




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

Alexander M. Frankell ¹, SriGanesh Jammula², Xiaodun Li¹, Gianmarco Contino¹, Sarah Killcoyne^{1,3}, Sujath Abbas¹, Juliane Perner², Lawrence Bower², Ginny Devonshire ², Emma Ococks¹, Nicola Grehan¹, James Mok¹, Maria O'Donovan⁴, Shona MacRae¹, Matthew D. Eldridge², Simon Tavaré², the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium⁵ and Rebecca C. Fitzgerald ^{1*}

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

Esophageal cancer is the eighth most common form of cancer worldwide and the sixth most common cause of cancer-related death¹. Esophageal adenocarcinoma (EAC) is the predominant subtype in the West, and its incidence has been rapidly rising². EAC is a highly aggressive neoplasm that usually presents at a late stage and is generally resistant to chemotherapy, thus leading to an overall 5-year survival of <15% (refs. 1,3). In comparison to other cancer types, it is characterized by very high mutation rates⁴ but also, paradoxically, by a paucity of recurrently mutated genes. EAC displays marked chromosomal instability and thus may be classified as a C-type neoplasm, which may be driven mainly by structural variation rather than mutations^{5,6}. Currently, the understanding of precisely which genetic events drive the development of EAC is limited, and consequently there are few available molecular biomarkers for prognosis or targeted therapeutics.

Methods to differentiate driver mutations from passenger mutations use features associated with known drivers to detect regions of the genome in which mutations are enriched in these features⁷. 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 MutSigCV⁸. MutSigCV has identified 12 known cancer genes as EAC drivers (*TP53*, *CDKN2A*, *SMAD4*, *ARID1A*, *ERBB2*, *KRAS*, *PIK3CA*, *SMARCA4*, *CTNNB1*, *ARID2*, *PBRM1* and *FBXW7*)^{6,9,10}. The Pancancer Analysis of Whole Genomes (PCA-WG) International Cancer Genome Consortium (ICGC) analysis has also identified a significantly mutated enhancer associated with

TP53TGI (ref. 11). However, these analyses leave most EAC cases with only one known driver mutation, usually *TP53*. Equivalent analyses in other cancer types have identified three or four drivers per case^{12,13}. Similarly, detection of copy number driver events in EAC has relied on identifying regions of the genome recurrently deleted or amplified by using GISTIC (genomic identification of significant targets in cancer)^{9,14–17}. GISTIC often identifies relatively large regions of the genome, and there is little indication of which specific gene copy number aberrations (CNAs) actually confer a selective advantage. There are also several non-selection-based mechanisms that can cause recurrent CNAs, such as genomic fragile sites, which have not been well differentiated from selection-based CNAs¹⁸. Epigenetic events such as methylation may also be important sources of driver events in EAC but are much more difficult to formally assess for selection.

To address these issues, by using our esophageal ICGC project, we accumulated a cohort of 551 genomically characterized EACs with high-quality clinical annotation and associated whole-genome sequencing (WGS) and RNA sequencing (RNA-seq) data on cases with sufficient material. We augmented our ICGC WGS cohort with publicly available whole-exome sequencing¹⁹ and WGS²⁰ data and applied several complementary driver-detection methods to produce a comprehensive assessment of mutations and CNAs under selection in EAC. We used these events to define functional cell processes that have been selectively dysregulated in EAC and identified new, verifiable and clinically relevant biomarkers for

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 list of members and affiliations appears at the end of the paper.

*e-mail: rcf29@mrc-cu.cam.ac.uk

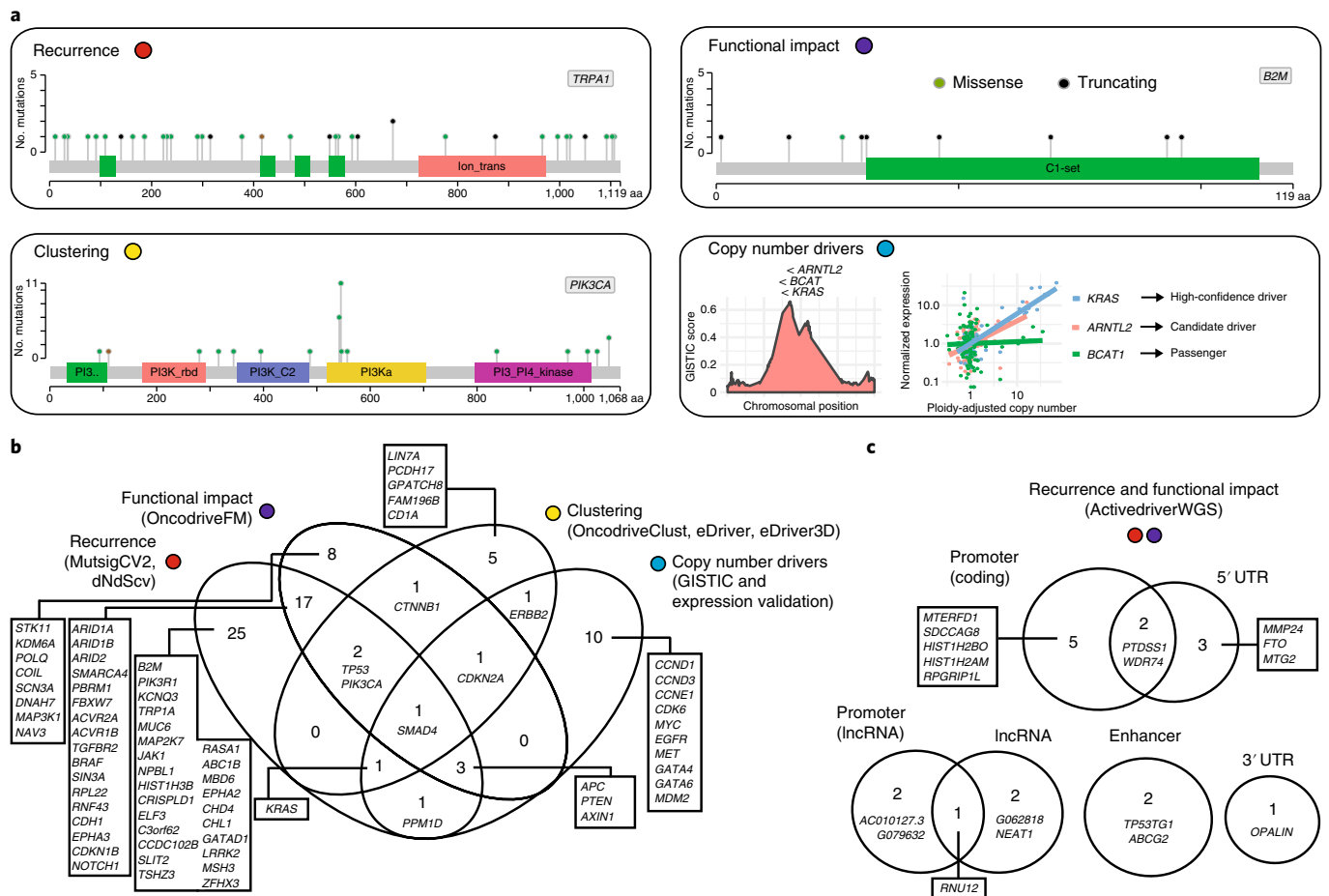


Fig. 1 | Detection of EAC driver genes. a, Types of driver-associated features used to detect positive selection in mutations and copy number events with examples of genes containing such features. **b**, Coding driver genes identified and their driver-associated features. **c**, Noncoding driver elements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

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

Results

Compendium of EAC driver events and their functional impact.

In 551 EACs, we identified a total of 11,813,333 single-nucleotide variants (SNVs) and small insertions or deletions (indels), with a median of 6.4 such mutations per megabase (Supplementary Fig. 1), and 286,965 CNAs. We also identified 134,697 structural variants in WGS cases. We use several complementary driver-detection tools to detect driver-associated features in mutations and CNAs (Fig. 1a). Each tool underwent quality control to ensure the reliability of the results (Methods). These features included highly recurrent mutations within a gene (dNdScv²¹, ActiveDriverWGS²² and MutSigCV2 (ref. ⁸)), high-functional-impact mutations within a gene (OncodriveFM²³ and ActiveDriverWGS²²), mutation clustering (OncodriveClust²⁴, eDriver²⁵ and eDriver3D²⁶) and recurrent amplification or deletion of genes (GISTIC¹⁴) undergoing concurrent over- or underexpression⁷ (Methods and Fig. 1a).

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

elements in the noncoding genome, we used ActiveDriverWGS²², a recently benchmarked method²⁹ using both functional impact and recurrence to determine driver status (Fig. 1c and Supplementary Fig. 3). We discovered 21 noncoding driver elements by using this method. We recovered several known noncoding driver elements from the pancancer PCAWG analysis¹¹, including the enhancer on chromosome 7, which is linked to *TP53TG1* and has been identified in EAC; the promoter and 5' untranslated regions of *PTDSS1* and *WDR74*. We also identified new noncoding cancer driver elements, including in the 5' untranslated region of *MMP24* and promoters of two related histone-encoding genes (*HIST1H2BO* and *HIST1H2AM*).

EAC is notable among cancer types for its high degree of chromosomal instability²⁰. Using GISTIC, we identified 149 recurrently deleted or amplified loci across the genome (Fig. 2a and Supplementary Tables 1 and 2). To determine which genes within these loci confer a selective advantage when they undergo CNAs, we used a subset of 116 cases with matched RNA-seq to detect genes in which homozygous deletion or amplification caused significant under- or overexpression, respectively (Supplementary Note and Supplementary Tables 3–6). Most genes in these regions showed no significant copy-number-associated expression change (74%), although work in larger cohorts suggests that we might have lacked the power to detect small expression changes³⁰. We observed highly significant expression changes in 17 known cancer genes within GISTIC loci, such as *ERBB2*, *KRAS* and *SMAD4*, which we designated high-confidence EAC drivers (Methods). We also found five

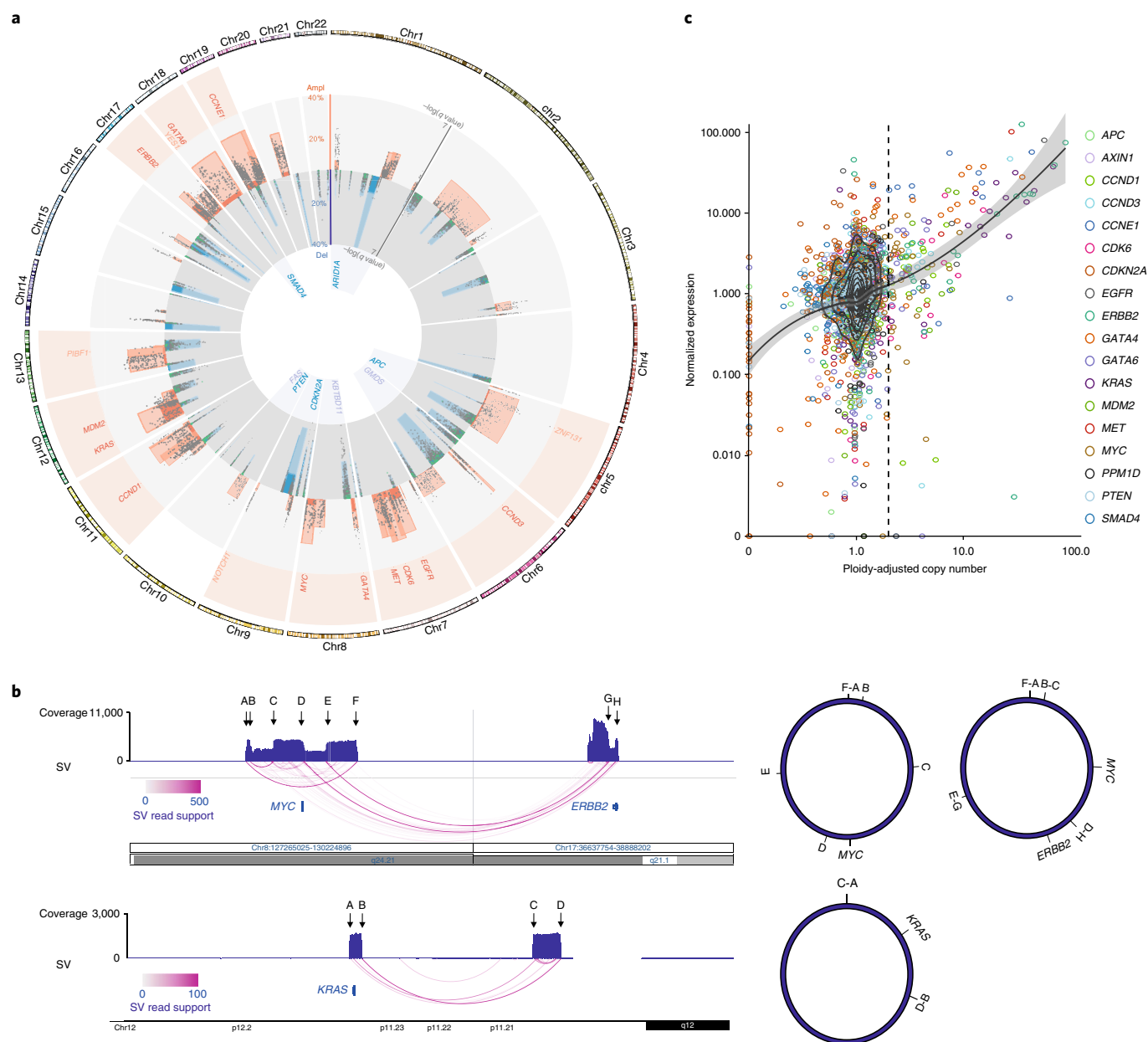


Fig. 2 | Copy number variation under positive selection. a, Recurrent copy number changes across the genome, identified by GISTIC in 551 EACs. Frequencies of different CNV types are indicated (dark blue, homozygous deletion; light blue, heterozygous deletion; dark red, extrachromosomal-like amplification;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

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

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

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

copy number alone. The ploidy of our samples varied from 1.4 to 6.2 (median 2.8), and hence a ploidy-adjusted copy number cutoff of >10 translated into >14 to 62 absolute copies (on average 28 copies). To discern a mechanism for these ultrahigh amplifications, we assessed structural variants associated with these events. For many of these events, the extreme amplification was produced largely from a single copy number step whose edges were linked by structural variants with ultrahigh read support. Two examples are in Fig. 2b, and further randomly selected examples are in Supplementary Fig. 5. In the first example, circularization and amplification initially occurred around *MYC* but subsequently incorporated *ERBB2* from an entirely different chromosome, and in the second, an inversion was followed by circularization and amplification of *KRAS*. Such a pattern of extrachromosomal amplification via double minutes has been noted in EAC²⁰ and other neoplasms³¹, and hence we refer to this amplification class with ultrahigh amplification (ploidy-adjusted copy number >10) as extrachromosomal-like amplifications.

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

Landscape of driver events in EAC. The overall landscape of driver-gene mutations and CNAs per case is depicted in Fig. 3a. These genes comprise both oncogenes and tumor-suppressor genes activated or repressed via different mechanisms. Passenger mutations occur by chance in most driver genes. For quantification, we used the observed/expected mutation ratios (calculated by dNdScv) to estimate the percentage of driver mutations in each gene and in different mutation classes. For many drivers, only specific mutation classes seemed to be under selection. Many tumor-suppressor genes (*ARID2*, *RNF43* and *ARID1B*, for example) are under selection for only truncating mutations, that is, splice-site, nonsense and frameshift indel mutations, but not missense mutations, which are passengers. However, oncogenes such as *ERBB2* contain only missense drivers that form clusters that activate gene function in a specific manner. When a mutation class is $<100\%$ driver mutations, mutational clustering can help to define the driver versus passenger status of a mutation (Supplementary Fig. 6). Mutational hotspots in EAC or other cancer types³² (Supplementary Table 7 and Supplementary Data) are indicated in Fig. 3a. Novel EAC drivers of particular interest include *B2M*, which encodes a core component of the MHC class I complex and is a marker of acquired resistance to immunotherapy³³; *MUC6*, which encodes a secreted glycoprotein involved in gastric acid resistance; and *ABCB1*, which encodes a channel pump protein associated with multiple instances of drug resistance³⁴. Notably, several of these drivers are associated with gastric and colorectal cancer^{13,35} (Supplementary Table 8).

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

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

To better understand the functional impact of driver mutations, we analyzed the expression of driver genes with different mutation types and compared their expression to normal tissue RNA (Fig. 3c and Supplementary Fig. 10). Because the surrounding squamous epithelium is a fundamentally different tissue from which EAC does not directly arise, we used duodenum and gastric cardia samples as gastrointestinal phenotype controls, which also have a columnar phenotype similar to EAC and Barrett's. Many driver genes had higher expression than that in normal controls; for example, *TP53* had upregulated RNA expression in wild-type tumor tissue and in cases with nontruncating mutations, but RNA expression was lost after gene truncation. In-depth analysis of different *TP53* mutation types revealed substantial heterogeneity within nontruncating mutations (Supplementary Fig. 9). The normal tissue expression of *CDKN2A* suggested that *CDKN2A* is generally activated in EAC, probably because of genotoxic or other cancer-associated cellular stresses³⁶, and returns to physiologically normal levels when deleted. Heterogeneous expression in wild-type *CDKN2A* cases suggested a different mechanism of inhibition, perhaps methylation, in some cases. Overexpression of some oncogenes occurs without genomic aberrations, such as *MYC*, which was overexpressed in *MYC*-wild-type EACs relative to normal tissues (Fig. 3c). Fewer driver genes were downregulated in EACs without genomic aberrations. Three-quarters of these genes (*GATA4*, *GATA6* and *MUC6*) are involved in the differentiated phenotype of gastrointestinal tissues and may be lost with tumor dedifferentiation.

Dysregulation of specific pathways and processes in EAC.

Selection preferentially dysregulates certain functionally related groups of genes and biological pathways in cancer³⁷. This phenomenon is highly evident in EAC, as shown in Fig. 4, which depicts the functional relationships between EAC drivers (Supplementary Note). Whereas *TP53* is the dominant driver in EAC, 28% of cases remain *TP53* wild type. MDM2 is an E3 ubiquitin ligase that targets *TP53* for degradation. Its selective amplification and overexpression is mutually exclusive with *TP53* mutation, thus suggesting that its degradation can functionally substitute for the effect of *TP53* mutation. Similar mutually exclusive relationships were observed among *KRAS* and *ERBB2*, *GATA4* and *GATA6*, and cyclin genes (*CCNE1*, *CCND1* and *CCND3*). Activation of the Wnt pathway occurred in 19% of cases, either by mutation of phosphorylated residues at the N terminus of β -catenin, preventing degradation, or loss of Wnt destruction-complex components such as APC. Many different chromatin-modifying genes, often belonging to the SWI-SNF complex, were also selectively mutated (28% of cases). In contrast to genes involved in other pathways, SWI-SNF genes were mutated significantly more often than expected by chance (two-sided Fisher's exact test, $q < 0.05$ for each gene; Methods), thus

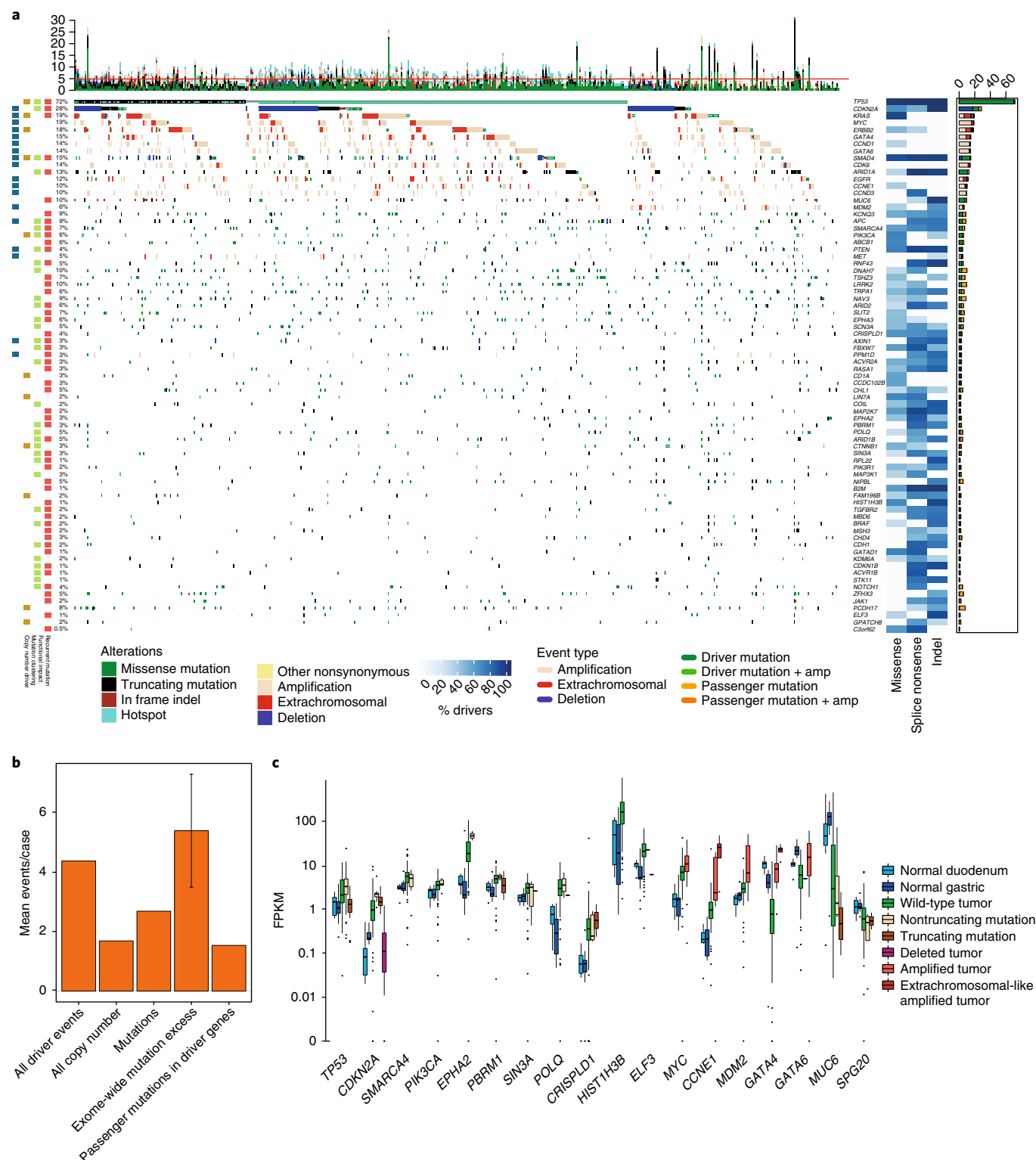


Fig. 3 | The driver-gene landscape of EAC. a, Driver mutations or CNVs are shown for each subject of 551 EACs. Amplification (amp) is defined as copy-number-adjusted ploidy >2 ($2\times$ ploidy of that case) and extrachromosomal amplification as >10 copy-number-adjusted ploidy ($10\times$ ploidy for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

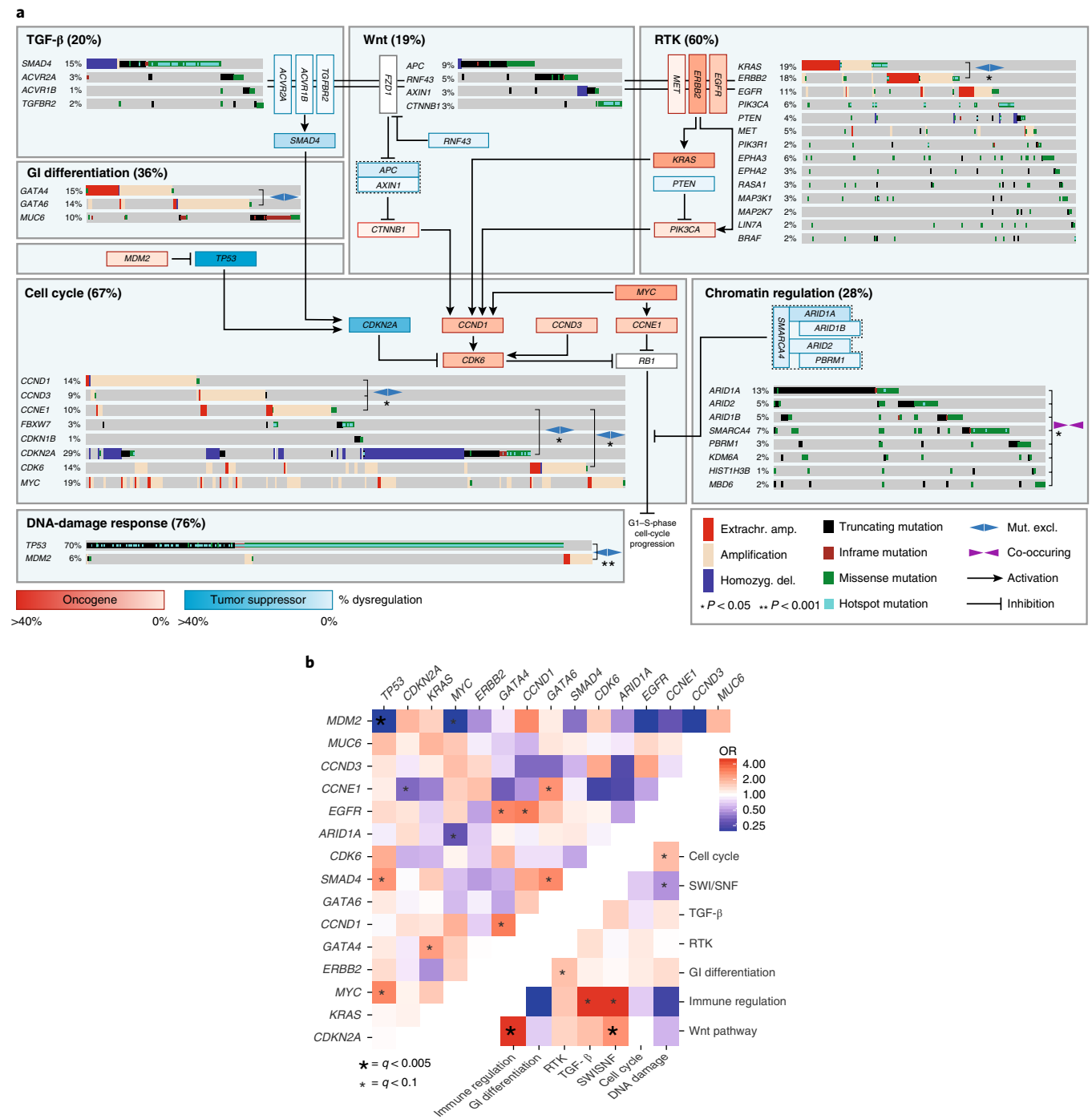


Fig. 4 | Biological pathways undergoing selective dysregulation in EAC. a, Biological pathways dysregulated by driver-gene mutation and/or CNVs in 551 cases. Wild-type cases for a pathway are not shown. Inter- and intrapathway interactions are described, and mutual exclusivities and/or associations between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

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

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

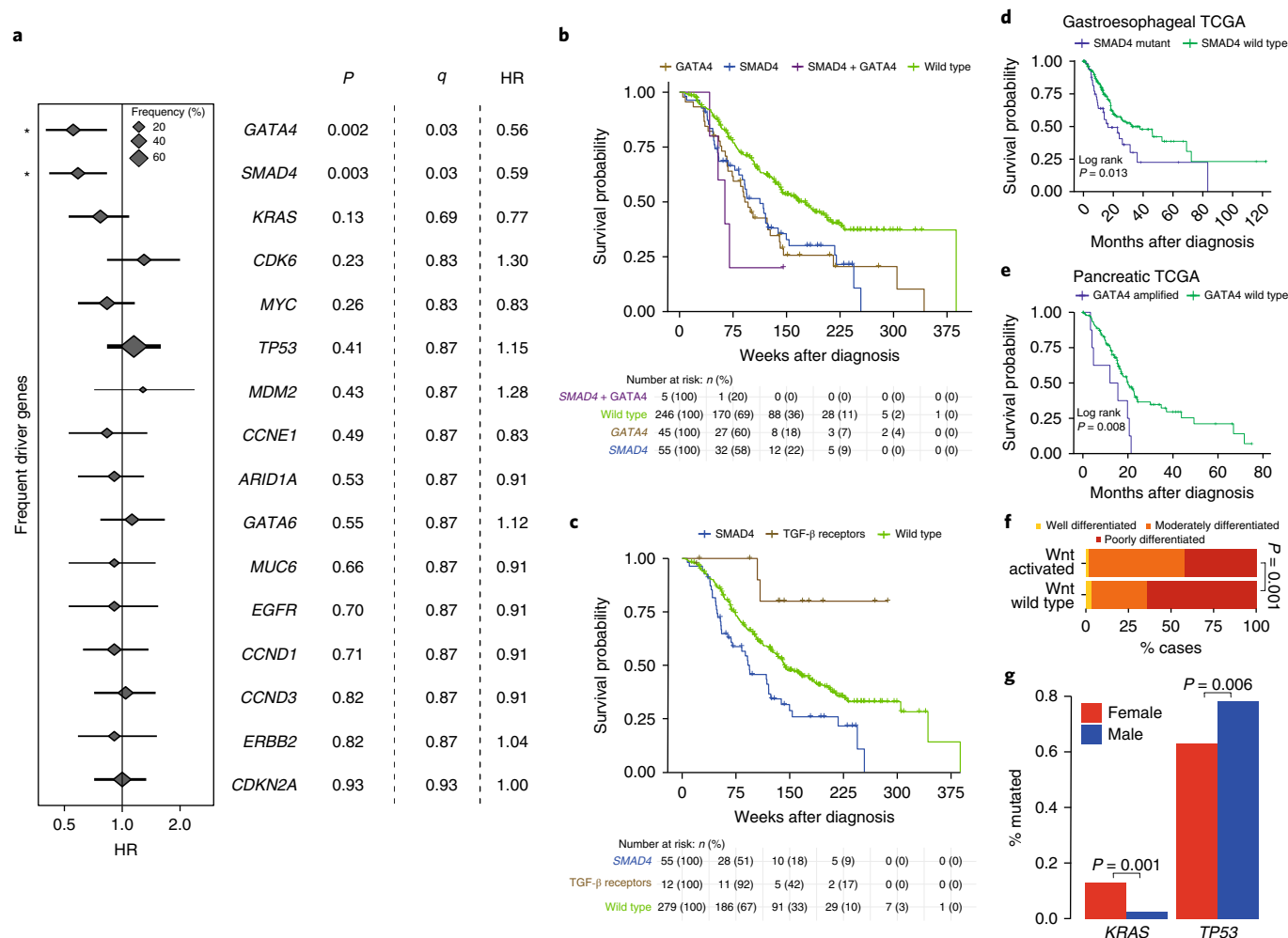


Fig. 5 | Clinical importance of driver events in 379 clinically annotated EACs. **a**, HRs and 95% CIs for Cox regression analysis across all driver genes with at least a 5% frequency of driver alterations. * $q < 0.05$ after Benjamini-Hochberg adjustment. **b**, Kaplan-Meier curves for EACs with different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

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

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

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

Genome Atlas (TCGA) gastroesophageal cohort (HR = 0.58, 95% CI = 0.37–0.90, $P = 0.014$) (Fig. 5c), and we also found that *GATA4* amplifications were prognostic in a cohort of TCGA pancreatic cancers (HR = 0.38, 95% CI = 0.18–0.80, $P = 0.011$) (Fig. 5d), the only available cohort containing a feasible number of *GATA4* amplifications. The prognostic effect of *GATA4* has been suggested in a previously published independent EAC cohort¹⁶, although it did not reach statistical significance after false discovery rate (FDR) correction in that study, and *SMAD4* expression loss has been linked to poor prognosis in EAC³⁸. We also noted stark survival differences between cases with *SMAD4* events and cases in which TGF- β receptors were mutated (Fig. 5e, HR = 5.6, 95% CI = 1.7–18.2, $P = 0.005$), in keeping with the biology of the TGF- β pathway, in which non-*SMAD* TGF- β signaling is oncogenic³⁹.

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

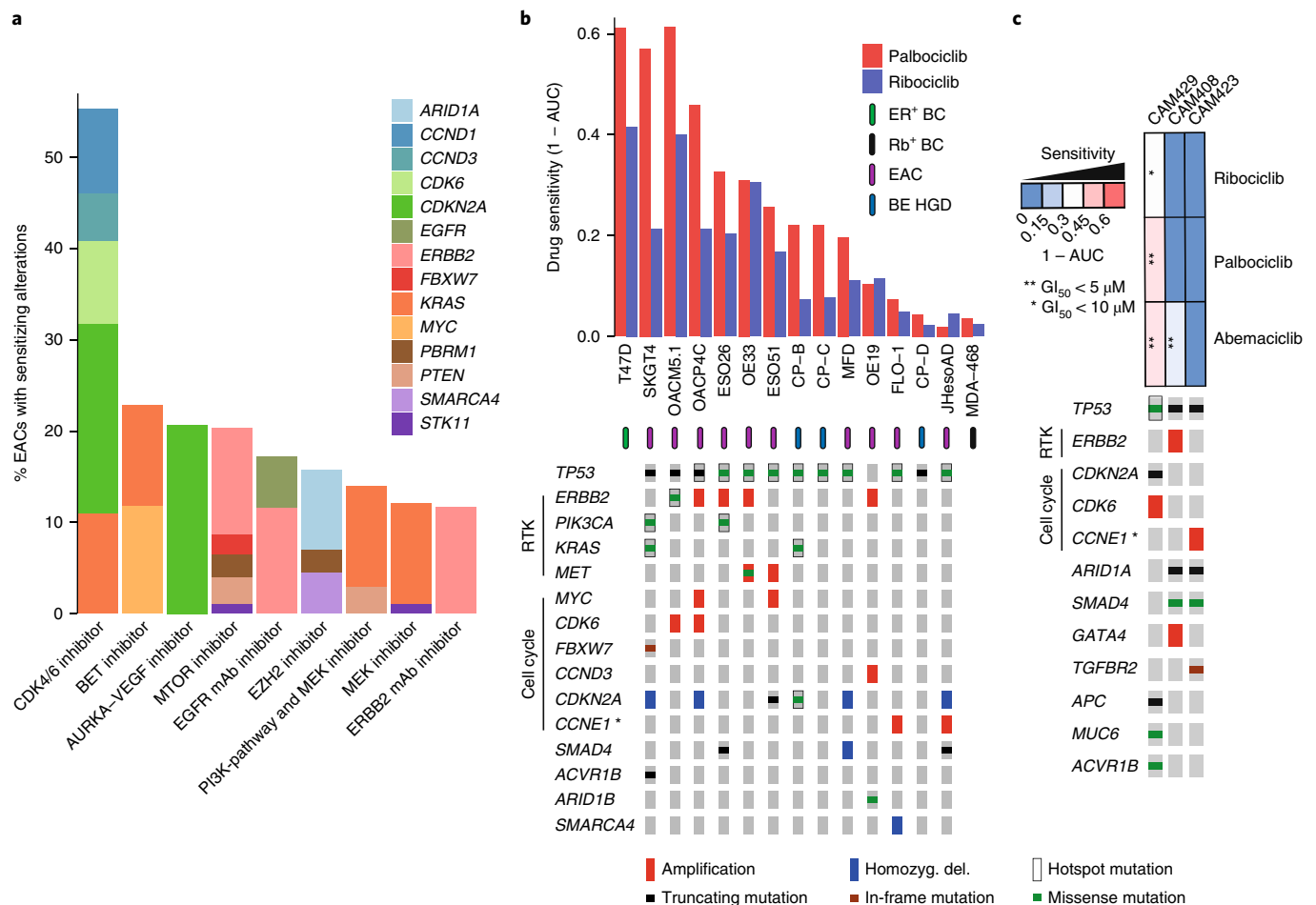


Fig. 6 | CDK4/CDK6 inhibitors in EAC. **a**, Drug classes for which sensitivity is indicated by EAC driver genes with data from the Cancer Biomarkers database³⁶. mAb, monoclonal antibody. **b**, AUC of sensitivity, shown for a panel of 13 EAC and Barrett's esophagus high-grade dysplasia (BE HGD) cell lines with associated WGS and their corresponding driver events, on the basis of primary tumor analysis. AUC is also shown for two control lines: T47D, an estrogen receptor (ER)-positive breast cancer (BC) line (positive control), and MDA-MB-468, a retinoblastoma-negative breast cancer (negative control). **CCNE1* is a known marker of resistance to CDK4/CDK6 inhibitors, owing to its regulation of retinoblastoma downstream of CDK4/CDK6, thus bypassing the need for CDK4/CDK6 activity (Fig. 4). **c**, Response of organoid cultures to three FDA-approved CDK4/CDK6 inhibitors and corresponding driver events.

This finding is of particular interest, given the male predominance of EAC³.

Targeted therapeutics based on EAC driver events. To investigate whether driver events, particularly genes and/or pathways, might sensitize EAC cells to certain targeted therapeutic agents, we used the Cancer Biomarkers database⁴⁰. We calculated the percentage of our cases that contained EAC-driver biomarkers of response to each drug class in the database (Fig. 6a and full data in Supplementary Table 11). Aside from *TP53*, which has been problematic to target clinically to date, we found several drugs with predicted sensitivity in >10% of EACs, including EZH2 inhibitors for some SWI-SNF-mutant cancers (23%, and 28% including all SWI-SNF EAC drivers), and BET inhibitors, which target *KRAS*-activated and *MYC*-amplified cases (23%). However, by far the most significantly effective class of drug was predicted to be inhibitors of CDK4 and CDK6 (CDK4/CDK6): >50% of cases had sensitivity-causing events in the RTK and core cell-cycle pathways (for example, in *CCND1*, *CCND3* and *KRAS*).

To verify that these driver events would also sensitize EAC tumors to such inhibitors, we used a panel of 13 EAC or Barrett's high-grade dysplasia cell lines that had undergone WGS⁴¹ and assessed them for the presence of EAC driver events (Fig. 6b).

The mutational landscape of these lines was broadly representative of EAC tumors. We found that the presence of cell-cycle and/or RTK-activating driver events was highly correlated with the response to two US Food and Drug Administration (FDA)-approved CDK4/CDK6 inhibitors, ribociclib and palbociclib, and several cell lines were sensitive below maximum tolerated blood concentrations in humans⁴² (Fig. 6b, Supplementary Table 12 and Supplementary Fig. 12). Such EAC cell lines had comparable sensitivity to T47D, which is derived from an estrogen-receptor-positive breast cancer in which CDK4/CDK6 inhibitors have been FDA approved. We noted three cell lines that were highly resistant, with little drug effect even at a concentration of 4,000 nM, similarly to a known retinoblastoma-mutant resistant breast cancer cell line (MDA-MB-468). Two of these three cell lines have amplification of *CCNE1*, which is known to drive resistance to CDK4/CDK6 inhibitors by bypassing CDK4/CDK6 and causing retinoblastoma phosphorylation via CDK2 activation⁴³. To verify these effects in a more representative model of EAC, we treated three whole-genome-sequenced EAC organoid cultures⁴⁴ with palbociclib and ribociclib, as well as a more recently approved CDK4/CDK6 inhibitor, abemaciclib. As observed in cell lines, cell-cycle and RTK driver events were present in only the more sensitive organoids, and *CCNE1* activation was present in only the most resistant organoid (Fig. 6c).

Discussion

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

Although the matched RNA-sequencing data are a strength of this study, we may not have been able to assess some uncommon variants for expression changes if these variants, detected in the full 551-patient cohort, were not well represented in the RNA-matched subcohort of 116 cases. Despite rigorous analyses to detect selected events, assessment of the global excess of mutations by dNdScv suggested that we could not detect all mutations selected in EAC, as in many other cancer types²¹. All driver-gene-detection methods that we used rely on driver-mutation recurrence in a genomic region to some degree. Many of these undetected driver mutations are hence probably spread across many genes, such that each is mutated at very low frequency across individuals with EAC. This tendency for low-frequency EAC drivers may be responsible for the low yield of MutSigCV in previous cohorts and may suggest that C-type cancers such as EAC are not less 'mutation driven' than M-type cancers but instead that their mutational drivers may be spread across a larger number of genes⁵. Copy number driver-gene identification is even more challenging because of the large size and lower frequency of these events, and hence many more EAC copy number drivers may remain to be discovered, some of which may have been identified as candidates here.

Although some previous reports have attempted to detect EAC drivers, they have had a limited yield per case. The first such study¹⁹ used methods that, despite being well regarded at the time, were subsequently discredited⁸. Since then, several reports, including our own, using MutSigCV^{9,10,17} on medium and large cohort sizes, have detected only a small number of mutational driver genes (7, 5 and 15 in each study, respectively). By using both a large cohort and more comprehensive methodologies, we markedly increased this figure to 66 mutational driver genes (excluding copy number drivers). Detection of driver CNAs has previously relied on GISTIC to detect regions with recurrent CNAs^{9,14–17}, but no analyses have been performed to determine which genes in these large regions are true drivers. Many of the genes annotated by such papers are unlikely to be copy number drivers, owing to their lack of expression modulation with CNAs (for example, *YEATS4* and *MCL1*), the role of recurrent heterozygous losses in driving LOH in some mutational drivers (*ARID1A* and *CDH11*) or their association with fragile sites (*PDE4D*, *WWOX* and *FHIT*). In contrast, we identified new EAC copy number drivers (for example, *CCND3*, *AXIN1*, *PPM1D* and *APC*).

We noted a three-way association among hypermutation, Wnt activation and loss of immune-signaling genes such as *B2M*. Microsatellite-instability-driven hypermutation has been associated with higher immune activity^{45,46}. However, Wnt dysregulation and mutation of immune-pathway genes such as *B2M*³³ have been linked to immunological escape⁴⁷, thus suggesting that this may be an acquired mechanism to prevent immune surveillance caused by hypermutation.

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

worthwhile. Although EAC is a poor-prognosis cancer type, substantial heterogeneity in survival outcomes makes triaging patients in treatment groups an important part of clinical practice that could be improved through better prognostication. Several targeted therapeutics may provide clinical benefit for EAC cases on the basis of individual genomic profiles. In particular, CDK4/CDK6 inhibitors deserve considerable attention as an option for EAC treatment because they are, by a large margin, the treatment for which the most EACs have sensitivity-causing driver events, excluding TP53 as an unlikely therapeutic biomarker at the current time. Previous work has noted the activity of the CDK4/CDK6 inhibitor palbociclib in a small number of EAC cell lines⁴⁸, but biomarkers were not investigated. The extensive in vitro validation of identified biomarkers for CDK4/CDK6 inhibitors in EAC across 16 cell lines and organoids suggests possible clinical benefit through use of a targeted approach.

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

Online content

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

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

Published online: 04 February 2019

References

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

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

Acknowledgements

We thank A. J. Bass and N. Waddell for providing data in Dulak et al.¹⁹ and Nones et al.²⁰, respectively, which were also included in our previous publication¹⁸. Inclusion of these data allowed for augmentation of our ICGC cohort and greater sensitivity for the detection of EAC driver variants. OCCAMS was funded by a Programme Grant from Cancer Research UK (RG66287), and the laboratory of R.C.F. is funded by a Core Programme Grant from the Medical Research Council. We thank the Human Research Tissue Bank, which is supported by the UK National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre, from Addenbrooke's Hospital. Additional infrastructure support was provided from the Cancer Research UK-funded Experimental Cancer Medicine Centre.

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

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

Reprints and permissions information is available at www.nature.com/reprints.

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

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

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

the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium

Rebecca C. Fitzgerald¹, Ayesha Noorani¹, Paul A. W. Edwards^{1,2}, Nicola Grehan¹, Barbara Nutzinger¹, Caitriona Hughes¹, Elwira Fidziukiewicz¹, Shona MacRae¹, Alex Northrop¹, Gianmarco Contino¹, Xiaodun Li¹, Rachel de la Rue¹, Annalise Katz-Summercorn¹, Sujath Abbas¹, Daniel Loureda¹, Maria O'Donovan^{1,4}, Ahmad Miremadi^{1,4}, Shalini Malhotra^{1,4}, Monika Tripathi^{1,4}, Simon Tavaré², Andy G. Lynch², Matthew Eldridge², Maria Secrier⁶, Ginny Devonshire², Juliane Perner², SriGanesh Jammula², Jim Davies⁷, Charles Crichton⁷, Nick Carroll⁸, Peter Safranek⁸, Andrew Hindmarsh⁸, Vijayendran Sujendran⁸, Stephen J. Hayes^{9,10}, Yeng Ang^{9,11,12}, Andrew Sharrocks¹², Shaun R. Preston¹³, Sarah Oakes¹³, Izhar Bagwan¹³, Vicki Save¹⁴, Richard J. E. Skipworth¹⁴, Ted R. Hupp¹⁴, J. Robert O'Neill^{14,15}, Olga Tucker^{16,17}, Andrew Beggs^{16,18}, Philippe Tanriere¹⁶, Sonia Puig¹⁶, Timothy J. Underwood^{19,20}, Robert C. Walker^{19,20}, Ben L. Grace¹⁹, Hugh Barr²¹, Neil Shepherd²¹, Oliver Old²¹, Jesper Lagergren^{22,23}, James Gossage^{22,24}, Andrew Davies^{22,24}, Fujun Chang^{22,24}, Janine Zylstra^{22,24}, Ula Mahadeva²², Vicky Goh²⁴, Francesca D. Ciccarelli²⁴, Grant Sanders²⁵, Richard Berrisford²⁵, Catherine Harden²⁵, Mike Lewis²⁶, Ed Cheong²⁶, Bhaskar Kumar²⁶, Simon L. Parsons²⁷, Irshad Soomro²⁷, Philip Kaye²⁷, John Saunders²⁷, Laurence Lovat²⁸, Rehan Haidry²⁸, Laszlo Igali²⁹, Michael Scott³⁰, Sharmila Sothi³¹, Sari Suortamo³¹, Suzy Lishman³², George B. Hanna³³, Krishna Moorthy³³, Christopher J. Peters³³, Anna Grabowska³⁴, Richard Turkington³⁵, Damian McManus³⁵, Helen Coleman³⁵, David Khoo³⁶ and Will Fickling³⁶

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

Methods

Cohort, sequencing and calling of genomic events. 379 cases (69%) of our EAC cohort were derived from the EAC WGS ICGC study, for which samples are collected through the UK-wide Oesophageal Cancer Classification and Molecular Stratification (OCCAMS) consortium. The procedures for obtaining the samples, quality-control processes, extractions and WGS were as previously described¹⁷. 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.¹⁹ (149 whole-exome sequencing samples; 27%) and Nones et al.²⁰ (22 WGS samples; 4%), for a total of 551 genome-characterized EACs. RNA-seq data were available from our ICGC WGS samples (116 of 379 samples). BAM files for all samples (including those from Dulak et al.¹⁹ and Nones et al.²⁰) were run through our alignment (BWA-MEM), mutation (Strelka), copy number (ASCAT) and structural-variant (Manta) calling pipelines, as described¹⁷. Our methods were benchmarked against various other available methods and have among the best sensitivity and specificity for variant calling (ICGC benchmarking exercise^{49,50}). Cell lines were subjected to WGS at 30× coverage with 150-bp paired-end reads on an Illumina HiSeq4000 instrument. Copy number calling was performed by Freec as described⁴¹. Mutations were called by GATK as described⁴¹ and filtered for germline variants in the 1000 Genomes Project, and any known oncogenic hotspots²³ were recovered. Amplifications were defined as genes with 2× the median copy number of the host chromosome or greater.

Total RNA was extracted with an All Prep DNA/RNA kit from Qiagen, and the quality was checked on an Agilent 2100 Bioanalyzer with an RNA 6000 nano kit (Agilent). A Qubit High Sensitivity RNA assay kit from Thermo Fisher was used for quantification. Libraries were prepared from 250 ng RNA, with a TruSeq Stranded Total RNA Library Prep Gold (Ribo-zero) kit, and ribosomal RNA (nuclear, cytoplasmic and mitochondrial rRNA) was depleted with biotinylated probes that selectively bind rRNA molecules, forming probe–rRNA hybrids. These hybrids were pulled down with magnetic beads, and rRNA-depleted total RNA was reverse transcribed. The libraries were prepared according to Illumina's protocol²¹. Paired-end 75-bp sequencing on a HiSeq4000 instrument generated the paired-end reads. For normal expression controls, we chose gastric cardia tissue, from which some hypothesize Barrett's esophagus may arise, and duodenum with intestinal histology, including goblet cells, which mimics the histology of Barrett's esophagus. We did not use Barrett's esophagus tissue itself as a normal control, given the heterogeneous and plentiful phenotypic and genomic changes that it undergoes early in its pathogenesis.

Analyzing EAC mutations for selection. To detect positively selected mutations in our EAC cohort, a multitool approach across various selection-related 'features' (recurrence, functional impact and clustering) was implemented to provide a comprehensive analysis. This procedure is broadly similar to those of several previous approaches^{7,11}, dNdScv²¹, MutSigCV², e-Driver²⁵, ActivedriverWGS²² and e-Driver3D²⁶ were run with the default parameters. To run OncodriverFM²³, Polyphen³² and SIFT³³ were used to score the functional impact of each missense nonsynonymous mutation (from 0, indicating nonimpactful, to 1, indicating highly impactful); synonymous mutations were given a score of 0 impact, and truncating mutations (nonsense and frameshift mutations) were given a score of 1. Any gene that had fewer than seven mutations and was unlikely to contain detectable drivers with this method was not considered to decrease the FDR. OncodriverClust was run with a minimum cluster distance of 3, a minimum number of mutations for a gene to be considered of 7 and a stringent probability cutoff to find cluster seeds of $P = 1 \times 10^{-13}$ to prevent infiltration of large numbers of probable false-positive genes. For all tool outputs, we undertook quality control including quantile–quantile plots to ensure that no tool produced inflated q values, and each tool produced at least 30% known cancer genes. Two tools were removed from the analysis, owing to the failure of both of these parameters in quality control in our hands (Activedriver²⁴ and Hotspot³³). For three of the quality-control-approved tools (dNdScv, OncodriverFM and MutSigCV) for which it was possible, we also undertook an additional FDR-reducing analysis by recalculating q values on the basis of analysis of known cancer genes only^{21,27,28}, as previously implemented^{21,55}. Significance cutoffs were set at $q < 0.1$ for coding genes. Tool outputs were then put through various filters to remove any further possible false-positive genes. Specifically, genes for which <50% of EAC cases had no expression (transcripts per kilobase million (TPM) <0.1) in our matched RNA-seq cohort were removed and, with dNdScv, genes with no or only a small mutation excess (observed/expected ratio >1.5:1) of any single mutation type were also removed. We also removed two mitochondrial genes (*MT-MD2* and *MT-MD4*) that were highly enriched in truncating mutations and were frequently called in OncodriverFM as well as other tools, possibly because of the different mutational dynamics caused by reactive oxygen species from the mitochondrial electron-transport chain and the high number of mitochondrial genomes per cell, which enables significantly more heterogeneity. These factors prevent the tools used from calculating an accurate null model for these genes, but they may be worthy of functional investigation. ActivedriverWGS calculates an expected background mutation rate on the basis of mutation rates of local, adjacent sequence for each tested element

while correcting for the differential mutation rates within each trinucleotide context; it thus tests observed mutation rates against this predicted background for each element. ActivedriverWGS also detects elements with mutations enriched in binding-site regions (high impact). For noncoding elements called by ActivedriverWGS, filtering for expression or dN/dS was not possible, and despite recent benchmarking²⁹, such methods are not well established. Hence, we took a more cautious approach with general significance cutoffs of $q < 0.001$ and $q < 0.1$ for previously identified elements in other cancer types¹¹. q values were not recalculated for previously identified elements alone, as with coding genes, but the $q < 0.1$ cutoff was calculated on the basis of P values for all assessed elements. To calculate exome-wide mutational excess, we removed hypermutated cases (>500 exonic mutations) and applied the global nonsynonymous dN/dS ratios to all dNdScv-annotated mutations, excluding 'synonymous' and 'no SNV' annotations, as described in Martincorena et al.²¹.

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

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

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

with Benjamini–Hochberg correction, and combined the resulting q values with the Fisher method; genes without $OR > 2$ for co-occurrence and < 0.5 for mutual exclusivity were removed. For this analysis, we also removed highly mutated cases (> 500 exonic mutations, 41 of 551), because they bias the distribution of mutations toward co-occurrence. To ensure that a nonrandom distribution of mutations across samples did not affect the strong co-occurrence of SWI–SNF genes (all genes $q < 0.05$ before q values were combined), we repeated the analysis, randomly iterating 30,000 times over all the other eight driver–gene combinations (excluding SWI–SNF genes) and found that only 0.01% (4 of 30,000) of random combinations had all genes $q < 0.05$, as found in SWI–SNF genes. We then performed these analyses across all pairs of driver genes with two-sided Fisher's exact tests and Benjamini–Hochberg multiple-hypothesis correction (q values < 0.1 are shown in Fig. 4b). We validated these relationships in independent TCGA cohorts of other gastrointestinal cancers in which we found cohorts with reasonable numbers of the genomic events in question (this procedure was not possible for *GATA4/GATA6*, for instance) with the cBioportal web interface tool³⁷.

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

Therapeutics. The cancer-biomarker database was filtered for drugs linked to biomarkers found in EAC drivers, and Supplementary Table 8 was constructed with the cohort frequencies of EAC biomarkers. Ten EAC cell lines (SKGT4, OACP4C, OACM5.1, ESO26, ESO51, OE33, MFD, OE19, Flo-1 and JHesoAD) and three Barrett's esophagus high-grade dysplasia cell lines (CP-B, CP-C and CP-D) with WGS data⁴¹ were used in proliferation assays to determine drug sensitivity to CDK4/CDK6 inhibitors, palbociclib (Biovision) and ribociclib (Selleckchem). Cell lines were grown in their normal growth media. Proliferation was measured with an Incucyte live-cell analysis system (Incucyte ZOOM Essen Biosciences). Each cell line was plated at a starting confluence of 10%, and the growth rate was measured over 4–7 d, depending on the basal proliferation rate (until 90% confluent in DMSO control). For each cell line–drug combination, concentrations of 16, 64, 250, 1,000 and 4,000 nM in 0.3% dimethylsulfoxide (DMSO) were used and compared to 0.3% DMSO only. Each condition was performed in at least triplicate (technical replicates) and for 12 of 12 randomly chosen cell lines, the drug combinations were successfully replicated with biological replicates (independent experiments). The time period of treatment to growth cessation in the control (0.3% DMSO) condition was used to calculate half-maximal growth inhibition (GI_{50}) and area under the curve (AUC). Accurate GI_{50} values could not be calculated in cases in which a cell line had $> 50\%$ proliferation inhibition even

with the highest drug concentration, and hence AUC was used to compare cell-line sensitivity. T47D had a highly similar GI_{50} for palbociclib to that previously calculated in other studies (112 nM versus 127 nM)³⁸. Primary organoid cultures were derived from EAC cases included in the OCCAMS–ICGC sequencing study. Detailed organoid culture and derivation methods have been described⁴⁴. Regarding the drug treatment, the seeding density for each organoid line was optimized to ensure cell growth in the logarithmic growth phase. Cells were seeded in complete medium for 24 h and then treated with compounds at five-point four-fold serial dilutions for 6 or 12 d. Cell viability was assessed with CellTiter–Glo (Promega) after drug incubation.

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

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

Code availability

Code associated with the analysis is available upon request.

Data availability

The WGS and RNA expression data can be found at the European Genome-phenome Archive under accession numbers [EGAD00001004417](https://www.ebi.ac.uk/ena/browser/view/EGAD00001004417) and [EGAD00001004423](https://www.ebi.ac.uk/ena/browser/view/EGAD00001004423), respectively.

References

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

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

vcp 86, BWA-MEM 0.7.10, Picard 1.115, Strelka 2.0.15, ASCAT 2.3, GATK 3.2-2, Manta 0.27.2

Data analysis

dNdScv 0.1.0, OncoDriverFM 1.0.3, OncodriveClust 1.0.0, MutsigCV 1.41, GISTIC 2.0.23, R 3.4.3, e-Driver (<https://github.com/eduardporta/e-Driver>), vcf2maf 1.6.14, ActiveDriverWGSv1.0, cBioportal v1.18.0 web interface accessed Sept 2018, Graph-pad prism 5, R 3.5.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size was chosen. Maximum number of available samples were used. Previous power calculations suggest our numbers provide significant power for detection of drivers even at a low frequency (even without using a wide range of detection methods) (Lawrence et al 2014 Nature)
Data exclusions	No data was excluded in Driver gene detection (Fig1). Exclusions were then pre-established based on the availability of specific data-types on samples, as is stated in the manuscript. Only cases with RNA-seq (116/551) were used in most analyses in figure 2 to detection CN drivers and Fig3C. For prognostic analyses (Fig 5) only ICGC cases were used where high quality clinical data was available (379/551). A
Replication	To ensure reproducibility of drug assays (Fig 6B) several measures were undertaken. We used known sensitive or resistance cell lines for our drugs (T47D and MDA-MB-468) to ensure our drugs worked as expected. We compared our GI50 for T47D to that previously reported and found it was highly similar (112 nM vs 127 nM). By using two different CDK4/6 inhibitors we controlled for off target effects and several cell-specific drug repeats were also undertaken successfully to ensure the drug assay was reproducible (data not shown). We also successfully validated our poor prognostic indicators in independent cohorts.
Randomization	Randomisation was not appropriate or required for any of the analyses
Blinding	Blinding was not appropriate or required for any of the analyses

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Included in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials All biological materials are available from the authors (Cell lines depending on MTAs - otherwise commercially available) or common commercial sources (Cell lines from sources below, CKD4/6 inhibitors Ribociclib, Palbociclib and Abemaciclib).

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	CP-D = ATCC, Flo-1 = Gift from David Beer, OE33 = ECACC, JHesoAD = gift from Anirban Maitra, ESO26 = ECACC, SK-GT-4 = ECACC,, OACp4C = ECACC,, OACM5.1 = ECACC,, T47D = A gift from Carlos Caldas ,MDA-MB-468= a gift from Carlos Caldas, ESO51 - ECACC, CP-B = ATCC, CP-C = ATCC, OE19 = ECACC, MFD = A kind gift from Tim Underwood
Authentication	STR testing was completed successfully for all cell lines
Mycoplasma contamination	All cell lines were confirmed mycoplasma negative
Commonly misidentified lines (See ICLAC register)	No cell lines used are commonly misidentified in the ICLAC register

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	EAC cases were broadly representative the wider population of EAC tumours containing a string gender bias towards male cases, late stage (mostly T3), See supplementary table 13.
Recruitment	Patients are recruited after diagnosis of oesophago-gastric cancer and samples taken at times of clinically indicated interventions either at the time of surgery or before using biopsies.