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

Alexander M Frankell¹, SriGanesh Jammula S², Xiaodun Li¹, Gianmarco Contino¹, Sarah Killcoyne¹,³, Sujath Abbas¹, Juliane Perner², Lawrence Bower², Ginny Devonshire², Emma Ococks¹, Nicola Grehan¹, James Mok¹, Maria O’Donovan⁴, Shona MacRae¹, Matthew D. Eldridge², Simon Tavaré², Rebecca C. Fitzgerald*¹ and the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium⁵

¹MRC cancer unit, Hutchison/MRC research centre, University of Cambridge, Cambridge, UK
²Cancer Research UK Cambridge institute, University of Cambridge, Cambridge, UK
³European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK
⁴Department of Histopathology, Cambridge University Hospital NHS Trust, Cambridge, UK
⁵A 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 case and compare driver mutation rates to the exome-wide mutational excess. We observe mutual exclusivity or co-occurrence of events within and between a number of EAC pathways 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 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. 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 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 genome in which mutations are enriched for these features\(^7\). 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\(^8\). MutsigCV has identified ten known cancer genes as EAC drivers (\textit{TP53, CDKN2A, SMAD4, ARID1A, ERBB2, KRAS, PIK3CA, SMARCA4, CTNNB1 and FBXW7})\(^6,9,10\). The PCAWG ICGC analysis also identified a significantly mutated enhancer associated with \textit{TP53TG1}\(^11\). However, these analyses leave most EAC cases with only one known driver mutation, usually TP53. Equivalent analyses in other cancer types have identified three or four drivers per case\(^12,13\). 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\(^9,14-17\). 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. 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 and whole genome sequencing 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 effects

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, ActivedriverWGS, MutsigCV2), high functional impact mutations within a gene (OncodriveFM, ActivedriverWGS), mutation clustering (OncodriveClust, eDriver and eDriver3D) and recurrent
amplification or deletion of genes (GISTIC\textsuperscript{14}) undergoing concurrent over or under-expression (see methods) (Fig 1A).  

These complementary methods produced highly significant agreement in calling EAC driver genes, particularly within the same feature-type (supplementary figure 2, Fig 1B). In total seventy six EAC driver genes were discovered, 86\% of which have not been detected in EAC previously\textsuperscript{9,10,15-17,19} and 69\% are known drivers in pan-cancer analyses \textsuperscript{21,27,28}. To detect driver elements in the non-coding genome we used ActiveDriverWGS\textsuperscript{22} a recently benchmarked\textsuperscript{29} 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\textsuperscript{11} including the enhancer on chr7 linked to \textit{TP53TG1} previously identified in EAC and the promoter/5’UTR regions of \textit{PTDSS1} and \textit{WRD74} were found in other cancer types. We also identified non-coding cancer driver elements novel to cancer including in the 5’UTR of \textit{MMP24} and promoters of two related histones (\textit{HIST1H2BO} and \textit{HIST1H2AM}).  

EAC is notable among cancer types for harbouring a high degree of chromosomal instability\textsuperscript{20}. 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 2 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\textsuperscript{30}. We observed highly significant expression changes in 17 known cancer genes within GISTIC peaks such as \textit{ERBB2}, \textit{KRAS} and \textit{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, $P = 4.97 \times 10^{-8}$) (supplementary figure 4). We use copy number adjusted ploidy to define amplifications as it produces superior correlation with expression data than absolute CN alone. Ploidy of our samples varies from 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 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 and other neoplasms, 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, $P = 9.62 \times 10^{-16}$ and $P = 7.64 \times 10^{-11}$ 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 = 2.5 normalised 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 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 6). Mutational hotspots in EAc or other cancer types (supplementary table 7) 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, MUC6 a secreted glycoprotein involved in gastric acid resistance and ABCB1 a channel pump protein which is associated with multiple instances of drug resistance. We note that several of these drivers have been previously associated
with gastric and colorectal cancer (supplementary table 8). 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 five events in driver genes per tumour (IQR = 3-7, Mean = 5.6) and only a very small fraction of cases has no such events detected (6 cases, 1%). When we remove the predicted percentage of passenger mutations using dnds ratios we find a mean of 4.4 true driver events per case which derive more commonly from mutations than CN events (Fig 3B, Supplementary table 9). Using hierarchal clustering of drivers we noted that TP53 mutant cases had significantly more CN drivers (Wilcox test, p = 0.0032, supplementary figures 7 and 8). dNdScv, one of the driver gene detection methods used, also analyses the genome-wide excess of non-synonymous mutations based on expected mutation rates to assess the total number of driver mutations across the exome which is calculated at 5.4 (95% CIs: 3.5-7.3) in comparison to 2.7 driver mutations which we calculate in our gene-centric analysis after passenger removal.

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 (Fig 3C, supplementary figure 10). Since surrounding squamous epithelium is a fundamentally different tissue, from which EAC does not directly arise, we have used duodenum and gastric cardia samples as gastrointestinal phenotype controls, likely to be similar to the, as yet unconfirmed, tissue of origin in EAC. A large number of driver genes have upregulated expression in comparison to normal controls, for example TP53 has upregulated RNA expression in WT tumour tissue and in cases with non-truncating mutations but RNA expression is lost upon gene truncation. In depth analysis of different TP53 mutation types reveals significant heterogeneity within non-truncating mutations (supplementary figure 9). Normal tissue expression of CDKN2A suggests that CDKN2A is generally activated in EAC, likely due to genotoxic or other cancer-associated cellular stresses 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. 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 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 10) suggesting some of these may be pan-cancer phenomenon. Wnt dysregulation was associated with hyper-mutated cases (> 50,000 SNVs or Indels, fisher’s exact test, p = 0.021, OR = 2.4) as was mutation in immune pathway genes (B2M and JAK1, > 50,000 SNVs or Indels, fisher’s exact test, p = 0.012, OR = 9.6).

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.

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 amplification (HR : 0.54, 95% CI : 0.38 – 0.78, P value = 0.0008) and SMAD4 mutation or homozygous deletion (HR : 0.60, 95% CI : 0.42 – 0.84, P value = 0.003). Both genes remained significant in multivariate Cox regression including pathological TNM staging, resection margin, curative vs palliative treatment intent and differentiation status (GATA4 = HR adjusted : 0.47, 95% CIs adjusted : 0.29 - 0.76, P value = 0.002 and SMAD4 = HR adjusted : 0.61, 95% CI adjusted : 0.40 – 0.94, P value = 0.026). 31% of EACs contain either SMAD4 mutation or homozygous deletion or GATA4 amplification (Fig 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\textsuperscript{16} although it did not reach statistical
significance after fdr correction and SMAD4 expression loss has been previously linked to poor
prognosis in EAC\textsuperscript{38}. 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\textsuperscript{39}.

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, 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\textsuperscript{3}.

Targeted therapeutics using EAC driver events.

To investigate whether the driver events in particular genes and/or pathways might sensitisie EAC
cells to certain targeted therapeutic agents we used the Cancer
Biomarkers database\textsuperscript{40}. 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 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 \textit{CCND1}, \textit{CCND3} and \textit{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 lines which have undergone whole genome sequencing\textsuperscript{41} 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)\textsuperscript{42}. 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 \textit{CCNE1} which is known to drive resistance to CDK4/6 inhibitors by bypassing CDK4/6 and causing Rb phosphorylation via CDK2 activation\textsuperscript{43}. To verify these effects in a more representative model of EAC we treated three whole genome sequenced EAC organoid cultures\textsuperscript{44} 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 \textit{CCNE1} activation in the most resistant (Fig 6C). We found Abemaciclib to be significantly more potent in comparison to both other CDK4/6 inhibitors, in organoids and cell lines (supplementary figure 10).

\textbf{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. 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 suggest 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.

While a number of previous reports have attempted to detect EAC drivers, they have had a limited yield per case. The first such study used methods which, despite being well regarded at the time, were subsequently discredited. Since then a number of reports, including our own, on medium and large cohort sizes using MutsigCV 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 recurrently CN aberrant regions 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 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, PPM1D and APC).

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

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. 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 EAC is a poor prognosis cancer type, significant heterogeneity of survival outcome makes triaging patients in treatment groups an important part of clinic practice which could be improved using better prognostication. 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.

Previously work has noted activity of the CDK4/6 inhibitor Palbociclib in a small number of EAC cell lines, however biomarkers were not investigated. The extensive \textit{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 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.

References


Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium:

Rebecca C. Fitzgerald, Ayesha Noorani, Paul A.W. Edwards, Nicola Grehan, Barbara Nutzinger, Caitriona Hughes, Elwira Fidziukiewicz, Jan Bornschein, Shona MacRae, Jason Crawte, Alex Northrop, Gianmarco Contino, Xiaodun Li, Rachel de la Rue, Maria O'Donovan, Ahmad Miremadi, Shalini Malhotra, Monika Tripathi, Simon Tavaré, Andy G. Lynch, Matthew Eldridge, Maria Secrrier, Lawrence Bower, Ginny Devonshire, Juliane Perner, Srijaneni, Jammula, Jim Davies, Charles Crichton, Nick Carroll, Peter Safranek, Andrew Hindmarsh, Vijayendran Sujendran, Stephen J. Hayes, Yeng Ang, Shaun R. Preston, Sarah Oakes, Izhar Bagwan, Vicki Save, Richard J.E. Skipworth, Ted R. Hupp, J. Robert O'Neill, Olga Tucker, Andrew Beggs, Philippe Taniere, Sonia Puig, Timothy J. Underwood, Fergus Noble, Jack Owsley, Hugh Barr, Neil Shepherd, Oliver Old, Jesper Lagergren, James Gossage, Andrew Davies, Fuji Chang, Janine Zylstra, Ula Mahadeva, Vicky Goh, Francesca D. Ciccarelli, Grant Sanders, Richard Berriosford, Catherine Harden, Mike Lewis, Ed Cheong, Bhaskar Kumar, Simon L Parsons, Irshad Soomro, Philip Kaye, John Saunders, Laurence Lovat, Rehan Haidry, Laszlo Igali, Michael Scott, Sharmila Sothi, Sari Suortamo, Suzy...
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 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 ratios and 95% confidence intervals for Cox regression analysis across all driver genes with at least a 5% frequency of driver alterations * = q < 0.05 after BH adjustment. b. Kaplan-Meier curves for 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. 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
Classification and Molecular Stratification) consortium. The procedures for obtaining the samples, quality control processes, extractions and whole genome sequencing are as previously described. Strict pathology consensus review was observed for these samples with a 70% cellularity requirement before inclusion. Comprehensive clinical information was available for the ICGC-OCCAMS cases (Supplementary Table 13). In addition, previously published samples were included in the analysis from Dulak et al 2013 – 139 WES and 10 WGS (total 27%) and Nones et al 2014 with 22 WGS samples (4%) to total 551 genome characterised EACs. RNA-seq data was available from our ICGC WGS samples (116/379). 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), copy number (ASCAT) and structural variant (Manta) calling pipelines, as previously described. Our methods were benchmarked against various other available methods and have among the best sensitivity and specificity for variant calling (ICGC benchmarking exercise). Mutation and copy number calling on cell lines was performed as previously described. Amplifications were defined as genes with 2x the average copy number of the host chromosome or greater.

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. Paired end 75bp sequencing on HiSeq4000 generated the paired end reads. For normal expression controls we chose gastric cardia tissue, from which some hypothesise Barrett’s may arise, and duodenum which contains intestinal histology, including goblet cells, which mimics that of Barrett’s. We did not use Barrett’s tissue itself as a normal control given the heterogeneous and plentiful phenotypic and genomic changes which it undergoes early in its pathogenesis.
Analysing 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.\[7,11\]

dNdScv, MutsigCV, e-Driver, ActivedriverWGS and e-Driver3D were run using the default parameters. To run OncodriverFM, Polyphen and SIFT 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 (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 $p = \text{Ex}10^{-13}$ to prevent infiltration of large numbers of, likely, false positive genes. For all tool outputs we undertook quality control including Q-Q plots to ensure no tool produces inflated q-values and each tool produced at least 30% known cancer genes. Two tools were removed from the analysis due to failure for both of these parameters at quality control (Activedriver and Hotspot). For three of the QC-approved tools (dNdScv, OncodriverFM, 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 as has been previously implemented. 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 dNdS was not
possible and despite recent benchmarking\textsuperscript{29} 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\textsuperscript{11}. 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 dnds ratios were applied to all dndscv annotated
mutations excluding “synonymous” and “no SNV” annotations as described in Martincorena et al\textsuperscript{21}.

**Detecting selection in CNVs**

ASCAT raw CN values were used to detected frequently deleted or amplified regions of the genome
using GISTIC2.0\textsuperscript{14}. 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
the copy number in a highly aberrant genomic regions. We found that the expression of genes in such regions matched well what we would expect given the surrounding copy number and hence we used the mean of the two adjacent copy number fragments to call copy number in the gene in question. We found amplification peak regions identified by GISTIC2.0 varied significantly in precise location both in analysis of different sub-cohorts and when comparing to published GISTIC data from EACs\textsuperscript{9,15,16}. A peak would often sit next to but not overlapping a well characterised oncogene or tumour suppressor. To account for this, we widened the amplification peak sizes upstream and downstream by twice the size of each peak to ensure we captured all possible drivers. Our expression analysis allows us to then remove false positives from this wider region and called drivers were still highly enriched for genes closer to the centre of GISTIC peak regions.

To detect genes in which amplification correlated with increased expression we compared expression of samples with a high CN for that gene (above 10\textsuperscript{th} percentile CN/Ploidy) with those which have a normal CN (median +/- 1) using the Wilcoxon rank-sum test and using the specific alternative hypothesis that high CN would lead to increased expression. Q-values were then generated based on Benjamini & Hochberg method, not considering genes without significant expression in amplified samples (at least 75% amplified samples with TPM > 0.1) and considering q<0.001 as significant. We also included an additional known driver gene only fdr reduction analysis as previously described for mutational drivers with q<0.1 considered as significant given the additional evidence for these genes in other cancer types. We also included MYC despite its p= 0.11 for expression correlation. This is due to frequent non-amplification associated overexpression of MYC when compared to normal controls and otherwise MYC is well evidence by a very close proximity to the peak centre (top 4 genes) and its high rate of amplification (19%). We took the same approach to detect genes in which homozygous deletion correlated with expression loss, comparing cases with CN = 0 to all others. Large expression modulation was a highly specific marker for known CN driver genes and was not a widespread feature in most recurrently copy number variant genes. However, while expression modulation is a requirement for selection of CNV only
drivers, it is not sufficient evidence alone and hence we grouped such genes into those which have
been characterised as drivers previously in other cancer types (high confidence EAC CN drivers) and
other genes (Candidate EAC CN drivers) which await functional validation. We used fragile site
regions detected in Wala et al 2017\textsuperscript{55}. We also defined regions which may be recurrently
heterozygous deleted, without any significant expression modulations, to allow LOH of tumour
suppressor gene mutations. To do this we analysed genes with at least 5 mutations for association
between LOH (ASCAT minor allele = 0) and mutation using fisher’s exact test and generated q values
using the Benjamini & Hochberg method. The analysis was repeated on known cancer genes only for
reduced fdr and q < 0.1 considered significant for both analyses. For those high confidence drivers
we chose to define amplification as CN/ploidy (referred to as Ploidy adjusted copy number) because
this produces superior correlation with expression. We chose a cut off for amplification at CN/ploidy
= 2 as has been previously used, and causes a highly significant increase in expression in our CN-
driver genes when amplified.

706

Pathways and relative distributions of genomic events

The relative distribution of driver events in each pathway was analysed using a fisher’s exact test in
the case of pair-wise comparisons including WT cases. In the case of multi-gene comparisons such as
the cyclins we calculate the p value and odds ratio for gene in the group using a two sided fisher’s
exact test, correct by Benjamini & Hochberg, and combine resulting q values using the Fisher
method, genes without odds ratios > 2 for cooccurrence and < 0.5 for mutual exclusivity were
removed. For this analysis we also remove highly mutated cases (>500 exonic mutations, 41/551) as
they bias distribution of genes towards co-occurrence. To ensure that a non-random distribution of
mutations across samples was still not affecting the strong co-occurrence of SWI/SNF genes (all
genes q < 0.05 before combining q values) we repeated the analysis randomly iterating 30,000 times
over other gene eight combinations (excluding SWI/SNF genes) and found only 0.01% (4/30,000) of
random combinations had all genes q < 0.05 as found in SWI/SNF genes. We then performed this
analyses across all pairs of driver genes using two sided Fisher’s exact tests and Benjamini &
Hochberg multiple hypothesis correction, q values < 0.1 are shown in Fig4B. We validated these
relationships in independent TGCA cohorts of other GI cancers where we could find cohorts with
reasonable numbers of the genomic events in question (not possible for GATA4/6 for instance) using
the cBioportal web interface tool\(^56\).

Correlating genomics with the clinical phenotype

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

Therapeutics
The cancer biomarker database was filtered for drugs linked to biomarkers found in EAC drivers and supplementary table 8 constructed using the cohort frequencies of EAC biomarkers. 10 EAC cell lines (SKGT4, OACP4C, OACM5.1, ESO26, ESO51, OE33, MFD, OE19, Flo-1 and JHesoAD) and 3 BE high grade dysplasia cell lines (CP-B, CP-C and CP-D) with WGS data were used in proliferation assays to determine drug sensitivity to CDK4/6 inhibitors, Palbociclib (Biovision) and Ribociclib (Selleckchem).

Cell lines were grown in their normal growth media. Proliferation was measured using the Incucyte live cell analysis system (Incucyte ZOOM Essen biosciences). Each cell line was plated at a starting confluence of 10% and growth rate measured across 4-7 days depending on basal proliferation rate (until 90% confluent). For each cell-line drug combination concentrations of 16, 64, 250, 1000 and 4000 nanomolar were used each in 0.3% DMSO and compared to 0.3% DMSO only. Each condition was performed in at least triplicate (technical replicates) and 12/12 randomly chosen cell line – drug combinations were successfully replicated with biological replicates (independent experiments). The time period of treatment to growth cessation in the control (0.3% DMSO) condition was used to calculate GI50 and AUC. Accurate GI50s could not be calculated in cases where a cell line had >50% proliferation inhibition even with the highest drug concentration and hence AUC was used to compare cell line sensitivity. T47D had a highly similar GI50 for Palbociclib to that previously calculated in other studies (112 nM vs 127 nM). Primary organoid cultures were derived from EAC cases included in the OCCAMS/ICGC sequencing study. Detailed organoid culture and derivation method have been previously described. Regarding the drug treatment, the seeding density for each organoid line was optimised to ensure cell growth in the logarithmic growth phase. Cells were seeded in complete medium for 24 hours then treated with compounds at a 5-point 4-fold serial dilutions for 6 days or 12 days. Cell viability was assessed using CellTiter-Glo (Promega) after drug incubation. The Life sciences reporting summary is also provided in supplementary materials detailing exact software and biological materials used and efforts made to ensure reproducibility results.


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 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 ploidy-adjusted CN) and deletion peak (wilcox test, one sided, expression compared between homozygous deleted and all other cases) respectively and occasions of significant association between LOH and mutation are indicated in green (fisher's exact test, one sided). Benjamini & Hochberg false discovery correction was applied in each of these cases. 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 loess 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. 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 = 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 (i.e., GATA4 amplification is more common in GATA6 WT cases) although this does not reach statistical significance (Fisher’s exact test, two-sided, \( p = 0.07 \), \( \text{OR} = 0.52 \)). 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 avoid bias towards co-occurrence, 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 shown 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.
**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.
Supplementary figure 3. Frequency and significance of EAC non-coding drivers from ActiveDriverWGS using 379 WGS ICGC EACs. **a.** The observed and expected mutation counts found on each element in ActiveDriverWGS. **b.** The FDR for each element in ActiveDriverWGS.
**Supplementary figure 5. Examples of Normal amplification (Ploidy-adjusted CN >2 & <10) and Extrachromosomal-like amplification (ploidy-adjusted CN >10) events.** 1-10 = Extrachromosomal-like amplifications and 11-20 = Normal amplification events. Events were picked at random using runif() function in R. SV and CNAs surrounding events are shown. Features indicative of extrachromosomal double minute (DM) formations include sharp, large CN steps, SVs with high read support at the edges of these steps and, when not derived from a continuous region of the genome, CN fragments in the DM may have the same CN status (taking into account other additional events which may have occurred in that region). These features are enriched in the extrachromosomal-like events, although example 20 may be a low-copy number extrachromosomal event. It should be noted that SV calling using short read sequencing techniques such as in this study has a relatively low sensitivity and accuracy for the precise localisation of many SV break points. Examples continue over four pages.
11) Copy number

Structural variants
0

RefSeq genes

12) Copy number

Structural variants
0

RefSeq genes

13) Copy number

Structural variants
0

RefSeq genes

14) Copy number

Structural variants
0

RefSeq genes

15) Copy number

Structural variants
0

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

Supplementary figure 6. 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. 8 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 hence most (2/7) 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 7.** Hierarchical Clustering of samples based on presence of driver variants with genes ordered by pathway membership.
Supplementary figure 8. A detailed breakdown of mutation and copy number types per case and the exome wide dn/ds excess for different mutation types in 551 EACs. Exome wide excess indel rates cannot be calculated as they have no synonymous mutation equivalent. However a null model based on indel rates in genes presumed to be under no selection is used in the per gene dn/ds method. Error bars indicate 95% confidence intervals for exome wide dn/ds mutation excess assessment and bars indicate the mean.
Supplementary figure 9. TP53 expression in different TP53 mutation types in comparison to TP53WT tumours, normal duodenum and gastric cardia tissues in 116 WGS EAC cases with matched RNAseq data. Boxplots represent the median and interquartile range.
Supplementary figure 10. Expression of all EAC driver genes across different genomic states for the gene in question in 116 EAC tumours with matched RNAseq data, and in comparison to duodenum and gastric cardia normal tissues. Boxplots represent median and interquartile range.
Supplementary figure 11. Growth inhibition responses of 15 EAC, high grade dysplasia BE and control cell lines to CDK4/6 inhibitors Palbociclib, Ribociclib. A subset of cell lines also received treatment with Abemaciclib which shows efficacy in cell lines as well as organoids (fig 6C). Data points indicate the mean and error bars the standard error of the mean (SEM) across technical replicates.
Supplementary Table 8. Q values for every driver detection tool across all mutational drivers calculated from 551 EACs. Driver novelty in EAC, gastric and colorectal cancers is shown.

<table>
<thead>
<tr>
<th>gene_name</th>
<th>Mutation Frequency (%)</th>
<th>alphaknock</th>
<th>alphaknockmph</th>
<th>alphaknockmhp</th>
<th>muteff</th>
<th>tumor profiling</th>
<th>Novel_GenotypicDrivers</th>
<th>Pan-cancer Drivers*</th>
<th>gastric_Drivers**</th>
<th>Colorectal_Drivers***</th>
<th>Cancerex</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPRD1</td>
<td>72.2%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>BRR5D</td>
<td>31.2%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MEX3C</td>
<td>5.0%</td>
<td>0</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>ARID3A</td>
<td>3.6%</td>
<td>0.0038</td>
<td>0.0038</td>
<td>0.0038</td>
<td>0</td>
<td>0.0038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>ARID2</td>
<td>12.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>EGR3</td>
<td>11.2%</td>
<td>0.0189</td>
<td>0.0189</td>
<td>0.0189</td>
<td>0</td>
<td>0.0189</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>APC</td>
<td>8.3%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MARIAC</td>
<td>7.6%</td>
<td>0.0038</td>
<td>0.0038</td>
<td>0.0038</td>
<td>0</td>
<td>0.0038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>ARID2</td>
<td>5.6%</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>PFKFB3</td>
<td>3.8%</td>
<td>0.0189</td>
<td>0.0189</td>
<td>0.0189</td>
<td>0</td>
<td>0.0189</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>TSHZ3</td>
<td>3.2%</td>
<td>0.0189</td>
<td>0.0189</td>
<td>0.0189</td>
<td>0</td>
<td>0.0189</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>RNF43</td>
<td>2.8%</td>
<td>0.0189</td>
<td>0.0189</td>
<td>0.0189</td>
<td>0</td>
<td>0.0189</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>SDH5</td>
<td>1.27</td>
<td>0.0038</td>
<td>0.0038</td>
<td>0.0038</td>
<td>0</td>
<td>0.0038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MUC6</td>
<td>9.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>NEDD4L</td>
<td>9.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>BTS3</td>
<td>6.1%</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>ARID3A</td>
<td>6.1%</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>TUBGCP5</td>
<td>2.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MAF</td>
<td>3</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>ESR1</td>
<td>3</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>ACVR1B</td>
<td>3.6</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>NDRG1</td>
<td>1</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MAP2K1</td>
<td>1.27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>TEP1</td>
<td>1.27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MST1R</td>
<td>1.27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MCL1</td>
<td>1.27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>RHOA</td>
<td>10.7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>10.7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>BAK1</td>
<td>9.8</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MDR1</td>
<td>9.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>FCAS</td>
<td>9.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>CLCN7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>CCDC5</td>
<td>8.3%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>TPRF1</td>
<td>7.6%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>TPRF2</td>
<td>7.6%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>RHOA</td>
<td>6.1%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>PAK1</td>
<td>5.6%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>DNM1</td>
<td>4.4%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>AKT1</td>
<td>4.4%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>NF1</td>
<td>3.2%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>PTPRQ</td>
<td>3.2%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>C3orf62</td>
<td>3.2%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>CASC7</td>
<td>2.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>C6orf26</td>
<td>2.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>VIM</td>
<td>2.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>VIM</td>
<td>2.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>ZNF503</td>
<td>2.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>CASC7</td>
<td>2.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>KIAA1429</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>JAK1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>AGT</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>ARID1A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>FBXW7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>KCNQ3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>36.34</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
</tbody>
</table>

*** Gastric cancer drivers defined from TCGA 2016 Nature.
**** Colorectal cancer drivers defined from Sladek et al 2016 Cancer Discovery.
***** Deletion or expression loss.
****** Deletion or expression loss. * Pan-cancer drivers** Gastric_Drivers? *** Colorectal_Drivers? ****
Supplementary Table 10. Validation of mutual exclusivity and cooccurrence of genes and pathways in independent cohorts.

P values and OR are calculated from Fisher’s exact test (two tailed)

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Gene A</th>
<th>Gene B</th>
<th>Neither</th>
<th>A Not B</th>
<th>B Not A</th>
<th>Both</th>
<th>Log OR</th>
<th>OR</th>
<th>p-Value</th>
<th>Tendency</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSKCC pan-cancer cohort (n= 10,945)</td>
<td>B2M*</td>
<td>CTNNB1*</td>
<td>10443</td>
<td>130</td>
<td>358</td>
<td>14</td>
<td>1.145</td>
<td>13.964</td>
<td>&lt;0.001</td>
<td>Co-occurrence</td>
</tr>
<tr>
<td>MSKCC pan-cancer cohort (n= 10,945)</td>
<td>TP53</td>
<td>SMAD4</td>
<td>6097</td>
<td>4286</td>
<td>204</td>
<td>358</td>
<td>0.915</td>
<td>8.2224</td>
<td>&lt;0.001</td>
<td>Co-occurrence</td>
</tr>
<tr>
<td>MSKCC pan-cancer cohort (n= 10,945)</td>
<td>TP53</td>
<td>MYC</td>
<td>6130</td>
<td>4293</td>
<td>171</td>
<td>351</td>
<td>1.075</td>
<td>11.885</td>
<td>&lt;0.001</td>
<td>Co-occurrence</td>
</tr>
<tr>
<td>MSKCC pan-cancer cohort (n= 10,945)</td>
<td>TP53</td>
<td>MDM2</td>
<td>5913</td>
<td>4570</td>
<td>388</td>
<td>74</td>
<td>-1.399</td>
<td>0.0399</td>
<td>&lt;0.001</td>
<td>Mutual exclusivity</td>
</tr>
<tr>
<td>MSKCC pan-cancer cohort (n= 10,945)</td>
<td>ERBB2</td>
<td>KRAS</td>
<td>8478</td>
<td>661</td>
<td>1741</td>
<td>65</td>
<td>-0.736</td>
<td>0.1837</td>
<td>&lt;0.001</td>
<td>Mutual exclusivity</td>
</tr>
<tr>
<td>Pancreatic adenocarcinoma**</td>
<td>SMAD4</td>
<td>GATA6</td>
<td>373</td>
<td>129</td>
<td>16</td>
<td>14</td>
<td>0.928</td>
<td>8.4723</td>
<td>0.013</td>
<td>Co-occurrence</td>
</tr>
</tbody>
</table>

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

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

Supplementary Table 12. GI50s for Palbociclib and Ribociclib across all lines. GI50s were only considered calculable when at least one concentration fell below 50% growth inhibition

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Palbociclib</th>
<th>Ribociclib</th>
</tr>
</thead>
<tbody>
<tr>
<td>T47D (+ve Cntrl)</td>
<td>112</td>
<td>400</td>
</tr>
<tr>
<td>OACM5.1</td>
<td>82</td>
<td>497</td>
</tr>
<tr>
<td>SKGT4</td>
<td>217</td>
<td>458</td>
</tr>
<tr>
<td>OACp4C</td>
<td>149</td>
<td>1870</td>
</tr>
<tr>
<td>ESO26</td>
<td>823</td>
<td>6114</td>
</tr>
<tr>
<td>OE33</td>
<td>547</td>
<td>1166</td>
</tr>
<tr>
<td>ESO51</td>
<td>1132</td>
<td>2567</td>
</tr>
<tr>
<td>CP-B</td>
<td>&gt;4000</td>
<td>&gt;4000</td>
</tr>
<tr>
<td>CP-C</td>
<td>2518</td>
<td>&gt;4000</td>
</tr>
<tr>
<td>OE19</td>
<td>&gt;4000</td>
<td>&gt;4000</td>
</tr>
<tr>
<td>MFD</td>
<td>2773</td>
<td>&gt;4000</td>
</tr>
<tr>
<td>Flo-1</td>
<td>5574</td>
<td>&gt;8000</td>
</tr>
<tr>
<td>CP-D</td>
<td>&gt;8000</td>
<td>&gt;8000</td>
</tr>
<tr>
<td>JHesoAD</td>
<td>5668</td>
<td>&gt;8000</td>
</tr>
<tr>
<td>MDA-468 (-ve Cntrl)</td>
<td>&gt;8000</td>
<td>&gt;8000</td>
</tr>
</tbody>
</table>
Supplementary Table 13. Clinical Characteristics of EACs used in study

<table>
<thead>
<tr>
<th>Item</th>
<th>Full cohort (N=551)</th>
<th>ICGC cohort with full clinical annotation used for Prognostication (n=379)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Years (median, IQR)</td>
<td>67.3 (59.3-75.0)</td>
</tr>
<tr>
<td>Sex</td>
<td>Male</td>
<td>81/551</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>470/551</td>
</tr>
<tr>
<td>Barrett’s associated</td>
<td>Positive</td>
<td>252/469</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>217/469</td>
</tr>
<tr>
<td>Siewert type</td>
<td>Esophageal</td>
<td>127/417</td>
</tr>
<tr>
<td></td>
<td>GEJ (unspec.)**</td>
<td>46/117</td>
</tr>
<tr>
<td></td>
<td>GEJ1</td>
<td>101/300</td>
</tr>
<tr>
<td></td>
<td>GEJ2</td>
<td>115/300</td>
</tr>
<tr>
<td></td>
<td>GEJ3</td>
<td>28/300</td>
</tr>
<tr>
<td>Histopathological grading</td>
<td>Well</td>
<td>22/482</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>185/482</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>275/482</td>
</tr>
<tr>
<td>Pre-treatment Tumour Stage</td>
<td>T0</td>
<td>5/483</td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>86/483</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>68/483</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>289/483</td>
</tr>
<tr>
<td></td>
<td>T4</td>
<td>35/483</td>
</tr>
<tr>
<td>Pretreatment nodal involvement</td>
<td>Positive</td>
<td>156/475</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>319/475</td>
</tr>
<tr>
<td>Pretreatment distant metastases</td>
<td>Positive</td>
<td>27/322</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>195/322</td>
</tr>
<tr>
<td>Treatment pathway***</td>
<td>Endoscopic Resection</td>
<td>8/358</td>
</tr>
<tr>
<td></td>
<td>Surgical Resection only</td>
<td>68/358</td>
</tr>
<tr>
<td></td>
<td>Chemoradiotherapy + Surgery</td>
<td>244/358</td>
</tr>
<tr>
<td></td>
<td>Definitive chemo-radiotherapy</td>
<td>6/358</td>
</tr>
<tr>
<td></td>
<td>Palliative treatment</td>
<td>23/358</td>
</tr>
<tr>
<td></td>
<td>Palliative support care</td>
<td>9/358</td>
</tr>
<tr>
<td>Overall Survival****</td>
<td>Weeks (median, IQR)</td>
<td>101 (54-161)</td>
</tr>
</tbody>
</table>

* Denominators smaller that total cohorts indicate missing data
** In publically available studies tumour location is reported as Esophageal vs GEJ only, the siewert type is not specified
***Not reported in publically available studies
****Not reported in Dulak et al 2013