Patient-specific detection of cancer genes reveals recurrently perturbed processes in esophageal adenocarcinoma

Thanos P. Mourikis[1,2,*], Lorena Benedetti[1,2,*], Elizabeth Foxall[1,2,*], Julianne Perner[3], Matteo Cereda[4], Jesper Lagergren[2], Michael Howell[5], Christopher Yau[6], Rebecca Fitzgerald[3], Paola Scaffidi[7,8] and Francesca D. Ciccarelli[1,2,], on behalf of the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium

[1]Cancer Systems Biology Laboratory, The Francis Crick Institute, London NW1 1AT, UK
[2]School of Cancer and Pharmaceutical Sciences, King's College London, London SE11UL, UK
[3]MRC Cancer Unit, Hutchison/MRC Research Centre, University of Cambridge, Cambridge, UK, CB2 OXZ
[4]Italian Institute for Genomic Medicine (IIGM), Turin, Italy, 10126
[5]High Throughput Screening Laboratory, The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK
[6]University of Birmingham, Birmingham, B15 2TT, UK
[7]Cancer Epigenetics Laboratory, The Francis Crick Institute, London NW1 1AT, UK
[8]UCL Cancer Institute, University College London, London WC1E 6DD, UK

* = equal contribution
¶ To whom correspondence should be addressed: Tel: +44 (0)20 7848 6616; Fax: +44 (0)20 7848 6220; Email: francesca.ciccarelli@crick.ac.uk

ABSTRACT

The identification of somatic alterations with a cancer promoting role is challenging in highly unstable and heterogeneous cancers, such as esophageal adenocarcinoma (EAC). Here we used a machine learning approach to identify cancer genes in individual patients considering all types of damaging alterations simultaneously (mutations, copy number alterations and structural rearrangements). Analysing 261 EACs from the OCCAMS Consortium, we discovered a large number of novel cancer genes that, together with well-known drivers, help promote cancer. Validation using 107 additional EACs confirmed the robustness of the approach. Unlike known drivers whose alterations recur across patients, the large majority of the newly discovered cancer genes are rare or patient-specific. Despite this, they converge towards perturbing similar biological processes, including cell cycle progression, proteasome activity, intracellular signalling, Toll-like receptor cascade and DNA replication. Recurrence of process perturbation, rather than individual genes, divides EACs into six clusters that differ in their molecular features and suggest patient stratifications for targeted treatments. Experimental validation of selected genes by mimicking the same alterations found in patients leads to cancer-related phenotypes, thus supporting their contribution to disease progression.

INTRODUCTION

Genome instability enables the onset of several hallmarks of cancer because acquired alterations can confer selective advantages to the mutated cells, thus driving their outgrowth and eventual dominance[1]. The identification of driver genes (the genes acquiring the driver alterations) is critical to fully understand the molecular determinants of cancer and to inform the development of precision oncology. Since driver genes are under positive selection during cancer progression, a reasonable assumption is that they mutate more frequently than expected. Over the past years, large-scale cancer genomic studies have provided the required power to detect driver events recurring across samples with good statistical confidence[2,3]. However, the full characterisation of driver events is particularly challenging when the genomic landscape of the cancer is highly variable and recurrent events are relatively few.

One such cancer is esophageal adenocarcinoma (EAC), which shows high mutational and chromosomal instability leading to widespread genetic heterogeneity. In recent years, EAC incidence has risen substantially in the western world[4,5], causing more than 400,000 deaths each year[6]. In over 400 EACs sequenced so far, mutations in *TP53*, *CDKN2A*, *SMARCA4*, *ARID1A, SMAD4*, *ERBB2* as well as amplifications of *VEGFA, ERBB2, EGFR, GATA4/6, CCNE1* are the only recurrent driver events[7-12]. This leaves up to 10-20% of patients without known genetic determinants and often the number of identified drivers per sample is too low to fully explain the disease. As a consequence of this heterogeneity, the molecular mechanisms that drive EAC have been difficult to characterise in full. This has a profound impact on the way in which EAC is currently diagnosed and treated. For example, EAC treatment is still largely based on conventional chemotherapy and phase III clinical trials with various targeted agents have failed to show benefits or reached inconclusive results[13-15].

Here we hypothesise that, alongside the critical role of recurrent and well-known drivers, complementary somatic alterations of several other genes help cancer progression in individual patients. Therefore, for a comprehensive characterisation of the molecular determinants relevant to cancer, both recurrent and patient-specific cancer genes should be considered. However, while recurrent drivers can be identified based on the frequency of alterations, genes that act to promote cancer in individual patients are difficult to recognise. To this aim, we

developed a machine-learning method, sysSVM, based on the rationale that somatic alterations sustaining cancer affect genes with specific properties[16]. Using these properties, rather than recurrence, allows patient-specific predictions.

We applied sysSVM to 261 EACs from the UK OCCAMS Consortium. We first trained the classifier using 34 properties specific to known cancer genes and then prioritised 908 genes that, together with the known drivers, help or promote cancer development in the whole EAC cohort. The large majority of these newly predicted 'helper' genes are rare or patient-specific, however they converge towards the perturbation of related biological processes. Several of these processes are well-known players in cancer development including cell cycle control, DNA replication, proteasome activity, Toll-like receptor signalling, and downstream effectors of intracellular signalling. We used the recurrence of process perturbation, rather than genes, to stratify the 261 EACs into six clusters that show distinct molecular features and suggest differential response to targeted treatment.

RESULTS

The landscape of patient-specific cancer genes in EAC

sysSVM applies machine learning to predict altered genes contributing to cancer based on the similarity of their molecular and systems-level properties to those of known cancer genes (Supplementary text). Molecular properties derive from cancer sequencing data and include somatic alterations with a predicted damaging effect on the protein function (gene gains and losses, translocations, inversions, insertions, truncating and non-truncating damaging alterations and gain of function mutations) as well as the overall mutation burden and the gene copy number (Supplementary Table 1). Systems-level properties are genomic, epigenomic, evolutionary and gene expression features that distinguish cancer genes from other genes. They include gene length and protein domain organisation[16,17], gene duplicability[18,19], chromatin state[20], connections and position of the encoded proteins in the protein-protein interaction network[18], number of associated regulatory miRNAs[19], gene evolutionary origin[19] and breadth of gene expression in human tissues[16,17] (Supplementary Table 1).

sysSVM is composed of three steps (Supplementary text, Figure 1A). In the first step, all altered genes in each cancer sample are annotated in terms of 10 molecular and 24 systems-level properties. In the second step, molecular and systems-level properties of the known cancer genes altered within the cancer sample cohort are used to train a one-class support vector machine using four kernels (linear, sigmoid, radial, polynomial). In the third step, the four best classifiers from the training (one per kernel) are used for prediction. All remaining altered genes are first scored in each patient individually by combining the corresponding decision values in the four kernels. Genes are then ranked in each patient according to the resulting combined score. Methods for the identification of cancer drivers usually apply statistical models and a well-defined cut-off (usually a p-value) to distinguish drivers from passenger events. Since our hypothesis is that the strength of the contribution of a gene to cancer depends on how similar its properties are to those of known cancer genes, our score reflects a gradient between driver and passenger activity. The contribution of a gene to cancer decreases with its decreasing similarity to known cancer genes and with its score. Therefore, we do not define a hard cut-off.

Rather, the top scoring genes are the most similar to known cancer genes and are the most likely contributors to cancer progression in that patient.

We applied sysSVM to 261 EACs from OCCAMS which are part of the International Cancer Genome Consortium (ICGC) dataset (Figure 1B, Supplementary Table 2). In the first step, we extracted 17,078 genes with predicted damaging alterations in the whole cohort (median of 382 damaged genes per patient, Supplementary Table 2). In the second step, we optimised the parameters and trained the four classifiers using 476 known cancer genes[21] altered in the cohort (Supplementary Table 3, Supplementary Table 4). As expected, known cancer genes tend to occupy specific regions of the feature space according to their molecular and systems-level properties (Supplementary Figure 1). In the third step, we scored and ranked the remaining 16,602 genes (Figure 1B). Overall, top scoring genes localise to the same high-density regions of known cancer genes (Figure 1C), with lower scoring genes being further away from them (Supplementary Figure 2). This indicates that the properties of top scoring genes indeed resemble those of known cancer genes. To evaluate how sysSVM classifier trained on the ICGC cohort performs on independent cohorts, we used 86 EACs from The Cancer Genome Atlas (TCGA) and 21 EACs from a previous study[22] (Supplementary Table 2). We scored all altered genes, including known cancer genes, in each of the 107 EACs independently, using the four classifiers trained on the ICGC cohort. In both datasets, known cancer genes (*e.g. TP53*, *ERBB2, EGFR*, *SMAD4*) have significantly higher scores than the rest of the altered genes (Supplementary Figure 2), indicating that sysSVM is able to recognise them as major cancer contributors.

Since we do not use a statistical cut-off to distinguish driver from passenger genes, we arbitrarily considered the top 10 scoring genes in each patient as the main cancer contributors for that patient. We verified that – as expected – the main findings of our study hold true if we apply higher or lower cut offs (see below). This led to a total of 908 genes (Supplementary Table 5). Since our hypothesis is that these genes help the known drivers altered in the same sample to promote cancer, we define them as 'helper genes'. Consistent with the prevalence of gene amplification in EAC (Supplementary Table 1), the vast majority of newly predicted helpers undergo copy number gain (Figure 1D) and are rare or patient-specific (Figure 1E). A few helper genes, however, are altered in more than 10% of EACs

(Supplementary Table 5) and are usually associated with frequently occurring amplification events[11] (Supplementary Figure 3). Interestingly, around 10% of helpers (86 genes) have been predicted as candidate cancer genes in previous studies[17], and 39 have recently been added to Tiers 1 and 2 of the Cancer Gene Census[23] (Supplementary Table 5). As a further assessment of sysSVM predictions, we searched for possible false positives among the helpers using two lists. The first is composed of 14 genes that have previously been considered as recurrent false positives of driver prediction[2] and only one is a predicted cancer helper (*PCLO*). *PCLO* is considered as a false positive because of its long coding sequence and biased base composition[2]. However, *PCLO* is amplified and not mutated in EAC (Supplementary Table 5) and recently it has been shown to exert an oncogenic role in esophageal cancer by interfering with the *EGFR* signalling pathway[24]. The second list is a manually curated set of 488 genes[3]. We found 37 helpers in this list (4% of the total), which is less than the known cancer genes from the Cancer Gene Census (46/719, 6.4%). Therefore, sysSVM robustly predicts cancer genes in multiple patient cohorts, with a minimal false positive rate.

Patient-specific helpers converge to perturb related biological processes

Alterations in the newly predicted helper genes in the OCCAMS cohort are mostly rare or patient-specific (Figure 1D). In order to gain insight into their cancer promoting functions and gather a comprehensive characterisation of the molecular determinants of EAC, we analysed the biological processes perturbed by helpers compared to drivers. To this aim, we manually reviewed all 476 known cancer genes[21] with damaging alterations in the OCCAMS cohort and retained 202 of them based on the concordance between the type of acquired modification and the literature evidence of their cancer role (Methods, Supplementary Table 3). The median number of drivers per EAC is in accordance with recent estimates[25,26] and the majority of them undergo gene amplification (Supplementary Figure 3). We performed two independent gene set enrichment analyses, one with the 202 known drivers and one with 908 helpers, to dissect their relative functional contribution to EAC. This led to 297 and 365 enriched pathways out of the 1,877 tested, respectively (FDR <0.01, Supplementary Table 6, Supplementary Figure 3). Approximately 30% of pathways (84) enriched in known drivers are perturbed in

more than 50% of EACs (Supplementary Table 6). These 'universal' pathways are involved in well-known cancer-related processes, such as intracellular signalling, cell cycle control, apoptosis, and DNA repair (Supplementary Figure 4) and are associated with the most recurrently altered known drivers (*TP53, CDKN2A, MYC, ERBB2, SMAD4, CDK6, KRAS*, Supplementary Table 3). Interestingly, 59 universal pathways (70% of the total) are also enriched in helpers, often in patients with no known drivers in that pathway (Figure 2A, Supplementary Table 6). This indicates that helpers also contribute to the perturbation of key cancer pathways and that their alteration may be sufficient for cancer development in some EACs.

Next, we clustered EACs according to the proportion of perturbed pathways that they have in common (Methods, Figure 2B). When using pathways enriched in known drivers, the clustering is dominated by the most recurrently altered genes and EACs are broadly divided in two groups depending on *TP53* mutated (75,5% of the total) or wild-type status (Figure 2C, D, Supplementary Table 2). Similarly, smaller clusters are driven by the mutational status of other recurrent drivers (for example *EGFR* and *ERBB2* in cluster 1D, *MYC, EGFR* and *ERBB2* in cluster 2D, *MYC* and *KRAS* in cluster 3D, Figure 2C, D). Samples with no alterations in *TP53* (clusters 5D and 6D) show an overall lower mutational burden ($p = 0.001$, Wilcoxon rank sum test), fewer known drivers and consequently a lower number of enriched pathways ($p = 2x10^{-6}$, Wilcoxon rank sum test, Supplementary Figure S3).

When grouping EACs according to the pathways enriched in helpers, six clusters can be identified (clusters 1H-6H, Figure 2C) where samples are brought together not by the recurrent alterations of specific genes, but by several altered helpers mapping to the same or related pathways (Supplementary Table 2). We refer to these related pathways that define the clusters as 'biological processes'. For example, 29 of 35 EACs in cluster 1H have helpers involved in cell cycle regulation, particularly S-phase entry (Supplementary Figure 5). These helpers include members of the E2F family of transcription factors, associated co-activators or downstream targets (Supplementary Table 2). EACs in this cluster show significant alterations of known drivers such as *GNAS*, *SS18L1*, and *TRRAP* (FDR <0.03, Fisher's exact test). Interestingly, TRRAP is a E2F cofactor and stimulates E2F-mediated transcription when overexpressed[27]. Alterations in the E2F family and associated co-factors may help cancer progression by contributing to increased

genomic instability through aberrant cell-cycle control[28,29]. Consistently with this, EACs in cluster 1H accumulate significantly more gene amplifications and damaging alterations as compared to other clusters (FDR $<2.2^{-16}$ and FDR <0.05, Wilcoxon rank-sum test, Figure 3A, B). They are also enriched in the mutational signature S17A (FDR <0.18, Wilcoxon rank-sum test, Figure 3C), typical of highly mutagenic EACs[11]. Cluster 2H contains EACs with perturbations in proteasome activity including ubiquitination, proteasome degradation and antigen presentation (Supplementary Figure 5). It also shows significant alterations of *ERBB2* and *RARA* (FDR $<2.2^{-16}$, Fisher's exact test) that, together with proteasome inhibition, may suggest new treatment stratifications. Patients in cluster 3H have recurrent perturbations in intracellular signalling pathways (Supplementary Figure 5) and show a significantly higher exposure to mutational signature S3 (FDR = 0.03, Wilcoxon rank-sum test, Figure 3C), which in EAC has been associated with defects in homologous recombination[11]. Cluster 4H shows perturbations in the Toll-like receptor signalling cascade (Supplementary Figure 5) that has recently been reported to be dysregulated in EAC[30]. Cluster 5H harbours perturbations in DNA replication, with recurrent alterations in members of the MCM complex (Supplementary Table 2). Interestingly, MCMs are targets of E2F1[31,32], and their dysregulation induces replicative stress and increases genomic instability[28]. This is consistent with the observation that EACs in cluster 5H also have significantly higher exposure to mutational signatures S3 and S2 (FDR <0.06 and 0.9, respectively, Figure 3C). Therefore, clusters 1H and 5H converge towards the disruption of related processes through several distinct genes. Overall, clusters 1H to 5H account for 151 EACs (58% of the total cohort). The remaining 110 EACs in cluster 6H share fewer perturbed pathways, although the majority of them have alterations in Rho GTPase activity (Supplementary Figure 5) with frequent modifications of Rho GTPases and their effectors such as and *CDC42*, *ROCK1*, *PAK1*, *NDE1* and *ABI2* (Supplementary Table 2). EACs in the six clusters obtained using helpers are broadly dispersed in the clustering of known drivers (Figure 2C) indicating that helpers bring together patients with similar perturbed processes that cannot be appreciated when focussing only on recurrent drivers.

To test whether the clustering is affected by considering only the top 10 helper genes in each patient, we performed the same analysis considering as helpers the

top five or top 15 scoring helper genes (487 and 1,282 unique genes, respectively). We found that the vast majority (92% and 78%) of the pathways enriched in these two datasets are also enriched when considering the top 10 helpers (Supplementary Figure 6). This indicates that the recurrently perturbed processes are highly overlapping. We then clustered EACs according to the proportion of common perturbed pathways and verified that the six clusters obtained using pathways enriched in top 10 genes recapitulated well the clusters obtained using pathways enriched in top five or 15 genes (Supplementary Figure 6). Therefore, the clustering is robust regardless of the applied score cut off.

Helper alteration contributes to cancer-related phenotypes

To test the contribution of EAC helper genes towards the perturbation of cellular processes, we experimentally assessed the consequences of altering representative helpers. We used EAC FLO-1 cells because all helper genes selected for validation are expressed in these cells and have no mutations or copy number alterations[33], allowing a clear evaluation of the effect of their alteration. As a main hallmark of cancer[3,34], we assessed the effect of helper alterations on cell proliferation but also performed gene-specific assays to further investigate the effect of specific helper perturbations.

First, we started from representatives of clusters 1H and 5H that are functionally related and selected the most commonly altered helpers, E2F1 (25 out of 35 samples in cluster 1H) and MCM7 (22 out of 26 samples in cluster 5H, Supplementary Table 2). The transcription factor E2F1 promotes G1-S transition and regulates DNA replication, while MCM7 is part of the MCM complex, which is loaded onto chromatin throughout G1 and is activated at the start of S phase (Figure 4A). Both *E2F1* and *MCM7* are amplified in EACs (Supplementary Table 5) leading to significant gene overexpression (median two-fold increase, $p=0.005$ and $p=0.001$, respectively Wilcoxon rank-sum test; Figure 4B). To mimic their amplification in EAC, we stably overexpressed *E2F1* and *MCM7* in FLO-1 cells to levels comparable to those observed in EAC patients (Figure 4C). In both cases we observed significantly higher proliferation of overexpressing cells as compared to control cells ($p<10^{-4}$ and $p=0.0006$, respectively, one-tailed T-test; Figure 4D). Since E2F1 promotes cell cycle progression, we assessed the replication rate by measuring EdU incorporation

during the cell cycle. We observed increased EdU intensity throughout S phase in *E2F1* overexpressing cells as compared to control cells ($p<10^{-4}$, Mann Whiney U test; Figure 4E). This suggests that E2F1 may help cancer growth by promoting S phase entry. To assess the functional consequence of *MCM7* overexpression on the MCM complex, we measured its loading onto chromatin. We observed that *MCM7* overexpressing cells display a lower MCM fluorescence intensity overall as compared to control cells when staining the chromatin-bound fraction for either MCM7 or MCM3 ($p<10^{-4}$, Mann-Whitney U test; Figure 4F, 4G). This suggests that less MCM complex is loaded onto chromatin by the start of S phase. Therefore, *MCM7* overexpression leads to both increased proliferation and perturbation of MCM complex activity, possibly through independent mechanisms[35].

E2F1 and *MCM7* are among the most frequently altered helpers in our cohort (~10% of EACs). To evaluate the role of rare helpers, we tested *NCOR2* that is altered in four EACs (1,5% of the total, Supplementary Table 5). NCOR2 is part of the nuclear receptor corepressor complex that favours global chromatin deacetylation and transcriptional repression[36,37] (Figure 5A). Consistently with the tumour suppressor role that has been suggested for *NCOR2* in lymphoma and prostate cancer[38,39], the most frequent *NCOR2* alterations in EAC lead to loss of function. Therefore, we edited the gene in FLO-1 cells using a vector-free CRISPR system[40]. Three pooled crRNAs were co-transfected with Cas9 and the tracrRNA (Methods, Supplementary Table 7) and the editing was confirmed and quantified using Miseq (Figure 5B). Supporting a tumour suppressor role for *NCOR2*, we observed that the edited cells proliferate more than control cells ($p<10^{-4}$, one-tailed T-test test; Figure 5C).

Finally, we tested the effect of altering members of the Rho GTPase effector pathway, which is pervasively perturbed in all six clusters, including 6H (Supplementary Table 2). This pathway contains several, often patient-specific helpers that modulate actin cytoskeletal dynamics, cell proliferation and migration (Figure 5D, Supplementary Table S5). Two such effectors, *ABI2* and *PAK1*, undergo damaging alterations and amplification in seven and nine EACs, respectively (Supplementary Table 5). We therefore edited *ABI2* and overexpressed *PAK1* as described above (Supplementary Table 7, Figure 5E) and in both cases observed

significantly increased proliferation as compared to control cells (ABI2: $p<10^{-4}$, PAK1: $p=10^{-3}$ one-tailed T-test; Figure 5F).

Taken together these data indicate that, independent of the frequency of the alteration, the modification of predicted helpers has an effect on promoting cell growth. Moreover, it is the perturbation of a biological process, rather than a specific gene, that acts to promote cancer.

DISCUSSION

Most state-of-the-art approaches to discovering cancer driver events rely on the detection of positively selected alterations of genes that are functionally beneficial for cancer development[3,25]. Even ratiometric methods based on gene properties[41] ultimately assess the effect of positive selection and distinguish the few selected drivers from the many passenger events. As a result, the discovery of cancer drivers is biased towards genes that are frequently altered across patients. This poses significant limitations for cancers such as EAC that have a highly variable but mostly flat (*i.e.* with few recurrent events) mutational landscape. In support of this, the overall selection acting on esophageal cancer genomes is among the lowest across cancer types[25], despite a median of 382 damaged genes per EAC (Supplementary Table 2). This indicates that the exclusive focus on genes under strong selection is likely to return only a partial representation of the genes important for EAC.

To overcome these limitations, our machine learning approach ranks somatically altered genes that are relevant for cancer based on their properties rather than mutation recurrence. Another advantage of not measuring selection is that sysSVM considers all types of gene alterations (SNVs, indels, CNVs, and structural variations) simultaneously. Therefore, sysSVM provides a comprehensive overview of the genetic modifications that play a cancer promoting role in each patient. When applied to 261 EACs from ICGC, sysSVM prioritises 908 altered genes that, together with known drivers, help cancer progression. This large number of helper genes is in agreement with the recent observation of a positive correlation between mutational burden and number of driver genes, which is only partially explained by a sample size effect[3]. We speculate that this positive correlation may indicate that the number of functionally relevant genes increases with the number of altered genes.

The heterogeneous landscape of EAC cancer genes is substantially reduced by considering the perturbed biological processes rather than individual genes (Figure 2C). Several of these processes are well-known contributors to cancer development, including intracellular signalling, cell cycle control, and DNA repair (Supplementary Table 6). Interestingly, while the known drivers tend to encode upstream players in these pathways, helpers are often downstream effectors. For example, we found several Rho GTPase effectors (Figure 5D, Supplementary Table

5) or genes downstream of previously reported EAC drivers in the Toll-like receptor cascade (Supplementary Table 5). This supports a more local role of helpers in contributing to cancer at the single patient level, possibly by sustaining or complementing the function of drivers. In this respect, helpers are conceptually similar to mini-drivers[42].

Analysing the pathways disrupted by helpers allows the division of the 261 EACs into six functional clusters that often have similar molecular features, such as the prevalence of the same mutational signatures or the association with known drivers (Figure 2D and Figure 3C). For example, two of these clusters (1H and 5H) are closely related in function, showing perturbations of processes involved in the control of S-phase entry and DNA replication. Consistent with this, they bring together the most genomically unstable samples. By experimentally mimicking the amplification of *E2F1* (representative of cluster 1H) and *MCM7* (representative of cluster 5H), we induced increased proliferation in EAC cells (Figure 4D). We also provide evidence that E2F1 increases proliferation by promoting S phase entry (Figure 4E). Interestingly, *MCM7* overexpression resulted in a reduction of MCM complex loading onto chromatin (Figure 4F and 4G), maybe due to a stoichiometric imbalance of complex subunits. This may indicate that MCM7 promotes cell growth through a separate mechanism besides its function in the MCM complex. For example, MCM7 interacts with the tumour suppressor protein Rb, a well-characterised inhibitor of E2F1[43]. It is possible that MCM7 overexpression may sequester Rb away from E2F1, thereby promoting E2F1-mediated cell cycle progression. While the majority of patients with MCM7 amplification are grouped in cluster 5H, several patients in cluster 1H also have amplification of MCM7 (Supplementary table 2), supporting an overlapping role for MCM7 and E2F1 towards the perturbation of similar pathways involved in cell cycle control.

We also confirmed the cancer promoting role of very rare helpers, such as *ABI2*, *NCOR2* and *PAK1* that are altered in 1.5% to 3.4% of EACs (Figure 5C, F). Therefore, irrespective of the frequency, helpers have a substantial impact on the progression of the cancer where their alteration occurs. This may indicate new, patient-specific gene dependencies and suggest possible stratifications that could inform the selection of targeted treatments. For example, cluster 2H is characterised by the pervasive alteration of several proteasome subunits (Figure 2C,

Supplementary Table 2). The proteasome can be targeted by small molecules such as bortezomib, which is currently used in myeloma and lymphoma and is in clinical trials for the treatment of solid tumours[44]. Approximately 7% of EACs in our cohort harbour alterations in proteasome-related helpers and it is tempting to speculate that these alterations could be used to stratify patients who may benefit from proteasome inhibition. Interestingly, proteasome inhibition has been shown to have a synergic effect in combination with ERBB2 inhibitors[45]. Since ERBB2 is also significantly altered in cluster 2H (Figure 2D), a combined therapy may be beneficial to patients in this cluster.

In summary, we provide one of the first attempts to extend the discovery of acquired perturbations contributing to cancer beyond those of recurrent drivers. Additional efforts are required to fully exploit the potential of these approaches to offer a more comprehensive view of the molecular mechanisms behind cancer and to guide novel clinical interventions.

METHODS

Annotation of molecular properties

Data on somatic single nucleotide variations (SNVs), small insertions and deletions (indels), copy number variations (CNVs), structural variations (SVs), and mutational signatures for 261 EACs were obtained from ICGC and analysed as previously described[11] (Supplementary Table 2). Briefly, SNVs and indels were called using Strelka v.1.0.13[46] and subsequently filtered as previously described[11]. For CNVs, the absolute copy number for each genomic region was obtained from ASCAT-NGS v.2.1[47] after correction for tumour content, using read counts at germline heterozygous positions as derived from GATK v.3.2-2[48]. To account for the high number of amplifications occurring in EAC, copy number gains were corrected by the ploidy of each sample as estimated by ASCAT-NGS. A gene was assigned with the copy number of a CNV region if at least 25% of its length was contained in that region. SVs (gene translocations, inversions, insertions) were identified from discordant read pairs using Manta[49] after excluding SVs that were also present in more than two normal samples of a panel of 15 esophagus and 50 blood samples[11]. In the case of TCGA validation cohort, SNVs, indels, and CNVs were derived from level 3 TCGA annotation data of 86 EACs (https://portal.gdc.cancer.gov/projects/TCGA-ESCA, Supplementary Table 2). In the case of 21 EACs from a previous study[22], SNVs, indels, and CNVs were called as described for the ICGC samples (Supplementary Table 2). The distribution of variant allele frequency of SNVs and indels across all samples was used to remove outliers likely indicating sequencing or calling artefacts. Variants with <10% frequency and indels longer than five base pairs were also removed. For CNVs, genomic regions were considered as amplified or deleted if their segment mean was higher than 0.3 or lower than -0.3, respectively, capping the segment mean to 1.5 to avoid hypersegmentation[50]. A gene was considered as amplified or deleted if at least 25% of its length was contained in a CNV region and the resulting copy number (CN) was estimated as:

$$CN = 2 \times 2^{segment\ mean}$$

No SV data were available for the validation cohorts.

Since only genes with predicted damaging alterations were used as input for sysSVM, further annotation for the variant damaging effect was performed. Stopgain,

stoploss, frameshift, nonframeshit, nonsynonymous, and splicing SNVs and indels were annotated using ANNOVAR (December 2015)[51]. All truncating alterations (stopgain, stoploss, and frameshift mutations) were considered as damaging. Nonframeshit and nonsynonymous mutations were considered as non-truncating damaging alterations if predicted by at least five of seven function-based methods (SIFT[52], PolyPhen-2 HDIV[53], PolyPhen-2 HVAR[53], MutationTaster[54], MutationAssessor[55], LRT[56] and FATHMM[57]) or by two out of three conservation-based methods (PhyloP[58], GERP++ RS[59], SiPhy[60]), using the scores from dbNSFP v.3.0[61]. Splicing modifications were considered as damaging if predicted by at least one of the two ensemble algorithms as implemented in dbNSFP v3.0. Putative gain of function alterations were predicted with OncodriveClust[62] with default parameters and applying a false discovery rate of 10%. The transcript lengths to estimate mutation clustering were derived from the refGene table of UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables). Gene gains, homozygous losses, translocations, inversions, insertions were always considered as putative damaging alterations.

Overall, 17,078 genes had at least one damaging alteration, for a total of 116,989 redundant damaged genes across 261 EACs (Supplementary Table 1). Of these, 476 were known cancer genes[21], corresponding to 4,091 redundant genes (Supplementary Table 2). For all 17,078 genes, the total number of exonic alterations (silent and nonsilent) and the somatic copy number were used as additional molecular features in sysSVM.

Annotation of systems-level properties

Protein sequences from RefSeq v.63[63] were aligned to the human reference genome assembly GRCh37 to define unique gene loci as previously described[17]. The length of the longest coding sequence was taken as the gene length. Genes aligning to more than one gene locus for at least 60% of the protein length were considered as duplicated genes[18]. Data on human ohnologs (gene duplicates retained after whole genome duplications) were collected from Makino et al., 2013[64]. The number of protein domains was derived from CDD[65]. The gene chromatin state based on Hi-C experiments[20] was retrieved from the covariate matrix of MutSigCV v1.2.01[2]. Data on protein-protein and miRNA-gene interactions, gene evolutionary origin and gene

expression were retrieved as described in An et al., 2016[17]. Briefly, human protein-protein interaction network was rebuilt from the integration of BioGRID v.3.4.125[66]; MIntAct v.190[67]; DIP (April 2015)[68]; HPRD v.9[69]; the miRNA-gene interactions were derived from miRTarBase v.4.5[70] and miRecords (April 2013)[60]; gene evolutionary origin was assessed as described in D'Antonio et al., 2011[19] using gene orthology from EggNOG v.4[71]; and gene expression in 30 normal tissues was retrieved from GTEx v.1.1.8[72]. Except gene length, duplication and ohnologs, all other systems-level properties had missing information for some of the 17,078 altered genes (Supplementary Table 1). To account for this, median imputation for continuous properties and mode imputation for categorical properties were implemented. Specifically, for each property median or mode values were calculated for known cancer genes and the rest of mutated genes. All missing values were replaced with their corresponding median or mode values.

Application of sysSVM to EACs

To train SysSVM, 476 known cancer genes with damaging alterations in 261 EACs (Supplementary Table 2) were used as positive training set. Each of the 34 features (Supplementary Table 1) was scaled to zero mean and unit variance to correct for the different numerical range across features. To optimise the parameters of the four kernels used in sysSVM (linear, radial, sigmoid and polynomial) a grid search was performed with a three-fold cross validation for 100 iterations (Supplementary Text). At each iteration, known cancer genes were randomly split into training set (around 317 genes) and test set (around 159 genes). The model with the maximum sensitivity and minimum variance was chosen as the best model in each kernel (Supplementary Table 4). The resulting four best models were trained using the whole training set and subsequently used to predict cancer helper genes among the 16,602 unique genes with damaging alterations in 261 EACs (corresponding to 112,898 redundant gene counts). To combine the results of the four models, a score was measured and used to rank all altered genes in each sample (Supplementary Text). After excluding the genes not expressed in normal esophagus according GTex annotation, the top 10 scoring genes were considered as the cancer helper genes for that sample (Supplementary Table 5).

Identification of perturbed processes and patient clustering

To identify the perturbed biological processes in the EAC cohort, both predicted cancer helper genes and known cancer driver genes were used. A manual revision of 476 known cancer genes altered in the ICGC cohort was performed and genes were considered as known drivers if (1) their somatic alteration had been previously associated with EAC, (2) they had a loss-of-function alteration and their tumour suppressor role had been reported in other cancer types[73], (3) they had a gain-of-function alteration and their oncogenic role had been reported in other cancer types[73]. The resulting 202 known cancer drivers (Supplementary Table 3) and 908 cancer helpers were used for the gene set enrichment analysis against Reactome v.58[74], composed of 1,877 pathways and 10,131 genes. After excluding pathways in levels 1 and 2 of Reactome hierarchy and those with less than 10 or more than 500 genes, 1,155 pathways were retained. These contained 9,061 genes, including 155 known drivers and 619 helpers. Gene set enrichment was assessed using one-tailed hypergeometric test and the resulting *P* values were corrected for multiple testing using the Benjamini & Hochberg method (Supplementary Table 6). Enriched pathways within the sets of known drivers or helpers were subsequently used to cluster samples taking into account the proportion of perturbed processes shared between samples. The Jaccard index (*A*) was calculated by deriving the proportion of shared perturbed processes between all possible sample pairs as:

$$Aij = (P_i \cap P_j)/(P_i \cup P_j)$$

where P_i *and* P_j are the perturbed processes in samples *i* and *j*, respectively.
Complete linkage hierarchical clustering using Euclidean distance was performed on the resulting matrix. Clusters were visualised using ComplexHeatmap R package[75].

Analysis of RNA sequencing data

Purified total RNA was extracted from 99 EACs from the ICGC cohort and sequenced as described previously[11]. RNA sequencing reads were then aligned to human reference genome hg19 and expression values were calculated using Gencode v19. The summariseOverlaps function in the R GenomicAlignments package was used to count any fragments overlapping with exons (parameters mode=Union, singleEnd, invertStrand and inter.feature were set according to the library protocol, fragments=TRUE, ignore.strand=FALSE). Gene length was derived as the number of base pairs in the exons after concatenating the exons per gene in

non-overlapping regions. FPKM (Fragments Per Kilobase Million) were calculated for each gene as:

$$FPKM = \frac{gene\ read\ count}{(library\ size/1000000)\times(gene\ length/1000)}$$

Gene overexpression

All experiments were carried out using the FLO-1 esophageal adenocarcinoma cell line obtained from the ECACC General Cell Collection. Cells were validated by short tandem repeat analysis and routinely checked for mycoplasma contamination. Cells were grown at 37°C and five per cent CO_2 in DMEM + 2mM Glutamine + 10% FBS (Biosera) + 1/10,000 units of penicillin–streptomycin. The vectors pCMVHA E2F1[76] (Item ID 24225, Addgene), pLX_TRC317 (TRCN0000481188, Sigma-Aldrich) and pcDNA3.1+/C-(K)-DYK (Clone ID: OHu19407D, Genscript) were used to induce *E2F1*, *MCM7*, and *PAK1* overexpression, respectively. Cells were transfected according to the manufacturer's protocol and selected with either G481/Geneticin (*E2F1*, *PAK1*) or Puromycin (*MCM7*). Empty vectors carrying G418 (pcDNA3.1+/C-(K)-DYK, Genscript) or Puromycin (Item ID 85966, Addgene) resistance were used as controls. The RNA from transfected cells was used to assess gene overexpression via quantitative RT-PCR using predesigned SYBR green primers (Sigma-Aldrich; Supplementary Table 7) and Brilliant III Ultra-Fast SYBR Green QRT-PCR Master Mix (Agilent Technologies). The average expression level across triplicates (*e*) was relativized to the average expression level of β-2-microglobulin (*c*):

$$r = e - c$$

where *r* is the relative gene expression. The fold change (*fc*) between the relative gene expression after overexpression and the relative gene expression in the control condition (r_c) was calculated as:

$$fc = 2^{(r_c - rKD)}$$

Each sample was assessed in triplicate and each experiment was repeated in biological duplicate.

Gene editing

To induce *ABI2* and *NCOR2* gene knock-out (KO), the vector-free CRISPR-mediated editing approach was used as previously described[40]. Briefly, cells were co-transfected using lipofectamine CRISPR max (Life technologies) with a 69-mer tracrRNA (Sigma-Aldrich), three gene-specific crRNAs (Sigma-Aldrich, Supplementary Table 7) and GeneArt Platinum Cas9 nuclease (Life technologies). To avoid off-target editing, all crRNAs used were verified to map only the gene of interest with a perfect match and additional hits in the genome with at least three mismatches. Control cells were transfected with the same protocol but using three non-targeting crRNAs. Gene editing was confirmed with Illumina Miseq sequencing. The regions surrounding the targeted sites were amplified from genomic DNA of edited cells with primers containing Illumina adapters (Supplementary Table 7) using Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs). DNA barcodes were added with a PCR reaction before pooling the samples for sequencing on Illumina MiSeq with the 250 base-pair paired-end protocol. Sequencing reads were merged into single reads and aligned to the human reference genome hg19 using BBMerge and BBMap functions of BBTools[77], obtaining an average of 78,864 aligned reads per experiment. SNVs and small indels in the regions corresponding to each crRNA (Supplementary Table 7) were called using the CrispRVariants package in R[78] and the percentage of edited alleles was estimated as the percentage of variant reads in each experiment.

Cell proliferation

Cell proliferation was measured every 24 hours for three days, starting three hours after seeding the cells (time zero) using either CellTiter 96 Non-Radioactive Cell Proliferation Assay (Promega) or CellTiter-Glo Luminescent Cell Viability Assay (Promega). Briefly, $4.5x10^3$ cells/well were seeded on 96-well plates in a final volume of 100µl per well. For the CellTiter 96 Non-Radioactive Cell Proliferation Assay, 15 µl of the dye solution was added into each well and cells were incubated at 37°C for two hours. The converted dye was released from the cells using 100µl of the solubilisation/Stop solution and absorbance was measured at 570nm after one hour using the Paradigm detection platform (Beckman Coulter). For the CellTiter-Glo Luminescent Cell Viability Assay, 100µl of the CellTiter-Glo reagent was added into each well and luminescence was measured after 30 minutes using the Paradigm detection platform (Beckman Coulter). Four replicates per condition were measured

at each time point and each measure was normalised to the average time zero measure for each condition. Each experiment was repeated at least twice. Conditions were compared using the one-tailed Student's t-test.

Flow cytometry

EdU incorporation and MCM loading were assessed using a modified version of the protocol described in Galanos et al., 2016[79]. Briefly, in each condition, $3x10^6$ cells were pulsed for 30 minutes with 10μM EdU (Invitrogen) before washing in 1% BSA/PBS. Chromatin fractionation was performed by incubating on ice for 10 minutes in CSK buffer (10mM HEPES, 100mM NaCl, 3mM $MgCl_2$, 1mM EGTA, 300mM sucrose, 1% BSA, 0.2% Triton-X100, 1mM DTT, cOmplete EDTA-free protease inhibitor cocktail tablets, Roche). Cells were then fixed in 4% formaldehyde/PBS for 10 minutes at room temperature before washing in 1% BSA/PBS. Cells were permeabilised and barcoded[80] by incubating in 70% ethanol containing 0-15μg/ml Alexa Fluor 488 (Thermo Fisher) for 15 minutes, then washed twice in 1% BSA/PBS. Barcoded cells were subsequently pooled before incubating in primary antibody (mouse monoclonal anti-MCM7: Santa Cruz Biotechnology sc-56324, or rabbit polyclonal anti-MCM3: Bethyl Laboratories A300-192A) diluted 1:100 in 1% BSA/PBS for 1 hour. After washing in 1% BSA/PBS, samples were incubated for 30 minutes in secondary antibody (Alexa Fluor 555-conjugated donkey anti-mouse: A-31570, or donkey anti-rabbit: A-31572) diluted 1:500 in 1% BSA/PBS, then washed again in 1% BSA/PBS. EdU labelling with Alexa Fluor 647 azide was performed using Click-iT EdU flow cytometry assay kit (Invitrogen, C10424) following the manufacturer's instructions before washing samples in 1% BSA/PBS. Samples were then incubated in 1% BSA/PBS containing RNase and 10mg/ml DAPI for 15 minutes before analysing with a BD LSR II Fortessa flow cytometer (BD Biosciences). Compensation was performed with single colour controls, using BD FACSDiva software (BD Biosciences). FlowJo 10.3 software was used to analyse MCM loading onto chromatin and EdU incorporation. The cells were gated to separate the barcoded populations using 488 intensity. Cells were then separated into cell cycle gates (G1, S1-4, G2) based on EdU-647 and DAPI intensity, and the geometric mean fluorescence intensity was obtained for each channel (MCM=555 or EdU=647).

FUNDING

This work was supported by Cancer Research UK [C43634/A25487] and by the Cancer Research UK King's Health Partners Centre at King's College London. Computational analyses were done using the High-Performance Computer at the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. CY is supported by the Medical Research Council [MR/L001411/1]. Funding for sample sequencing was through the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium as part of the International Cancer Genome Consortium and was funded by a programme grant from Cancer Research UK (RG66287). 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.

AUTHOR CONTRIBUTIONS

FDC conceived and directed the study. AM developed all the bioinformatics methods under the supervision of FDC, MC and CY. LB and EF performed the experiments with the help of MH and PS. AM, LB, EF, FDC analysed the data. JP helped with the analysis of RNA-Seq data. FDC, AM, LB and EF wrote the manuscript. RCF, PS, MH, JL edited the manuscript. All authors approved the manuscript.

ACKNOWLEDGMENTS

We thank Stephanie Hills and John Diffley (The Francis Crick Institute) for their help with the experiments on E2F1 and MCM7 and for the discussion of results. We thank Valdone Maciulyte (The Francis Crick Institute) for help with the MiSeq library preparation. We are grateful for the support of the High Throughput Screening Facility, Flow Cytometry Facility, Advanced Sequencing Facility and Cell Services of the Francis Crick Institute.

Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium:

Rebecca C. Fitzgerald[1], Ayesha Noorani[1], Paul A.W. Edwards[1,2], Rachael Fels Elliott[1], Nicola Grehan[1], Barbara Nutzinger[1], Caitriona Hughes[1], Elwira Fidziukiewicz[1], Jan Bornschein[1], Shona MacRae[1], Jason Crawte[1], Alex Northrop[1], Gianmarco Contino[1], Xiaodun Li[1], Rachel de la Rue[1], Annalise Katz-Summercorn[1], Sujath Abbas[1], Daniel Loureda[1], Maria O'Donovan[1,3], Ahmad Miremadi[1,3], Shalini Malhotra[1,3], Monika Tripathi[1,3], Simon Tavaré[2], Andy G. Lynch[2], Matthew Eldridge[2], Maria Secrier[2], Lawrence Bower[2], Ginny Devonshire[2], Juliane Perner[2], Sriganesh Jammula[2], Jim Davies[5], Charles Crichton[5], Nick Carroll[6], Peter Safranek[6], Andrew Hindmarsh[6], Vijayendran Sujendran[6], Stephen J. Hayes[7,14], Yeng Ang[7,8,29], Andrew Sharrocks[29], Shaun R. Preston[9], Sarah Oakes[9], Izhar Bagwan[9], Vicki Save[10], Richard J.E. Skipworth[10], Ted R. Hupp[23], J. Robert O'Neill[10,23], Olga Tucker[11,33], Andrew Beggs[11,28], Philippe Taniere[11], Sonia Puig[11],Timothy J. Underwood[12,13], Robert C. Walker[12,13], Ben L. Grace[12], Hugh Barr[15], Neil Shepherd[15], Oliver Old[15], Jesper Lagergren[16,25], James Gossage[16,24], Andrew Davies[16,24,] , Fuju Chang[16,24], Janine Zylstra[16,24], Ula Mahadeva[16], Vicky Goh[24], Francesca D. Ciccarelli[24], Grant Sanders[17], Richard Berrisford[17], Catherine Harden[17], Mike Lewis[18], Ed Cheong[18], Bhaskar Kumar[18],Simon L Parsons[19], Irshad Soomro[19], Philip Kaye[19], John Saunders[19], Laurence Lovat[20], Rehan Haidry[20], Laszlo Igali[21], Michael Scott[22], Sharmila Sothi[26], Sari Suortamo[26], Suzy Lishman[27], George B. Hanna[31], Christopher J. Peters[31],Krishna Moorthy[31], Anna Grabowska[32], Richard Turkington[34], Damian McManus[34], David Khoo[35], Will Fickling[35].

[1]Medical Research Council Cancer Unit, Hutchison/Medical Research Council Research Centre, University of Cambridge, Cambridge, UK

[2]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

[3]Department of Histopathology, Addenbrooke's Hospital, Cambridge, UK

[4]Oxford ComLab, University of Oxford, UK, OX1 2JD

[5]Department of Computer Science, University of Oxford, UK, OX1 3QD

[6]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK, CB2 0QQ

[7]Salford Royal NHS Foundation Trust, Salford, UK, M6 8HD

[8]Wigan and Leigh NHS Foundation Trust, Wigan, Manchester, UK, WN1 2NN

[9]Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK, GU2 7XX
[10]Edinburgh Royal Infirmary, Edinburgh, UK, EH16 4SA
[11]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK, B15 2GW
[12]University Hospital Southampton NHS Foundation Trust, Southampton, UK, SO16 6YD
[13]Cancer Sciences Division, University of Southampton, Southampton, UK, SO17 1BJ
[14]Faculty of Medical and Human Sciences, University of Manchester, UK, M13 9PL
[15]Gloucester Royal Hospital, Gloucester, UK, GL1 3NN
[16] Guy's and St Thomas's NHS Foundation Trust, London, UK, SE1 7EH
[17]Plymouth Hospitals NHS Trust, Plymouth, UK, PL6 8DH
[18]Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, UK, NR4 7UY
[19]Nottingham University Hospitals NHS Trust, Nottingham, UK, NG7 2UH
[20]University College London, London, UK, WC1E 6BT
[21]Norfolk and Waveney Cellular Pathology Network, Norwich, UK, NR4 7UY
[22]Wythenshawe Hospital, Manchester, UK, M23 9LT
[23]Edinburgh University, Edinburgh, UK, EH8 9YL
[24]King's College London, London, UK, WC2R 2LS
[25]Karolinska Institute, Stockholm, Sweden, SE-171 77
[26]University Hospitals Coventry and Warwickshire NHS, Trust, Coventry, UK, CV2 2DX
[27]Peterborough Hospitals NHS Trust, Peterborough City Hospital, Peterborough, UK, PE3 9GZ
[28]Institute of Cancer and Genomic sciences, University of Birmingham, B15 2TT
[29]GI science centre, University of Manchester, UK, M13 9PL.
[30]Queen's Medical Centre, University of Nottingham, Nottingham, UK, NG7 2UH
[31]Department of Surgery and Cancer, Imperial College, London, UK, W2 1NY
[32]Queen's Medical Centre, University of Nottingham, Nottingham, UK
[33]Heart of England NHS Foundation Trust, Birmingham, UK, B9 5SS.
[34]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Northern Ireland, UK, BT7 1NN.
[35]Queen's Hospital, Romford, RM7 0AG

REFERENCES

1. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-74 (2011).
2. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
3. Bailey, M.H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e18 (2018).
4. Edgren, G., Adami, H.-O., Weiderpass, E. & Nyrén, O. A global assessment of the oesophageal adenocarcinoma epidemic. *Gut* **62**, 1406-1414 (2013).
5. Lagergren J & P., L. Recent developments in esophageal adenocarcinoma. *CA Cancer J Clin* **63**, 232-48 (2013).
6. Torre, L.A. *et al.* Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians* **65**, 87-108 (2015).
7. Agrawal, N. *et al.* Comparative Genomic Analysis of Esophageal Adenocarcinoma and Squamous Cell Carcinoma. *Cancer Discovery* **2**, 899-905 (2012).
8. 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).
9. Weaver, J.M.J. *et al.* Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat Genet* **46**, 837-843 (2014).
10. Ross-Innes, C.S. *et al.* Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat Genet* **47**, 1038-46 (2015).
11. Secrier, M. *et al.* Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat Genet* **48**, 1131-1141 (2016).
12. The Cancer Genome Atlas Research, N. Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169-175 (2017).
13. Woo, J., Cohen, S.A. & Grim, J.E. Targeted therapy in gastroesophageal cancers: past, present and future. *Gastroenterology Report* **3**, 316-329 (2015).
14. Young, K. & Chau, I. Targeted Therapies for Advanced Oesophagogastric Cancer: Recent Progress and Future Directions. *Drugs* **76**, 13-26 (2016).
15. Kopp, H.G. & Hofheinz, R.D. Targeted Treatment of Esophagogastric Cancer. *Oncology Research and Treatment* **39**, 788-794 (2016).
16. D'Antonio, M. & Ciccarelli, F.D. Integrated analysis of recurrent properties of cancer genes to identify novel drivers *Genome Biol* **14**, R52 (2013).
17. An, O., Dall'Olio, G.M., Mourikis, T.P. & Ciccarelli, F.D. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic Acids Research* **44**, D992-D999. (2016).
18. Rambaldi, D., Giorgi, F.M., Capuani, F., Ciliberto, A. & Ciccarelli, F.D. Low duplicability and network fragility of cancer genes. *Trends in Genetics* **24**, 427-430 (2008).
19. D'Antonio, M. & Ciccarelli, F.D. Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput. Biol.* **7**, e1002029 (2011).
20. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
21. Futreal, P.A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83 (2004).
22. Nones, K. *et al.* Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nat Commun* **5**, 5224 (2014).

23. Wellcome Sanger Institute. Catalogue of somatic mutations in cancer. At <https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=NCOR2#overview> (2018)..
24. Zhang, W. *et al.* Piccolo mediates EGFR signaling and acts as a prognostic biomarker in esophageal squamous cell carcinoma. *Oncogene* **36**, 3890-3902 (2017).
25. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e21 (2017).
26. Sabarinathan, R. *et al.* The whole-genome panorama of cancer drivers. *bioRxiv* (2017).
27. Lang, S.E., McMahon, S.B., Cole, M.D. & Hearing, P. E2F transcriptional activation requires TRRAP and GCN5 cofactors. *J Biol Chem* **276**, 32627-34 (2001).
28. Hills, S.A. & Diffley, J.F.X. DNA Replication and Oncogene-Induced Replicative Stress (vol 24, pg R435, 2014). *Current Biology* **24**, 1563-1563 (2014).
29. Nath, S. *et al.* Deregulation of Rb-E2F1 axis causes chromosomal instability by engaging the transactivation function of Cdc20-anaphase-promoting complex/cyclosome. *Mol Cell Biol* **35**, 356-69 (2015).
30. Fels Elliott, D.R. *et al.* Impact of mutations in Toll-like receptor pathway genes on esophageal carcinogenesis. *PLoS Genet* **13**, e1006808 (2017).
31. Yoshida, K. & Inoue, I. Regulation of Geminin and Cdt1 expression by E2F transcription factors. *Oncogene* **23**, 3802-3812 (2004).
32. Ohtani, K. *et al.* Cell growth-regulated expression of mammalian MCM5 and MCM6 genes mediated by the transcription factor E2F. *Oncogene* **18**, 2299-2309 (1999).
33. Wellcome Sanger Institute. Cell lines project v84. At <http://cancer.sanger.ac.uk/cell_lines> (2018).
34. Ng, P.K. *et al.* Systematic Functional Annotation of Somatic Mutations in Cancer. *Cancer Cell* **33**, 450-462 e10 (2018).
35. Neves, H. & Kwok, H.F. In sickness and in health: The many roles of the minichromosome maintenance proteins. *Biochim Biophys Acta* **1868**, 295-308 (2017).
36. Altucci, L. & Gronemeyer, H. The promise of retinoids to fight against cancer. *Nat Rev Cancer* **1**, 181-93 (2001).
37. Privalsky, M.L. The role of corepressors in transcriptional regulation by nuclear hormone receptors. *Annu Rev Physiol* **66**, 315-60 (2004).
38. Godoy, A.S. *et al.* Altered corepressor SMRT expression and recruitment to target genes as a mechanism that change the response to androgens in prostate cancer progression. *Biochem Biophys Res Commun* **423**, 564-70 (2012).
39. Song, L. *et al.* Alteration of SMRT tumor suppressor function in transformed non-Hodgkin lymphomas. *Cancer Res* **65**, 4554-61 (2005).
40. Benedetti, L., Cereda, M., Monteverde, L., Desai, N. & Ciccarelli, F.D. Synthetic lethal interaction between the tumour suppressor STAG2 and its paralog STAG1. *Oncotarget* **8**, 37619-37632 (2017).
41. Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A* **113**, 14330-14335 (2016).
42. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat Rev Cancer* **15**, 680-5 (2015).

43. Sterner, J.M., Dew-Knight, S., Musahl, C., Kornbluth, S. & Horowitz, J.M. Negative regulation of DNA replication by the retinoblastoma protein is mediated by its association with MCM7. *Mol Cell Biol* **18**, 2748-57 (1998).
44. Manasanch, E.E. & Orlowski, R.Z. Proteasome inhibitors in cancer therapy. *Nat Rev Clin Oncol* **14**, 417-433 (2017).
45. Issaenko, O.A., Bitterman, P.B., Polunovsky, V.A. & Dahlberg, P.S. Cap-dependent mRNA translation and the ubiquitin-proteasome system cooperate to promote ERBB2-dependent esophageal cancer phenotype. *Cancer Gene Ther* **19**, 609-18 (2012).
46. Saunders, C.T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-7 (2012).
47. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-5 (2010).
48. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
49. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-2 (2016).
50. Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-7 (2012).
51. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
52. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).
53. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
54. Schwarz, J.M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575-6 (2010).
55. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118 (2011).
56. Chun, S. & Fay, J.C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553-61 (2009).
57. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N. & Gaunt, T.R. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**, 1504-10 (2013).
58. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-21 (2010).
59. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).
60. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62 (2009).
61. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235-41 (2016).

62. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238-44 (2013).
63. Pruitt, K.D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42**, D756-63 (2014).
64. Makino, T., McLysaght, A. & Kawata, M. Genome-wide deserts for copy number variation in vertebrates. *Nat Commun* **4**, 2283 (2013).
65. Marchler-Bauer, A. *et al.* CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* **41**, D348-52 (2013).
66. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369-D379 (2017).
67. Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-63 (2014).
68. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-51 (2004).
69. Keshava Prasad, T.S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-72 (2009).
70. Hsu, S.D. *et al.* miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* **42**, D78-85 (2014).
71. Powell, S. *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* **42**, D231-9 (2014).
72. Mele, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-5 (2015).
73. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).
74. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-7 (2016).
75. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-9 (2016).
76. Lukas, J., Petersen, B.O., Holm, K., Bartek, J. & Helin, K. Deregulated expression of E2F family members induces S-phase entry and overcomes p16INK4A-mediated growth suppression. *Mol Cell Biol* **16**, 1047-57 (1996).
77. Joint Genome Institute. BBTools. At <https://jgi.doe.gov/data-and-tools/bbtools/> (2018).
78. Lindsay, H. *et al.* CrispRVariants charts the mutation spectrum of genome engineering experiments. *Nat Biotechnol* **34**, 701-2 (2016).
79. Galanos, P. *et al.* Chronic p53-independent p21 expression causes genomic instability by deregulating replication licensing. *Nat Cell Biol* **18**, 777-89 (2016).
80. Krutzik, P.O. & Nolan, G.P. Fluorescent cell barcoding in flow cytometry allows high-throughput drug screening and signaling profiling. *Nat Methods* **3**, 361-8 (2006).

FIGURE LEGENDS

Figure 1. Cancer helper genes in 261 EACs

a. Identification of cancer genes using sysSVM.

b. Application of sysSVM to EAC. Genes with somatic damaging alterations (n=116,989) were extracted from 261 EACs and divided into training (known cancer genes, blue) and prediction (rest of altered genes, red) sets. sysSVM was trained on the properties of known drivers and the best models were used for prediction. All altered genes were scored in each patient individually and the top 10 hits were considered as the cancer helper genes in that patient, for a total of 2,608 helper genes, corresponding to 908 unique hits.

c. t-distributed Stochastic Neighbour Embedding (t-SNE) plot of 116,989 altered genes in 261 EACs. Starting from the 34 properties used in sysSVM, a 2-D map of the high-dimensional data was built using Rtsne package (https://github.com/jkrijthe/Rtsne) in R. Curves are coloured according to the density of 4,091 known cancer genes (blue dots) used as a training set and the rest of altered genes are coloured according to their sysSVM score.

d. Distribution of damaging alterations in 908 cancer helpers. Overall, these genes acquire 2,804 damaging alterations.

e. Recurrence of cancer helpers across 261 EACs. Only samples acquiring alterations with a damaging effect are considered.

Figure 2. Perturbed processes in 261 EACs

a. Scatterplot of 59 universal pathways enriched in known drivers and helpers. For each pathway, the number of EACs with altered drivers and the number of EACs with altered drivers and helpers is shown. The size of dots is proportional to the additional EACs with perturbations in these pathways because of altered helpers only.

b. Schematic of the procedure to cluster EACs according to pathways enriched in known drivers or helpers. Enriched pathways are mapped to individual EACs and the Jaccard index is calculated as the proportion of shared pathways over the total pathways in each pair of samples (*i, j*). Hierarchical clustering is then performed.

c. Clustering of 261 EACs according to pathways enriched in known drivers and helpers. Six clusters were identified using known drivers (1D-6D) and six using

helpers (1H-6H). Cluster-matching coloured lines show where EACs clustered by pathways enriched in helpers map in the driver clusters.

d. Mutational status of five known drivers across 261 EACs.

Figure 3. Molecular features of EAC clusters driven by pathways enriched in helpers

For each cluster (1H-6H) indicated are (**a**) the number of genes with damaging mutations (truncating, non-truncating damaging and gain-of-function alterations), (**b**) the number of genes undergoing amplification and (**c**) the exposure to three mutational signatures (S2, S3, S17A). Heatmaps below each cluster represents the median value for that cluster. Only molecular features for which clusters significantly differ are shown. Enrichment in number of altered genes was assessed using Fisher's exact test, while distribution of mutational signatures where compared using Wilcoxon rank-sum test. FDR = false discovery rate after correction for multiple testing.

Figure 4. Cancer helper role of *E2F1* and *MCM7*

a. Schematic of the functional role of E2F1 and MCM7 as representatives of cluster 1H and 5H respectively.

b. *E2F1* and *MCM7* expression in EACs where they are amplified (11 and 13) as compared to EACs where they are copy number neutral (88 and 86). Significance was assessed using the Wilcoxon rank-sum test.

c. E2F1 and MCM7 mRNA expression in FLO-1 cells assessed by qRT-PCR. Expression was relativized to β-2-microglobulin and normalized to control cells. Two biological replicates were performed, with qRT-PCR reactions performed in triplicate.

d. Proliferation curve of FLO-1 cells overexpressing *E2F1* or *MCM7* as compared to the corresponding control cells. Three biological replicates were performed, each with four technical replicates. Proliferation was assessed every 24 hours and each time point was normalized to time zero. Mean values at 72 hours were compared by one-tailed Student's t-test.

e. Assessment of EdU (5-ethynyl-2'-deoxyuridine) incorporation by flow cytometry in *E2F1* overexpressing cells as compared to control cells. Cells were separated into G1, S and G2 phases, and S phase cells were subdivided into 4 gates from early to late S phase (S1-S4). The geometric mean fluorescence intensity of EdU was

measured for the cells in each gate and differences between EdU intensity were assessed using the Mann-Whitney U test. Three biological replicates were performed and a representative experiment is shown.

Quantification of MCM complex loading onto chromatin in *MCM7* overexpressing or control cells via staining of MCM7 (**f**) or MCM3 (**g**). Cells were pulsed with EdU, and chromatin fractionation was performed before staining for MCM7 or MCM3 to detect the MCM complex bound to chromatin. Cells were separated into cell cycle phases using EdU and DAPI intensity (see Methods). MCM7 or MCM3 fluorescence intensity during S phase illustrates the unloading of the MCM complex from chromatin. The geometric mean fluorescence intensity of MCM staining was measured for the cells in each cell cycle gate and differences in MCM intensity were assessed using Mann-Whitney U test. Representative data from one of three biological replicates are shown.

Figure 5. Cancer helper role of *NCOR2*, *ABI2*, and *PAK1*

a. Function of NCOR2 as part of the nuclear receptor co-repressor complex, whose activity results in chromatin deacetylation and transcriptional repression.

b. Editing of the *NCOR2* gene using three pooled crRNAs where cells are transiently co-transfected with Cas9 protein, crRNAs and tracrRNA[40]. The editing efficiency was measured using Miseq and the range of edited alleles and cells was derived considering the two opposite scenarios where all three crRNAs edit the same alleles/cells or different alleles/cells, respectively.

c. Proliferation curve of *NCOR2* or NTC edited FLO-1 cells. Proliferation was assessed every 24 hours and each time point was normalized to time zero. Mean values at 72 hours were compared by one-tailed Student's t-test. Three biological replicates were performed, each with four technical replicates.

d. Manual curation of the helpers contributing to the Rho-GTPase effectors pathway. Heatmap indicates the number of samples with alterations in each gene. ABI2 (blue) and PAK1 (red) were selected for experimental validation.

e. Induced alterations in *ABI2* and *PAK1* genes. Editing of *ABI2* was performed and assessed as described for *NCOR2*. PAK1 mRNA expression in FLO-1 cells was assessed by qRT-PCR, relativized to β-2-microglobulin and normalised to control cells. Experiments were done in triplicate in two biological replicates.

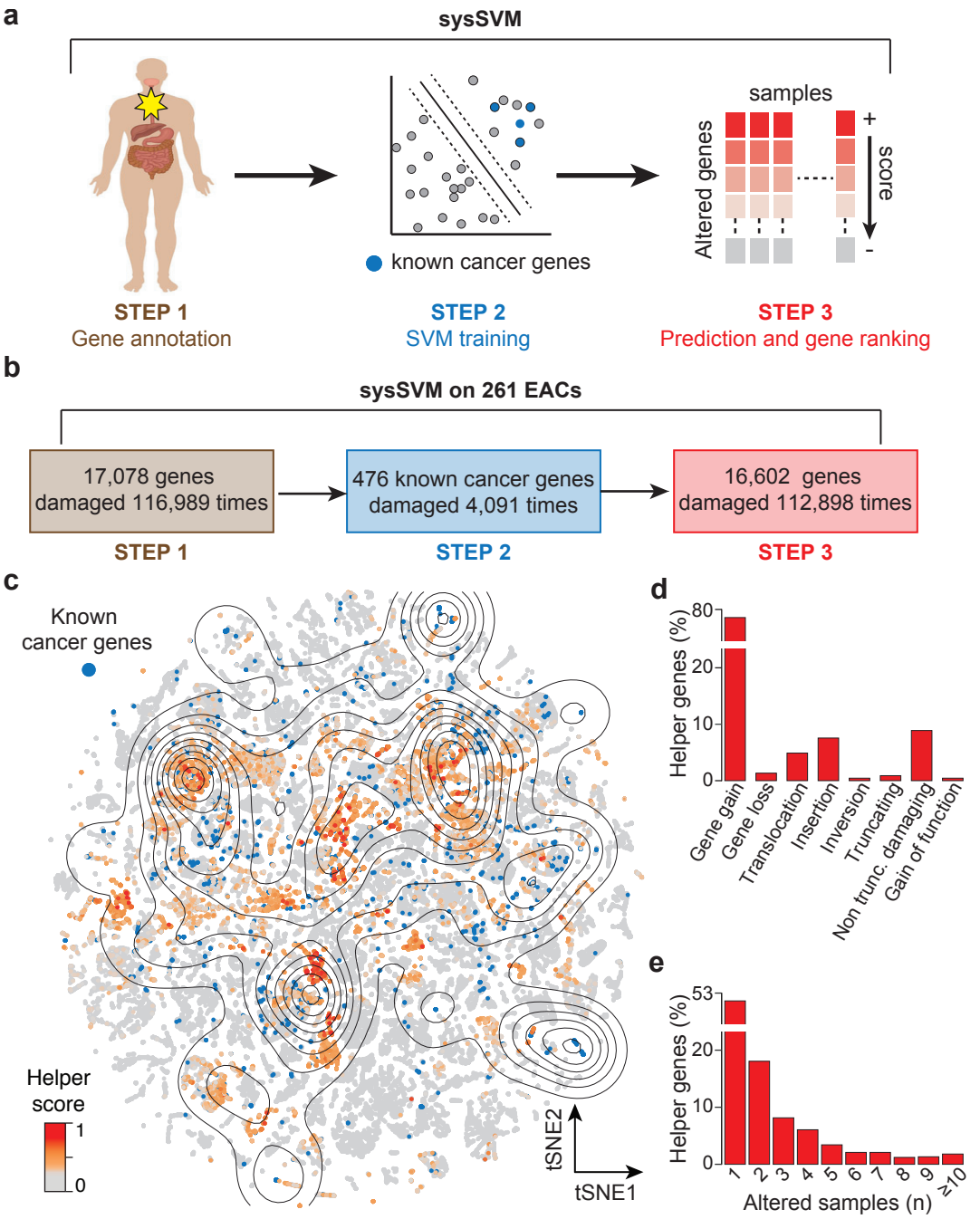
f. Proliferation curves of FLO-1 cells after *ABI2* editing or PAK1 overexpression. Three biological replicates were performed, each with four technical replicates. Proliferation was assessed every 24 hours and each time point was normalized to time zero. Mean values at 72 hours were compared by one-tailed Student's t-test.

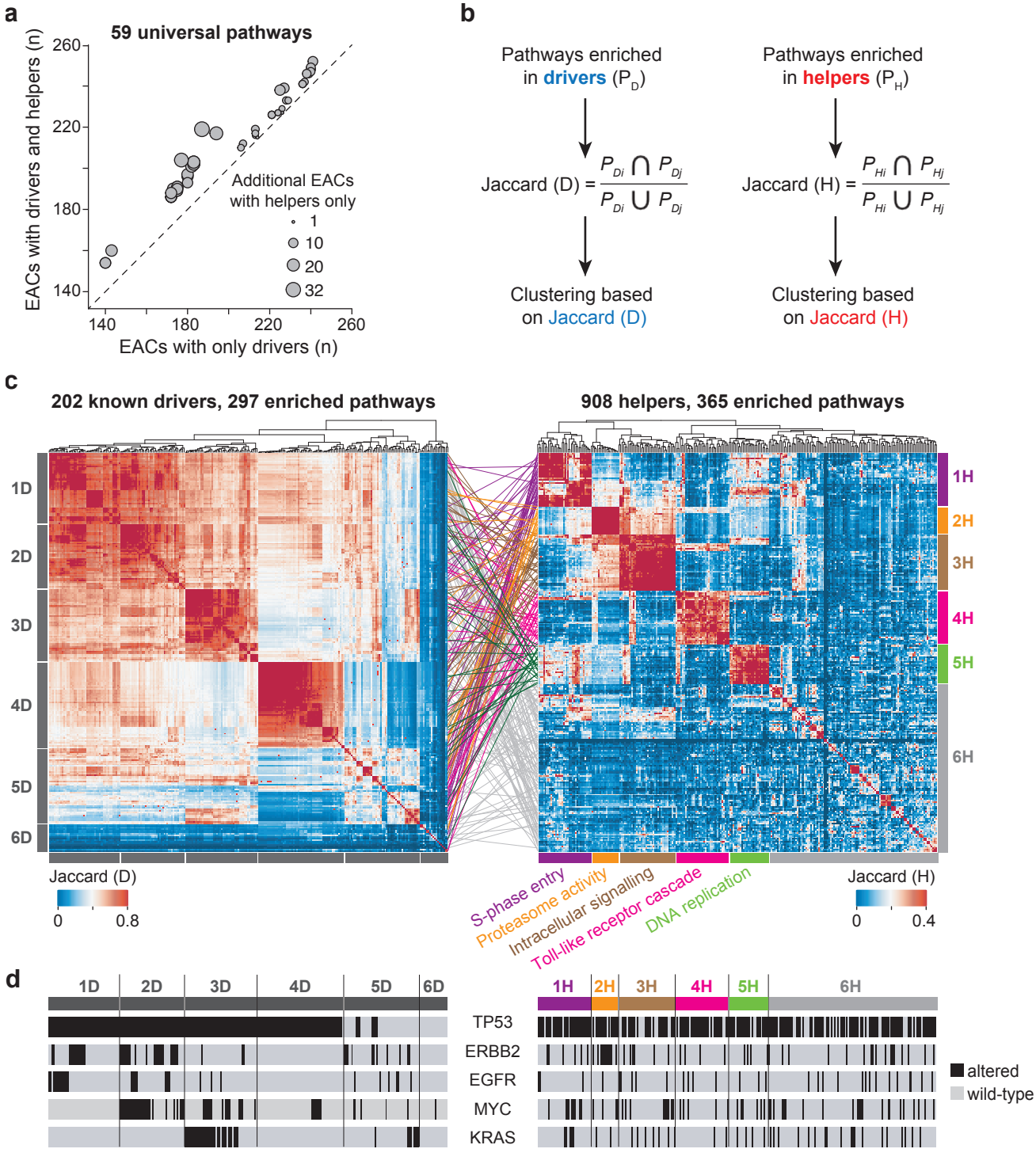
LIST OF SUPPLEMENTARY DATA

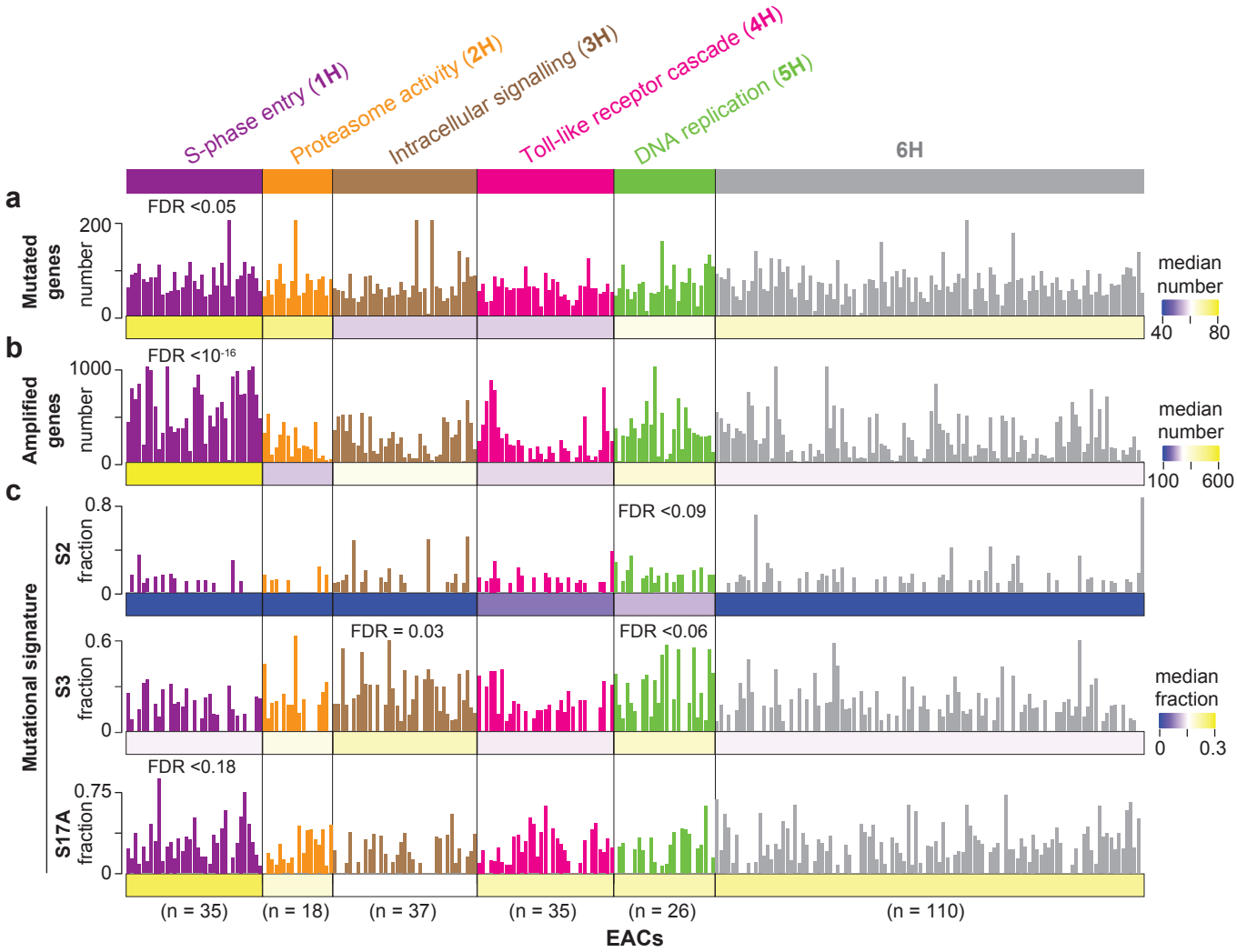
Supplementary Text. The sysSVM algorithm

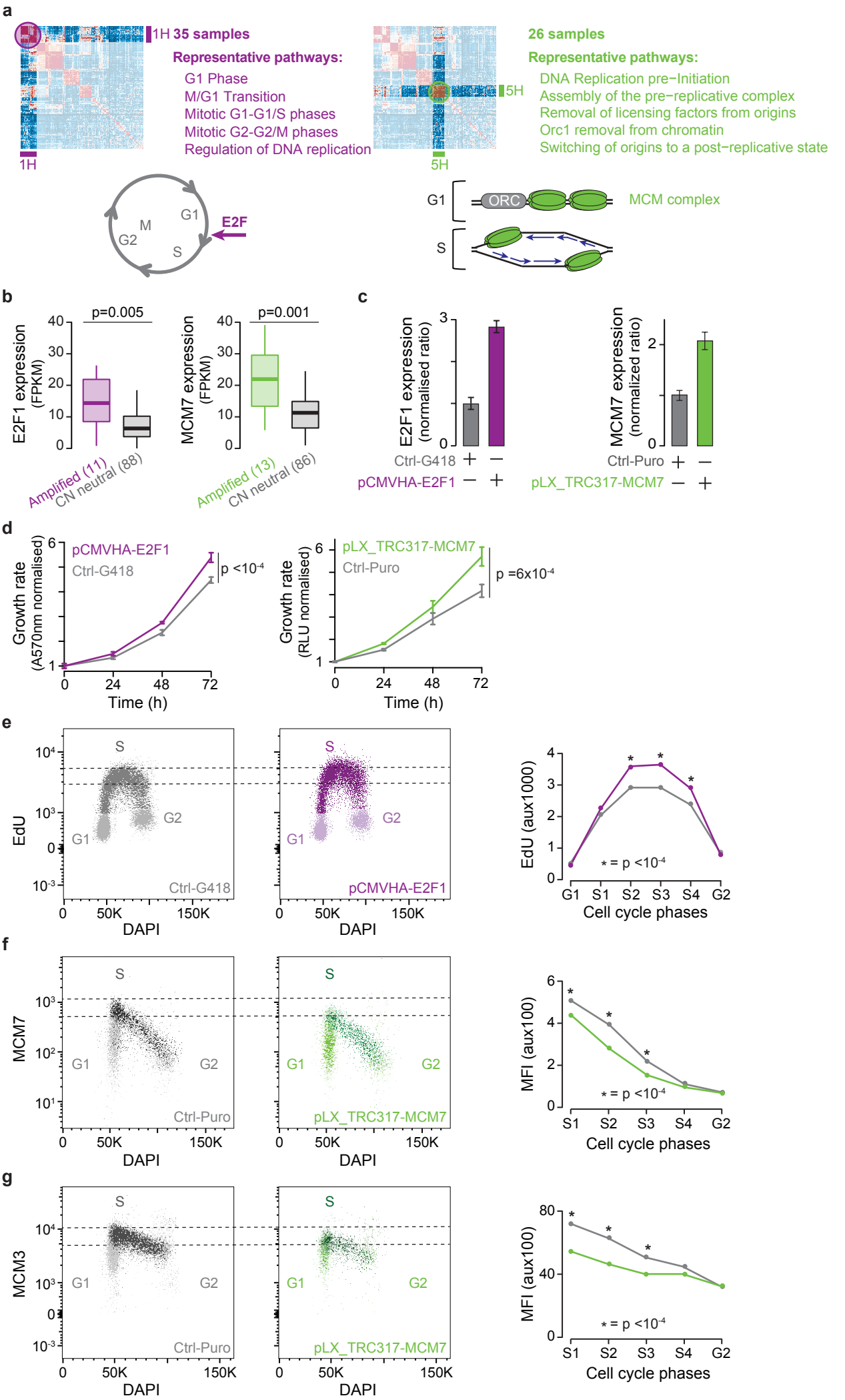
Supplementary Figure 1. tSNE plots of 34 properties of known cancer genes
Supplementary Figure 2. sysSVM validation
Supplementary Figure 3. Co-amplifications and features of known cancer drivers
Supplementary Figure 4. Clustering of 261 EACs using 297 pathways enriched in known drivers
Supplementary Figure 5. Clustering of 261 EACs using 365 pathways enriched in helpers
Supplementary Figure 6. Comparison of helpers using different ranking cut offs

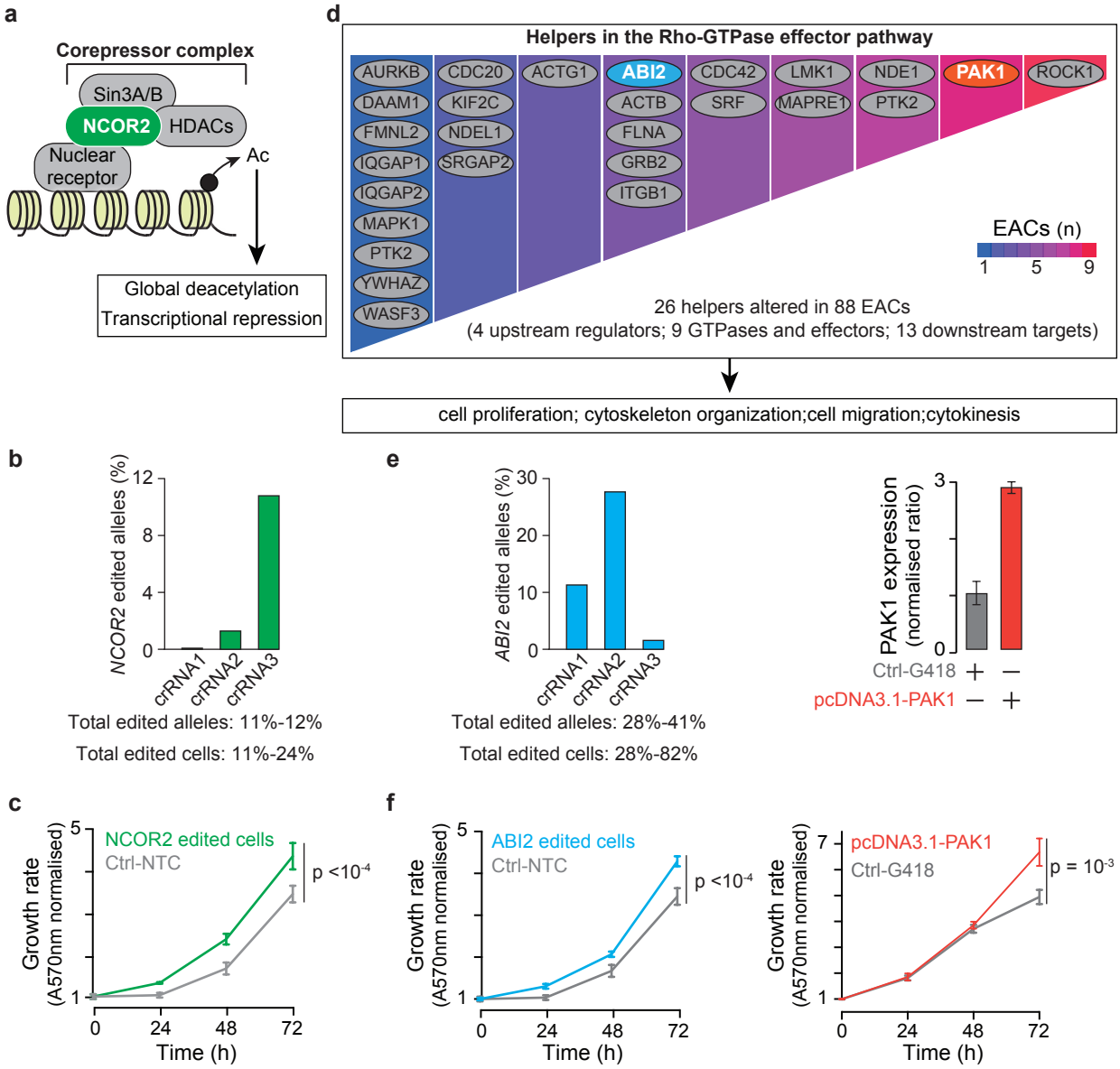
Supplementary Table 1. Molecular and systems-level properties used in sysSVM
Supplementary Table 2. EACs used in this study
Supplementary Table 3. Known cancer genes with driver alterations in the 261 EACs used as training set
Supplementary Table 4. Parameters and performance of the four best models used to predict EAC cancer genes
Supplementary Table 5. Cancer helper genes in 261 EACs
Supplementary Table 6. Results of the gene set enrichment analysis of EAC drivers or helpers
Supplementary Table 7. List of oligos used in the study


a
sysSVM
known cancer genes
samples
Altered genes
score
+
-
STEP 1
Gene annotation
STEP 2
SVM training
STEP 3
Prediction and gene ranking
b
sysSVM on 261 EACs
17,078 genes damaged 116,989 times
476 known cancer genes damaged 4,091 times
16,602 genes damaged 112,898 times
STEP 1
STEP 2
STEP 3
c
Known cancer genes
Helper score
1
0
tSNE2
tSNE1
d
Helper genes (%)
80
20
10
0
Gene gain
Gene loss
Translocation
Insertion
Inversion
Truncating
Non trunc. damaging
Gain of function
e
Helper genes (%)
53
20
10
0
1
2
3
4
5
6
7
8
9
≥10
Altered samples (n)


a
59 universal pathways
EACs with drivers and helpers (n)
EACs with only drivers (n)
Additional EACs with helpers only
1
10
20
32
b
Pathways enriched in drivers (P_D)
Pathways enriched in helpers (P_H)
Jaccard (D) = P_Di ∩ P_Dj / P_Di ∪ P_Dj
Jaccard (H) = P_Hi ∩ P_Hj / P_Hi ∪ P_Hj
Clustering based on Jaccard (D)
Clustering based on Jaccard (H)
c
202 known drivers, 297 enriched pathways
908 helpers, 365 enriched pathways
1D
2D
3D
4D
5D
6D
1H
2H
3H
4H
5H
6H
Jaccard (D)
0
0.8
Jaccard (H)
0
0.4
S-phase entry
Proteasome activity
Intracellular signalling
Toll-like receptor cascade
DNA replication
d
TP53
ERBB2
EGFR
MYC
KRAS
altered
wild-type


S-phase entry (1H)
Proteasome activity (2H)
Intracellular signalling (3H)
Toll-like receptor cascade (4H)
DNA replication (5H)
6H
a
Mutated genes
number
200
0
FDR <0.05
median number
40
80
b
Amplified genes
number
1000
0
FDR <10-16
median number
100
600
c
Mutational signature
S2
fraction
0.8
0
FDR <0.09
S3
fraction
0.6
0
FDR = 0.03
FDR <0.06
median fraction
0
0.3
S17A
fraction
0.75
0
FDR <0.18
(n = 35)
(n = 18)
(n = 37)
(n = 35)
(n = 26)
(n = 110)
EACs


a
1H 35 samples
Representative pathways:
G1 Phase
M/G1 Transition
Mitotic G1-G1/S phases
Mitotic G2-G2/M phases
Regulation of DNA replication
1H
26 samples
Representative pathways:
DNA Replication pre−Initiation
Assembly of the pre−replicative complex
Removal of licensing factors from origins
Orc1 removal from chromatin
Switching of origins to a post−replicative state
5H
5H
G1
M
G2
S
E2F
G1
ORC
MCM complex
S
b
p=0.005
E2F1 expression (FPKM)
Amplified (11)
CN neutral (88)
p=0.001
MCM7 expression (FPKM)
Amplified (13)
CN neutral (86)
c
E2F1 expression (normalised ratio)
Ctrl-G418
pCMVHA-E2F1
MCM7 expression (normalized ratio)
Ctrl-Puro
pLX_TRC317-MCM7
d
pCMVHA-E2F1
Ctrl-G418
p <10-4
Growth rate (A570nm normalised)
Time (h)
pLX_TRC317-MCM7
Ctrl-Puro
p =6x10-4
Growth rate (RLU normalised)
Time (h)
e
EdU
DAPI
Ctrl-G418
pCMVHA-E2F1
EdU (aux1000)
* = p <10-4
Cell cycle phases
f
MCM7
DAPI
Ctrl-Puro
pLX_TRC317-MCM7
MFI (aux100)
* = p <10-4
Cell cycle phases
g
MCM3
DAPI
Ctrl-Puro
pLX_TRC317-MCM7
MFI (aux100)
* = p <10-4
Cell cycle phases


a
Corepressor complex
Sin3A/B
NCOR2
HDACs
Nuclear receptor
Ac
Global deacetylation
Transcriptional repression
d
Helpers in the Rho-GTPase effector pathway
AURKB
DAAM1
FMNL2
IQGAP1
IQGAP2
MAPK1
PTK2
YWHAZ
WASF3
CDC20
KIF2C
NDEL1
SRGAP2
ACTG1
ABI2
ACTB
FLNA
GRB2
ITGB1
CDC42
SRF
LMK1
MAPRE1
NDE1
PTK2
PAK1
ROCK1
EACs (n)
1
5
9
26 helpers altered in 88 EACs
(4 upstream regulators; 9 GTPases and effectors; 13 downstream targets)
cell proliferation; cytoskeleton organization;cell migration;cytokinesis
b
NCOR2 edited alleles (%)
crRNA1
crRNA2
crRNA3
Total edited alleles: 11%-12%
Total edited cells: 11%-24%
e
ABI2 edited alleles (%)
crRNA1
crRNA2
crRNA3
Total edited alleles: 28%-41%
Total edited cells: 28%-82%
PAK1 expression (normalised ratio)
Ctrl-G418
pcDNA3.1-PAK1
c
Growth rate (A570nm normalised)
NCOR2 edited cells
Ctrl-NTC
p <10-4
Time (h)
f
Growth rate (A570nm normalised)
ABI2 edited cells
Ctrl-NTC
p <10-4
Time (h)
pcDNA3.1-PAK1
Ctrl-G418
p = 10-3
Time (h)