNA
194 level with truncation, for instance ARID1A (supplementary figure 7).

195

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

197 It is known that selection preferentially dysregulates certain functionally related groups of genes and
198 biological pathways in cancer[33]. This phenomenon is highly evident in EAC, as shown in Figure 4
199 which depicts the functional relationships between EAC drivers. This provides greater functional
200 homogeneity to the landscape of driver events.

While TP53 is the dominant driver in EAC, 30% 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 (31% 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), 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 4B). 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. We were able to confirm some of these relationships in independent cohorts in different cancer types (supplementary table 4) suggesting some of these may be pan-cancer phenomenon. As shown in figure 3, 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[29]), 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).

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 EACs (ie after removal of predicted passengers, 16 genes) to detect prognostic biomarkers (Fig 5A). Events in two genes conferred significantly poorer prognosis after multiple hypothesis correction, GATA4 (HR : 0.54 , 95% CI : 0.38 – 0.78, *P* value = 0.0008) and SMAD4 (HR : 0.60 , 95% CI : 0.42 – 0.84, *P* value = 0.003). Both genes remained significant in multivariate Cox regression including pathological tumour stage (GATA4 = HR adjusted : 0.63, 95% CIs adjusted : 0.40 - 0.98, *P* value = 0.042 and SMAD4 = HR adjusted : 0.63, 95% CI adjusted : 0.41 – 0.97, *P* value = 0.038). 31% of EACs contain either SMAD4 mutation or homozygous deletion or GATA4 amplification and cases with both genes altered had a poorer prognosis (figure 5B). We validated the poor prognostic impact of SMAD4 events in an independent TCGA gastroesophageal cohort (HR = 0.58, 95% CI = 0.37 – 0.90, *P* value =0.014) (Fig 5C) 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) (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[16] although it did not reach statistical significance after FDR correction and SMAD4 expression loss has been previously linked to poor prognosis in EAC[34]. 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[35].

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, see methods, Fig 5F). We 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) (Fig 5G). This is of particular interest given the male predominance of EAC[3].

Targeted therapeutics using EAC 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 EAC cells to certain targeted therapeutic agents we used the Cancer Biomarkers database[36]. We calculated the percentage of our cases which contain EAC-driver biomarkers of response to each drug class in the database (summary shown Fig 6A, and full data supplementary table 5). 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 33% 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 eight EAC cell lines which have undergone whole genome sequencing[37] and assessed them for presence of EAC driver events (Figure 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 (Figure 6B, supplementary table 6, Supplementary Figure 8)[38]. Such EAC cell lines had comparable sensitivity to T47D which is derived from an ER +ve breast cancer where CDK4/6is have been FDA approved. We noted three cell lines without sensitising events which were highly resistant, with little drug effect even at 4000nM 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 CDK4/6i by bypassing CDK4/6 and causing Rb phosphorylation via CDK2 activation[39].

Discussion

We present here a detailed catalogue of 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, producing insights into EAC biology. 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 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 EAC, similar to many other cancer types[21]. 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 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]. The identification of these very low frequency mutations will require substantially different detection techniques to those which are currently in wide spread use and such methods are in development[40] although they require validation. Undoubtedly many copy drivers are also left undiscovered and validation of candidates identified here is an important avenue of future work.

While a number of previous reports have attempted to detect EAC drivers, they have had a limited yield per case for a variety of reasons. The first such study[19] used methods which, despite being well regarded at the time, were subsequently discredited[9]. Hence a number of known false positive genes (EYS, SYNE1 and CNTTAP5) were erroneously reported as drivers, along with an additional unknown number of genes. Since then a number of reports, including our own, on medium and large cohort sizes using MutsigCV[10,11,17] 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 52 mutational driver genes (excluding CN drivers). Detection of driver CNAs has previously relied on GISTIC to detect recurrently mutated regions[10,14-17] but no analyses have been performed to evidence which genes in these large regions are true drivers. Many of the genes annotated by such papers are unlikely to be CN drivers from this analysis due to their lack of expression modulation with CNAs (eg YEATS4 and MCL1), the role of recurrent heterozygous losses to drive LOH in some mutational drivers (ARID1A and CDH11) or their association with fragile sites (PDE4D, WWOX, FHIT). Conversely, we have been able to identify novel EAC copy number drivers (eg CCND3, AXIN1 and APC).

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 EAC tumours and how they modify EAC biology. Particularly interesting are the GI specific genes GATA4/6 and MUC6 which modulate prognosis and 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. Prospective clinical work to verify and implement SMAD4 and GATA4 biomarkers in this study would be worthwhile. While whole genome or whole exome sequencing may be impractical for use in the clinic, 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[41]. In EAC 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 EAC cases based on their individual genomic profile. In particular CDK4/6 inhibitors deserve considerable attention as an option for EAC treatment as they are, by a significant margin, the treatment to which the most EACs harbour sensitivity-causing driver events, excluding TP53 as an unlikely therapeutic biomarker. The in vitro validation of these biomarkers for CDK4/6 inhibitors in EAC is also 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 functional and clinical 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.

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[1,3], Ahmad Miremadi[1,3], Shalini Malhotra[1,3], Monika Tripathi[1,3], Simon Tavaré[2], Andy G. Lynch[2], Matthew Eldridge[2], Maria Secrier[2], Lawrence Bower[2], Ginny Devonshire[2], Juliane Perner[2], Sriganesh Jammula[2], Jim Davies[5], Charles Crichton[5], Nick Carroll[6], Peter Safranek[6], Andrew Hindmarsh[6], Vijayendran Sujendran[6], Stephen J. Hayes[7,14], Yeng Ang[7,8,29], Shaun R. Preston[9], Sarah Oakes[9], Izhar Bagwan[9], Vicki Save[10], Richard J.E. Skipworth[10], Ted R. Hupp[10], J. Robert O'Neill[10,23], Olga Tucker[11,33], Andrew Beggs[11,28], Philippe Taniere[11], Sonia Puig[11],Timothy J. Underwood[12,13], Fergus Noble[12], Jack Owsley[12], Hugh Barr[15], Neil Shepherd[15], Oliver Old[15], Jesper Lagergren[16,25], James Gossage[16,24], Andrew Davies[16,24,] , Fuju Chang[16,24], Janine Zylstra[16,24], Ula Mahadeva[16], Vicky Goh[24], Francesca D. Ciccarelli[24], Grant Sanders[17], Richard Berrisford[17], Catherine Harden[17], Mike Lewis[18], Ed Cheong[18], Bhaskar Kumar[18],Simon L Parsons[19], Irshad Soomro[19], Philip Kaye[19], John Saunders[19], Laurence Lovat[20], Rehan Haidry[20], Laszlo Igali[21], Michael Scott[22], Sharmila Sothi[26], Sari Suortamo[26], Suzy Lishman[27], George B. Hanna[31], Christopher J. Peters[31], Anna Grabowska[32], Richard Turkington[34.]

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

[33]Heart of England NHS Foundation Trust, Birmingham, UK, B9 5SS.
[34]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 data and performed statistical analyses. RCF and AMF designed the experiments. AMF 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. SM and NG coordinated the clinical centres and were responsible for sample collections. MO led the pathological sample QC for sequencing. LB and GD ran 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.

Acknowledgements

We would like to thank Dr. Adam Bass and Dr. Nic Waddel 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 genes.

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.** Driver genes identified and their driver-associated features.

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 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 (bi) two populations of extrachromosomal DNA are apparent (biii), one amplifying only MYC and the second also incorporating ERBB2 from a different chromosome. In the second example (bii) an inversion has occurred before circularization and amplification around KRAS (biv). **c.** Relationship between copy number and expression in CN driver genes.

Figure 3. The driver gene landscape of Esophageal Adenocarcinoma. 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 >10Copy 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). **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. GATA4/6 amplifications have a mutually exclusive relationship although this does not reach statistical significance (fisher's exact test 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 drivers 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 TGFbeta 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 utility 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 8 EAC cell lines with associated WGS and driver event, based in tumour analysis, in these cell lines indicated. 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 CDK4/6is due to its regulation of Rb downstream of CDK4/6 hence bypassing the need for CDK4/6 activity (see figure 5).

Supplementary figure legends

Supplementary figure 1. Distribution of small scale mutations (SNVs and Indels) across the 551 EAC cohort. Red line indicates the median mutations per case (6.4)

Supplementary Figure 2. Concordance between driver gene detection methods. A. Hierarchical clustering between tools based on gene identified**. B** Genes identified by each tool**.**

S**upplementary Figure 3.** Frequency of Extrachromosomal like events (CN adjusted Ploidy >10) in GISTIC amplification peaks and presence of high confidence drivers in those peaks indicated**.**

Supplementary Figure 4. A scheme demonstrating how to use mutational clustering along with dnds ratios to estimate the probability of a particular mutation being a driver. In this case the dnds ratio suggests 2/3 of missense mutations are drivers hence 10/15. 7 missense mutation 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 a most (5/7) other mutations are passengers. Similarly, mutations on amino acids known to be hyper mutated in other cancer types (see Supplementary table 3, for instance if we found a single KRAS G12 mutation) can be considered likely drivers.

Supplementary Figure 5. A detailed breakdown of mutation and copy number types per case and a breakdown of exome wide dnds excess for different mutation types (note that exome wide indel cannot be calculated excess as they have no synonymous mutation equivalent, although a null model is used in the per gene dnds method to use them to detect driver genes).

Supplementary Figure 6. TP53 expression in different TP53 mutation types in comparison to TP53 WT tumours and normal duodenum and gastric cardia tissues.

Supplementary Figure 7. Expression of all EAC driver genes across different genomic states for the gene in question in 119 EAC tumours, and in comparison to duodenum and gastric cardia tissues.

Supplementary Figure 8. Growth inhibition responses of EAC cell lines and control lines to CDK4/6 inhibitors Palbociclib and Ribociclib.

Methods

Cohort, sequencing and calling of genomic events

380 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 Classification and Molecular Stratification) consortium. The procedures for obtaining the samples, quality control processes, extractions and whole genome sequencing are as previously described[17]. 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. In addition, previously published samples were included in the analysis from Dulak et al 2013[19] – 139 WES and 10 WGS (total 27%) and Nones et al 2014[20] with 22 WGS samples (4%) to total 551 genome characterised EACs. RNA-seq data was available from our ICGC WGS samples (119/380). BAM files for all samples (include those from Dulak et al 2013 and Nones et al 2014) were run through our alignment (BWA-MEM), mutation (Strelka) and copy number (ASCAT) and structural variant (Manta) calling pipelines, as previously described[17]. Our methods were benchmarked against various other available methods and have among the best sensitivity and specificity for variant calling (ICGC benchmarking excerise[42]). Mutation and copy number calling on cell lines was performed as previously described[37].

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[43]. Paired end 75bp sequencing on HiSeq4000 generated the paired end reads.

Analysing EAC mutations for selection

To detect positively selected mutations in our EAC cohort, a multi-tool approach across various selection related 'Features' (Recurrance, Functional impact, Clustering) was implemented in order to provide a comprehensive analysis. This is broadly similar to several previous approaches[8,44]. dNdScv[21], MutsigCV[9], e-Driver[24] and e-Driver3D[25] were run using the default parameters. To run OncodriverFM[22], Polyphen[45] and SIFT[46] were used to score the functional impact of each missense non-synonomous mutation (from 0, non-impactful to 1 highly impactful), synonymous mutation

were given a score of 0 impact and truncating mutations (Non-sense 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[47] and Hotspot[28]). 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[21,26,27] as has been previously implemented[21,48]. 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 > 2:1) of any single mutation type were also removed (8 genes). 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.

564 **Detecting selection in CNVs**

565 ASCAT raw CN values were used to detected frequently deleted or amplified regions of the genome
566 using GISTIC2.0[14]. To determine which genes in these regions confer a selective advantage, CNVs
567 from each gene within a GISTIC identified loci were correlated with FPKM from matched RNA-seq in
568 a sub-cohort of 119 samples and with mutations across all 551 samples. To call copy number in
569 genes which spanned multiple copy number segments in ASCAT we considered the total number of
570 full copies of the gene (ie the lowest total copy number). Occasionally ASCAT is unable to confidently
571 call the copy number in a highly aberrant genomic regions. We found that the expression of genes in
572 such regions matched well what we would expect given the surrounding copy number and hence we
573 used the mean of the two adjacent copy number fragments to call copy number in the gene in
574 question. We found amplification peak regions identified by GISTIC2.0 varied significantly in precise
575 location both in analysis of different sub-cohorts and when comparing to published GISTIC data from
576 EACs[10,15,16]. A peak would often sit next to but not overlapping a well characterised oncogene or
577 tumour suppressor. To account for this, we widened the amplification peak sizes upstream and
578 downstream by twice the size of each peak to ensure we captured all possible drivers. Our
579 expression analysis allows us to then remove false positives from this wider region and called drivers
580 were still highly enriched for genes closer to the centre of GISTIC peak regions.

581 To detect genes in which amplification correlated with increased expression we compared
582 expression of samples with a high CN for that gene (top 25% percentile of CN) with those which have
583 a normal CN (median +/- 1) using the Wilcox rank-sum test and using the specific alternative
584 hypothesis that high CN would lead to increased expression. Q-values were then generated based on
585 Benjamini & Hochberg method, not considering genes without significant expression in amplified
586 samples (at least 80% amplified samples with FPKM > 1) and considering $q<0.001$ as significant. We
587 also included an additional known driver gene only FDR reduction analysis as previously described
588 for mutational drivers with $q<0.05$ considered as significant given the additional evidence for these

589 genes in other cancer types. We took the same approach to detect genes in which homozygous
590 deletion correlated with expression loss. Expression modulation was a highly specific marker for
591 known CN driver genes and was not a widespread feature in most recurrently copy number variant
592 genes. However, while expression modulation is a requirement for selection of CNV only drivers, it is
593 not sufficient evidence alone and hence we grouped such genes into those which have been
594 characterised as drivers previously in other cancer types (high confidence EAC CN drivers) and other
595 genes (Candidate EAC CN drivers) which await functional validation. We used fragile site regions
596 detected in Wala et al 2017[49]. We also defined regions which may be recurrently heterozygous
597 deleted, without any significant expression modulations, to allow LOH of tumour suppressor gene
598 mutations. To do this we analyses genes with at least 5 mutations in the matched RNA cohort for
599 association between LOH (ASCAT minor allele = 0) and mutation using fisher's exact test and
600 generated q values using the Benjamini & Hochberg method. The analysis was repeated on known
601 cancer genes only for reduced FDR and $q < 0.05$ considered significant for both analyses. For those
602 high confidence drivers we chose to define amplification as CN/ploidy (referred to as Ploidy adjusted
603 copy number) this produces superior correlation with expression. We chose a cut off for
604 amplification at CN/ploidy = 2 as has been previously used, and as causes a highly significant
605 increase in expression in our CN-driver genes.

606

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

608 The relative distribution of driver events in each pathway was analysed using a fisher's exact test in
609 the case of pair-wise comparisons including WT cases. In the case of multi-gene comparisons such as
610 the Cyclins we calculate the p value and odds ratio for each pair in the group by fisher's exact test
611 and combine p values using the Fisher method, Genes without comparable Odds ratios to the rest of
612 the genes in question were removed. For this analysis we also remove highly mutated cases (>500
613 exonic mutations, 41/551) as they bias distribution of genes towards co-occurrence. We repeated

614 this analyses across all pairs of driver genes using BH multiple hypothesis correction. We validated
615 these relationships in independent TGCA cohorts of other GI cancers where we could find cohorts
616 with reasonable numbers of the genomic events in question (not possible for GATA4/6 for instance)
617 using the cBioportal web interface tool[50].

618

619 **Correlating genomics with the clinical phenotype**

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

632

633 **Therapeutics**

634 The cancer biomarker database was filtered for drugs linked to biomarkers found in EAC drivers and
635 supplementary table 6 constructed using the cohort frequencies of EAC biomarkers. 8 EAC cell lines
636 with WGS data[37] were used in proliferation assays to determine drug sensitivity to CDK4/6 inhibitors,
637 Palbociclib (Biovision) and Ribociclib (Selleckchem). Cell lines were grown in their normal growth

media (methods table 1). 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 4000nM 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)[51].

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. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature.* Mar 20 2014;507(7492):315-322.

13. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature.* Sep 11 2014;513(7517):202-209.

14. 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.

15. 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.

16. 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.

17. 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.

18. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet.* Oct 2013;45(10):1134-1140.

19. 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.

20. 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.

21. 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.

22. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* Nov 2012;40(21):e169.

23. 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.

24. Porta-Pardo E, Godzik A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics.* Nov 1 2014;30(21):3109-3114.

25. Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res.* Jan 2015;43(Database issue):D968-973.

26. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer.* Mar 2004;4(3):177-183.

27. 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.

28. 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.

29. Zaretsky JM, Garcia-Diaz A, Shin DS, et al. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N Engl J Med.* Sep 1 2016;375(9):819-829.

30. 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.

31. Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* Jul 18 2012;487(7407):330-337.

32. 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.

33. 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.

34. 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.

35. 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.

36. Tamborero D, Rubio-Perez C, Deu-Pons J, et al. Cancer Genome Interpreter Annotates The Biological And Clinical Relevance Of Tumor Alterations. *bioRxiv.* 2017.

37. Contino G, Eldridge MD, Secrier M, et al. Whole-genome sequencing of nine esophageal adenocarcinoma cell lines. *F1000Res.* 2016;5:1336.

38. 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.

39. 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.

40. 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.

41. 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.

42. Ding J, McConechy MK, Horlings HM, et al. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat Commun.* Oct 5 2015;6:8554.

43. Nagai K, Kohno K, Chiba M, et al. Differential expression profiles of sense and antisense transcripts between HCV-associated hepatocellular carcinoma and corresponding non-cancerous liver tissue. *Int J Oncol.* Jun 2012;40(6):1813-1820.

44. 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.

45. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet.* Jan 2013;Chapter 7:Unit7 20.

46. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet.* 2006;7:61-80.

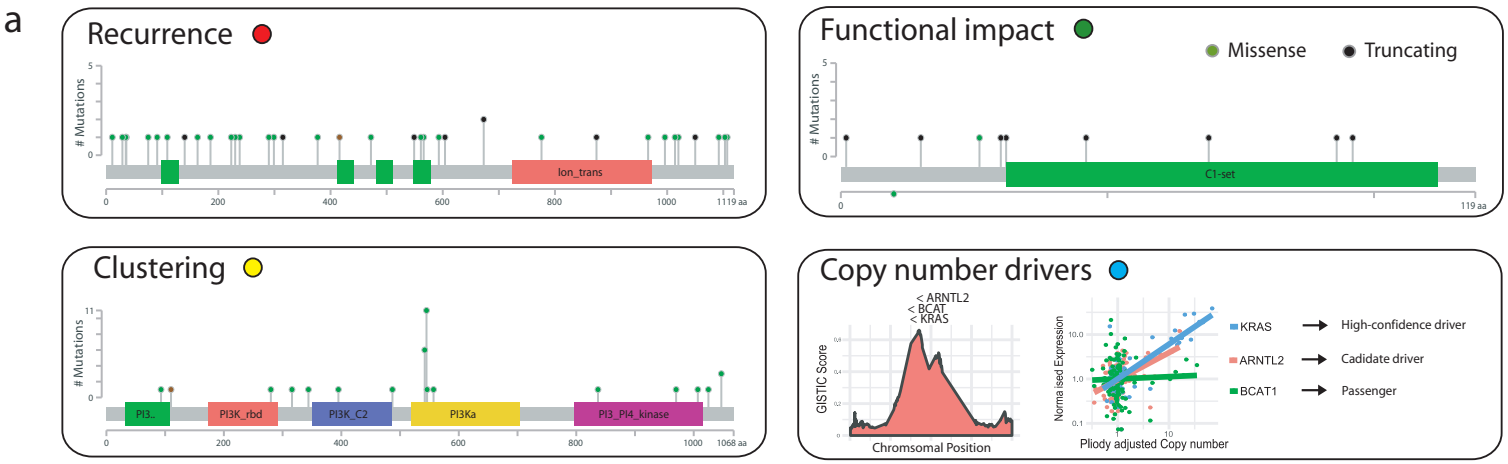
47. Reimand J, Wagih O, Bader GD. The mutational landscape of phosphorylation signaling in cancer. *Sci Rep.* Oct 2 2013;3:2651.

48. Northcott PA, Buchhalter I, Morrissy AS, et al. The whole-genome landscape of medulloblastoma subtypes. *Nature.* Jul 19 2017;547(7663):311-317.

49. Wala JA, Shapira O, Li Y, et al. Selective and mechanistic sources of recurrent rearrangements across the cancer genome. *bioRxiv.* 2017.

50. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* Apr 2 2013;6(269):pl1.

51. Finn RS, Dering J, Conklin D, 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.* 2009;11(5):R77.



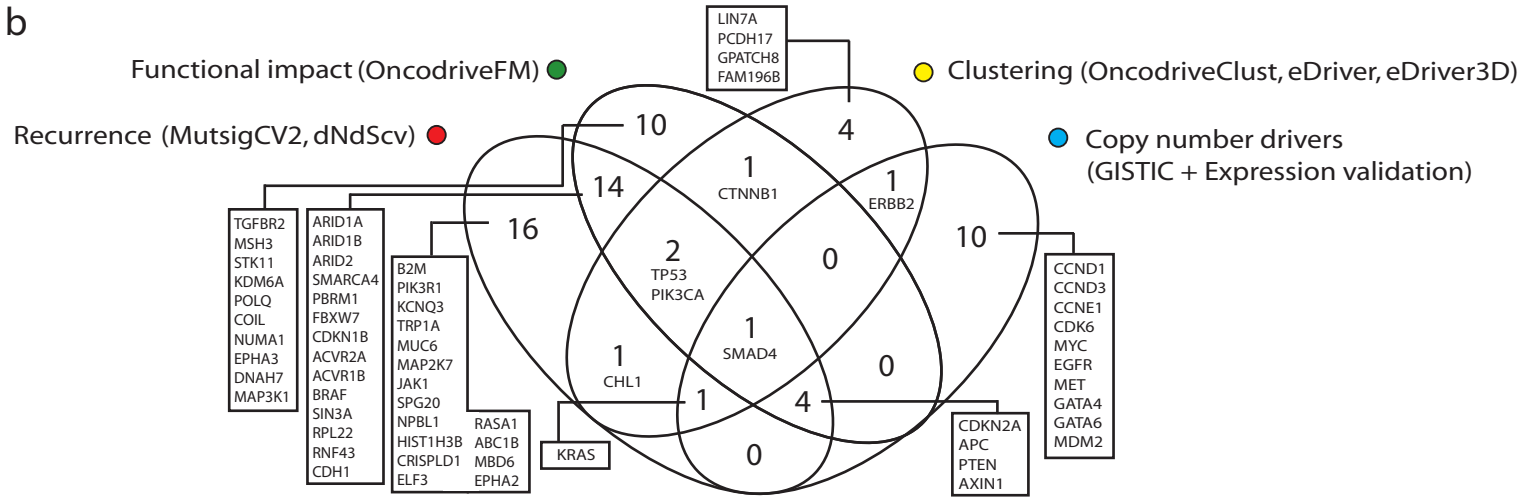
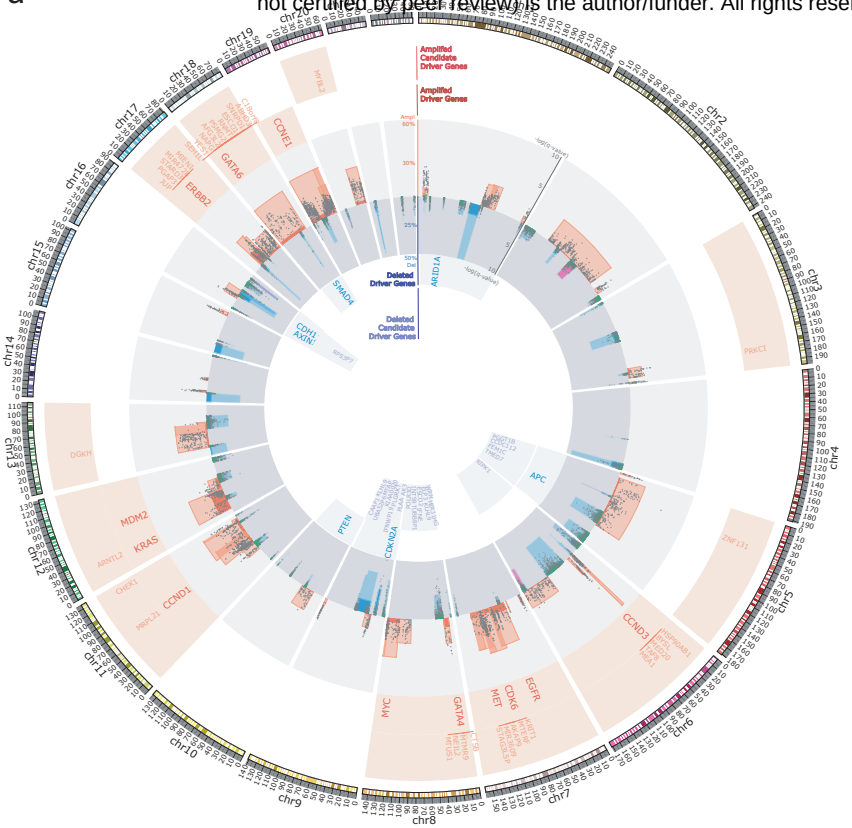
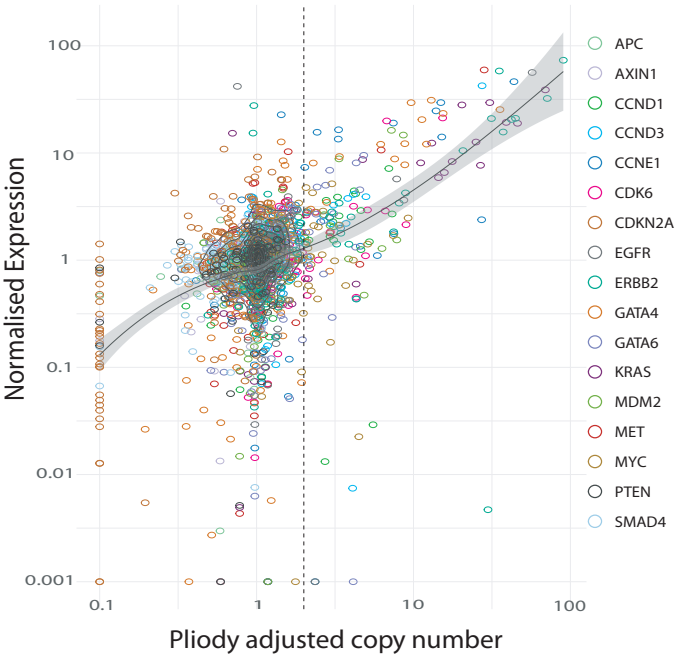


Figure 1 Detection of EAC driver Genes. a. Types of driver-associated features used to detect positive selection in mutations and copy number events with examples of genes containing such features **b**. Driver genes identified and their driver-associated features.





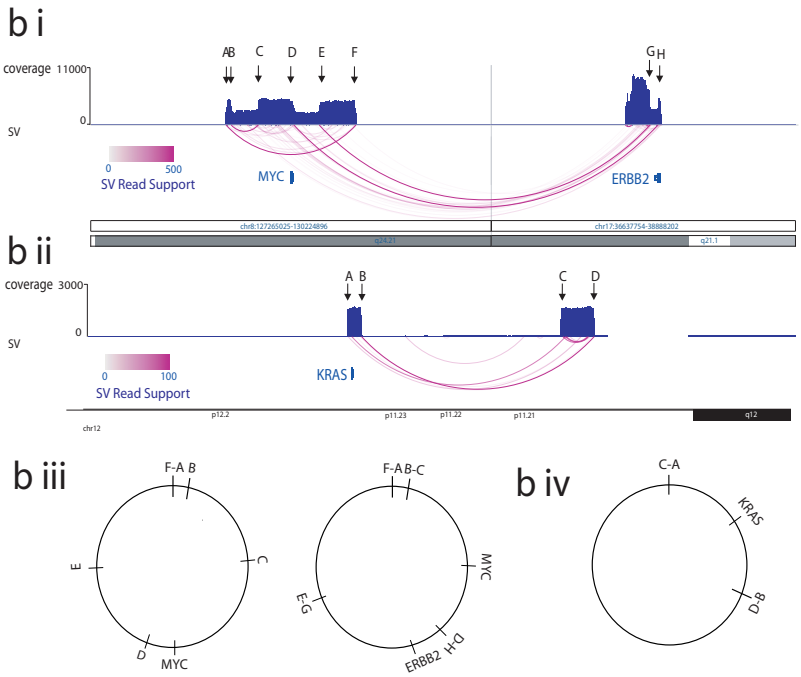


Figure 2. Copy number variation under positive selection. **a.** Recurrent copy number changes across the genome identified by GISTIC. Frequency of different CNV types are indicated as well as the position of CNV high confidence driver genes and candidate driver genes. The q value for expression correlation 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 (bi) two populations of extrachromosomal DNA are apparent (biii), one amplifying only MYC and the second also incorporating ERBB2 from a different chromosome. In the second example (bii) an inversion has occurred before circularization and amplification around KRAS (biv). **c.** Relationship between copy number and expression in CN driver genes.

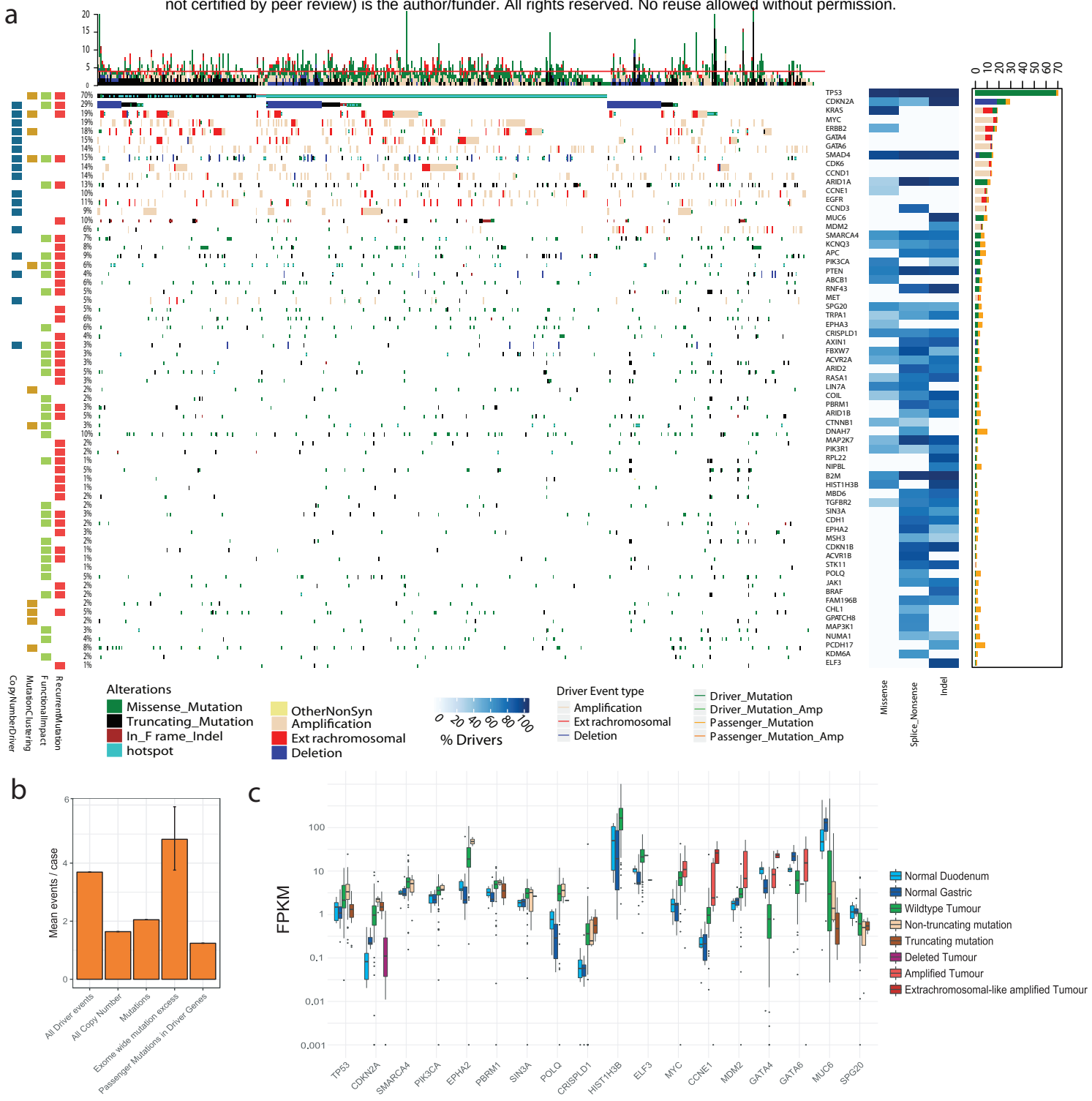
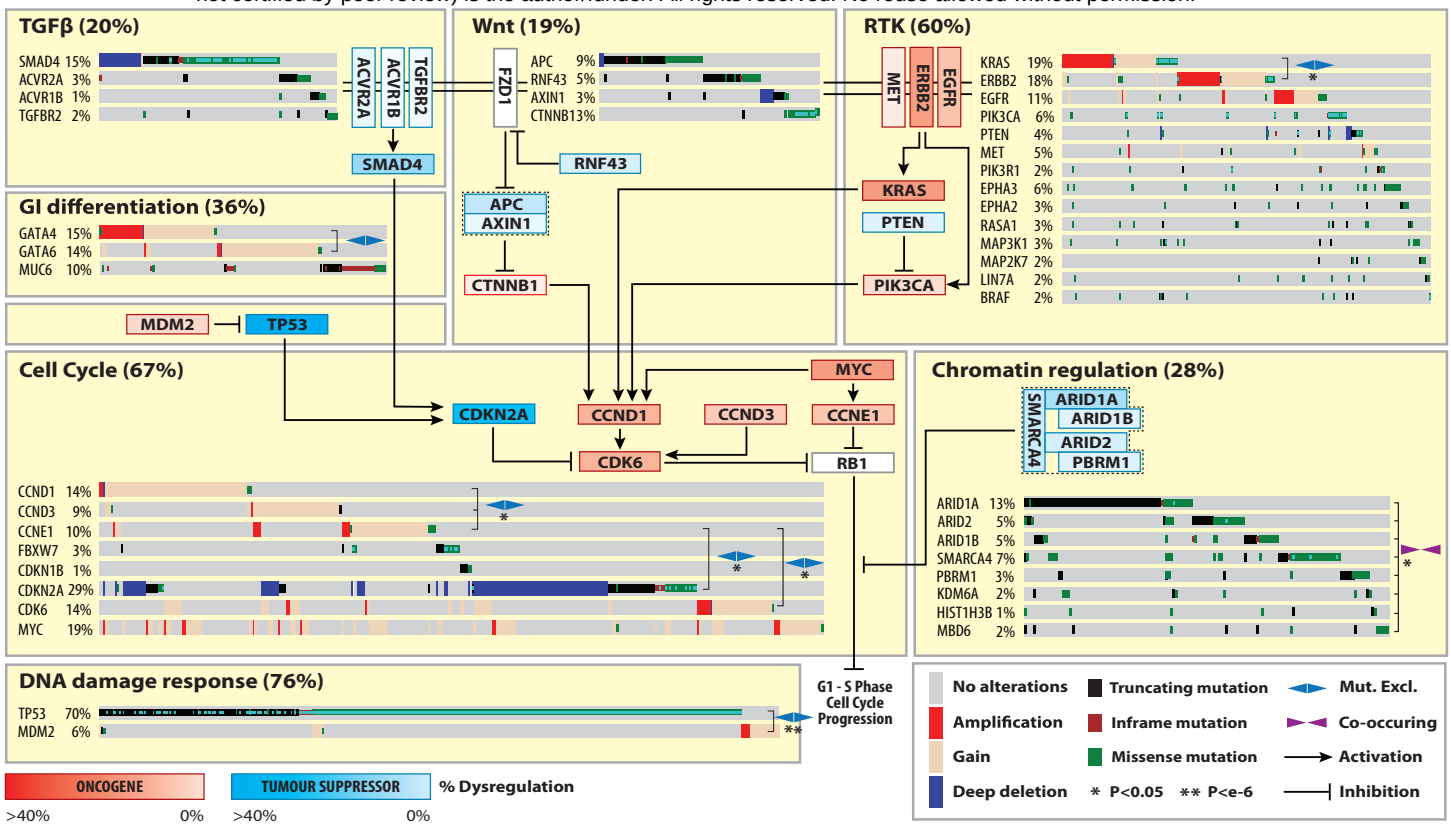


Figure 3. The driver gene landscape of Esophageal Adenocarcinoma. a Driver mutations or CNVs are shown for each patient. Amplification is defined as >2 Copy number adjusted ploidy (2 x ploidy of that case) and extrachromosomal amplification as >10 Copy number adjusted ploidy (10 x ploidy for that case). 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 median (red line = 4). **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. Only genes with signifcant expression changes of note are shown.

a



b

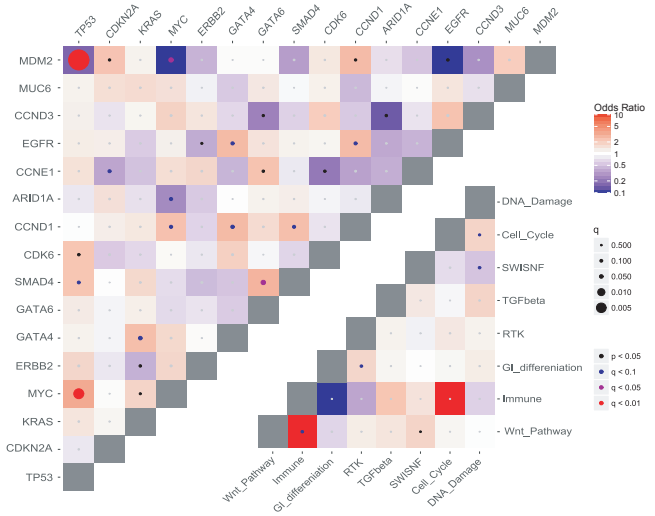


Figure 4. Biological pathways undergoing selective dysregulation in EAC. a. Biological Pathways dysregulated by driver gene mutation and/or CNVs. WT cases for a pathway are not shown. Mutual exclusivities and/or associations between genes in a pathway are annotated. GATA4/6 amplifications have a mutually exclusive relationship (ie GATA4 ampification is more common in GATA6 WT cases) although this does not reach statistical significance (fisher's exact test p=0.07 OR =0.52). **b.** Pairwise assessment of mutual exclusivity and association in EAC driver genes and pathways.

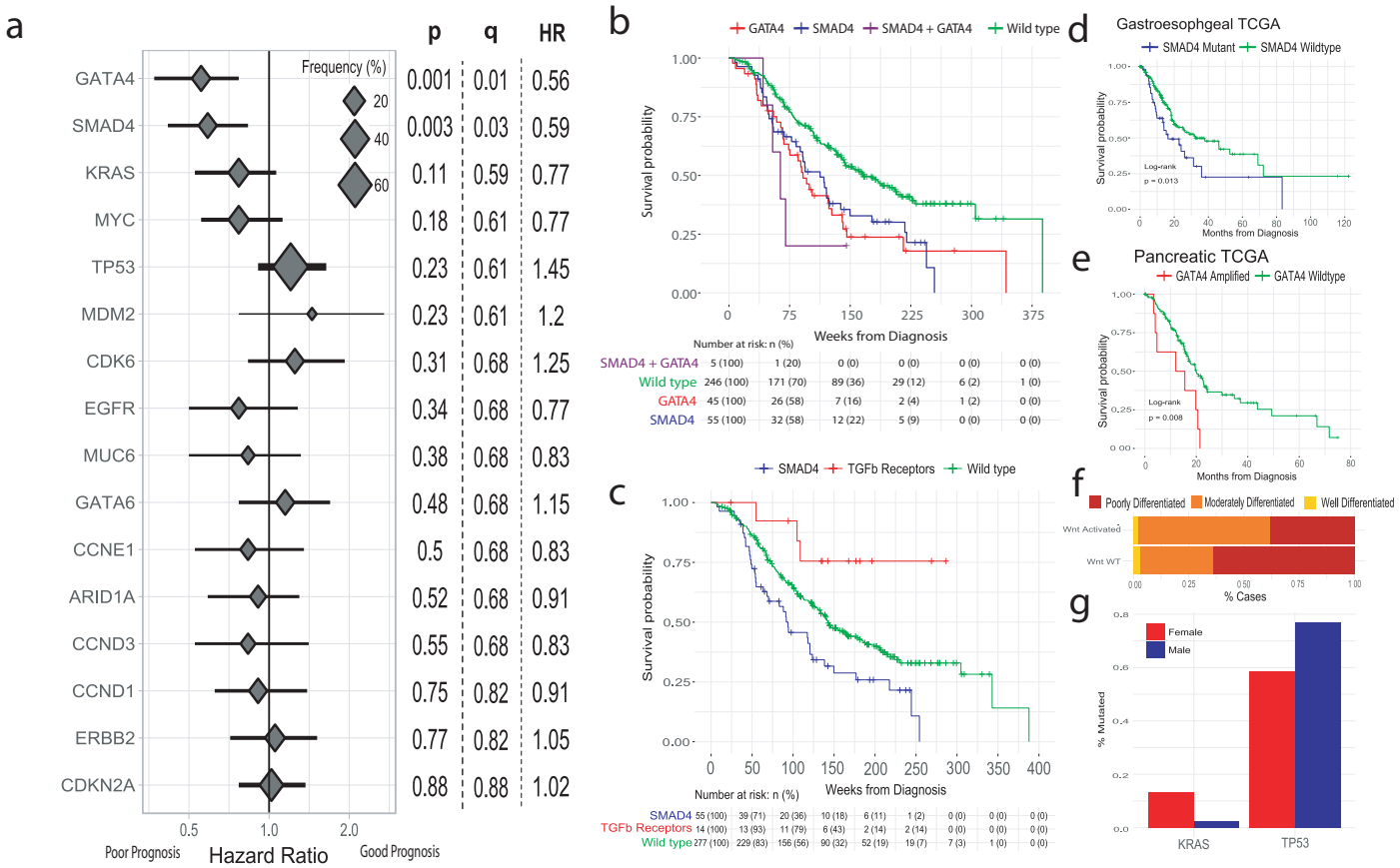


Figure 5. Clinical significance of Driver events in EAC. a. Hazard rations and 95% confidence intervals for Cox regression analysis across all drivers genes with at least a 5% frequency of driver alterations. P values are generated from the wald test and q values generated using BH correction. **b.** Kaplan-Meier curves for EACs with different status of significant prognostic indicators (GATA4 and SMAD4). **c.** Kaplan-Meier curves for different alterations in the TGFbeta 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).

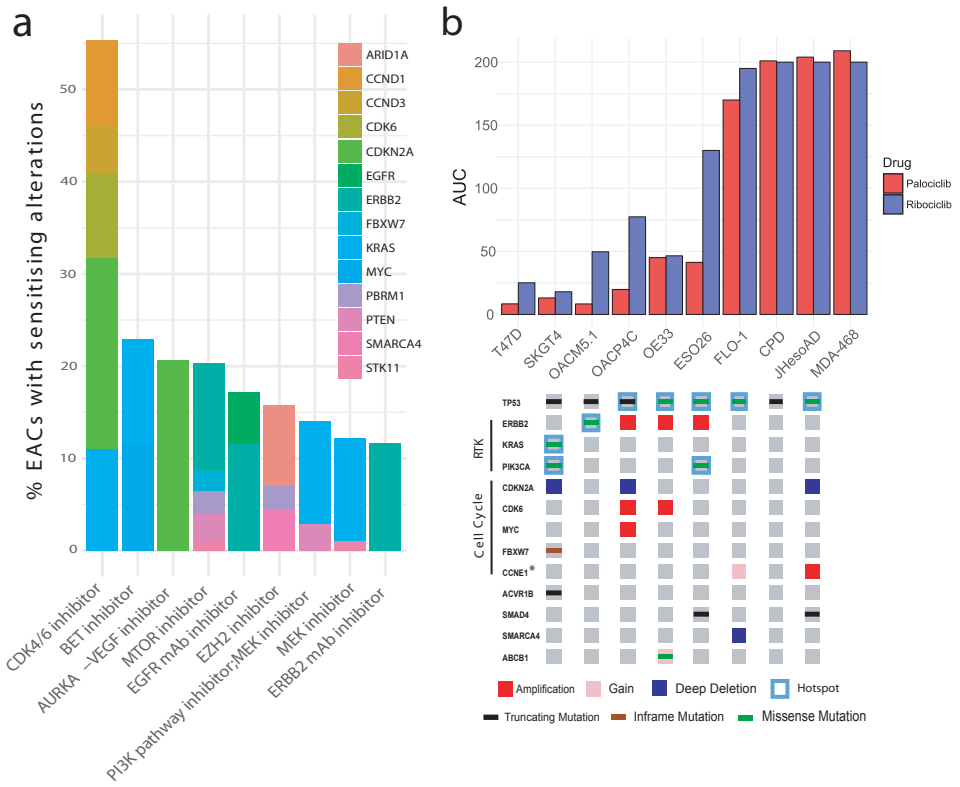


Figure 6. CDK4/6 inhibitors utility in EAC. a. Drug classes for which sensitivity is indicated by EAC driver genes with data from the Cancer Biomarkers database36. **b.** Area under the curve (AUC) of sensitivity is shown in a panel of 8 EAC cell lines with associated WGS and driver event, based in tumour analysis, in these cell lines indicated. 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 CDK4/6is due to its regulation of Rb downstream of CDK4/6 hence bypassing the need for CDK4/6 activity (see figure 4).