Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance

Maria Secrier[1, 12], Xiaodun Li[2,12] , Nadeera de Silva[2], Matthew D. Eldridge[1], Gianmarco Contino[2], Jan Bornschein[2], Shona MacRae[2], Nicola Grehan[2], Maria O'Donovan[2,3], Ahmad Miremadi[2,3], Tsun-Po Yang[2], Lawrence Bower[1], Hamza Chettouh[2], Jason Crawte[2], Núria Galeano-Dalmau[2], Anna Grabowska[4], John Saunders[5], Tim Underwood[6,25], Nicola Waddell[7], Andrew P. Barbour[8, 9], Barbara Nutzinger[2], Achilleas Achilleos[1], Paul A. W. Edwards[10], Andy G. Lynch[1], Simon Tavaré[1], Rebecca C. Fitzgerald[2,] on behalf of the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium[11].

[1] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
[2] Medical Research Council Cancer Unit, Hutchison/Medical Research Council Research Centre, University of Cambridge, Cambridge, UK
[3] Department of Histopathology, Cambridge University Hospital NHS Trust, Cambridge, UK
[4] Queen's Medical Centre, University of Nottingham, Nottingham, UK
[5] Department of Oesophagogastric Surgery, Nottingham University Hospitals NHS Trust, Nottingham, UK
[6] Cancer Sciences Division, University of Southampton, Southampton, UK
[7] Department of Genetics and Computational Biology, QIMR Berghofer, Herston, Queensland, Australia
[8]Surgical Oncology Group, School of Medicine, The University of Queensland, Translational Research Institute at the Princess Alexandra Hospital, Woolloongabba, Brisbane, Queensland, Australia
[9]Department of Surgery, School of Medicine, The University of Queensland, Princess Alexandra Hospital, Woolloongabba, Brisbane, Queensland, Australia
[10]Department of Pathology, University of Cambridge, Cambridge, UK
[11] A full list of contributers from the OCCAMS Consortium is available at the end of the manuscript
[12] These quthors contributed equally
Correspondance should be addressed to R.C.F (rcf29@mrc-cu.cam.ac.uk).

ABSTRACT

Esophageal adenocarcinoma (EAC) has a poor outcome, and targeted therapy trials have thus far been disappointing due to a lack of robust stratification methods. Whole-genome sequencing (WGS) analysis of 129 cases demonstrates that this is a heterogeneous cancer dominated by copy number alterations with frequent large scale rearrangements. Co-amplification of receptor tyrosine kinases (RTKs) and/or downstream mitogenic activation is almost ubiquitous; thus tailored combination RTKi therapy might be required, as we demonstrate *in vitro*. However, mutational signatures reveal three distinct molecular subtypes with potential therapeutic relevance, which we verify in an independent cohort (n=87): i) enriched for BRCA signature with prevalent defects in the homologous recombination pathway; ii) dominant T>G mutational pattern associated with a high mutational load and neoantigen burden; iii) C>A/T mutational pattern with evidence of an ageing imprint. These subtypes could be ascertained using a clinically applicable sequencing strategy (low coverage) as a basis for therapy selection.

INTRODUCTION

Esophageal cancer is the eighth most common cancer world-wide, and the sixth most common cause of cancer-related deaths [1]. There are two main subtypes, squamous and adenocarcinoma, and the incidence of EAC has increased 4.6-fold amongst white males in the US over the past three decades [2]. It is an aggressive disease, with early loco-regional spread, resulting in a median overall survival of less than a year [3].

Curative treatment has been based on esophagectomy, with the addition of peri-operative chemotherapy or chemoradiotherapy improving survival [4–6]. The use of molecularly targeted agents has lagged behind that of other cancers and the results so far have been disappointing. Indeed, only Trastuzumab treatment has led to any improvement in outcomes, and this was only in ERBB2 positive cases with metastatic disease [7]. Advances in this area have been hampered by the lack of understanding of the molecular drivers of this cancer.

Major sequencing efforts have enabled new classifications of cancers based on their molecular parameters [8, 9]. The emerging genomic biomarkers are based on single nucleotide mutations, structural rearrangements and mutational signatures [10–14], and in some instances these have led to the development of stratified trials with the promise of improved patient outcomes [15].

Exome sequencing and a small number of whole-genome sequences have uncovered a limited number of potential driver mutations in EAC. However, as many of the mutations occur in tumor suppressor genes (*TP53*, *SMAD4*, *ARID1A*), actionable oncogenic mutations have remained elusive [16, 17]. What is emerging is a picture of genomic instability with complex rearrangements leading to significant heterogeneity between patients [18]. What is still lacking is an understanding of how to use these complex molecular data to stratify patients to help inform clinical decision making.

Here, we present WGS data for over 100 cases as part of the International Cancer Genome Consortium, with verification of key findings in independent cohorts. We have used genomic information coupled with expression data and *in vitro* experiments to better understand the failure of targeted therapies and to uncover mechanisms of disease pathogenesis that may inform tumor classification and therapy selection.

RESULTS

Large-scale alterations dominate the EAC landscape

WGS data from 129 EAC patients (including tumors from the gastroesophageal junction, Siewert type 1 and 2) have allowed us to comprehensively catalog the genomic alterations in this cancer, including the large-scale structural rearrangements not detectable from exome sequencing. The clinical characteristics of the cohort are typical for the disease (Supplementary Table 1).

As previously noted, point mutations are abundant in this cancer [16]. However, the overall genomic landscape suggests a disease driven by structural variation and copy number changes (Fig. 1 and Supplementary Figure 1). Analysis of a combined cohort of 111 EAC cases from TCGA [19] and Nones et al [18] confirms a dominance of copy number alterations, compared to point mutations, in the majority of cases (Supplementary Figure 2).

When examining the specific loci affected, potential gene driver events were highly heterogeneous between cases, and structural changes again dominated (Fig. 1). Among the genes altered in 10% or more of cases, many more were rearranged, amplified or deleted than were affected by indels or nonsynonymous point mutations. We observed novel recurrently rearranged genes, including *SMYD3* in 39% of cases, *RUNX1* 27%, *CTNNA3* 22%, *RBFOX1* 21%, the *CDKN2A/2B* locus 18%, *CDK14* 16% (important transcriptional, signalling and cell communication regulators), and fragile sites (*FHIT* 95%, *WWOX* 84%). Somatic L1 mobile element insertions were also abundant. Detecting inserts that had transduced unique flanking sequences identified an average of 25 inserts/tumor (range 0–1127), including those already known to transduce [20, 21] and novel examples. These numbers are substantially higher than previously reported [20] because of improved sensitivity. Mobile element insertions were found in signalling, cell cycle and cell adhesion regulators: *ERBB4* - 6/129, *CTNNA3* – 5/129, *CTNNA2* – 4/129, *CDH18* – 3/129, *SOX5* – 2/129.

Significantly amplified loci according to GISTIC2.0 [22] (7q22, 13q14, 18q11 etc – residual q-value<<0.0001) comprised genes like *ERBB2*, *EFGR*, *RB1*, *GATA4/6*, *CCND1*, *MDM2* among others, while the top significantly deleted loci in the cohort (9p21, 21p11, 3p14, etc – residual q value<<0.0001) showed losses of e.g. *CLDN22*, *CDKN2A*, *CKN2B*, as well as several fragile sites (Supplementary Figure 3 and Supplementary Tables 2 and 3).

The most frequent somatic mutation/indel events included a number of known driver genes with roles in DNA damage, signal transduction, cell cycle and chromatin remodelling. Seven of these reached statistical significance (adjusted to P<0.1) as likely driver genes, as inferred by MutSigCV [23] (Fig. 1e and Supplementary Table 4): *TP53* (81%, P<<0.0001), *ARID1A* (17%, P<<0.0001), *SMAD4* (16%, P<0.0001), *CDKN2A* (15%, P<0.0001), *KCNQ3* (12%, P<0.001), *CCDC102B* (9%, P=0.031), *CYP7B1* (7%, P=0.054), largely as previously described [16, 17]. In addition *SYNE1* was mutated in 23% of cases, but did not reach significance by MutSigCV.

The high frequency of genomic catastrophes observed was consistent with a significant role of larger-scale events in this disease - chromothripsis: 39/129 patients (30%), kataegis: 40/129 (31%), complex rearrangement events: 41/129 (32%), (Methods, Figure 1f and Supplementary Figures 4–7). The complex rearrangements included: focal amplifications with BFB pattern (11/129, 9%); focal amplifications <5Mb-wide with irregular copy number amplification steps (26/129, 20%); focal amplifications 5–10 Mb-wide with symmetric copy number amplification steps (10/129, 8%); double minute-like patterns (3/129, 2%); and subtelomeric BFBs (1/129, 1%) (Supplementary Figure 7). The chromothripsis and BFB/complex rearrangement event frequencies were in a similar range to that described by Nones et al [18] – 33% and 27%, respectively. Kataegis rates were lower than that previously reported (19/22 = 86%), likely due to our more stringent criteria for calling (Methods). An enrichment of C>T and C>G mutations was observed in kataegis regions, as previously reported [24] (Supplementary Figure 5).

Hence, this is a heterogeneous cancer dominated by copy number alterations and large scale rearrangements. Clinically meaningful genomic subgroups relevant for therapy are not immediately apparent from these analyses.

RTK receptors and their targets are pervasively disrupted in EAC

Next we examined the genomic data to understand possible reasons for the disappointing results seen with many of the trials targeting growth factor receptors. Resistance to RTK therapy generally results from co-amplifications of alternative RTKs or amplification/activation of downstream mitogenic pathways. In our cohort we observed widespread gene amplification across multiple RTKs, as well as downstream within the MAPK and PI3K pathways. Such patterns were similar among endoreduplicated and non-endoreduplicated samples, as well as in a panel of cell models (Fig. 2a, 2b).

When considering high level amplifications (GISTIC cut-off greater than 2), we observe similar rates to those reported previously for *EGFR* and *ERBB2* [25, 26]. *ERBB2* was the most amplified RTK (22/129 patients = 17%), followed by *EGFR* (14/129 patients = 11%). Other commonly over-expressed RTKs included *MET* and *FGFR*. All these receptors are targeted in clinical trials with ongoing recruitment (see URLs). When considering lower level amplifications across these RTKs and downstream signaling pathways (GISTIC > 1), these are highly prevalent and may still have relevance for disappointing trial results.

We used expression data for available cases to check the consequences of the observed gains/losses at the transcriptional level for key amplified genes. The genes falling in amplified/gained regions show an increased expression compared to those in lost/deleted regions, confirming the observations from the WGS data (Fig. 2c). This, together with results from IHC staining of matched cases, suggests phenotypic relevance of the genome-level findings (Fig. 2d).

Overall, 40% of the samples have both receptor gain and downstream activation of at least one gene, 43% RTK gain alone, and 2% have downstream activation alone (Fig. 2e). We only see a single RTK gain, without gains or amplifications in the MAPK or PI3K pathways, in 9% of tumors. The observed co-amplification patterns are unlikely to be biased by locus positioning, as the inspected RTKs have a varied distribution on chromosomes; hence they appear to be selected for.

We therefore surmised that tailored RTKi combination therapy might be beneficial in some cases and decided to explore this in *in vitro* model systems. Since copy number gain events were seen most commonly in *ERBB2*, *EGFR*, *MET* and *FGFR*s, a panel of small molecular inhibitors was selected to target these RTKs. As expected, a single agent did trigger a cytotoxic effect in cell lines with a gain at that locus, but only in the micromolar range (Fig. 2g). In cell lines with an *ERBB2* and a *MET* amplification, a significant reduction in cell proliferation was observed when both RTKs were inhibited with a GI50 down in the nanomolar range, for example OE33 (Fig. 2f, 2g, Table 1). A similar finding was observed in FLO-1 (*EGFR*/*MET* copy gain) and OAC-P4C (*ERBB2*/*FGFR2* amplification) when treated with EGFRi/METi and ERBB2i/FRFGi combinations, respectively. These results suggest that a combination of RTK inhibitors tailored to the amplification profile might offer a clinical therapeutic strategy. Nevertheless, the complexity and diffuse patterns of these alterations provide a distinct challenge in the stratification of patients for therapy.

Mutational signatures uncover distinct etiology in EAC

In view of the heterogeneity and RTK-resistance mechanisms, we sought alternative therapeutic insights into the data using mutational signature analysis in a three-base context via the non-negative matrix factorization (NMF) methodology described by Alexandrov et al [27]. We also used the recently described pmsignature [28] and SomaticSignatures [29] for comparison. These methods are based on different statistical frameworks and therefore some differences are to be expected; nevertheless the same key signature patterns were observed with similar-sized patient subgroups expressing the dominant signature types (Supplementary Notes, Supplementary Figures 8–12). Six signatures were prominent (Supplementary Figures 13–14): S17, the hallmark signature of EAC [16, 17] dominated by T>G substitutions in a CTT context and possibly associated with gastric acid reflux – here renamed S17A; a previously uncharacterized variant of this signature combining a relatively higher frequency of T>C substitutions with the classical T>G pattern found in S17, which we call S17B; S3, a complex pattern caused by defects in the BRCA1/2-led homologous recombination pathway; S2, C>T mutations in a TCA/TCT context, an APOBEC-driven hypermutated phenotype; S1, C>T in a *CG context, associated with aging processes; and an S18-like signature, C>A/T dominant in a GCA/TCT context, formerly described in neuroblastoma, breast and stomach cancers (Fig. 3a). The exploration of a seven-base signature context using pmsignature yielded an A/T base dominance at the -3 and -2 positions for the S17 signature, but no other striking preferences for nucleotide combinations at the 2nd and 3rd bases for any of the other signatures (Supplementary Figure 15). Overall, this suggests that the bases immediately adjacent to the position where the mutation occurs exert the main bias, with a potentially more complex mechanism for the S17 signature.

When considering the dominant mutation signatures on a per-patient basis, three subgroups of patients became apparent: *C>A/T dominant* (age, S18-like), *DNA Damage Repair* (*DDR) impaired* (BRCA), and *mutagenic* (predominantly S17A or S17B) (Fig. 3a). We chose the descriptor *mutagenic* because the mutation rate was significantly higher in this subgroup (Welch's t-test p = 0.0007; Supplementary Figure 16). The robustness of the subgroups was ensured through consensus clustering and confirmed by silhouette statistics (Methods, Supplementary Figures 17–18). We also validated our findings in an independent cohort of 87 samples [18] and show that: when we apply the NMF method the same dominant signatures (S1, S2, S3, S17, S18-like) are observed; and when we perform clustering three subgroups emerge which are of similar composition and proportions to those seen in the original cohort (Methods, Fig. 3b compared with Fig. 3a). Furthermore, the total mutational burden is again consistently higher in the mutagenic subgroup of the validation cohort. No cellularity bias or batch effect was observed among subgroups (Supplementary Figure 19).

To test whether spatial sampling might have induced a bias in the predicted signatures, we inspected three additional patients who had multiple samples taken. The mutational patterns showed remarkable consistency across all three biopsies, especially regarding the dominant signature (Fig. 3c).

We next examined whether the defined subgroups presented similarities in terms of genomic characteristics. All three subgroups showed a similar degree of heterogeneity in copy number alterations by chromosomal arm (Supplementary Figure 20), and the RTK co-amplification profiles were fairly similar among subgroups (Supplementary Figure 21). Of note, the C>A/T dominant subgroup had a two-fold higher frequency of *ERBB2*/*MET* co-amplifications, but this did not reach statistical significance.

The rearrangement patterns in the three subgroups denoted differences in genomic stability. In particular, unstable genomes were less frequent in the C>A/T dominant subgroup and most frequent in the DDR impaired subgroup [11, 18] (Supplementary Figure 22). When examining SV signatures using the NMF framework (Methods), the C>A/T dominant subgroup also had lower levels of large-scale duplications and an increased frequency of focal interchromosomal translocations, which suggest mobile element insertion events (Supplementary Figure 23). The DDR impaired subgroup seemed to have the largest degree of genomic instability, though SV signatures were overall rather heterogeneous. No recurrently altered genes (in >10% of the cohort) were over-represented in any of the three subgroups after multiple testing correction, nor were there any differences in *TP53* or *ERBB2* status among the subgroups to account for the differences in genomic stability.

The clinical characteristics of the three subgroups did not differ significantly (Supplementary Table 5, Supplementary Figure 24), implying that the classification, and hence spectrum of mutation patterns, does not vary with smoking, age, sex, tumor histopathological grade, tumor stage, response to chemotherapy, overall or recurrence–free survival etc. Hence, the mutation signature profiles seem to be capturing a different type of information compared with current clinical classification methods.

Evidence of DNA damage repair deficiency in EAC

Next we investigated what aspects of the DNA damage response were defective in the DDR impaired subgroup. Although a BRCA signature was recovered, there were only 3 nonsynonymous mutations and 3 germline variants (non-intronic) in either BRCA1 or 2 in a total of 5 out of 18 patients, suggesting that other mechanisms were largely responsible for this signature (Supplementary Tables 6 and 7). We thus assessed the mutation rates across more than 450 genes associated with DDR, as previously described in a pan-cancer analysis [30] (Fig 4, Methods). We found that there was a 4.3-fold enrichment of samples with alterations in homologous recombination (HR) pathways in the DDR impaired subgroup compared to the others (95% CI [1.47, 12.56]). It is therefore likely that a pathway-level disruption of HR contributes to the BRCA-like mutational signature rather than mutations of BRCA genes.

The analysis of DDR genes in the whole cohort unsurprisingly showed that the most mutated pathway was *TP53* (Supplementary Figure 25), and this was consistent among subgroups (Fig. 4a), as were the amplification and deletion patterns (Supplementary Figure 26). In addition, more than 24% of the genomes had defects in chromatin remodelling, comprising recurrently mutated genes like *ARID1A* (8%) and *SMARCA4* (8%) (Fig. 4b). *ARID1A* is also recruited to DNA double strand breaks (DSB), where it facilitates processing to single strand ends [31]. Defects in *ARID1A* impair this process and may sensitise cells *in vitro* and *in vivo* to PARP inhibition (PARPi) [31].

Neoantigen and CD8 profiles in the mutagenic subgroup

Modulation of the cytotoxic T cell response using monoclonal antibodies against the Programmed Death Receptor or Ligand (PD-1 and PD-L1 inhibitors), as well as those targeting CTLA4 (Ipilimumab) have shown promise in the treatment of solid tumors [32–34]. The recent literature suggests that both numbers of mutations and total neoantigen burden have been coupled with significantly better clinical responses to immunotherapy [35–37].

We found that the mutagenic subgroup, whose observed signature may be due to gastric acid reflux, harbored a significantly higher nonsynonymous mutational burden, as well as higher levels of neoantigen presentation (Welch's t-test $p = 0.0007$ and Wilcoxon rank-sum test $p << 0.0001$, respectively; Fig 5a and Supplementary Figure 16). This is in keeping with that observed for lung cancer and metastatic melanoma, with a 1.5-fold higher median neoantigen burden in this subgroup versus the rest – similar to the two-fold ratio reported by Rizvi et al [35, 38]. Using available RNA expression data we observed a significantly higher number of neoantigens expressed in this subgroup compared to the rest (Wilcoxon rank-sum test p-value = 0.042, Fig. 5a).

In recent studies, an enriched population of pre-existing CD8+ T cells was shown to predict a favorable outcome from PD-1 blockade therapy [39, 40]. We found a higher density of CD8+ T cells in a subset of available samples from the mutagenic signature subgroup compared with samples from the other subgroups (Fig. 5a, 5b).

Treatment responses in mutational signature subgroups

Given the complexity of the RTK landscape and the apparent need to profile each patient to determine the optimal combination of RTK inhibitors, we hypothesised that the more homogeneous profile of mutational signatures might be a more clinically applicable starting point to guide therapy decisions. To start to test this hypothesis, we used newly derived cell line models from patients in the OCCAMS consortium with an available germline reference sequence from which we could derive the signatures: OES127, DDR impaired profile; MFD, mutagenic profile; CAM02 C>A/T dominant profile (Fig. 6a). For the DDR impaired profile we hypothesised that PARPi, with or without a DNA-damaging agent such as Topotecan, might be beneficial [31, 41, 42]. Topoisomerase I (Topo1) is an enzyme required for DNA replication and when inhibited in combination with Olaparib it has been shown to generate synthetic lethality in BRCA deficient cases [43, 44]. Unexpectedly, no cytotoxic effect was observed when Olaparib or Topotecan was used as single reagent, however, a marked synergistic effect was shown when Topotecan was combined with Olaparib for OES127 (DDR impaired group), but not for the other primary cell lines (Fig. 6b, Supplementary Table 8).

Next we tested the efficacy of Wee1/Chk1 inhibitors given the high frequency of *TP53* mutation in this disease [45, 46]. Several recent studies revealed that pharmacological inhibition of G2/M-phase checkpoint regulators Wee1 and Chk1/2 resulted in an antitumorigenic effect in some highly mutated cancers [47, 48]. We therefore hypothesised that inhibition of mitotic checkpoints would be cytotoxic in EAC and that this might be more apparent in cells with a high mutation burden [49, 50]. As expected, a cytotoxic effect for these drugs was observed to some extent in all of our primary cell lines, but the sensitivity was increased in the CAM02 and MFD lines in comparison with the wild-type *TP53* line OES127 (Fig. 6c, Supplementary Table 9). In the MFD cells with a mutagenic signature, there was a 25-fold and 10-fold increased sensitivity in response to the Wee1 and Chk1/2 inhibitor, respectively, compared with the CAM02 cells from the C>A/T dominant subgroup.

These experimental data provide a starting point from which to evaluate therapeutic options derived from mutational signatures, especially as primary model systems more closely resembling human disease and with stromal components become available [51, 52].

DISCUSSION

Whole-genome sequencing of 129 EAC patients has unveiled a high prevalence of large-scale alterations that may play an important role in the development of this cancer. Similarly to ovarian, breast and lung cancers which have been described as ‘copy number driven’ [53], relatively few genes were recurrently point-mutated (except TP53), but there were frequent recurrent amplifications in sites harbouring oncogenes, deletions of important cell cycle components (*CDKN2A*, *CDKN2B*) and rearrangements of genes like *RUNX1*, frequently translocated in leukemias [54]. The highly heterogeneous landscape explains the difficulties encountered to date in finding suitable avenues for tailored therapies. Currently 88 of 262 registered esophageal trials (see URLs) target RTKs and mitogenic signalling pathways with remarkably little clinical efficacy. The genomic and *in vitro* analyses performed here suggest that the high prevalence of co-amplification of RTKs and downstream mitogenic pathway genes is likely to explain these disappointing results.

Although all six mutational signatures are seen to some extent in most patient tumors, three distinct dominant subtypes, namely *DDR impaired*, *C>A/T dominant*, and *mutagenic*, point to specific etiological factors or genetic instabilities dominating the development of any individual's EAC. We hypothesise that the insights obtained from mutational signatures could be harnessed for future studies to investigate the potential of tailored therapies to complement the current treatment options as summarized in Figure 7.

In the DDR impaired subgroup with an enrichment for HR dysfunction, a synthetic lethality approach may prove useful. Indeed, HR scarring is a good a biomarker for DDR targeted treatment [55], being well established in breast and ovarian cancer and more recently also reported in gastric tumors [56]. HR dysfunction renders tumors sensitive to platinum-based chemotherapy and PARPi, which has started to make a survival impact in other BRCA-related tumors [57]. Indeed, we also observe some increased sensitivity to platinum-based chemotherapy in the DDR impaired subgroup (Supplementary Figure 27). PARPi in combination with irradiation has shown to be potent in HR scarred tumors [58] and our data from a primary line with a DDR signature suggests that PARPi in combination with a DNA damaging agent might be beneficial.

Expression of PD-L1 has been demonstrated in gastroesophageal tumors at all stages, and therefore PD-L1 based immunotherapy might be an attractive therapeutic avenue to explore [59]. Both the nonsynonymous mutation burden and the neoantigen level, as well as CD8+ cell infiltration, have been shown to be good biomarkers in predicting response to immunotherapy in both smoking-related non-small cell lung cancer and melanoma [35, 36, 40, 59]. In keeping with these tumors which result from chronic exposure to mutagens (smoking and UV irradiation, respectively), we observe similar features in our mutagenic cohort containing an 'acid' signature. This type of genomic classification has also been proposed in other tumor types for patient stratification for immunotherapy [60] and warrants further investigation in this cancer. Similarly, Chk/Wee1 inbitors may be promising tools for future studies in highly mutated, p53-inactive tumours [47, 48].

Patients in the C>A/T dominant subgroup would continue to be treated with conventional chemotherapy until more progress is made, e.g. with synthetic lethality approaches combined with radiotherapy or mutant p53 reactivating drugs [61,63]. Alternatively, combined RTK inhibitors (especially ERBB2 and MET, given their prevalence in this subgroup) may be beneficial and combined MEK and Akt inhibition might be worthy of consideration given the low levels of amplifications/activation seen downstream in the MAPK and PI3K pathways [64].

One practical question that arises is how this approach could be implemented clinically. Despite the decreasing costs of WGS, it is still expensive and signatures are problematic to derive from whole-exome data [27]. However, lower coverage whole-genome (10x), or even shallow (1x) genome sequencing could provide a cost-effective, high-throughput alternative for signature-based stratification and we have shown using simulations down to 10x that we can confidently retrieve dominant signatures at lower coverage (Supplementary Figure 28). Moreover, while designing custom gene panels would pose serious difficulties in such a heterogeneous disease, mutational signature-based classification would enable us to bypass the tumor heterogeneity bottleneck by providing a genome-wide, spatially-independent classification strategy (Fig. 3c).

For subsequent individual patient classification, we propose a quadratic programming approach whereby we predict exposures to the six mutational signatures without having to estimate a large set of parameters (as with the classical NMF algorithm) and use the dominant signature pattern for patient assignment (Supplementary Notes). Figure 7 illustrates this fast and effective way of classifying new patients. This methodology is of course not without limitation: the age, S18-like and APOBEC signatures are currently grouped together, but in a much larger cohort a distinct 'age' or 'APOBEC' subgroup might emerge. Similarly, signatures S17A and S17B may merge in a much larger cohort, as was the case for signatures S1A and S1B [27]. It should be noted that algorithms for defining signatures are evolving with improved speed of computation [28] and there is inherent variation in sample categorization between methods. Methodology is also being developed to accurately identify signatures de-novo in single patients, which we expect will offer promising alternatives for patient stratification.

In summary, we have uncovered possible reasons for the lack of efficacy in molecularly targeted trials and present a novel genomic classification which links etiology to patient stratification with potential therapeutic relevance. Further studies will be needed for pre-clinical validation prior to implementation in trials, as well as to understand the extent to which this genomic distinction is maintained downstream, at the level of the transcriptome, proteome and cellular phenotype.

[12] Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium:

Ayesha Noorani[2], Rachael Fels Elliott[2], Jamie Weaver[2], Caryn Ross-Innes[2], Laura Smith[2], Zarah Abdullahi[2], Rachel de la Rue[2], Alison Cluroe[3], Shalini Malhotra[3], Richard Hardwick[14], Hugo Ford[14],Mike L Smith[1], Jim Davies[15], Richard Turkington[16],Stephen J. Hayes[17,18], Yeng Ang[17,19,20], Shaun R. Preston[21], Sarah Oakes[21], Izhar Bagwan[21], Vicki Save[22], Richard J.E. Skipworth[22], Ted R. Hupp[22], J. Robert O'Neill[22,23], Olga Tucker[24,25], Philippe Taniere[24], Fergus Noble[26], Jack Owsley[26], Laurence Lovat[27], Rehan Haidry[27], Victor Eneh[27], Charles Crichton[28], Hugh Barr[29], Neil Shepherd[29], Oliver Old[29], Jesper Lagergren[30,31,32], James Gossage[30,31], Andrew Davies[30,31], Fuju Chang[30,31], Janine Zylstra[30,31],Grant Sanders[33], Richard Berrisford[33], Catherine Harden[33], David Bunting[33], Mike Lewis[34], Ed Cheong[34], Bhaskar Kumar[34], Simon L Parsons[5], Irshad Soomro[5], Philip Kaye[5], Pamela Collier[5], Laszlo Igali[35], Ian Welch[36], Michael Scott[36], Shamila Sothi[37], Sari Suortamo[37], Suzy Lishman[38], Duncan Beardsmore[39], Hayley E. Francies[40], Mathew J. Garnett[40], John V. Pearson[7,40], Katia Nones[7,40], Ann-Marie Patch[7,40], Sean M. Grimmond[40,41]

[13]Oesophago-Gastric Unit, Addenbrooke's Hospital, Cambridge, UK
[14]Oxford ComLab, University of Oxford, UK
[15]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Northern Ireland, UK
[16]Salford Royal NHS Foundation Trust, Salford, UK
[17]Faculty of Medical and Human Sciences, University of Manchester, UK
[18]Wigan and Leigh NHS Foundation Trust, Wigan, Manchester, UK
[19]GI science centre, University of Manchester, UK
[20]Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK
[21] The Royal Infirmary of Edinburgh (NHS Lothian), Edinburgh, UK
[22]Edinburgh University, Edinburgh, UK
[23]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[24]Institute of Cancer and Genomic Sciences, University of Birmingham
[25]University Hospital Southampton NHS Foundation Trust, Southampton, UK
[26]University College London, London, UK
[27]Department of Computer Science, University of Oxford, UK
[28]Gloucester Royal Hospital, Gloucester, UK
[29]St Thomas's Hospital, London, UK
[30]King's College London, London, UK
[31]Karolinska Institutet, Stockholm, Sweden
[32]Plymouth Hospitals NHS Trust, Plymouth, UK
[33]Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, UK
[34]Norfolk and Waveney Cellular Pathology Network, Norwich, UK
[35]University Hospital of South Manchester NHS Foundation Trust, Wythenshawe, Manchester, UK
[36]University Hospitals Coventry and Warwickshire NHS, Trust, Coventry, UK
[37]Peterborough Hospitals NHS Trust, Peterborough City Hospital, Peterborough, UK
[38]Royal Stoke University Hospital, UHNM NHS Trust, UK
[39]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK
[40]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, Queensland, Australia
[41]Victorian Comprehensive Cancer Centre, University of Melbourne, Melbourne, Australia

URLs

UKCRN Trial Portfolio [cited 2015 22/11/15]: https://www.ukctg.nihr.ac.uk/
US National Institutes of Health trial registry [cited 22/06/15]: https://clinicaltrials.gov/
Picard 1.105: http://broadinstitute.github.io/picard
BWA-MEM: http://arxiv.org/abs/1303.3997

Accession codes

The whole-genome sequencing and RNA expression data can be found at the European Genome-phenome Archive (EGA) under accession EGAD00001002218 and EGAD00001002260.

ACKNOWLEDGEMENTS

This paper is dedicated to Nadeera de Silva who tragically and unexpectedly died whilst this paper was undergoing revision. He made an important contribution to this research, particularly bringing his clinical oncology perspective to bear on the translational relevance of the findings.

Whole-genome sequencing of esophageal adenocarcinoma samples was performed as part of the International Cancer Genome Consortium (ICGC) through the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium and was funded by a programme grant from Cancer Research UK (RG66287, RG81771, RD84119).We thank the ICGC members for their input on verification standards as part of the benchmarking exercise. 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 and UCL. Also the University Hospital of Southampton Trust and the Southampton, Birmingham, Edinburgh and UCL Experimental Cancer Medicine Centres and the QEHB charities. R.C.F. is funded by an NIHR Professorship (RG67258) and receives core funding from the Medical Research Council (RG84369) and infrastructure support from the Biomedical Research Centre (RG64237) and the Experimental Cancer Medicine Centre (RG62923). We acknowledge the support of The University of Cambridge, Cancer Research UK (C14303/A17197) and Hutchison Whampoa Limited. We would like to thank Dr. Peter Van Loo for providing the NGS version of ASCAT for copy number calling. We are grateful to all the patients who provided written consent for participation in this study and the staff at all participating centres.

Some of the work was undertaken at UCLH/UCL who received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health. The work at UCLH/UCL was also supported by the CRUK UCL Early Cancer Medicine Centre.

Author Contributions

R.C.F. conceived the overall study. M.S., X.L. and P.A.W.E. analysed the data. R.C.F., M.S., X.L., N.S., P.A.W.E. and A.G.L. conceived and designed the experiments. M.S. performed the statistical analysis. X.L., G.C., S.M., M.O., A.M., J.C. and N.G.D. performed the experiments. M.E. performed benchmarking studies on the variant calls, implemented and ran several variant calling and analysis pipelines. G.C. contributed to the structural variant analysis. J.B. contributed expression data and curated the clinical data collection. S.M. and N.G. coordinated sample processing with clinical centers and was responsible for sample collections. T.P.Y. performed the BFB analysis. L.B. ran the variant calling pipelines. H.C. contributed to the RTK analysis. A.G., J.S. and T.U. contributed cell lines. N.W. and A.P.B. contributed sequencing data for validation. B.N. coordinated data and tissue collection from centres for the study. A.A. helped develop the copy number calling pipeline. R.C.F. and S.T. jointly supervised the research. M.S., N.S., X.L. and R.C.F. wrote the manuscript. All authors approved the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

References

1. Ferlay, J., et al., *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.* International Journal of Cancer, 2015. **136**(5): p. E359-86.
2. Brown, L.M., S.S. Devesa, and W.H. Chow, *Incidence of adenocarcinoma of the esophagus among white Americans by sex, stage, and age.* J Natl Cancer Inst, 2008. **100**(16): p. 1184-7.
3. Cunningham, D., A.F. Okines, and S. Ashley, *Capecitabine and oxaliplatin for advanced esophagogastric cancer.* N Engl J Med, 2010. **362**(9): p. 858-9.
4. Allum, W.H., et al., *Long-term results of a randomized trial of surgery with or without preoperative chemotherapy in esophageal cancer.* J Clin Oncol, 2009. **27**(30): p. 5062-7.
5. Cunningham, D., et al., *Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer.* N Engl J Med, 2006. **355**(1): p. 11-20.
6. van Hagen, P., et al., *Preoperative chemoradiotherapy for esophageal or junctional cancer.* N Engl J Med, 2012. **366**(22): p. 2074-84.
7. Bang, Y.J., et al., *Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial.* Lancet, 2010. **376**(9742): p. 687-97.
8. Gao, Y.B., et al., *Genetic landscape of esophageal squamous cell carcinoma.* Nat Genet, 2014. **46**(10): p. 1097-102.
9. Schulze, K., et al., *Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets.* Nat Genet, 2015. **47**(5): p. 505-11.
10. *Genomic Classification of Cutaneous Melanoma.* Cell, 2015. **161**(7): p. 1681-96.
11. Waddell, N., et al., *Whole genomes redefine the mutational landscape of pancreatic cancer.* Nature, 2015. **518**(7540): p. 495-501.

12. Totoki, Y., et al., *Trans-ancestry mutational landscape of hepatocellular carcinoma genomes.* Nat Genet, 2014. **46**(12): p. 1267-73.
13. *Comprehensive molecular characterization of gastric adenocarcinoma.* Nature, 2014. **513**(7517): p. 202-9.
14. *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-50.
15. Chantrill, L.A., et al., *Precision Medicine for Advanced Pancreas Cancer: The Individualized Molecular Pancreatic Cancer Therapy (IMPaCT) Trial.* Clin Cancer Res, 2015. **21**(9): p. 2029-37.
16. Dulak, A.M., et al., *Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity.* Nat Genet, 2013. **45**(5): p. 478-86.
17. Weaver, J.M., et al., *Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis.* Nat Genet, 2014. **46**(8): p. 837-43.
18. Nones, K., et al., *Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis.* Nat Commun, 2014. **5**: p. 5224.
19. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.
20. Paterson, A.L., et al., *Mobile element insertions are frequent in oesophageal adenocarcinomas and can mislead paired-end sequencing analysis.* BMC Genomics, 2015. **16**: p. 473.
21. Tubio, J.M., et al., *Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes.* Science, 2014. **345**(6196): p. 1251343.
22. 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, 2011. **12**(4): p. R41.
23. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
24. Nik-Zainal, S., et al., *The life history of 21 breast cancers.* Cell, 2012. **149**(5): p. 994-1007.
25. Paterson, A.L., et al., *Characterization of the timing and prevalence of receptor tyrosine kinase expression changes in oesophageal carcinogenesis.* Journal of Pathology, 2013. **230**(1): p. 118-28.
26. Van Cutsem, E., et al., *HER2 screening data from ToGA: targeting HER2 in gastric and gastroesophageal junction cancer.* Gastric Cancer, 2014.
27. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
28. Shiraishi, Y., et al., *A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures.* PLoS Genet, 2015. **11**(12): p. e1005657.
29. Gehring, J.S., et al., *SomaticSignatures: inferring mutational signatures from single-nucleotide variants.* Bioinformatics, 2015. **31**(22): p. 3673-5.
30. Pearl, L.H., et al., *Therapeutic opportunities within the DNA damage response.* Nat Rev Cancer, 2015. **15**(3): p. 166-80.
31. Shen, J., et al., *ARID1A Deficiency Impairs the DNA Damage Checkpoint and Sensitizes Cells to PARP Inhibitors.* Cancer Discov, 2015. **5**(7): p. 752-67.
32. Hodi, F.S., et al., *Improved survival with ipilimumab in patients with metastatic melanoma.* N Engl J Med, 2010. **363**(8): p. 711-23.

33. Larkin, J., et al., *Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma.* N Engl J Med, 2015. **373**(1): p. 23-34.
34. Herbst, R.S., et al., *Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial.* Lancet, 2015.
35. Rizvi, N.A., et al., *Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer.* Science, 2015. **348**(6230): p. 124-8.
36. Snyder, A., et al., *Genetic basis for clinical response to CTLA-4 blockade in melanoma.* N Engl J Med, 2014. **371**(23): p. 2189-99.
37. McGranahan, N., et al., *Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade.* Science, 2016. **351**(6280): p. 1463-9.
38. Van Allen, E.M., et al., *Genomic correlates of response to CTLA-4 blockade in metastatic melanoma.* Science, 2015. **350**(6257): p. 207-11.
39. Tumeh, P.C., et al., *PD-1 blockade induces responses by inhibiting adaptive immune resistance.* Nature, 2014. **515**(7528): p. 568-71.
40. Hamanishi, J., et al., *Programmed cell death 1 ligand 1 and tumor-infiltrating CD8+ T lymphocytes are prognostic factors of human ovarian cancer.* Proc Natl Acad Sci U S A, 2007. **104**(9): p. 3360-5.
41. Benafif, S. and M. Hall, *An update on PARP inhibitors for the treatment of cancer.* Onco Targets Ther, 2015. **8**: p. 519-28.
42. Oza, A.M., et al., *Olaparib combined with chemotherapy for recurrent platinum-sensitive ovarian cancer: a randomised phase 2 trial.* Lancet Oncol, 2015. **16**(1): p. 87-97.
43. Demel, H.R., et al., *Effects of topoisomerase inhibitors that induce DNA damage response on glucose metabolism and PI3K/Akt/mTOR signaling in multiple myeloma cells.* Am J Cancer Res, 2015. **5**(5): p. 1649-64.
44. Farmer, H., et al., *Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy.* Nature, 2005. **434**(7035): p. 917-21.
45. Di Leonardo, A., et al., *DNA damage triggers a prolonged p53-dependent G1 arrest and long-term induction of Cip1 in normal human fibroblasts.* Genes Dev, 1994. **8**(21): p. 2540-51.
46. Agarwal, M.L., et al., *A p53-dependent S-phase checkpoint helps to protect cells from DNA damage in response to starvation for pyrimidine nucleotides.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14775-80.
47. Brooks, K., et al., *A potent Chk1 inhibitor is selectively cytotoxic in melanomas with high levels of replicative stress.* Oncogene, 2013. **32**(6): p. 788-96.
48. Vera, J., et al., *Chk1 and Wee1 control genotoxic-stress induced G2-M arrest in melanoma cells.* Cell Signal, 2015. **27**(5): p. 951-60.
49. Liu, Q., et al., *Chk1 is an essential kinase that is regulated by Atr and required for the G(2)/M DNA damage checkpoint.* Genes Dev, 2000. **14**(12): p. 1448-59.
50. Watanabe, N., M. Broome, and T. Hunter, *Regulation of the human WEE1Hu CDK tyrosine 15-kinase during the cell cycle.* EMBO J, 1995. **14**(9): p. 1878-91.
51. van de Wetering, M., et al., *Prospective derivation of a living organoid biobank of colorectal cancer patients.* Cell, 2015. **161**(4): p. 933-45.
52. Sato, T., et al., *Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche.* Nature, 2009. **459**(7244): p. 262-5.
53. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers.* Nat Genet, 2013. **45**(10): p. 1127-33.

54. Osato, M., *Point mutations in the RUNX1/AML1 gene: another actor in RUNX leukemia.* Oncogene, 2004. **23**(24): p. 4284-96.

55. Watkins, J.A., et al., *Genomic scars as biomarkers of homologous recombination deficiency and drug response in breast and ovarian cancers.* Breast Cancer Res, 2014. **16**(3): p. 211.

56. Alexandrov, L.B., et al., *A mutational signature in gastric cancer suggests therapeutic strategies.* Nat Commun, 2015. **6**: p. 8683.

57. Ledermann, J., et al., *Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial.* Lancet Oncol, 2014. **15**(8): p. 852-61.

58. Verhagen, C.V., et al., *Extent of radiosensitization by the PARP inhibitor olaparib depends on its dose, the radiation dose and the integrity of the homologous recombination pathway of tumor cells.* Radiother Oncol, 2015. **116**(3): p. 358-65.

59. Kelly RJ, T.E., Zahurak. M, Cornish. T, Cuka. N, Abdelfatah. E, Taube. JM, Yang. S, Duncan. M, Ahuja. N, Murphy. A, Anders. RA, *Adaptive immune resistance in gastro-esophageal cancer: Correlating tumoral/stromal PDL1 expression with CD8+ cell count.* J Clin Oncol 2015. **33**((suppl; abstr 4031)).

60. Nakamura, H., et al., *Genomic spectra of biliary tract cancer.* Nat Genet, 2015. **47**(9): p. 1003-10.

61. Bridges, K.A., et al., *MK-1775, a novel Wee1 kinase inhibitor, radiosensitizes p53-defective human tumor cells.* Clin Cancer Res, 2011. **17**(17): p. 5638-48.

62. Wang, Y., et al., *Radiosensitization of p53 mutant cells by PD0166285, a novel G(2) checkpoint abrogator.* Cancer Res, 2001. **61**(22): p. 8211-7.

63. Liu, D.S., et al., *APR-246 potently inhibits tumour growth and overcomes chemoresistance in preclinical models of oesophageal adenocarcinoma.* Gut, 2015. **64**(10): p. 1506-16.

64. Stewart, A., et al., *Titration of signalling output: insights into clinical combinations of MEK and AKT inhibitors.* Annals of Oncology, 2015. **26**(7): p. 1504-10.

65. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

66. Saunders, C.T., et al., *Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs.* Bioinformatics, 2012. **28**(14): p. 1811-7.

67. McLaren, W., et al., *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.* Bioinformatics, 2010. **26**(16): p. 2069-70.

68. Van Loo, P., et al., *Allele-specific copy number analysis of tumors.* Proc Natl Acad Sci U S A, 2010. **107**(39): p. 16910-5.

69. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.

70. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-40.

71. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.* Bioinformatics, 2012. **28**(3): p. 423-5.

72. Chen, X., et al., *Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications.* Bioinformatics, 2016. **32**(8): p. 1220-2.

73. Schulte, I., et al., *Structural analysis of the genome of breast cancer cell line ZR-75-30 identifies twelve expressed fusion genes.* BMC Genomics, 2012. **13**: p. 719.

74. Le Tallec, B., et al., *Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes.* Cell Rep, 2013. **4**(3): p. 420-8.

75. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

76. Wilkerson, M.D. and D.N. Hayes, *ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.* Bioinformatics, 2010. **26**(12): p. 1572-3.

77. Nilsen, G., et al., *Copynumber: Efficient algorithms for single- and multi-track copy number segmentation.* BMC Genomics, 2012. **13**: p. 591.

78. Korbel, J.O. and P.J. Campbell, *Criteria for inference of chromothripsis in cancer genomes.* Cell, 2013. **152**(6): p. 1226-36.

79. Puente, X.S., et al., *Non-coding recurrent mutations in chronic lymphocytic leukaemia.* Nature, 2015. **526**(7574): p. 519-24.

80. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat Protoc, 2009. **4**(7): p. 1073-81.

81. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.

82. Lundegaard, C., et al., *NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W509-12.

83. Adiconis, X., et al., *Comparative analysis of RNA sequencing methods for degraded or low-input samples.* Nat Methods, 2013. **10**(7): p. 623-9.

Figure 1. Recurrent genomic events in the cohort (n = 129). The top panel highlights the total number of protein-coding genes affected by copy number or structural changes (above the 0 axis), and point mutations or indels (below the 0 axis), respectively, for every patient (depicted on the X-axis). (a) The top rearranged genes, excluding fragile sites, containing structural variant hotspots and recurrent in >10% of patients. *INK4/ARF comprises the *CDKN2A/2B* locus. 'Interchr trans' = interchoromosomal translocation. (b) Fragile sites rearranged in at least 20% of the patients. (c) Mobile element (ME) insertions detected by structural variant analysis, plotted on a log2 scale. Grey tiles correspond to cases without any evidence of ME insertions. (d) Loci that are significantly amplified/deleted according to GISTIC2.0 and that are recurrent in >10% of the patients. The most extreme copy number alteration within the locus is shown for each patient (see Supplementary Tables 2 and 3 for lists of genes in such loci). Only amplification and deletions are counted for the frequency histogram. (e) Genes altered by nonsynonymous SNVs/indels, deemed significantly mutated by MutSigCV. Loss of heterozygosity (LOH) regions are indicated in black rectangles when the gene also presents a mutation, indicating likely loss of function. (f) Presence of genomic catastrophes. (g) Cellularities, estimated by histopathology (H) or computationally using ASCAT (A). All samples sequenced have passed the histopathological cellularity cut-off of 70%. The total frequency of a specific gene alteration or event in the cohort is shown on the right-hand side for each panel.

Figure 2. RTK copy number profiling and responses to targeted RTK therapy (n=129). (a) RTK copy number gains/losses in the patient cohort and cell models. The score refers to: amplifications (2), homozygous deletions (–2), relative gains/losses (+1/–1) (Methods). Columns correspond to samples, ordered by the average ploidy. Samples with average ploidy ≥ 3 are highlighted as potentially whole-genome duplicated. (b) Copy number alterations in key genes of downstream pathways (c) Expression of RTKs and downstream key genes in samples with gains (light red) versus losses (light blue) of respective genes. The number of samples varies depending on the availability of cases with gain/loss (indicated in brackets). * marks p-values <0.05 after multiple testing correction. The solid horizontal line within the box represents the median. The interquartile range (IQR) is defined as Q3–Q1 with whiskers that extend 1.5 times the IQR from the box edges. (d) IHC staining of selected samples displaying consequences of copy number loss/gain in ERBB2 and MET. The GISTIC score (CN) is marked. (e) Breakdown of major resistance mechanisms to RTK-based monotherapy. "Amplification" denotes anything with a score ≥1. (f) Growth curve of OE33 cells after 72-hour exposure to Lapatinib, Crizotinib and in combination. Mean values as percentage of DMSO treated cells and ±SD for three experiments. Olaparib in combination was 1μM. (g) The effects of Lapatinib, Crizotinib and in combination on the cell lines with varying RTK status. Error bars represent the standard deviation. * indicates p-values <0.05.

Figure 3. Mutational signature-based clustering reveals differences in disease etiology in the cohort and is spatially consistent within a single tumor. (a) The heat map highlights the sample exposures to six main mutational signatures, as identified in the cohort (n=120) using the NMF methodology. The strength of exposure to a certain signature may vary from 0% to 100% (on a color scale from grey to red). Three main subgroups can be observed from the clustering based on the predominant signature: C>A/T dominant (S18-like/S1 age) – orange, 32% samples; DDR impaired (S3-BRCA) – purple, 15% samples; and mutagenic (S17A/B dominant) – green, 53% samples. The

TP53, *ERBB2* status, and catastrophic event distribution in the corresponding genomes are highlighted below (no significant difference observed among subgroups). The total mutational burden is significantly higher in the mutagenic subgroup. Consensus clustering was used for the heat map (Methods). b) Validation of the mutational signature-based clustering in an independent cohort (n=87). Unsupervised hierarchical clustering (Pearson correlation distance, Ward linkage method) reveals three main subgroups, similar to the ones in the discovery cohort: (1) DDR impaired (S3-BRCA) dominant – purple, 22% of the cohort; (2) C>A/T dominant (S18-like/S1 age) – orange, 25% of the cohort; (3) mutagenic (S17A/B dominant) – green, 53% of the cohort. The total SNV burden is also highlighted, confirming higher abundance in the mutagenic subgroup. c) Mutational signature contributions in three cases with multiple sampling from the same tumor. The relative exposures to the 6 signatures are highlighted on a grey-to-red gradient for each case. The group assignment is based on the dominant signature.

Figure 4. DNA damage repair pathways altered through nonsynonymous mutations/indels in the cohort. (a) For each of the three defined subgroups, the percentage of patients harboring defects in the different DDR-related pathways is shown. Only nonsynoymous mutations in genes mutated in the cohort significantly more compared to the expected background rate and predicted to be potentially damaging to the protein structure (Methods) have been considered in the analysis. (b) HR, CR and CPF genes altered in the three subgroups (the numbers in the gradients indicate how many patients have mutations in the respective gene). AM, alternative mechanism for telomere maintenance; BER, base excision repair; CPF, checkpoint factor; CR, chromatin remodelling; CS, chromosome segregation; FA, Fanconi anaemia pathway; HR, homologous recombination; MMR, mismatch repair; NER, nucleotide excision repair; NHEJ, non-homologous end joining; OD, other double-strand break repair; TLS, translesion synthesis; TM, telomere maintenance; UR, ubiquitylation response.

Figure 5. Neoantigen burden is significantly higher in the mutagenic subgroup and associates with an increased CD8+ T-cell density. (a) From left to right: Neoantigen burden compared among the 3 mutational signature subgroups shows significant differences. A two-sided Welch's t-test was used to compare the mutagenic group to the rest; Expression data available for a subset of the samples (25 from the mutagenic subgroup and 21 from the others) reveals that the number of expressed potential neoantigens is significantly higher in the mutagenic subgroup (Wilcoxon rank-sum test p = 0.042); Numbers of CD8+ T cells per mm^2 observed in patients. Patients were grouped into the mutagenic group and BRCA+C>A/T dominant group (n = 10 for each group). (b) Two representative images of CD8 IHC staining from each group (magnification 200x, scale bar, 100μm).

Figure 6. Treatment response in different mutational signature groups. (a) Three cell lines, OES127, MFD and CAM02 have been derived, each representative of a distinct signature-dominant subgroup: DDR impaired (OES127), mutagenic (MFD) and C>A/T dominant (CAM02). (b) Growth curves of OES127 cell lines after 72-hour exposure to Olaparib, Topotecan and in combination. Mean values as a percentage of DMSO treated cells and ±SD for three experiments are shown. Olaparib used in combination was kept at 1μM. (c) Growth curve of MFD cell lines after 72-hour exposure to MK-1775 and in

AZD-7762. Mean values as a percentage of DMSO treated cells and ±SD for three experiments are shown.

Figure 7. Proposed subclassification of EAC based on mutational signatures informs etiology and, consequently, potential tailored therapies to be further investigated for the disease. Patients are currently treated uniformly, but classification based on mutational signatures may enable targeted treatments that would complement classical therapy routes and potentially achieve more durable responses. The highlighted box (right) exemplifies classifying new patients into the defined etiological categories based on mutational signatures using a quadratic programming approach (see Methods). The bars highlight the relative contributions of the six expected signatures to the observed mutations in 7 new tumors (not part of the 129 sample cohort). The dominant signature is indicative of the group to which the sample should be assigned.

Table 1. *In vitro* cytotoxicity of RTKi as single or combined reagents in EAC cell lines. Key RTK amplification status and drug targets are shown. Bold text indicates that a synergistic effect of the combination treatment was observed.

Cell line	RTK status	RTKi	GI50 (95% CI) (nM)	AUC
OE33	*ERBB2*/*MET* Amp	Lapatinib (EGFR/ERBB2)	3.92 x10^3 (3.16–4.87 x10^3)	195.7
		Crizotinib (MET)	317.3 (166.3–605.4)	108.8
		Lapatinib + Crizotinib	**6.56 (2.42–17.84)**	**47.0**
SK-GT-4	*ERBB2* Amp/*MET* Gain	Lapatinib (EGFR/ERBB2)	3.72 x10^3 (2.27–6.08 x10^3)	173.9
		Crizotinib (MET)	3.47 x10^3 (2.90–4.15 x10^3)	183.2
		Lapatinib + Crizotinib	**530 (273.1–1029)**	**120.0**
OAC-P4C	*ERBB2*/*FGFR2* Amp	Lapatinib (EGFR/ERBB2)	2.28 x10^3 (1.34–3.90 x10^3)	159.1
		AZD-4547(FGFR1/2/3)	3.82 x10^3 (3.32–4.40 x10^3)	194.7
		Lapatinib + AZD-4547	**373.2 (260.9–533.7)**	**104.8**
FLO-1	*EGFR*/*MET* Gain	Lapatinib (EGFR/ERBB2)	11.64 x10^3 (7.80–17.39 x10^3)	212.0
		Crizotinib (MET)	1.90 x10^3 (1.51–2.39 x10^3)	159.3
		Lapatinib + Crizotinib	**243.4 (78.0–759.5)**	**109.0**
OES127	*ERBB2* Amp/*MET* Gain	Lapatinib (EGFR/ERBB2)	1.14 x10^3 (0.68–1.90 x10^3)	139.6
		Crizotinib (MET)	3.09 x10^3 (2.35–4.05 x10^3)	173.4
		Lapatinib + Crizotinib	**587.7 (450.5–766.7)**	**117.5**

ONLINE METHODS

Ethical approval, sample collection and DNA extraction

The study was registered (UKCRNID 8880), approved by the Institutional Ethics Committees (REC 07/H0305/52 and 10/H0305/1), and all subjects gave individual informed consent. Samples were obtained from surgical resection or by biopsy at endoscopic ultrasound. Blood or normal squamous esophageal samples at least 5 cm from the tumor were used as a germline reference. All tissue samples were snap frozen and before DNA extraction, a hematoxylin and eosin stained section was sent for cellularity review by two expert pathologists. Cancer samples with a cellularity ≥ 70% were submitted for whole-genome sequencing. DNA was extracted from frozen esophageal tissue using the AllPrep kit (Qiagen) and from blood samples using the QIAamp DNA Blood Maxi kit (Qiagen).

A total of 129 cases (matched tumor-normal) were sequenced. True esophageal and gastroesophageal (GOJ) type 1 and 2 tumors (according to Siewert classification) were used. All GOJ type 3 tumors (14 in total) were excluded from the analysis.

Whole-genome sequencing analysis

A single library was created for each sample, and 100-bp paired-end sequencing was performed under contracts by Illumina and the Broad Institute to a typical depth of at least 50x for tumors and 30x for matched normals, with 94% of the known genome being sequenced to at least 8x coverage and achieving a Phred quality of at least 30 for at least 80% of mapping bases. Read sequences were mapped to the human reference genome (GRCh37) using Burrows-Wheeler Alignment (BWA) 0.5.9 [65], and duplicates were marked and discarded using Picard 1.105 (see URLs). As part of an extensive quality assurance process, quality control metrics and alignment statistics were computed on a per-lane basis.

The FastQC package was used to assess the quality score distribution of the sequencing reads and perform trimming if necessary.

Samples were examined for potential microsatellite instability (MSI) using computational tools, and five cases with potential MSI were subsequently excluded from the analysis, as previously performed in other studies [16] (Supplementary Notes and Supplementary Table 10).

Somatic mutation and indel calling

Somatic mutations and indels were called using Strelka 1.0.13 [66]. SNVs were filtered as described in Supplementary Table 11. Functional annotation of the resulting variants was performed using Variant Effect Predictor (VEP release 75) [67].

Significantly mutated genes were identified using MutSigCV [23].

Copy number and loss of heterozygosity analysis

For patient-derived samples, absolute genome copy number after correction for estimated normal-cell contamination was called with ASCAT-NGS v2.1 [68], using read counts at germline heterozygous positions estimated by GATK 3.2-2 [69].

Cellularity, expressed as the relative proportion of tumor and normal nuclei, was also obtained using ASCAT. It was distributed as follows: 18% of samples had cellularity <0.3; 71% of samples between 0.3 and 0.7; 11% of samples ≥ 0.7.

Significantly amplified/deleted regions in the cohort were identified using GISTIC2.0 [22], after correcting the copy numbers for ploidy (total copy number of the segment divided by the average estimated ploidy of each sample). GISTIC2.0 was run on an input defined as the log2 of such corrected copy number values, with gain (-ta) and loss (-td) thresholds of 0.1 and sample centering prior to analysis. Copy number change thresholds considered for downstream analysis were: amplifications, GISTIC score ≥2; deletions, ≤–2. Loss of heterozygosity (LOH) was defined as ASCAT-estimated minor allele copy number of 0.

A whole-genome duplication event was considered to have occurred in a sample if the average estimated ploidy by ASCAT was ≥ 3, similar to the cut-offs suggested in [70].

For cell lines, copy number calling was performed using Control-FREEC [71].

RTK copy number profiling

To examine the landscape of copy number alterations in RTKs and downstream key genes (Fig. 2), a score from -2 to 2 was used to denote: deletions (-2), losses (-1), gains (+1), amplifications (+2). For the patient derived samples, copy numbers estimated using ASCAT were subsequently classified according to GISTIC2.0 using the same scoring scheme. For the cell models, a GISTIC-equivalent score was derived by dividing the estimated copy numbers by Control-FREEC by the average ploidy of each cell line, and classifying regions ≥2 as amplified (equivalent score = 2), regions ≤–2 as deleted (equivalent score = –2), and regions >1 or <1 as gained or lost, respectively (equivalent scores +1/-1). For the MFD line only the parent tumour was sequenced, so the copy numbers were inferred using ASCAT and GISTIC2.0 as described above.

In Figure 2b, the average copy number value of downstream key genes is highlighted for each representative gene (e.g. *RAS* summarizes the copy number landscape of *HRAS*, *KRAS*, *NRAS*), hence the scores take continuous rather than discrete values as in panel 2a.

Structural variant and mobile element insertion calling and annotation

Structural variants were called using BWA-mem for alignment (see URLs), against the GRCh37 reference human genome, followed by clustering of putative breakpoint junctions identified by discordant read pairs and split reads using Manta [72]. We then discarded: SVs overlapping gaps, satellite sequences, simple repeats >1000 basepairs or extreme read depth regions; and deletions of < 1000bp that were not supported by at least one split read defining the deletion junction. Small inversions up to 10 kb were also discarded as they are generated artefactually in some libraries [73]. Breakpoints in genes were annotated against Ensembl GRCh37, version 75 [18]. Fragile sites were annotated from Le Tallec et al [74], and potential additional sites to be excluded from gene recurrence analysis were determined as in Supplementary Table 12. Mobile element insertions and gene rearrangement hotspots were determined as described in the Supplementary Notes.

Structural variant-based classification of genomes

The structural variant-based classification was used to annotate unstable, stable, locally rearranged and scattered genomes as previously described [11], but with different cut-offs for stable and unstable genomes, to account for the different genomic instability landscape in EAC compared to pancreatic cancer: genomes were deemed “stable” if the total number of SVs was less than the 5% quantile in the cohort, and unstable if the number of SVs exceeded the 95% quantile. The criteria for locally rearranged and scattered genomes were as previously described.

Mutational signature analysis

Discovery

Mutational signatures were identified using the NMF methodology described by Alexandrov et al [27]. Before running the software, common variants in the 1000 genomes database [75] appearing in at least 0.5% of the population were removed, and samples with cellularity <25% (from ASCAT estimates) were not included, leaving a total of 120 samples for the analysis. The optimal number of signatures in the dataset was chosen to balance the signature stability against the Frobenius reconstruction error (Supplementary Figure 13). To increase confidence in the findings, two other methods were also used: the R packages pmsignature [28] and SomaticSignatures [29] (Supplementary Notes and Supplementary Figures 9–12).

To establish which of the two C[T>G]T signatures resembled most the classical S17 signature recorded in the COSMIC database, we used the cosine similarity distance measure between the probability vectors of these signatures. The signature which we termed S17A had a higher cosine similarity distance compared to S17B (0.98 versus 0.92), and we hence considered it to be more reflective of the signature reported in the literature.

Samples in the discovery cohort were clustered by their signature exposures using a consensus clustering approach [76] (based on Pearson correlation distance with complete linkage) in order to increase the robustness of the subgroup assignment.

Validation

The three mutational signature subgroups were validated in an independent cohort of 87 EAC samples (21 from [18] and 66 independent patients in our ICGC study post-neoadjuvant therapy and surgery). These had been selected from a slightly larger cohort after removing low cellularity and MSI positive samples. Within the validation cohort, the same dominant signatures were inferred using the NMF method, as above. The signature contributions were estimated based on the six main processes inferred in the test cohort using quadratic programming (described later in the Methods).

Multiple sampling

To test the differences in mutational exposures, we used three available cases for which multiple samples had been collected from the same tumour. We obtained the mutational exposures for the six described signatures using quadratic programming.

Structural variant signature analysis

Similar to inferring mutational signatures, we used the methodology by Alexandrov et al [27] to discover structural variant signatures in EAC genomes. We classified structural variants (deletions, inversions, insertions, interchromosomal translocations) by their size and distribution along the genome. SVs were grouped by size into “small” and “large”, defined with respect to the 25% quantile length in the cohort for the respective

SV type). To determine the SV distribution along the genome, we assessed the degree of clustering within 10 Mb windows along the genome. If the SV of interest fell within a window of clustered events (where the total number of SVs exceeded 1.5x the 75% quantile of the total number of events in that genome), then it was deemed a "focal" event. Otherwise, it was catalogued as "genomically distributed". These characteristics defined a total of 14 features to be used for signature discovery (Supplementary Figure 23).

Identification of catastrophic events

Kataegis was called in a similar manner to Nones et al [18], by calculating the distance between consecutive mutations and segmenting the resulting genome-wide signal using piecewise constant fitting as implemented in the *copynumber* Bioconductor package [77] (Supplementary Figure 5). However, acknowledging that the intermutational distance distribution varies from genome to genome, we did not use a fixed cutoff of 1000 bases for the mean distance between mutations in kataegis loci, but instead applied a variable cutoff that was determined as the 1% quantile of the intermutational distances within the respective genome.

Chromothripsis events were identified in chromosomes containing >10 CN steps, according to the criteria described by Korbel and Campbell [78] and Nones et al [18]: (a) clustering of breakpoints; (b) regularity of oscillating CN steps; (c) interspersed loss and retention of heterozygosity; (d) randomness of DNA segment order and fragment joins; (e) ability to walk the derivative chromosome. Scripts were developed to assess these criteria, and the final chromothripsis calls were prioritized through visual inspection (Supplementary Figure 6).

Regions of clustered inversions were identified as a proxy for BFB and complex rearrangement events. These were defined by scanning for enrichments of inversions (1.5x the upper quantile of the total number of events in the genome) within 5-Mb windows throughout the genome. Visual inspection was used to prioritize those regions that displayed BFB-like characteristics. Several types of complex rearrangement events were identified: focal amplifications with BFB pattern (clustered inversions along with progressive amplification steps primarily on one side of the inversion cluster, i.e. asymmetric); other focal amplifications within narrow regions <5 Mb (clustered inversions coupled with copy number amplifications displaying an irregular pattern), focal amplifications within wider 5–10 Mb regions (clustered inversions and progressive copy number amplification steps, often with multiple peaks); double minute-like patterns (clustered inversions at high copy number amplification regions without evidence of a progressive mechanism); potential subtelomeric BFBs (amplifications located close to the ends of the chromosomes, coupled with inversion clusters and distal deletions). See Supplementary Figure 7 for sample illustrations of the patterns described.

DNA damage repair (DDR) analysis

To assess the alterations in DNA damage-related pathways, we performed an analysis similar to the one described by Pearl et al [30]. Among the genes involved in defined DNA damage pathways as described in the paper, we only selected those affected more often than the expected background of synonymous mutations, similar to the method described by Puente et al [79]. The probability of a gene being affected by M nonsynonymous mutations in the cohort follows a poisson binomial distribution and is calculated relative to a basal probability depending on the number of nonsynonymous (n_{ns}) and synonymous (n_s) mutations, gene size (L), local mutational density for the locus (d) and total length of coding regions in the genome (E) as follows: $P_{ns} = \frac{n_{ns}Ld}{(n_{ns}+n_s)E}$ Subsequently, we catalogued those that harboured nonsynonymous somatic mutations/indels with possible deleterious effect (as predicted by SIFT [80]/PolyPhen [81]) or copy number alterations (amplifications and deletions using the defined GISTIC cut-offs) in our cohort. We then compared the mutational load in 16 main pathways among the defined mutational signature subgroups.

Neoantigen predictions and analysis

In order to quantify the neoantigen load in the tumors, we performed the analysis as described in [35]. We first collected all peptides defined by a 17 amino-acid region centered on the amino acid which changes upon the mutation. We identified mutant nonamers with ≤500 nM binding affinity for patient-specific class I human lymphocyte antigen (HLA) alleles, constituting potential candidate neoantigens. Binding affinities were predicted using NetMHC-3.4 [82]. We then quantified the peptides that displayed high affinity binding in tumor, but low or no binding in the respective matched normal and obtained total counts for each defined mutational subgroup. The neoantigen burden in tumours belonging to the different subgroups varied as follows: DDR impaired - an average of 77 (s.d. = 42.2); C>A/T dominant - an average of 86 (s.d. = 41.3); mutagenic - an average of 111 (s.d. = 43.9). The three groups presented unequal variance in terms of nonsynonymous mutation burden, as shown by pairwise F-tests (p<0.05 after multiple testing correction using the Benjamini-Hochberg method). To adjust for this, the mutation burden among subgroups was compared using Welch's t-test. The neoantigen load, on the other hand, had similar variance between the mutagenic group and the other two groups combined (F-test p>0.05), so the Wilcoxon rank-sum test was used to compare the predicted neoantigen presence in tumors.

To verify that the predicted neoantigens were indeed expressed in the samples, expression Z-scores were investigated and all peptides with a score higher than the average in the respective sample were considered expressed.

Expression profiling

Purified Total RNA was extracted using the AllPrep DNA/RNA Mini Kit from Qiagen. Quality of RNA was assessed using the NanoDrop and the Agilent Bioanalyser, and only samples with RIN>7 were accepted. The Illumina HTv4.0 beadchip was used as platform for expression analysis. Bead level readings were corrected for spatial artefacts and the signal per probe ratio was computed. Relative array weights were applied before quantile normalization for gene expression analysis.

For sequencing, purified total RNA was subject to ribosomal depletion using methods already published [83]. In brief, 195 DNA oligonucleotides (Sigma Life Sciences) were pooled together in equal molar amounts and incubated with total RNA Hybridase Thermostable RNase H (Epicentre). RNaseH-treated RNA was purified using 2.2x RNAClean SPRI beads (Beckman Coulter LifeSciences) and oligonucleotides removed using TURBO DNase rigorous treatment. A further purification of the DNase-treated RNA with 2.2x RNAClean SPRI beads was followed by library preparation using the TruSeq HT Stranded mRNA kit according to the manufacturers instructions (Illumina) and generated single end reads using the HiSeq 2500.

For the validation of RTK gains/losses and neoantigen expression, available expression data for a total of 42 samples were used. To evaluate expression levels for selected genes, Z-scores were obtained relative to the average expression in the sample or of the specific investigated gene.

For the validation of neoantigen expression, available RNA-Seq data for a total of 18 samples were used. To evaluate expression levels for selected genes, Z-scores were obtained relative to the average expression in the sample.

Cell lines and reagents

The primary cell line panel was derived from EAC cases included in the ICGC sequencing study , including MFD (Tim Underwood, Southampton, OCCAMS consortium member), OES127 (Anna Grabowska, Nottingham, OCCAMS consortium member) and CAM02 (organoid, Mathew Garnett, Cambridge). The MFD line required 10% fetal calf serum (PAA) in DMEM medium (Invitrogen, ThermoFisher Scientific) and the CAM02 culture method was as previously described [51]. The feeder layer system was used to expand OES127 lines. The established EAC lines, SK-GT-4, OAC-P4C, OACM5.1C, and OE33 were cultured in RPMI medium (Sigma) with 10% fetal calf serum, except for FLO-1, which was grown in DMEM with 10% fetal calf serum. The identity of all cell lines was verified by short tandem repeat (STR) profiling and routinely examined for mycoplasma contamination.

Small molecular inhibitors used for treatment were: Lapatinib, AZD-4547, Olaparib, MK-1775 and AZD-7762 (BioVision), Crizotinib (LKT Labs) and Topotecan (Cayman Chemical). Inhibitors were diluted to working concentrations in DMSO (Sigma).

Immunohistochemistry

Sections of 3.5μm were stained by a Bond Max autostainer according to the manufacturer's instruction (Leica Microsystems). Primary antibodies ERBB2 (#2165, 1:300, Cell Signaling Technology), MET (#8198, 1:300, Cell Signaling Technology), CD8 (#M7103, 1:100, Dako) were optimised and applied with negative controls.

CD8+ cells were counted manually in two tumour areas of 1 mm^2 each (except in one case where there was sufficient material for one count only) and an average was calculated.

Drug sensitivity assays

The seeding density for each line was optimised to ensure cell growth in the logarithmic growth phase. Cells were seeded in complete medium for 24 hours then treated with compounds at 4-fold serial dilutions for 72 hours. Cell proliferation was assessed using CellTiter-Glo (Promega). The anchor inhibitors were kept constant at 1M in combination studies.

The concentrations of a compound causing 50% growth inhibition relative to the vehicle control (GI50) were determined by nonlinear regression dose-response analysis and the area under the curve (AUC) was calculated using GraphPad Prism.

Statistics

All statistical tests were performed using a Wilcoxon rank-sum test or ANOVA (for continuous data), and a Fisher exact test or Chi-square test (for count data). Welch's t-test was used when comparing groups of unequal variance. Multiple testing corrections were performed where necessary using the Benjamini-Hochberg method. All reported p-values were two-sided.

Code availability

The scripts used to perform the analysis are available upon request.

Methods-only references

65. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
66. Saunders, C.T., et al., *Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs.* Bioinformatics, 2012. **28**(14): p. 1811-7.
67. McLaren, W., et al., *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.* Bioinformatics, 2010. **26**(16): p. 2069-70.
68. Van Loo, P., et al., *Allele-specific copy number analysis of tumors.* Proc Natl Acad Sci U S A, 2010. **107**(39): p. 16910-5.
69. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
70. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-40.
71. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.* Bioinformatics, 2012. **28**(3): p. 423-5.
72. Chen, X., et al., *Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications.* Bioinformatics, 2016. **32**(8): p. 1220-2.
73. Schulte, I., et al., *Structural analysis of the genome of breast cancer cell line ZR-75-30 identifies twelve expressed fusion genes.* BMC Genomics, 2012. **13**: p. 719.
74. Le Tallec, B., et al., *Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes.* Cell Rep, 2013. **4**(3): p. 420-8.
75. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
76. Wilkerson, M.D. and D.N. Hayes, *ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.* Bioinformatics, 2010. **26**(12): p. 1572-3.

77. Nilsen, G., et al., *Copynumber: Efficient algorithms for single- and multi-track copy number segmentation.* BMC Genomics, 2012. **13**: p. 591.

78. Korbel, J.O. and P.J. Campbell, *Criteria for inference of chromothripsis in cancer genomes.* Cell, 2013. **152**(6): p. 1226-36.

79. Puente, X.S., et al., *Non-coding recurrent mutations in chronic lymphocytic leukaemia.* Nature, 2015. **526**(7574): p. 519-24.

80. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat Protoc, 2009. **4**(7): p. 1073-81.

81. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.

82. Lundegaard, C., et al., *NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W509-12.

83. Adiconis, X., et al., *Comparative analysis of RNA sequencing methods for degraded or low-input samples.* Nat Methods, 2013. **10**(7): p. 623-9