

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

Maria Secrier^{1, 11}, Xiaodun Li^{2, 11}, Nadeera de Silva², Matthew D. Eldridge¹, Gianmarco Contino², Jan Bornschein², Shona MacRae², Nicola Grehan², Maria O'Donovan², Ahmad Miremadi³, Tsun-Po Yang², Lawrence Bower¹, Hamza Chettouh², Jason Crawte², N ria Galeano-Dalmau², Anna Grabowska⁴, John Saunders⁵, Tim Underwood⁶, Nicola Waddell⁷, Andrew P. Barbour^{8, 9}, Barbara Nutzinger², Achilleas Achilleos¹, Paul A. W. Edwards¹⁰, Andy G. Lynch¹, Simon Tavar ¹, Rebecca C. Fitzgerald^{2, 12} on behalf of the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium¹³.

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, Addenbrooke's Hospital, 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 Faculty of Medicine, University of Southampton, Southampton General Hospital, 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 These authors contributed equally

¹² Corresponding author: rcf29@mrc-cu.cam.ac.uk

¹³ A full list of contributors from the OCCAMS Consortium is available at the end of the manuscript

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

53 males in the US over the past three decades [2]. It is an aggressive disease, with early
54 loco-regional spread, resulting in a median overall survival of less than a year [3].

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

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

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

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

81 82 **RESULTS**

83 84 **Large-scale alterations dominate the EAC landscape**

85
86 WGS data from 129 EAC patients (including tumors from the gastroesophageal junction,
87 Siewert type 1 and 2) have allowed us to comprehensively catalog the genomic
88 alterations in this cancer, including the large-scale structural rearrangements not
89 detectable from exome sequencing. The clinical characteristics of the cohort are typical
90 for the disease (Supplementary Table 1).

91 As previously noted, point mutations are abundant in this cancer [16]. However, the
92 overall genomic landscape suggests a disease driven by structural variation and copy
93 number changes (Fig. 1 and Supplementary Figure 1). Analysis of a combined cohort of
94 111 EAC cases from TCGA [19] and Nones et al [18] confirms a dominance of copy
95 number alterations, compared to point mutations, in the majority of cases
96 (Supplementary Figure 2).

97 When examining the specific loci affected, potential gene driver events were highly
98 heterogeneous between cases, and structural changes again dominated (Fig. 1). Among
99 the genes altered in 10% or more of cases, many more were rearranged, amplified or
100 deleted than were affected by indels or nonsynonymous point mutations. We observed
101 novel recurrently rearranged genes, including *SMYD3* in 39% of cases, *RUNX1* 27%,

102 *CTNNA3* 22%, *RBFOX1* 21%, the *CDKN2A/2B* locus 18%, *CDK14* 16% (important
103 transcriptional, signalling and cell communication regulators), and fragile sites (*FHIT*
104 95%, *WWOX* 84%). Somatic L1 mobile element insertions were also abundant. Detecting
105 inserts that had transduced unique flanking sequences identified an average of 25
106 inserts/tumor (range 0–1127), including those already known to transduce [20, 21] and
107 novel examples. These numbers are substantially higher than previously reported [20]
108 because of improved sensitivity. Mobile element insertions were found in signalling, cell
109 cycle and cell adhesion regulators: *ERBB4* - 6/129, - 5/129, *CTNNA2* - 4/129, *CDH18* -
110 3/129, *SOX5* - 2/129.

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

116 The most frequent somatic mutation/indel events included a number of known driver
117 genes with roles in DNA damage, signal transduction, cell cycle and chromatin
118 remodelling. Seven of these reached statistical significance as likely driver genes, as
119 inferred by MutSigCV [23] (Fig. 1e and Supplementary Table 4): *TP53* (81%), *ARID1A*
120 (17%), *SMAD4* (16%), *CDKN2A* (15%), *KCNQ3* (12%), *CCDC102B* (9%), *CYP7B1* (7%),
121 largely as previously described [16, 17]. In addition *SYNE1* was mutated in 23% of cases,
122 but did not reach significance by MutSigCV.

123 The high frequency of genomic catastrophes observed was consistent with a
124 significant role of larger-scale events in this disease - chromothripsis: 39/129 patients
125 (30%), kataegis: 40/129 (31%), complex rearrangement events: 41/129 (32%),
126 (Methods, Figure 1f and Supplementary Figures 4–7). The complex rearrangements
127 included: focal amplifications with BFB pattern (11/129, 9%); focal amplifications
128 <5Mb-wide with irregular copy number amplification steps (26/129, 20%); focal
129 amplifications 5–10 Mb-wide with symmetric copy number amplification steps (10/129,
130 8%); double minute-like patterns (3/129, 2%); and subtelomeric BFBs (1/129, 1%)
131 (Supplementary Figure 7). The chromothripsis and BFB/complex rearrangement event
132 frequencies were in a similar range to that described by Nones et al [18] - 33% and 27%,
133 respectively. Kataegis rates were lower than that previously reported (19/22 = 86%),
134 likely due to our more stringent criteria for calling (Methods). An enrichment of C>T and
135 C>G mutations was observed in kataegis regions, as previously reported [24]
136 (Supplementary Figure 5).

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

140
141

142 **RTK receptors and their targets are pervasively disrupted in EAC**

143

144 Next we examined the genomic data to understand possible reasons for the
145 disappointing results seen with many of the trials targeting growth factor receptors.
146 Resistance to RTK therapy generally results from co-amplifications of alternative RTKs
147 or amplification/activation of downstream mitogenic pathways. In our cohort we
148 observed widespread gene amplification across multiple RTKs, as well as downstream
149 within the MAPK and PI3K pathways. Such patterns were similar among

150 endoreduplicated and non-endoreduplicated samples, as well as in a panel of cell
151 models (Fig. 2a, 2b).

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

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

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

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

186

187 **Mutational signatures uncover distinct etiology in EAC**

188

189 In view of the heterogeneity and RTK-resistance mechanisms, we sought alternative
190 therapeutic insights into the data using mutational signature analysis in a three-base
191 context via the non-negative matrix factorization (NMF) methodology described by
192 Alexandrov et al [27]. We also used the recently described pmsignature [28] and
193 SomaticSignatures [29] for comparison. These methods are based on different statistical
194 frameworks and therefore some differences are to be expected; nevertheless the same
195 key signature patterns were observed with similar-sized patient subgroups expressing
196 the dominant signature types (Supplementary Notes, Supplementary Figures 8–12). Six
197 signatures were prominent (Supplementary Figures 13–14): S17, the hallmark signature
198 of EAC [16, 17] dominated by T>G substitutions in a CTT context and possibly associated

199 with gastric acid reflux – here renamed S17A; a previously uncharacterized variant of
200 this signature combining a relatively higher frequency of T>C substitutions with the
201 classical T>G pattern found in S17, which we call S17B; S3, a complex pattern caused by
202 defects in the BRCA1/2-led homologous recombination pathway; S2, C>T mutations in a
203 TCA/TCT context, an APOBEC-driven hypermutated phenotype; S1, C>T in a *CG context,
204 associated with aging processes; and an S18-like signature, C>A/T dominant in a
205 GCA/TCT context, formerly described in neuroblastoma, breast and stomach cancers
206 (Fig. 3a). The exploration of a seven-base signature context using pmsignature yielded
207 an A/T base dominance at the -3 and -2 positions for the S17 signature, but no other
208 striking preferences for nucleotide combinations at the 2nd and 3rd bases for any of the
209 other signatures (Supplementary Figure 15). Overall, this suggests that the bases
210 immediately adjacent to the position where the mutation occurs exert the main bias,
211 with a potentially more complex mechanism for the S17 signature.

212 When considering the dominant mutation signatures on a per-patient basis, three
213 subgroups of patients became apparent: *C>A/T dominant* (age, S18-like), *DNA Damage*
214 *Repair (DDR) impaired* (BRCA), and *mutagenic* (predominantly S17A or S17B) (Fig. 3a).
215 We chose the descriptor *mutagenic* because the mutation rate was significantly higher in
216 this subgroup (Welch's t-test $p = 0.0007$; Supplementary Figure 16). The robustness of
217 the subgroups was ensured through consensus clustering and confirmed by silhouette
218 statistics (Methods, Supplementary Figures 17–18). We also validated our findings in an
219 independent cohort of 87 samples [18] and show that: when we apply the NMF method
220 the same dominant signatures (S1, S2, S3, S17, S18-like) are observed; and when we
221 perform clustering three subgroups emerge which are of similar composition and
222 proportions to those seen in the original cohort (Methods, Fig. 3b compared with Fig.
223 3a). Furthermore, the total mutational burden is again consistently higher in the
224 mutagenic subgroup of the validation cohort. No cellularity bias or batch effect was
225 observed among subgroups (Supplementary Figure 19).

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

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

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

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

254

255 **Evidence of DNA damage repair deficiency in EAC**

256

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

268 The analysis of DDR genes in the whole cohort unsurprisingly showed that the most
269 mutated pathway was *TP53* (Supplementary Figure 25), and this was consistent among
270 subgroups (Fig. 4a), as were the amplification and deletion patterns (Supplementary
271 Figure 26). In addition, more than 24% of the genomes had defects in chromatin
272 remodelling, comprising recurrently mutated genes like *ARID1A* (8%) and *SMARCA4*
273 (8%) (Fig. 4b). *ARID1A* is also recruited to DNA double strand breaks (DSB), where it
274 facilitates processing to single strand ends [31]. Defects in *ARID1A* impair this process
275 and may sensitise cells *in vitro* and *in vivo* to PARP inhibition (PARPi) [31].

276

277

278 **Neoantigen and CD8 profiles in the mutagenic subgroup**

279

280 Modulation of the cytotoxic T cell response using monoclonal antibodies against the
281 Programmed Death Receptor or Ligand (PD-1 and PD-L1 inhibitors), as well as those
282 targeting CTLA4 (Ipilimumab) have shown promise in the treatment of solid tumors
283 [32–34]. The recent literature suggests that both numbers of mutations and total
284 neoantigen burden have been coupled with significantly better clinical responses to
285 immunotherapy [35–37].

286 We found that the mutagenic subgroup, whose observed signature may be due to
287 gastric acid reflux, harbored a significantly higher nonsynonymous mutational burden,
288 as well as higher levels of neoantigen presentation (Welch's t-test $p = 0.0007$ and
289 Wilcoxon rank-sum test $p \ll 0.0001$, respectively; Fig 5a and Supplementary Figure 16).
290 This is in keeping with that observed for lung cancer and metastatic melanoma, with a
291 1.5-fold higher median neoantigen burden in this subgroup versus the rest – similar to
292 the two-fold ratio reported by Rizvi et al [35, 38]. Using available RNA expression data
293 we observed a significantly higher number of neoantigens expressed in this subgroup
294 compared to the rest (Wilcoxon rank-sum test p -value = 0.042, Fig. 5a).

295 In recent studies, an enriched population of pre-existing CD8+ T cells was shown to
296 predict a favorable outcome from PD-1 blockade therapy [39, 40]. We found a higher

297 density of CD8+ T cells in a subset of available samples from the mutagenic signature
298 subgroup compared with samples from the other subgroups (Fig. 5a, 5b).

299

300 **Treatment responses in mutational signature subgroups**

301

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

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

329 These experimental data provide a starting point from which to evaluate therapeutic
330 options derived from mutational signatures, especially as primary model systems more
331 closely resembling human disease and with stromal components become available [51,
332 52].

333

334

335

336 **DISCUSSION**

337

338 Whole-genome sequencing of 129 EAC patients has unveiled a high prevalence of large-
339 scale alterations that may play an important role in the development of this cancer.
340 Similarly to ovarian, breast and lung cancers which have been described as ‘copy
341 number driven’ [53], relatively few genes were recurrently point-mutated (except
342 *TP53*), but there were frequent recurrent amplifications in sites harbouring oncogenes,
343 deletions of important cell cycle components (*CDKN2A*, *CDKN2B*) and rearrangements of
344 genes like *RUNX1*, frequently translocated in leukemias [54]. The highly heterogeneous
345 landscape explains the difficulties encountered to date in finding suitable avenues for

346 tailored therapies. 88/262 registered esophageal trials (see URLs) target RTKs and
347 mitogenic signalling pathways with remarkably little clinical efficacy. The genomic and
348 *in vitro* analyses performed here suggest that the high prevalence of co-amplification of
349 RTKs and downstream mitogenic pathway genes is likely to explain these disappointing
350 results.

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

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

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

380 Patients in the C>A/T dominant subgroup would continue to be treated with
381 conventional chemotherapy until more progress is made, e.g. with synthetic lethality
382 approaches combined with radiotherapy or mutant TP53 reactivating drugs [61-63].
383 Alternatively, combined RTK inhibitors (especially ERBB2 and MET, given their
384 prevalence in this subgroup) may be beneficial and combined MEK and Akt inhibition
385 might be worthy of consideration given the low levels of amplifications/activation seen
386 downstream in the MAPK and PI3K pathways [64].

387 One practical question that arises is how this approach could be implemented
388 clinically. Despite the decreasing costs of WGS, it is still expensive and signatures are
389 problematic to derive from whole-exome data [27]. However, lower coverage whole-
390 genome (10x), or even shallow (1x) genome sequencing could provide a cost-effective,
391 high-throughput alternative for signature-based stratification and we have shown using
392 simulations down to 10x that we can confidently retrieve dominant signatures at lower
393 coverage (Supplementary Figure 28). Moreover, while designing custom gene panels
394 would pose serious difficulties in such a heterogeneous disease, mutational signature-

395 based classification would enable us to bypass the tumor heterogeneity bottleneck by
396 providing a genome-wide, spatially-independent classification strategy (Fig. 3c).

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

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

417
418

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

420 Ayesha Noorani², Rachael Fels Elliott², Jamie Weaver², Laura Smith², Zarah Abdullahi²,
421 Rachel de la Rue², Alison Cluroe³, Shalini Malhotra³, Richard Hardwick¹⁴, Hugo
422 Ford¹⁴, Mike L Smith¹, Jim Davies¹⁵, Richard Turkington¹⁶, Stephen J. Hayes^{17,18}, Yeng
423 Ang^{17,19,20}, Shaun R. Preston²¹, Sarah Oakes²¹, Izhar Bagwan²¹, Vicki Save²², Richard J.E.
424 Skipworth²², Ted R. Hupp²², J. Robert O'Neill^{22,23}, Olga Tucker^{24,25}, Philippe Taniere²⁴,
425 Fergus Noble²⁶, Jack Owsley²⁶, Laurence Lovat²⁷, Rehan Haidry²⁷, Victor Eneh²⁷, Charles
426 Crichton²⁸, Hugh Barr²⁹, Neil Shepherd²⁹, Oliver Old²⁹, Jesper Lagergren^{30,31,32}, James
427 Gossage^{30,31}, Andrew Davies^{30,31}, Fujun Chang^{30,31}, Janine Zylstra^{30,31}, Grant Sanders³³,
428 Richard Berrisford³³, Catherine Harden³³, David Bunting³³, Mike Lewis³⁴, Ed Cheong³⁴,
429 Bhaskar Kumar³⁴, Simon L Parsons⁵, Irshad Soomro⁵, Philip Kaye⁵, Pamela Collier⁵,
430 Laszlo Igali³⁵, Ian Welch³⁶, Michael Scott³⁶, Shamila Sothi³⁷, Sari Suortamo³⁷, Suzy
431 Lishman³⁸, Duncan Beardsmore³⁹, Hayley E. Francies⁴⁰, Mathew J. Garnett⁴⁰, John V.
432 Pearson^{7,41}, Katia Nones^{7,41}, Ann-Marie Patch^{7,41}, Sean M. Grimmond^{41,42}

433 ¹⁴Oesophago-Gastric Unit, Addenbrooke's Hospital, Cambridge, UK

434 ¹⁵Oxford ComLab, University of Oxford, UK

435 ¹⁶Centre for Cancer Research and Cell Biology, Queen's University Belfast, Northern Ireland, UK

436 ¹⁷Salford Royal NHS Foundation Trust, Salford, UK

437 ¹⁸Faculty of Medical and Human Sciences, University of Manchester, UK

438 ¹⁹Wigan and Leigh NHS Foundation Trust, Wigan, Manchester, UK

439 ²⁰GI science centre, University of Manchester, UK

440 ²¹Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK

441 ²²Edinburgh Royal Infirmary, Edinburgh, UK

442 ²³Edinburgh University, Edinburgh, UK

443 ²⁴University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

444 ²⁵Institute of Cancer and Genomic Sciences, University of Birmingham

445 ²⁶University Hospital Southampton NHS Foundation Trust, Southampton, UK

446 ²⁷University College London, London, UK

447 ²⁸Department of Computer Science, University of Oxford, UK

448 ²⁹Gloucester Royal Hospital, Gloucester, UK

449 ³⁰St Thomas's Hospital, London, UK

450 ³¹King's College London, London, UK

451 ³²Karolinska Institutet, Stockholm, Sweden

452 ³³Plymouth Hospitals NHS Trust, Plymouth, UK

453 ³⁴Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, UK

454 ³⁵Norfolk and Waveney Cellular Pathology Network, Norwich, UK

455 ³⁶Wythenshawe Hospital, Manchester, UK

456 ³⁷University Hospitals Coventry and Warwickshire NHS, Trust, Coventry, UK

457 ³⁸Peterborough Hospitals NHS Trust, Peterborough City Hospital, Peterborough, UK

458 ³⁹Royal Stoke University Hospital, UHNM NHS Trust, UK

459 ⁴⁰Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

460 ⁴¹Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland,
461 Queensland, Australia

462 ⁴²Victorian Comprehensive Cancer Centre, University of Melbourne, Melbourne, Australia

463

464

465

466

467 **URLs**

468 UKCRN Trial Portfolio [cited 2015 22/11/15]: <http://public.ukcrn.org.uk/search/>
469 US National Institutes of Health trial registry [cited 22/06/15]:
470 <https://clinicaltrials.gov/>
471 Picard 1.105: <http://broadinstitute.github.io/picard>
472 BWA-MEM: <http://arxiv.org/abs/1303.3997>

473

474 **Accession codes**

475

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

479

480 **ACKNOWLEDGEMENTS**

481

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

486

487 Whole-genome sequencing of esophageal adenocarcinoma samples was performed as
488 part of the International Cancer Genome Consortium (ICGC) through the Oesophageal
489 Cancer Clinical and Molecular Stratification (OCCAMS) Consortium and was funded by a
490 programme grant from Cancer Research UK. We thank the ICGC members for their input
491 on verification standards as part of the benchmarking exercise. We thank the Human
492 Research Tissue Bank, which is supported by the National Institute for Health Research
493 (NIHR) Cambridge Biomedical Research Centre, from Addenbrooke's Hospital and UCL.
494 Also the University Hospital of Southampton Trust and the Southampton, Birmingham,
495 Edinburgh and UCL Experimental Cancer Medicine Centres and the QEHB charities. This
496 study was partly funded by a project grant from Cancer Research UK. R.C.F. is funded by
497 an NIHR Professorship and receives core funding from the Medical Research Council and
498 infrastructure support from the Biomedical Research Centre and the Experimental
499 Cancer Medicine Centre. We acknowledge the support of The University of Cambridge,
500 Cancer Research UK (C14303/A17197) and Hutchison Whampoa Limited. We would
501 like to thank Dr. Peter Van Loo for providing the NGS version of ASCAT for copy number
502 calling. We are grateful to all the patients who provided written consent for
503 participation in this study and the staff at all participating centres.

504

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

510

511 **Author Contributions**

512

513 R.C.F. conceived the overall study. M.S., X.L. and P.A.W.E. analysed the data. R.C.F., M.S.,
514 X.L., N.S., P.A.W.E. and A.G.L. conceived and designed the experiments. M.S. performed
515 the statistical analysis. X.L., G.C., S.M., M.O., A.M., J.C. and N.G.D. performed the

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

526
527

528

529 **COMPETING FINANCIAL INTERESTS**

530

531 The authors declare no competing financial interests.

532

533

534

References

535

536 1. Ferlay, J., et al., *Cancer incidence and mortality worldwide: sources, methods and*

537 *major patterns in GLOBOCAN 2012*. International Journal of Cancer, 2015. **136**(5):
538 p. E359-86.

539 2. Brown, L.M., S.S. Devesa, and W.H. Chow, *Incidence of adenocarcinoma of the*
540 *esophagus among white Americans by sex, stage, and age*. J Natl Cancer Inst, 2008.
541 **100**(16): p. 1184-7.

542 3. Cunningham, D., A.F. Okines, and S. Ashley, *Capecitabine and oxaliplatin for*
543 *advanced esophagogastric cancer*. N Engl J Med, 2010. **362**(9): p. 858-9.

544 4. Allum, W.H., et al., *Long-term results of a randomized trial of surgery with or*
545 *without preoperative chemotherapy in esophageal cancer*. J Clin Oncol, 2009.
546 **27**(30): p. 5062-7.

547 5. Cunningham, D., et al., *Perioperative chemotherapy versus surgery alone for*
548 *resectable gastroesophageal cancer*. N Engl J Med, 2006. **355**(1): p. 11-20.

549 6. van Hagen, P., et al., *Preoperative chemoradiotherapy for esophageal or junctional*
550 *cancer*. N Engl J Med, 2012. **366**(22): p. 2074-84.

551 7. Bang, Y.J., et al., *Trastuzumab in combination with chemotherapy versus*
552 *chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-*
553 *oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled*
554 *trial*. Lancet, 2010. **376**(9742): p. 687-97.

555 8. Gao, Y.B., et al., *Genetic landscape of esophageal squamous cell carcinoma*. Nat
556 Genet, 2014. **46**(10): p. 1097-102.

557 9. Schulze, K., et al., *Exome sequencing of hepatocellular carcinomas identifies new*
558 *mutational signatures and potential therapeutic targets*. Nat Genet, 2015. **47**(5): p.
559 505-11.

560 10. *Genomic Classification of Cutaneous Melanoma*. Cell, 2015. **161**(7): p. 1681-96.

561 11. Waddell, N., et al., *Whole genomes redefine the mutational landscape of pancreatic*
562 *cancer*. Nature, 2015. **518**(7540): p. 495-501.

563 12. Totoki, Y., et al., *Trans-ancestry mutational landscape of hepatocellular carcinoma*
564 *genomes*. Nat Genet, 2014. **46**(12): p. 1267-73.

- 565 13. *Comprehensive molecular characterization of gastric adenocarcinoma*. Nature, 2014. **513**(7517): p. 202-9.
- 566
- 567 14. *Comprehensive molecular profiling of lung adenocarcinoma*. Nature, 2014. **511**(7511): p. 543-50.
- 568
- 569 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.
- 570
- 571
- 572 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.
- 573
- 574
- 575 17. Weaver, J.M., et al., *Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis*. Nat Genet, 2014. **46**(8): p. 837-43.
- 576
- 577 18. Nones, K., et al., *Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis*. Nat Commun, 2014. **5**: p. 5224.
- 578
- 579 19. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project*. Nat Genet, 2013. **45**(10): p. 1113-20.
- 580
- 581 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.
- 582
- 583
- 584 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.
- 585
- 586
- 587 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.
- 588
- 589
- 590 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.
- 591
- 592 24. Nik-Zainal, S., et al., *The life history of 21 breast cancers*. Cell, 2012. **149**(5): p. 994-1007.
- 593
- 594 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.
- 595
- 596
- 597 26. Van Cutsem, E., et al., *HER2 screening data from ToGA: targeting HER2 in gastric and gastroesophageal junction cancer*. Gastric Cancer, 2014.
- 598
- 599 27. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer*. Nature, 2013. **500**(7463): p. 415-21.
- 600
- 601 28. Shiraishi, Y., et al., *A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures*. PLoS Genet, 2015. **11**(12): p. e1005657.
- 602
- 603 29. Gehring, J.S., et al., *SomaticSignatures: inferring mutational signatures from single-nucleotide variants*. Bioinformatics, 2015. **31**(22): p. 3673-5.
- 604
- 605 30. Pearl, L.H., et al., *Therapeutic opportunities within the DNA damage response*. Nat Rev Cancer, 2015. **15**(3): p. 166-80.
- 606
- 607 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.
- 608
- 609 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.
- 610
- 611 33. Larkin, J., et al., *Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma*. N Engl J Med, 2015. **373**(1): p. 23-34.
- 612

- 613 34. Herbst, R.S., et al., *Pembrolizumab versus docetaxel for previously treated, PD-L1-*
614 *positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised*
615 *controlled trial*. Lancet, 2015.
- 616 35. Rizvi, N.A., et al., *Cancer immunology. Mutational landscape determines sensitivity*
617 *to PD-1 blockade in non-small cell lung cancer*. Science, 2015. **348**(6230): p. 124-8.
- 618 36. Snyder, A., et al., *Genetic basis for clinical response to CTLA-4 blockade in*
619 *melanoma*. N Engl J Med, 2014. **371**(23): p. 2189-99.
- 620 37. McGranahan, N., et al., *Clonal neoantigens elicit T cell immunoreactivity and*
621 *sensitivity to immune checkpoint blockade*. Science, 2016. **351**(6280): p. 1463-9.
- 622 38. Van Allen, E.M., et al., *Genomic correlates of response to CTLA-4 blockade in*
623 *metastatic melanoma*. Science, 2015. **350**(6257): p. 207-11.
- 624 39. Tumeu, P.C., et al., *PD-1 blockade induces responses by inhibiting adaptive immune*
625 *resistance*. Nature, 2014. **515**(7528): p. 568-71.
- 626 40. Hamanishi, J., et al., *Programmed cell death 1 ligand 1 and tumor-infiltrating CD8+*
627 *T lymphocytes are prognostic factors of human ovarian cancer*. Proc Natl Acad Sci
628 U S A, 2007. **104**(9): p. 3360-5.
- 629 41. Benafif, S. and M. Hall, *An update on PARP inhibitors for the treatment of cancer*.
630 Onco Targets Ther, 2015. **8**: p. 519-28.
- 631 42. Oza, A.M., et al., *Olaparib combined with chemotherapy for recurrent platinum-*
632 *sensitive ovarian cancer: a randomised phase 2 trial*. Lancet Oncol, 2015. **16**(1): p.
633 87-97.
- 634 43. Demel, H.R., et al., *Effects of topoisomerase inhibitors that induce DNA damage*
635 *response on glucose metabolism and PI3K/Akt/mTOR signaling in multiple*
636 *myeloma cells*. Am J Cancer Res, 2015. **5**(5): p. 1649-64.
- 637 44. Farmer, H., et al., *Targeting the DNA repair defect in BRCA mutant cells as a*
638 *therapeutic strategy*. Nature, 2005. **434**(7035): p. 917-21.
- 639 45. Di Leonardo, A., et al., *DNA damage triggers a prolonged p53-dependent G1 arrest*
640 *and long-term induction of Cip1 in normal human fibroblasts*. Genes Dev, 1994.
641 **8**(21): p. 2540-51.
- 642 46. Agarwal, M.L., et al., *A p53-dependent S-phase checkpoint helps to protect cells*
643 *from DNA damage in response to starvation for pyrimidine nucleotides*. Proc Natl
644 Acad Sci U S A, 1998. **95**(25): p. 14775-80.
- 645 47. Brooks, K., et al., *A potent Chk1 inhibitor is selectively cytotoxic in melanomas with*
646 *high levels of replicative stress*. Oncogene, 2013. **32**(6): p. 788-96.
- 647 48. Vera, J., et al., *Chk1 and Wee1 control genotoxic-stress induced G2-M arrest in*
648 *melanoma cells*. Cell Signal, 2015. **27**(5): p. 951-60.
- 649 49. Liu, Q., et al., *Chk1 is an essential kinase that is regulated by Atr and required for*
650 *the G(2)/M DNA damage checkpoint*. Genes Dev, 2000. **14**(12): p. 1448-59.
- 651 50. Watanabe, N., M. Broome, and T. Hunter, *Regulation of the human WEE1Hu CDK*
652 *tyrosine 15-kinase during the cell cycle*. EMBO J, 1995. **14**(9): p. 1878-91.
- 653 51. van de Wetering, M., et al., *Prospective derivation of a living organoid biobank of*
654 *colorectal cancer patients*. Cell, 2015. **161**(4): p. 933-45.
- 655 52. Sato, T., et al., *Single Lgr5 stem cells build crypt-villus structures in vitro without a*
656 *mesenchymal niche*. Nature, 2009. **459**(7244): p. 262-5.
- 657 53. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human*
658 *cancers*. Nat Genet, 2013. **45**(10): p. 1127-33.
- 659 54. Osato, M., *Point mutations in the RUNX1/AML1 gene: another actor in RUNX*
660 *leukemia*. Oncogene, 2004. **23**(24): p. 4284-96.

- 661 55. Watkins, J.A., et al., *Genomic scars as biomarkers of homologous recombination*
662 *deficiency and drug response in breast and ovarian cancers*. Breast Cancer Res,
663 2014. **16**(3): p. 211.
- 664 56. Alexandrov, L.B., et al., *A mutational signature in gastric cancer suggests*
665 *therapeutic strategies*. Nat Commun, 2015. **6**: p. 8683.
- 666 57. Ledermann, J., et al., *Olaparib maintenance therapy in patients with platinum-*
667 *sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of*
668 *outcomes by BRCA status in a randomised phase 2 trial*. Lancet Oncol, 2014. **15**(8):
669 p. 852-61.
- 670 58. Verhagen, C.V., et al., *Extent of radiosensitization by the PARP inhibitor olaparib*
671 *depends on its dose, the radiation dose and the integrity of the homologous*
672 *recombination pathway of tumor cells*. Radiother Oncol, 2015. **116**(3): p. 358-65.
- 673 59. Kelly RJ, T.E., Zahurak. M, Cornish. T, Cuka. N, Abdelfatah. E, Taube. JM, Yang. S,
674 Duncan. M, Ahuja. N, Murphy. A, Anders. RA, *Adaptive immune resistance in*
675 *gastro-esophageal cancer: Correlating tumoral/stromal PDL1 expression with*
676 *CD8+ cell count*. J Clin Oncol 2015. **33**(suppl; abstr 4031)).
- 677 60. Nakamura, H., et al., *Genomic spectra of biliary tract cancer*. Nat Genet, 2015.
678 **47**(9): p. 1003-10.
- 679 61. Bridges, K.A., et al., *MK-1775, a novel Wee1 kinase inhibitor, radiosensitizes p53-*
680 *defective human tumor cells*. Clin Cancer Res, 2011. **17**(17): p. 5638-48.
- 681 62. Wang, Y., et al., *Radiosensitization of p53 mutant cells by PD0166285, a novel G(2)*
682 *checkpoint abrogator*. Cancer Res, 2001. **61**(22): p. 8211-7.
- 683 63. Liu, D.S., et al., *APR-246 potently inhibits tumour growth and overcomes*
684 *chemoresistance in preclinical models of oesophageal adenocarcinoma*. Gut, 2015.
685 **64**(10): p. 1506-16.
- 686 64. Stewart, A., et al., *Titration of signalling output: insights into clinical combinations*
687 *of MEK and AKT inhibitors*. Annals of Oncology, 2015. **26**(7): p. 1504-10.
- 688 65. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-*
689 *Wheeler transform*. Bioinformatics, 2009. **25**(14): p. 1754-60.
- 690 66. Saunders, C.T., et al., *Strelka: accurate somatic small-variant calling from*
691 *sequenced tumor-normal sample pairs*. Bioinformatics, 2012. **28**(14): p. 1811-7.
- 692 67. McLaren, W., et al., *Deriving the consequences of genomic variants with the*
693 *Ensembl API and SNP Effect Predictor*. Bioinformatics, 2010. **26**(16): p. 2069-70.
- 694 68. Van Loo, P., et al., *Allele-specific copy number analysis of tumors*. Proc Natl Acad Sci
695 U S A, 2010. **107**(39): p. 16910-5.
- 696 69. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for*
697 *analyzing next-generation DNA sequencing data*. Genome Res, 2010. **20**(9): p.
698 1297-303.
- 699 70. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration*. Nat Genet,
700 2013. **45**(10): p. 1134-40.
- 701 71. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic*
702 *content using next-generation sequencing data*. Bioinformatics, 2012. **28**(3): p.
703 423-5.
- 704 72. Chen, X., et al., *Manta: rapid detection of structural variants and indels for germline*
705 *and cancer sequencing applications*. Bioinformatics, 2016. **32**(8): p. 1220-2.
- 706 73. Schulte, I., et al., *Structural analysis of the genome of breast cancer cell line ZR-75-*
707 *30 identifies twelve expressed fusion genes*. BMC Genomics, 2012. **13**: p. 719.

- 708 74. Le Tallec, B., et al., *Common fragile site profiling in epithelial and erythroid cells*
709 *reveals that most recurrent cancer deletions lie in fragile sites hosting large genes.*
710 *Cell Rep*, 2013. **4**(3): p. 420-8.
- 711 75. Auton, A., et al., *A global reference for human genetic variation.* *Nature*, 2015.
712 **526**(7571): p. 68-74.
- 713 76. Wilkerson, M.D. and D.N. Hayes, *ConsensusClusterPlus: a class discovery tool with*
714 *confidence assessments and item tracking.* *Bioinformatics*, 2010. **26**(12): p. 1572-3.
- 715 77. Nilsen, G., et al., *Copynumber: Efficient algorithms for single- and multi-track copy*
716 *number segmentation.* *BMC Genomics*, 2012. **13**: p. 591.
- 717 78. Korbel, J.O. and P.J. Campbell, *Criteria for inference of chromothripsis in cancer*
718 *genomes.* *Cell*, 2013. **152**(6): p. 1226-36.
- 719 79. Puente, X.S., et al., *Non-coding recurrent mutations in chronic lymphocytic*
720 *leukaemia.* *Nature*, 2015. **526**(7574): p. 519-24.
- 721 80. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous*
722 *variants on protein function using the SIFT algorithm.* *Nat Protoc*, 2009. **4**(7): p.
723 1073-81.
- 724 81. Adzhubei, I.A., et al., *A method and server for predicting damaging missense*
725 *mutations.* *Nat Methods*, 2010. **7**(4): p. 248-9.
- 726 82. Lundegaard, C., et al., *NetMHC-3.0: accurate web accessible predictions of human,*
727 *mouse and monkey MHC class I affinities for peptides of length 8-11.* *Nucleic Acids*
728 *Res*, 2008. **36**(Web Server issue): p. W509-12.
- 729 83. Adiconis, X., et al., *Comparative analysis of RNA sequencing methods for degraded*
730 *or low-input samples.* *Nat Methods*, 2013. **10**(7): p. 623-9.
731
732

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

753
754 **Figure 2. RTK copy number profiling and responses to targeted RTK therapy**
755 **(n=129).** (a) RTK copy number gains/losses in the patient cohort and cell models. The
756 score refers to: amplifications (2), homozygous deletions (-2), relative gains/losses
757 (+1/-1) (Methods). Columns correspond to samples, ordered by the average ploidy.
758 Samples with average ploidy ≥ 3 are highlighted as potentially whole-genome duplicated.
759 (b) Copy number alterations in key genes of downstream pathways (c) Expression of
760 RTKs and downstream key genes in samples with gains (light red) versus losses (light
761 blue) of respective genes. The number of samples varies depending on the availability of
762 cases with gain/loss (indicated in brackets). * marks p-values <0.05 after multiple
763 testing correction. The solid horizontal line within the box represents the median. The
764 interquartile range (IQR) is defined as Q₃-Q₁ with whiskers that extend 1.5 times the
765 IQR from the box edges. (d) IHC staining of selected samples displaying consequences of
766 copy number loss/gain in ERBB2 and MET. The GISTIC score (CN) is marked. (e)
767 Breakdown of major resistance mechanisms to RTK-based monotherapy. “Amplification”
768 denotes anything with a score ≥ 1 . (f) Growth curve of OE33 cells after 72-hour exposure
769 to Lapatinib, Crizotinib and in combination. Mean values as percentage of DMSO treated
770 cells and \pm SD for three experiments. Olaparib in combination was 1 μ M. (g) The effects of
771 Lapatinib, Crizotinib and in combination on the cell lines with varying RTK status. Error
772 bars represent the standard deviation. * indicates p-values <0.05.

773
774 **Figure 3. Mutational signature-based clustering reveals differences in disease**
775 **etiology in the cohort and is spatially consistent within a single tumor.** (a) The heat
776 map highlights the sample exposures to six main mutational signatures, as identified in
777 the cohort (n=120) using the NMF methodology. The strength of exposure to a certain
778 signature may vary from 0% to 100% (on a color scale from grey to red). Three main
779 subgroups can be observed from the clustering based on the predominant signature:
780 C>A/T dominant (S18-like/S1 age) – orange, 32% samples; DDR impaired (S3-BRCA) –
781 purple, 15% samples; and mutagenic (S17A/B dominant) – green, 53% samples. The

782 *TP53*, *ERBB2* status, and catastrophic event distribution in the corresponding genomes
783 are highlighted below (no significant difference observed among subgroups). The total
784 mutational burden is significantly higher in the mutagenic subgroup. Consensus
785 clustering was used for the heat map (Methods). b) Validation of the mutational
786 signature-based clustering in an independent cohort (n=87). Unsupervised hierarchical
787 clustering (Pearson correlation distance, Ward linkage method) reveals three main
788 subgroups, similar to the ones in the discovery cohort: (1) DDR impaired (S3-BRCA)
789 dominant – purple, 22% of the cohort; (2) C>A/T dominant (S18-like/S1 age) – orange,
790 25% of the cohort; (3) mutagenic (S17A/B dominant) – green, 53% of the cohort. The
791 total SNV burden is also highlighted, confirming higher abundance in the mutagenic
792 subgroup. c) Mutational signature contributions in three cases with multiple sampling
793 from the same tumor. The relative exposures to the 6 signatures are highlighted on a
794 grey-to-red gradient for each case. The group assignment is based on the dominant
795 signature.

796
797 **Figure 4. DNA damage repair pathways altered through nonsynonymous**
798 **mutations/indels in the cohort.** (a) For each of the three defined subgroups, the
799 percentage of patients harboring defects in the different DDR-related pathways is shown.
800 Only nonsynonymous mutations in genes mutated in the cohort significantly more
801 compared to the expected background rate and predicted to be potentially damaging to
802 the protein structure (Methods) have been considered in the analysis. (b) HR, CR and
803 CPF genes altered in the three subgroups (the numbers in the gradients indicate how
804 many patients have mutations in the respective gene). AM, alternative mechanism for
805 telomere maintenance; BER, base excision repair; CPF, checkpoint factor; CR, chromatin
806 remodelling; CS, chromosome segregation; FA, Fanconi anaemia pathway; HR,
807 homologous recombination; MMR, mismatch repair; NER, nucleotide excision repair;
808 NHEJ, non-homologous end joining; OD, other double-strand break repair; TLS,
809 translesion synthesis; TM, telomere maintenance; UR, ubiquitylation response.

810
811 **Figure 5. Neoantigen burden is significantly higher in the mutagenic subgroup and**
812 **associates with an increased CD8+ T-cell density.** (a) From left to right: Neoantigen
813 burden compared among the 3 mutational signature subgroups shows significant
814 differences. A two-sided Welch's t-test was used to compare the mutagenic group to the
815 rest; Expression data available for a subset of the samples (25 from the mutagenic
816 subgroup and 21 from the others) reveals that the number of expressed potential
817 neoantigens is significantly higher in the mutagenic subgroup (Wilcoxon rank-sum test
818 $p = 0.042$); Numbers of CD8+ T cells per mm^2 observed in patients. Patients were
819 grouped into the mutagenic group and BRCA+C>A/T dominant group (n = 10 for each
820 group). (b) Two representative images of CD8 IHC staining from each group
821 (magnification 200x, scale bar, 100 μm).

822
823 **Figure 6. Treatment response in different mutational signature groups.** (a) Three
824 cell lines, OES127, MFD and CAM02 have been derived, each representative of a distinct
825 signature-dominant subgroup: DDR impaired (OES127), mutagenic (MFD) and C>A/T
826 dominant (CAM02). (b) Growth curves of OES127 cell lines after 72-hour exposure to
827 Olaparib, Topotecan and in combination. Mean values as a percentage of DMSO treated
828 cells and $\pm\text{SD}$ for three experiments are shown. Olaparib used in combination was kept
829 at 1 μM . (c) Growth curve of MFD cell lines after 72-hour exposure to MK-1775 and in

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

832

833 **Figure 7. Proposed subclassification of EAC based on mutational signatures**
834 **informs etiology and, consequently, potential tailored therapies to be further**
835 **investigated for the disease.** Patients are currently treated uniformly, but
836 classification based on mutational signatures may enable targeted treatments that
837 would complement classical therapy routes and potentially achieve more durable
838 responses. The highlighted box (right) exemplifies classifying new patients into the
839 defined etiological categories based on mutational signatures using a quadratic
840 programming approach (see Methods). The bars highlight the relative contributions of
841 the six expected signatures to the observed mutations in 7 new tumors (not part of the
842 129 sample cohort). The dominant signature is indicative of the group to which the
843 sample should be assigned.

844 **Table 1. *In vitro* cytotoxicity of RTKi as single or combined reagents in EAC cell**
845 **lines.** Key RTK amplification status and drug targets are shown. Bold text indicates that
846 a synergistic effect of the combination treatment was observed.
847

Cell line	RTK status	RTKi	GI50 (95% CI) (nM)	AUC
OE33	<i>ERBB2/MET</i> 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	<i>ERBB2</i> Amp/ <i>MET</i> 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	<i>ERBB2/FGFR2</i> 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	<i>EGFR/MET</i> 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	<i>ERBB2</i> Amp/ <i>MET</i> 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

848

1 **ONLINE METHODS**

3 **Ethical approval, sample collection and DNA extraction**

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

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

19 **Whole-genome sequencing analysis**

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

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

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

37 **Somatic mutation and indel calling**

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

42 Significantly mutated genes were identified using MutSigCV [23].

44 **Copy number and loss of heterozygosity analysis**

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

49 Cellularity, expressed as the relative proportion of tumor and normal nuclei, was also
50 obtained using ASCAT. It was distributed as follows: 18% of samples had cellularity
51 <0.3 ; 71% of samples between 0.3 and 0.7; 11% of samples ≥ 0.7 .

52 Significantly amplified/deleted regions in the cohort were identified using GISTIC2.0
53 [22], after correcting the copy numbers for ploidy (total copy number of the segment
54 divided by the average estimated ploidy of each sample). GISTIC2.0 was run on an input
55 defined as the log2 of such corrected copy number values, with gain (-ta) and loss (-td)
56 thresholds of 0.1 and sample centering prior to analysis. Copy number change
57 thresholds considered for downstream analysis were: amplifications, GISTIC score ≥ 2 ;
58 deletions, ≤ -2 . Loss of heterozygosity (LOH) was defined as ASCAT-estimated minor
59 allele copy number of 0.

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

62 For cell lines, copy number calling was performed using Control-FREEC [71].
63

64 *RTK copy number profiling*

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

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

78

79 **Structural variant and mobile element insertion calling and annotation**

80

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

91

92 **Structural variant-based classification of genomes**

93

94 The structural variant-based classification was used to annotate unstable, stable,
95 locally rearranged and scattered genomes as previously described [11], but with
96 different cut-offs for stable and unstable genomes, to account for the different genomic
97 instability landscape in EAC compared to pancreatic cancer: genomes were deemed

98 “stable” if the total number of SVs was less than the 5% quantile in the cohort, and
99 unstable if the number of SVs exceeded the 95% quantile. The criteria for locally
100 rearranged and scattered genomes were as previously described.

102 **Mutational signature analysis**

104 *Discovery*

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

114 To establish which of the two C[T>G]T signatures resembled most the classical S17
115 signature recorded in the COSMIC database, we used the cosine similarity distance
116 measure between the probability vectors of these signatures. The signature which we
117 termed S17A had a higher cosine similarity distance compared to S17B (0.98 versus
118 0.92), and we hence considered it to be more reflective of the signature reported in the
119 literature.

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

124 *Validation*

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

133 *Multiple sampling*

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

138 **Structural variant signature analysis**

139
140 Similar to inferring mutational signatures, we used the methodology by Alexandrov et al
141 [27] to discover structural variant signatures in EAC genomes. We classified structural
142 variants (deletions, inversions, insertions, interchromosomal translocations) by their
143 size and distribution along the genome. SVs were grouped by size into “small” and
144 “large”, defined with respect to the 25% quantile length in the cohort for the respective
145 SV type). To determine the SV distribution along the genome, we assessed the degree of
146 clustering within 10 Mb windows along the genome. If the SV of interest fell within a

147 window of clustered events (where the total number of SVs exceeded 1.5x the 75%
148 quantile of the total number of events in that genome), then it was deemed a “focal”
149 event. Otherwise, it was catalogued as “genomically distributed”. These characteristics
150 defined a total of 14 features to be used for signature discovery (Supplementary Figure
151 23).

152

153 **Identification of catastrophic events**

154

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

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

170 Regions of clustered inversions were identified as a proxy for BFB and complex
171 rearrangement events. These were defined by scanning for enrichments of inversions
172 (1.5x the upper quantile of the total number of events in the genome) within 5-Mb
173 windows throughout the genome. Visual inspection was used to prioritize those regions
174 that displayed BFB-like characteristics. Several types of complex rearrangement events
175 were identified: focal amplifications with BFB pattern (clustered inversions along with
176 progressive amplification steps primarily on one side of the inversion cluster, i.e.
177 asymmetric); other focal amplifications within narrow regions <5 Mb (clustered
178 inversions coupled with copy number amplifications displaying an irregular pattern),
179 focal amplifications within wider 5–10 Mb regions (clustered inversions and progressive
180 copy number amplification steps, often with multiple peaks); double minute-like
181 patterns (clustered inversions at high copy number amplification regions without
182 evidence of a progressive mechanism); potential subtelomeric BFBs (amplifications
183 located close to the ends of the chromosomes, coupled with inversion clusters and distal
184 deletions). See Supplementary Figure 7 for sample illustrations of the patterns
185 described.

186

187

188 **DNA damage repair (DDR) analysis**

189

190 To assess the alterations in DNA damage-related pathways, we performed an analysis
191 similar to the one described by Pearl et al [30]. Among the genes involved in defined
192 DNA damage pathways as described in the paper, we only selected those affected more
193 often than the expected background of synonymous mutations, similar to the method
194 described by Puente et al [79]. The probability of a gene being affected by M
195 nonsynonymous mutations in the cohort follows a poisson binomial distribution and is

196 calculated relative to a basal probability depending on the number of nonsynonymous
197 (n_{ns}) and synonymous (n_s) mutations, gene size (L), local mutational density for the
198 locus (d) and total length of coding regions in the genome (E) as follows: $P_{ns} = \frac{n_{ns}Ld}{(n_{ns}+n_s)E}$

199 Subsequently, we catalogued those that harboured nonsynonymous somatic
200 mutations/indels with possible deleterious effect (as predicted by SIFT [80]/PolyPhen
201 [81]) or copy number alterations (amplifications and deletions using the defined GISTIC
202 cut-offs) in our cohort. We then compared the mutational load in 16 main pathways
203 among the defined mutational signature subgroups.

204

205 **Neoantigen predictions and analysis**

206

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

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

227

228 **Expression profiling**

229

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

236

237

238

239

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

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

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

256

257 **Cell lines and reagents**

258

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

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

273

274 **Immunohistochemistry**

275

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

280 CD8+ cells were counted manually in two tumour areas of 1 mm² each (except in one
281 case where there was sufficient material for one count only) and an average was
282 calculated.

283

284 **Drug sensitivity assays**

285

286 The seeding density for each line was optimised to ensure cell growth in the logarithmic
287 growth phase. Cells were seeded in complete medium for 24 hours then treated with

288 compounds at 4-fold serial dilutions for 72 hours. Cell proliferation was assessed using
289 CellTiter-Glo (Promega). The anchor inhibitors were kept constant at 1M in combination
290 studies.

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

294

295 **Statistics**

296

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

302

303 **Code availability**

304

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

306

307

308 **Methods-only references**

309

- 310 65. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-*
311 *Wheeler transform*. *Bioinformatics*, 2009. **25**(14): p. 1754-60.
- 312 66. Saunders, C.T., et al., *Strelka: accurate somatic small-variant calling from*
313 *sequenced tumor-normal sample pairs*. *Bioinformatics*, 2012. **28**(14): p. 1811-7.
- 314 67. McLaren, W., et al., *Deriving the consequences of genomic variants with the*
315 *Ensembl API and SNP Effect Predictor*. *Bioinformatics*, 2010. **26**(16): p. 2069-70.
- 316 68. Van Loo, P., et al., *Allele-specific copy number analysis of tumors*. *Proc Natl Acad Sci*
317 *U S A*, 2010. **107**(39): p. 16910-5.
- 318 69. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for*
319 *analyzing next-generation DNA sequencing data*. *Genome Res*, 2010. **20**(9): p.
320 1297-303.
- 321 70. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration*. *Nat Genet*,
322 2013. **45**(10): p. 1134-40.
- 323 71. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic*
324 *content using next-generation sequencing data*. *Bioinformatics*, 2012. **28**(3): p.
325 423-5.
- 326 72. Chen, X., et al., *Manta: rapid detection of structural variants and indels for germline*
327 *and cancer sequencing applications*. *Bioinformatics*, 2016. **32**(8): p. 1220-2.
- 328 73. Schulte, I., et al., *Structural analysis of the genome of breast cancer cell line ZR-75-*
329 *30 identifies twelve expressed fusion genes*. *BMC Genomics*, 2012. **13**: p. 719.
- 330 74. Le Tallec, B., et al., *Common fragile site profiling in epithelial and erythroid cells*
331 *reveals that most recurrent cancer deletions lie in fragile sites hosting large genes*.
332 *Cell Rep*, 2013. **4**(3): p. 420-8.
- 333 75. Auton, A., et al., *A global reference for human genetic variation*. *Nature*, 2015.
334 **526**(7571): p. 68-74.
- 335 76. Wilkerson, M.D. and D.N. Hayes, *ConsensusClusterPlus: a class discovery tool with*
336 *confidence assessments and item tracking*. *Bioinformatics*, 2010. **26**(12): p. 1572-3.

- 337 77. Nilsen, G., et al., *Copynumber: Efficient algorithms for single- and multi-track copy*
338 *number segmentation*. BMC Genomics, 2012. **13**: p. 591.
- 339 78. Korbel, J.O. and P.J. Campbell, *Criteria for inference of chromothripsis in cancer*
340 *genomes*. Cell, 2013. **152**(6): p. 1226-36.
- 341 79. Puente, X.S., et al., *Non-coding recurrent mutations in chronic lymphocytic*
342 *leukaemia*. Nature, 2015. **526**(7574): p. 519-24.
- 343 80. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous*
344 *variants on protein function using the SIFT algorithm*. Nat Protoc, 2009. **4**(7): p.
345 1073-81.
- 346 81. Adzhubei, I.A., et al., *A method and server for predicting damaging missense*
347 *mutations*. Nat Methods, 2010. **7**(4): p. 248-9.
- 348 82. Lundegaard, C., et al., *NetMHC-3.0: accurate web accessible predictions of human,*
349 *mouse and monkey MHC class I affinities for peptides of length 8-11*. Nucleic Acids
350 Res, 2008. **36**(Web Server issue): p. W509-12.
- 351 83. Adiconis, X., et al., *Comparative analysis of RNA sequencing methods for degraded*
352 *or low-input samples*. Nat Methods, 2013. **10**(7): p. 623-9