A comparative analysis of whole genome sequencing of oesophageal adenocarcinoma pre- and post-chemotherapy.

A Noorani[1]*, J Bornschein[1]*, AG Lynch[2]*, M Secrier[2], A Achilleos[2], M Eldridge[2], L Bower[2], JMJ Weaver[1], J Crawte[1], CA Ong[1], N Shannon[1], S MacRae[1], N Grehan[1], B Nutzinger[1], M O'Donovan[1,3], R Hardwick[4], S Tavaré[2], RC Fitzgerald[1] on behalf of the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium[5].

[1] Medical Research Council Cancer Unit, Hutchison/Medical Research Council Research Centre, University of Cambridge, Cambridge, UK
[2] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
[3] Department of Histopathology, Addenbrooke's Hospital, Cambridge, UK
[4] Oesophago-Gastric Unit, Addenbrooke's Hospital, Cambridge, UK
[5] A full list of contributers from the OCCAMS Consortium is available at the end of the manuscript

***denotes equal contribution**

Keywords: Whole-genome sequencing, esophageal adenocarcinoma, copy number aberrations, single nucleotide variations, mutational signatures, structural variants, mutational burden, neoadjuvant chemotherapy.

Correspondence
Prof. Rebecca C. Fitzgerald
MRC Cancer Unit
University of Cambridge
Hutchison/MRC Research Centre
Box 197, Biomedical Campus
Cambridge, CB2 0XZ
United Kingdom
email: rcf29@mrc-cu.cam.ac.uk

ABSTRACT

The scientific community has avoided using tissue samples from patients that have been exposed to systemic chemotherapy to infer the genomic landscape of a given cancer. Esophageal adenocarcinoma is a heterogeneous, chemo-resistant tumor for which the availability and size of pre-treatment endoscopic samples are limiting. This study compares whole-genome sequencing data obtained from chemo-naive and chemo-treated samples.

The quality of whole genomic sequencing data is comparable across all samples regardless of chemotherapy status. Inclusion of samples collected post-chemotherapy increased the proportion of late stage tumors. When comparing matched pre- and post-chemotherapy samples from 10 cases the mutational signatures, copy number and SNV mutational profiles reflect the expected heterogeneity in this disease. Analysis of SNVs in relation to allele specific copy number changes pinpoints the common ancestor to a point prior to chemotherapy. For cases in which pre- and post-chemotherapy samples do show substantial differences, the timing of the divergence is near-synchronous with endoreduplication. Comparison across a large prospective cohort (62 treatment-naive, 58 chemotherapy-treated samples), reveals no significant differences in the overall mutation rate, mutation signatures, specific recurrent point mutations or copy number events in respect to chemotherapy status.

In conclusion, whole-genome sequencing of samples obtained following neoadjuvant chemotherapy is representative of the genomic landscape of esophageal adenocarcinoma. Excluding these samples reduces the material available for cataloguing and introduces a bias towards the earlier stages of cancer.

INTRODUCTION

The incidence of esophageal adenocarcinoma (EAC) has increased six-fold in the last 30 years (Lepage et al. 2013). The majority of patients present with advanced disease and the overall survival is under 15% despite advances in multi-modal therapy (Jemal et al. 2011; Masclee et al. 2014). Patients who do not have distant nodal or organ metastases are considered suitable for treatment with curative intent. This generally comprises systemic chemotherapy followed by surgical excision. Chemotherapy has been shown to improve survival to over 30% for those entering a curative pathway and is now an integral part of standard care either alone or in combination with radiotherapy, although the benefits of radiotherapy are greater in esophageal squamous cell carcinoma (Medical Research Council Oesophageal Cancer Working 2002); (Cunningham et al. 2006). Complete pathological response after neoadjuvant chemotherapy is rare and constitutes less than 15% of all cases, highlighting that residual cancer cells often remain after this treatment (Sjoquist et al. 2011; Orditura et al. 2014).

Chemotherapeutic agents exert their effect by directly or indirectly inducing DNA damage and cell death. In EAC three distinct classes of drugs are mainly used in combination: an intercalating agent, a platinum-derivative and an anti-metabolite (Allum et al. 2011). Drugs such as epirubicine intercalate directly with the DNA strand and thereby disrupt further replication in rapidly dividing cells. Platinum drugs directly modify DNA through coordinate-covalent bonds between DNA and the platinum moiety and the gross DNA damage is repaired via the nucleotide excision repair pathway (NER) if intact. 5'-fluorouracil and derivatives target DNA metabolism and result in DNA adducts, strand breaks or stalled/collapsed DNA replication forks. In addition, many of these drugs result in an increase of reactive oxygen species (ROS) which can in turn induce DNA damage including single strand DNA breaks (Woods and Turchi 2013). Hence, one might expect to see direct effects of chemotherapeutic agents on the DNA sequence and the extent might depend on the tumor responsiveness to treatment (Rebucci and Michiels 2013).

The mutation burden in EAC is high with 8.0 mutations/Mb (range 1.53–34.56/Mb) per haploid genome (Alexandrov et al. 2013). The genomic landscape appears to be complex and heterogeneous with a large

number of point mutations occurring at very low frequency apart from *TP53* mutations, which are present in 70-80% cases (Dulak et al. 2013; Weaver et al. 2014). Whole-genome sequencing studies and SNP arrays are providing more detail on large-scale chromosomal rearrangements that are common with evidence of catastrophic events such as chromothripsis and breakage-bridge-fusion (BFB) occuring in around one third of patients (Dulak et al. 2013; Nones et al. 2014).

In the current study we performed whole-genome sequencing in highly clinically annotated samples of EAC that included chemo-naive and chemo-treated samples as part of the International Cancer Genome Consortium. We took the opportunity to critically evaluate the impact of chemotherapy on the genomic landscape. It has recently been reported from exome data that chemotherapy imposes a bottleneck on tumor evolution (Findlay et al. 2016). We therefore first sought to establish the genetic relationship between 10 matched pre- and post- chemotherapy samples and the point at which the samples diverged. After perfoming this initial analysis, we examined the single nucleotide variant (SNV) spectrum, mutational/tri-nucleotide context and copy number aberrations in a larger cohort of 58 chemotherapy-treated and 62 chemotherapy-naive samples.

RESULTS

Whole-genome sequencing of paired samples pre and post-chemotherapy

Whole-genome sequencing data were first analysed for 10 cases from which samples were taken pre- and post- chemotherapy. The clinical details of this cohort are shown in Supplemental Table S1. Of these 10 cases, eight had a single sample taken before and after neoadjuvant chemotherapy and three had multiple samples taken before and after treatment.

Overall the matched samples showed the expected range of estimated tumor cellularity, overall ploidy, mutational signature composition, SNV burden, and copy number variation including LOH, as well as focal amplifications and deletions (Supplemental Table S3). Regions of LOH, amplifications and deletions are mostly the same pre-and post-chemotherapy (with LOH always observed on the same allele for paired samples). Paired samples range from being almost identical (patient 001 – 97% of the genome in the same copy number state, 95% of SNVs called in both samples) to very altered (patient 008 – 27% of the genome in the same copy number state, 23% of SNVs called in both samples).

For each patient we observe copy number features present in all tumor cells pre-chemotherapy that are not present post-chemotherapy, and *vice versa*. The key question is whether these differences are a consequence of the chemotherapy or simply a reflection of heterogeneity. In seven out of nine cases (the cellularity in one case is too low to call), we identify regions that have lost heterozygosity in the pre-chemotherapy samples, but have retained heterozygosity in the post-chemotherapy samples. This implies that the post-chemotherapy sample cannot have evolved from the pre-chemotherapy sample, but rather they have a shared antecedent.

It is informative to discuss the two extreme cases indicated above in more detail (Figures 1 and 2). In patient 008 a minority of the genome has the same copy number state pre- and post-chemotherapy (Figure 1 A-C), and in addition a minority of SNVs are observed both pre- and post-chemotherapy (Figure 1 D). Events that are known to be early, e.g. mutations of TP53 and LOH of key genes (Supplemental Fig S1) are seen to be shared, and indeed the majority of the genome that does exhibit LOH is common to both samples

(Supplemental Table S3), and always occurs on the same allele when it is common. While different mutations are observed in the pre- and post-chemotherapy samples, the same mutational processes appear to be present (Figure 1 E, Figure 3), and the AAB copy number state is the most common (Supplemental Table S3). Unlike most of the pre and post samples in this cohort, the majority of focal amplifications are not shared, but convergent amplification of the *FGF* region is observed in both samples (Figure 1 H-I). One can therefore infer that clonal divergence occurred shortly after endoreduplication (Figure 1 F-G, J) and hence the differences between the two samples are attributable to events that pre-date chemotherapy.

Patient 001 is a very different case, with virtually no differences pre- and post-chemotherapy (Figure 2 A-D, Supplemental Table S3). We can see some 'clonal' differences between the two samples (Figure 2 F). 'Sub-clonal' behavior pre-chemotherapy appears to be a recent change from a clonal state that matched the post-chemotherapy sample, indicating that although the two samples have diverged only recently (Figure 2 H, Supplemental Table S2), the differences such as they are cannot be attributed to the chemotherapy regime.

In general, when the pre- and post-chemotherapy samples show substantial differences, the timing of the divergence of the samples can be traced to being near-synchronous to endoreduplication. For the samples that show little difference, we can still trace their common ancestry to a point prior to the chemotherapy, and see no evidence in the mutational signatures (Figure 3), copy numbers (Supplemental Table S3) or key genes (Supplemental Fig S1) to suggest that we are seeing anything other than the heterogeneity.

It is clear that the mutational signatures change over time (Figure 1 E, Figure 2 E and Figure 3), and that the more recent mutations are disproportionately affected by factors affecting the power to detect SNVs (including sequencing depth, genomic complexity and cellularity). Caution must therefore be taken in concluding that chemotherapy has had an effect on the observed mutational signatures, and we do not draw such a conclusion.

Our data suggest that differences seen pre- and post-chemotherapy are reflective of tumor heterogeneity and that either sample could be considered equally representative of the case. However, from these data on a small patient cohort, we cannot rule out the possibility of a subtle selective pressure and in order to

address this we require larger cohorts of pre- and post-chemotherapy samples, which thus form the second part of this analysis.

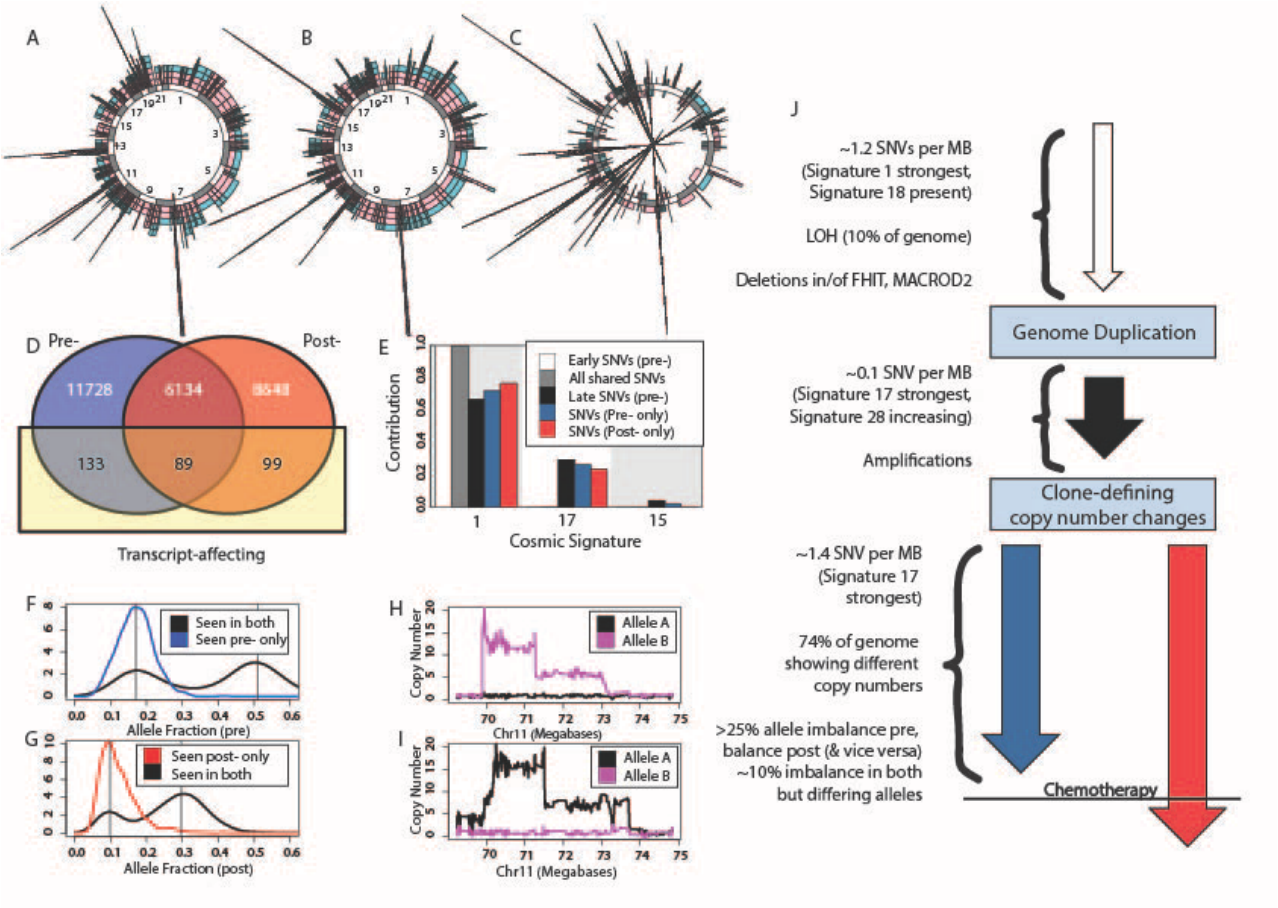

Figure 1: Profiling case 008 where the pre and post-chemotherapy samples are different A-C) Illustrating allele-specific copy number states for the 22 autosomes: A) pre-chemotherapy (alleles represented by colours), B) post-chemotherapy, C) the difference between the two allele-specific copy number profiles pre and post-chemotherapy. Copy number increases post- to pre-chemotherapy are shown outside the circle, decreases are shown inside the circle. D) Venn diagram showing the numbers of SNV calls shared pre- and post- chemotherapy, classified also by whether they affect coding genes. E) The mutational process signatures (reported in the Catalogue of Somatic Mutations in Cancer) that contribute substantially to the called SNVs are shown. Of the shared SNVs, approximately 6,000 lie within copy number states AA, AAB, AABB, AAA, or AAAA, and can confidently be categorized as early or late (relative to their copy number changes). The contributions for these subsets are shown also. F) For regions that in the pre-chemotherapy sample have copy-number status AAA, we see that no SNVs unique to this sample have three copies. G) For regions that in the post-chemotherapy sample have copy-number status AAA, we see that no SNVs unique to this sample that have three copies. H) Illustrating allele specific copy numbers for a region of chromosome 11 in the pre-chemotherapy sample. I) Illustrating allele-specific copy numbers for a region of chromosome 11 in the post-chemotherapy sample. J) A sketched likely timeline for this sample, although inherent to this type of data the timings of losses is supposition.

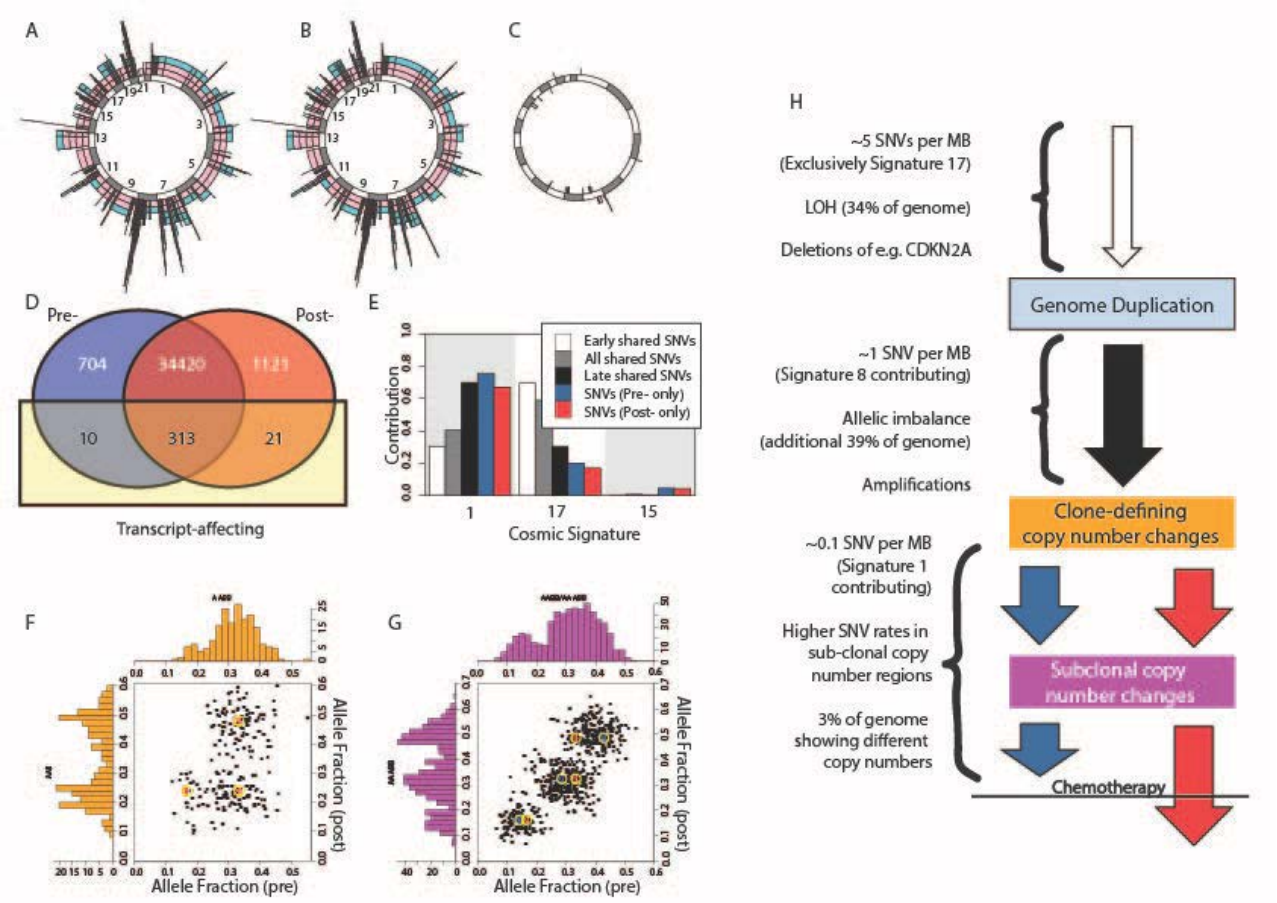

Figure 2: Profiling case 001 where the pre and post-chemotherapy samples are similar. A) Illustrating allele-specific copy number states for the 22 autosomes pre-chemotherapy (alleles represented by colours). B) Illustrating allele-specific copy number states for the 22 autosomes post-chemotherapy. C) Illustrating the difference between the two allele-specific copy number profiles. Copy number increases post- to pre-chemotherapy are shown outside the circle, decreases are shown inside the circle. D) Venn diagram showing the numbers of SNV calls shared pre- and post- chemotherapy, classified also by whether they are transcript-affecting. E) The mutational process signatures (reported at the Catalogue of Somatic Mutations in Cancer) that contribute substantially to the called SNVs are shown. Of the shared SNVs, approximately 14,000 lie within copy number states AA, AAB, AABB, AAAA, or AAAAB and can confidently be categorized as early or late (relative to their copy number changes). The contributions for these subsets are shown also. F) Illustrating SNVs for a region that exhibits different copy number states pre- (AABB) and post- (AAB) chemotherapy. The centres of predicted clusters for these states are indicated. Within each sample, the copy number state appears to be consistent in 100% of tumor cells. G) Illustrating SNVs for a region that demonstrates sub-clonal copy number behaviour pre- chemotherapy. The two sets of expected cluster centres for clonal AABB (red) and AAABB (blue) solutions pre-chemotherapy, against AAABB post-chemotherapy, are illustrated. The lack of a fourth SNV cluster is a strong indicator that this is a sub-clonal loss of one copy from a previously clonal AAABB state. H) A sketched likely timeline for this sample, although naturally the timing of losses is supposition.

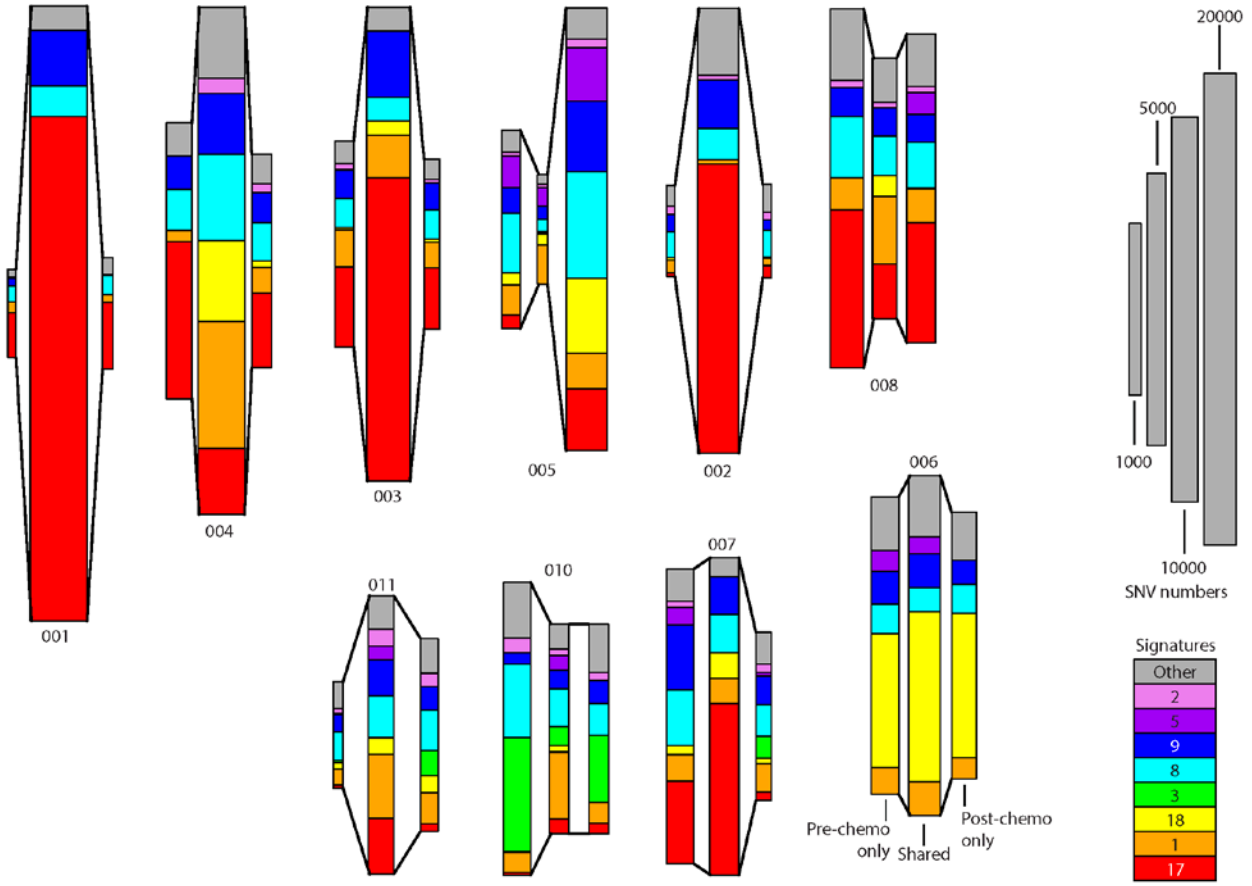

Figure 3. Mutational Signatures of mutational context for paired samples pre and post neaodjuvant chemotherapy.

Illustrating the numbers of shared and unique single nucleotide variants, and their breakdown into 30 known signatures from COSMIC. Since inference of signatures de novo is complicated by the non-independence and small numbers of samples, we do not attempt to do so, but rather infer the breakdown using quadratic programming methods (see Methods). For each patient, three rectangles are presented showing SNVs called pre-chemotherapy only (left), shared SNV calls (centre) and SNVs called post-chemotherapy only (right). The size of the rectangle indicates the number of SNVs, and the proportion of colour the breakdown into signatures, as indicated in the key. Only eight signatures that make sizeable contributions are individually identified.

Systematic comparison of whole-genome sequencing data on a large cohort of chemotherapy-naive or chemotherapy treated samples.

Our large cohort (n=120), shown in Supplemental Fig S2a, comprised 314 patients from whom there were 138 chemotherapy-naive samples taken at endoscopic diagnosis, prior to any treatment, or at the time of surgical resection if no neoadjuvant systemic chemotherapy was given; and 176 samples taken at surgery following systemic chemotherapy. For the patients receiving chemotherapy samples were not available both before and after treatment in this cohort. A further breakdown of the samples selected for the large cohort analysis is shown in Supplemental Fig S2b. The chemotherapy given at all centres was in line with the UK recommendation, comprising a platinum compound as a backbone generally combined with Epirubicin and a 5-fluoruracil derivative. Patients receiving radiotherapy were excluded in order to maintain consistency across the cohorts. The details of the study cohort are shown in Table 1.

Table 1: Demographic and pathological data of the large patient cohort (N=120)

		naive n=62		**treated n=58**		**total N=120**		**p-value**
Age *	**Years (IQR)**	71.9	(62.0-76.7)	65.1	(57.7-69.3)	66.6	(59.8-74.6)	0.002
Gender	**(male)**	54	87.1%	50	86.2%	104	86.7%	1.000
UICC Stage *	**1**	16	25.8%	9	15.5%	25	20.8%	0.024
	2	10	16.1%	7	12.1%	17	14.2%	
	3	31	50.0%	42	72.4%	73	60.8%	
	4	5	8.1%	0	0.0%	5	4.2%	
Grading[#]	**well**	1	1.7%	0	0.0%	1	0.8%	0.166
	moderate	30	50.0%	22	37.9%	52	44.1%	
	poor	29	48.3%	36	62.1%	65	55.1%	
Recurrence[#]		22	36.7%	29	50.9%	51	43.6%	0.139
Alive		34	54.8%	26	44.8%	60	50.0%	0.361

Abbreviations: IQR: interquartile range; Neoadj. CTx: neoadjuvant chemotherapy; age is given as median and IQR. Statistical analysis for homogeneity: Mann-Whitney's U-test for comparison of age, Fisher's exact test for categorical variables, $p<0.05$. Significant categories are marked by an asterisk. Items marked [#] represent incomplete data in selected cases.

Patients for which chemotherapy-treated samples were sequenced were significantly younger (p=0.002) and, as expected presented at a more advanced stage of disease (p=0.024) since for those patients going down curative pathways chemotherapy is not required for early stage tumours and patients have to be fit

enough to endure toxic therapy. Thus twenty-five patients (40%) with chemotherapy-naive samples went straight for surgery without neoadjuvant systemic treatment and were thus earlier stage. Histological response to neoadjuvant chemotherapy as assessed by the Mandard regression score was documented in 78 of the 95 patients (Supplemental Table S7). Of these, 16 (21%) had Mandard scores of 1-3 indicating some degree of histological response. A score of 4-5 (present in the remaining 79%) indicates poor response to neoadjuvant treatment, as expected for this particular cancer. Although the chemotherapy-treated group showed higher recurrence rates, this was not statistically significant (p=0.139). At the time of analysis, 50% of the patients were alive, with no significant difference between the treated and the chemotherapy-naive group (p=0.361). Please note that these statistics reflect the earlier stage cases in the chemotherapy naive group and are thus not reflective of the known benefit of chemotherapy shown in randomized trials for this disease.

All cases underwent a stringent pathological review of a frozen H&E section from the same sample that would be submitted for sequencing to confirm the diagnosis and ensure that the histopathological estimate of tumor cellularity exceeded 70%. Of the total cohort of n=314 samples (n=176 chemotherapy-treated, n=138 treatment-naive), significantly more samples that were exposed to chemotherapy failed this pathology review and were therefore excluded (n=98, 55.7% vs n=35, 25.4%; p<0.001, Supplemental Fig S2b). Of the treatment-naive samples, a higher proportion of endoscopic biopsies failed the pathology review compared to surgical specimens as would be expected from their small size (n=33, 30.6% vs n=2, 6.7%; p=0.01).

Genomic metrics of chemo-naive and chemo-treated samples of the large cohort

The group of treatment-naive samples contained a median of 24,449 SNVs and indels (combined), with a median absolute deviation (MAD) of 16,355, while the chemo-treated group had a median of 20,071 SNVs and indels (MAD = 12,223). The mutation rates in the chemo-naive and treated groups were similar, with a mean of 8.7 mutations/Mb for the former and 7.5 mutations/Mb for the latter (Wilcoxon rank-sum test p-value = 0.4).

Some genes were only recurrently mutated in the chemotherapy treated samples, e.g. *PTGES3L-AARSD1, RN7SL332P, AC011893.3, OR4D12P, TSPAN10, PPFIA3, C15orf39, SLC27A4, NAA30* in at least 15% of this group. However, the top recurrently mutated genes that have been previously characterized for esophageal adenocarcinoma, which are more likely drivers in this cancer, were generally mutated in a similar proportion of cases across the two cohorts (Figure 4) (Dulak et al. 2013; Weaver et al. 2014).

The tissue samples in the two groups displayed similar proportions of amplifications, deletions and LOH regions (Wilcoxon rank-sum test p-values >0.05, Figure 5, Supplemental Table S4, Supplemental Table S8, Supplemental Fig S3). Furthermore, in each group the defined genomic characteristics were similar regardless of patient age, disease stage, resection margin status (positive or negative for tumor cells) or sample source (biopsy or resection specimen), (Supplemental Table S5).

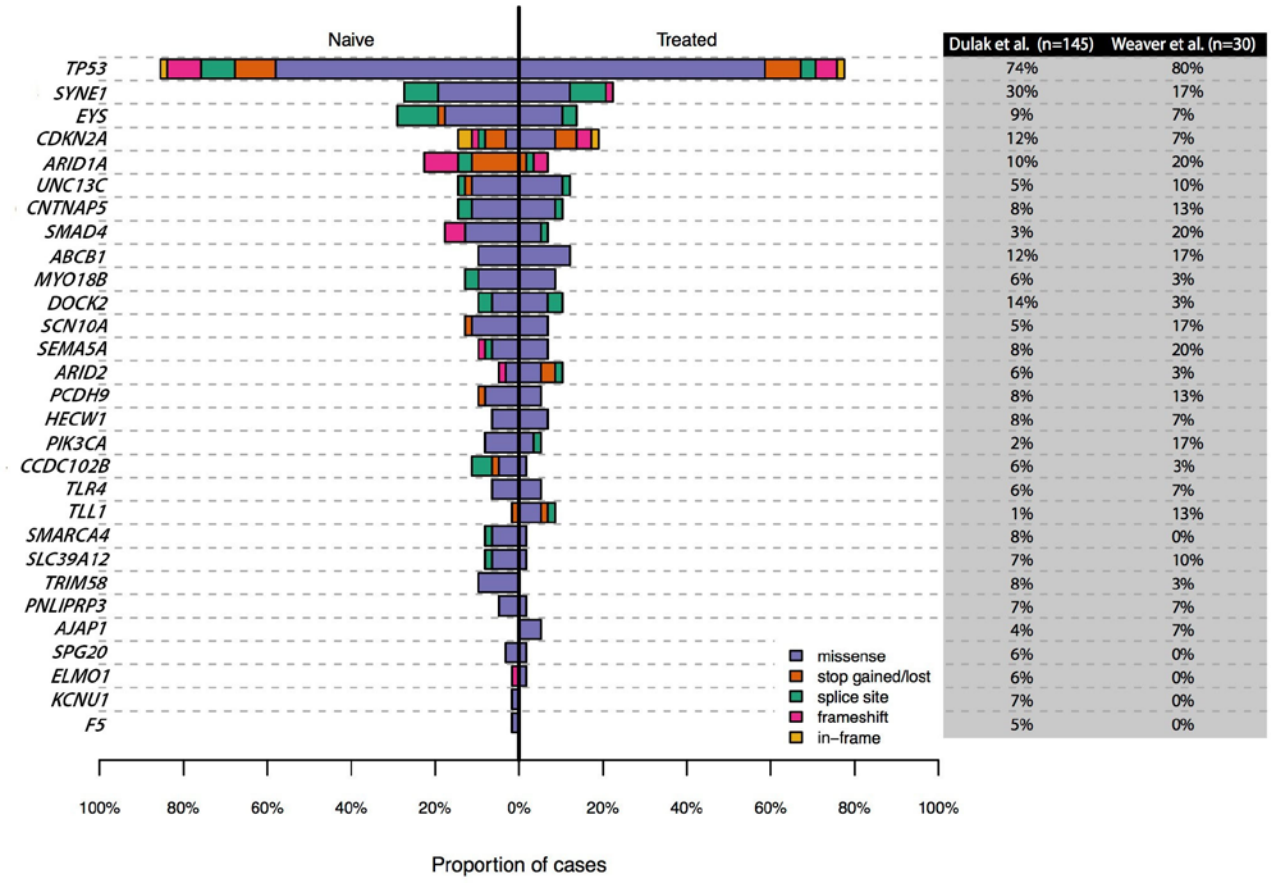

Figure 4. Proportion of non-synonymous SNVs and indels in recurrently mutated genes in chemotherapy treated and chemotherapy naive cohorts.

The genes were selected from the top ranking genes described in either of the Dulak *et al.* or Weaver *et al.* studies. The corresponding table demonstrates the percentage of samples that had mutations in these selected genes (Dulak et al. 2013; Weaver et al. 2014).

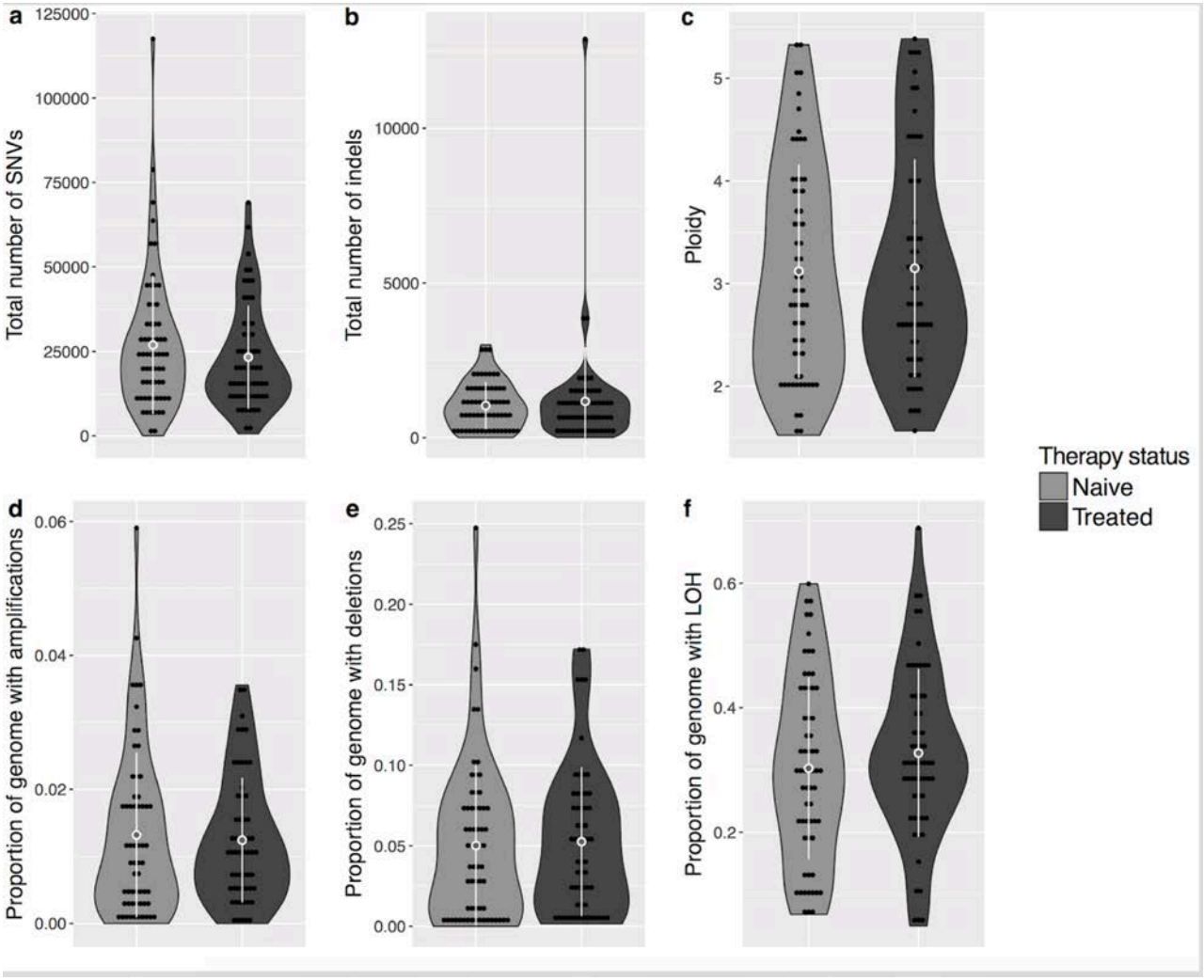

Figure 5. Genomic architecture in chemotherapy-naive (n=62) and chemotherapy-treated (n=58) samples. (a) Total number of SNVs, (b) Total number of indels, (c) Average ploidy, (d) Percentage of the genome that is amplified (defined as copy number ≥2x the average ploidy), (e) Percentage of the genome with deletions (defined as copy number ≤0.5x the average ploidy), (f) Percentage of the genome with LOH. No significant difference between naive and chemo-treated groups is observed in any case. The mean ± 1 standard deviation are highlighted in each case.

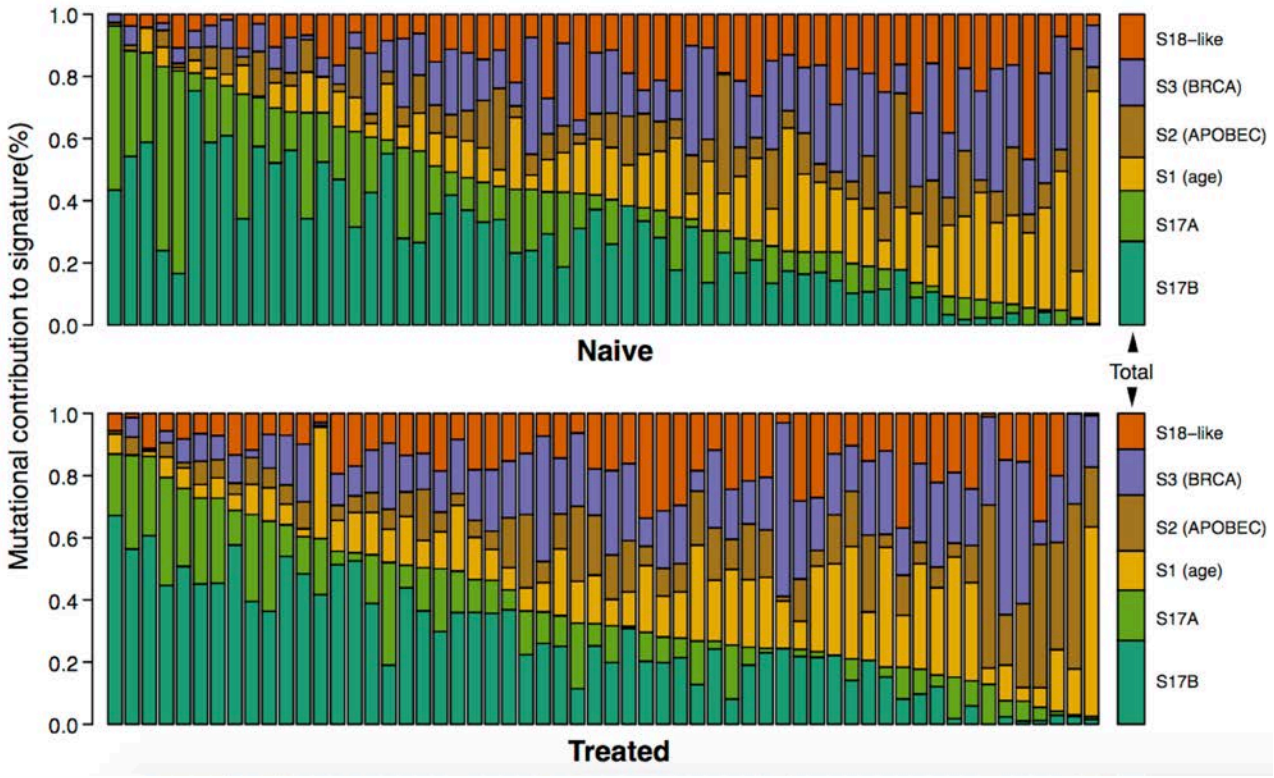

Figure 6: Mutational spectra analysis in chemotherapy-treated and chemotherapy-naive patient groups. The most relevant signatures for each group have been identified according to the method published by Alexandrov *et al.* (Alexandrov et al. 2013). Each bar shows the proportion of calls for the relevant signature per sample, the bars on the far right the cumulative proportion for each group.

The effect of chemotherapy on mutational spectrum analysis in the large cohort

EAC mutational signatures were extracted using the method presented in Alexandrov et al. 2013. A total of six mutational signatures were identified, of which five have been previously identified in EAC and other cancer types (Dulak et al. 2013; Weaver et al. 2014). None of these five signatures have been previously associated with exposure to chemotherapy. We therefore compared the number of mutations generated by each signature within the two cohorts and did not observe any clear difference (Figure 6). A comparison of the mutational signatures for the 10 pairs of samples pre and post-chemotherapy again showed similar non-significant differences (Supplemental Fig S6).

Since there was no significant overrepresentation of a particular trinucleotide in either group, (Wilcoxon rank-sum test adjusted p-values >0.05, Supplemental Fig S4), this prompted us to repeat the analysis with a focus on C>A substitutions mutations occurring at CpC dinucleotides that have been previously reported to be associated with systemic treatment with Cisplatin (Meier et al. 2014). There was a significant enrichment for the cisplatin induced mutational signature in the chemotherapy treated cohort (Fisher exact test p-value <0.0001, Supplemental Fig S5), in agreement with observations by (Murugaesu et al. 2015)

DISCUSSION

In this study we have used whole-genome sequencing, incorporating a comprehensive analysis of copy number, single nucleotide variants and mutational signatures from prospectively collected samples with stringent pathology QC but without imposing any restriction on including samples collected from patients who had already been treated with chemotherapy.

The first aim was to ensure that inclusion of chemotherapy exposed tissues did not result in poor quality samples in terms of low cellularity, DNA quality or sequencing metrics and we have demonstrated that the quality metrics were generally favourable and resulted in the inclusion of a greater proportion of late stage tumors that would otherwise have been excluded. We then examined a small cohort of patients with samples collected pre and post chemotherapy (n=10) and observed a range in the degree of genomic concordance given the degree of heterogeneity expected in this disease (Dulak et al. 2013; Nones et al. 2014; Weaver et al. 2014). For the samples within a case which showed a high degree of similarity (e.g. IDs 002 and 001) we can trace their common ancestry to a point prior to the chemotherapy. On the other hand, when the pre- and post-chemotherapy samples show substantial differences (e.g. IDs 007, 008 and 005), the timing of the divergence of the samples can be traced to being near-synchronous with endoreduplication. When investigating the effect of chemotherapy on a larger scale, in a cohort of 120 patients, we observe that the genome of EAC is remarkably resistant to the effect of neoadjuvant chemotherapy. Indeed, there was a striking similarity noted between chemotherapy-naive samples, and those treated with neoadjuvant chemotherapy at the level of copy number aberrations, single nucleotide variants and mutational spectra. This study was not designed to examine the genetic predictors of response to chemotherapy, which requires a different experimental approach given that it is generally a chemo-resistant disease. Only 20% of patients in our study showed a histopathological response (Mandard score 1-3, based on the degree of fibrosis and proportion of tumor cells remaining) to chemotherapy, which is consistent with the treatment response expected from the literature (Cunningham et al. 2006; Alderson et al. 2015).

To date, most large scale sequencing efforts including The Cancer Genome Atlas (TCGA) and other International Cancer Genome Consortium (ICGC) projects have been confined to patients who are naive to systemic treatment. Hence, for cancers treated with chemotherapy prior to surgical resection (e.g. cancers of the stomach, oesophagus, breast, bladder, cervix, and lung) this has restricted the samples available for analysis to pre-treatment diagnostic biopsies which are generally obtained via endoscopy or laparoscopy and are challenging to work with due to their small size. The main reason for exclusion of samples in our cohort was low cellularity (<70%) as determined by expert pathology review (3 independent pathologists) of a frozen section taken from the samples used for DNA extraction. The proportion of chemotherapy-treated samples excluded at this stage was more than twice as high as the proportion of treatment-naive samples and so this will potentially bias selection away from those who show a good histopathological response to systemic neoadjuvant treatment. However, apart from cellularity there was no further difference in the quality or quantity of DNA, library or sequence obtained. In the future as technology improves sequencing of single cells in cases that are highly responsive to neo-adjuvant therapy maybe informative.

Our observation that the majority of EAC genomes remained rather stable following chemotherapy is consistent with breast cancer studies when considering those patients with chemo-resistant disease. For example, in a candidate gene study of 47 breast cancer patients Almendro *et al* found that intratumor genetic diversity was indicative of the tumor subtype and remained stable in patients with only partial or no response to treatment (Almendro et al. 2014). Yates *et al.* interrogated the sub-clonal architecture of breast cancer in 50 patients, of which 18 had samples taken before and after neoadjuvant chemotherapy (Yates et al. 2015). In five of these patients, new clones were seen in the post chemotherapy samples with potential driver events such as amplifications in *MYC* and *FGFR2* and deletions in *RUNX1*. Detailed phylogenetic reconstruction of these five cases suggested that the treatment resistant clones they observed were likely to have been missed at the time of pre chemotherapy sampling, and were unlikely to be the result of new sub-clones arising during treatment.

In the context of EAC Murugaesu *et al.* performed exome sequencing on samples from eight cases taken before and after chemotherapy (Murugaesu et al. 2015). The extensive multi-region sampling was a strength of this small study and they found a positive correlation between the degree of intra-tumoral heterogeneity and a poor response to neo-adjuvant chemotherapy, which in turn correlated with a worse survival. Our study was performed as part of the ICGC which is designed to examine the landscape by virtue of examining a large number of tumor: normal pairs; and hence we were generally unable to perform multi-regional sampling. Findlay *et al.* recently reported results from their exome anlaysis of 30 pre and post chemotherapy EAC samples and in this study they purposefully selected cases showing a range of responses to chemotherapy (Findlay et al. 2016). They associated good clinical response, as determined by the histopathological Mandard score generated from the post chemotherapy surgical resection specimen, with evidence for genomic bottlenecking as a result of chemotherapy. This is at odds with our interpretation.

We cannot, from such a limited number of cases with pre and post chemotherapy samples, in such a diverse disease, separate the potential sources of heterogeneity arising from spatial sampling, temporal sampling, and chemotherapy, unless we can make some inference about the timing of events. It has been reported previously that some point mutations, losses of heterozygosity and genome duplication events occur early in the cancer progression, and that genomic catastrophes and the accumulation of clonal diversity may play a role. Our paired cases support these prior observations (Nones et al. 2014).

On average, we noted that approximately a quarter of the genome had undergone LOH both pre- and post-chemotherapy in our samples and in all cases with paired samples the same allele was lost pre- and post-chemotherapy. Therefore, we infer that LOH, and then genome doubling, occur early in the life history of the cancers. The high point mutation rate associated with EAC allows us to say something about the timing of genomic catastrophes and the establishment of clonal diversity. If large-scale genomic rearrangements predate clonal diversity, then we expect to see SNVs that occur after the copy number changes, but which are shared pre-and-post chemotherapy. If the clonal diversity occurs before the copy number changes then we would expect to see SNVs that are unique to one sample but which predate local copy number changes. We see neither of these, strongly suggesting that the establishment of clonal diversity and the copy number

changes are roughly concurrent. This suggests that it is not just localized catastrophes but genome-wide changes that seem to occur near-simultaneously. Therefore, the divergence of the clones observed pre- and post- chemotherapy must have occurred substantially before treatment was administered and thus chemotherapy cannot be responsible for the divergence. An alternative explantion would be selective pressures for one clone out of those available, but the larger cohorts discussed above revealed little evidence of systematic selection of this kind.

Regarding the mutational signature analysis, we used the methods of Alexandrov et al which identified six main SNV signatures in our data, five of which have been previously described in EAC datasets (Dulak et al. 2013; Nones et al. 2014). When comparing samples taken pre and post chemotherapy we observe that the signature patterns are often different between those occurring before the CN changes and those timed as occurring after, but the more recent signature is consistent between both the pre- and post-chemotherapy samples. Thus any apparent differences in the mutational signatures pre- and post-chemotherapy are likely attributable to cellularity induced differences in the power to detect the recent SNVs that, by definition, have lower allele fractions.

While our study was not designed to determine the prognostic value of genomic response to chemotherapy, we acknowledge that some of the samples for which pre- and post- chemotherapy profiles differ the most (e.g. 007 and 008) are some of those with the best survival. However, we also note they are two of the cases with the best pathological TNM staging. Any approach to prognosticate based on genomic factors (e.g. perhaps following the results of Findlay et al. 2016) should at most temper established prognostic factors such as these fundamental phenotypic characteristics. Moreover, as discussed, some mutations were found to be more recurrent following chemotherapy and this is an area ripe for further research as the appropriate cohorts become available.

In conclusion, the overall genomic profile of EAC remains similar before and after chemotherapy. The poor survival in EAC would support our findings that this cancer is resistant to chemotherapy with remarkable consistency in the genome of the primary tumor over time. Based on our findings, we would suggest that inclusion of neoadjuvant treated samples for large scale sequencing efforts should be considered by the

sequencing community. Such an approach will avoid biasing cohorts towards the earlier stages of the disease and increase the number of samples available for analysis particularly in tumor types with neo-adjuvant therapy regimens. With the increasing recognition of the extent of epithelial tumor heterogenity large scale efforts are essential to maximize the power of uncovering the full spectrum of mechanisms driving tumorigenesis.

METHODS

Sample collection and processing

Esophageal adenocarcinoma (EAC) patients were recruited prospectively from 11 sites across the UK as part of the OCCAMS (Oesophageal Clinical And Molecular Stratification) consortium. Patients on a palliative treatment pathway as well as those treated with radiotherapy were excluded. The study was approved by the Institutional Ethics Committees (REC Ns 07/H0305/52, 10/H0305/1) and all patients gave written informed consent.

Samples were obtained either during the diagnostic esophagogastroduodenoscopy or endoscpic ultrasound procedure used for staging and/or from the surgical resection specimen (Figure 1). For each patient, blood or normal squamous esophageal samples, at least 5cm distant from the tumor, was used as a germline reference. In 10 cases tumor samples were taken from multiple spatially distinct sites at surgery and in 2 cases, also at EGD.

All tissue samples were snap-frozen in liquid nitrogen immediately after collection and stored at -80°C. H&E stained sections from cancer samples were reviewed independently by two expert Hisptopathologists and DNA was extracted and sequenced if tumour cellularity was ≥70%. DNA was extracted from frozen esophageal tissue using the All PrepDNA/RNA Mini Kit (Qiagen, Hilden, Germany) and from blood samples using the QIAamp DNA Blood Maxi kit (Qiagen, Hilden, Germany) according to manufacturer's instructions.

Whole-genome sequencing

As part of the International Cancer Genome Consortium, 100-125bp paired-end sequencing was performed under contract by Illumina to a typical depth of at least 50x, with 94% of the known genome being sequenced to at least 8x coverage while achieving a PHRED quality of at least 30 for at least 80% of mapping bases. QC metrics were computed on a per lane basis using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and in-house tools, enabling the

identification of sequence reads that required trimming. Technical details of the sequencing metrics are given in Supplemental Table S2.

Mutation calling

Sequence reads were aligned to the human reference genome (GRCh37 from Ensembl release 71, (Yates et al. 2016) using BWA 0.5.9 (Li and Durbin 2009). Aligned reads were then sorted into genome coordinate order and duplicate reads marked using Picard 1.115 (FixMateInformation and MarkDuplicates tools respectively, http://broadinstitute.github.io/picard). Somatic SNVs and indels were detected using Strelka 1.0.13 (Saunders et al. 2012). To increase accuracy, additional filters were applied to high-confidence calls (those passing Stelka's filters); details are given in Supplemental Table S6. Functional annotation of the resulting variants was performed using Variant Effect Predictor (VEP release 75) (McLaren et al. 2016).

Copy Number Calling

For the large cohort absolute copy number alterations, cellularities and ploidies for each sample were estimated using ASCAT-NGS v.2.1, using read counts at germline heterozygous positions estimated by GATK 3.2-2 (Van Loo et al. 2010; Nik-Zainal et al. 2012). Segments were considered amplified if the ratio of absolute copy number to ploidy exceeded two and deleted if the ratio was less than 0.5. Losses of heterozigosity (LOH) regions were defined as regions in the genome where the minor copy number was 0.

Mutational signature analysis

Mutational signatures were identified using the methodology described by Alexandrov et al (Alexandrov et al. 2013). Before running the software, common variants in the 1000 Genomes database (The 1000 Genomes Project 2015) appearing in at least 0.5% of the population were removed. The optimal number of signatures in the dataset was chosen to balance the signature stability against the Frobenius reconstruction error. The cisplatin signature enrichment analysis was performed as described by Murugaesu et al (Murugaesu et al. 2015).

Multiple Sample Analysis

A GATK walker was used to identify a set of germline-heterozygous loci for each trio. The search was restricted to the autosomes, sites with no more than 20 germline reads were filtered by GATK (McKenna et al. 2010), sites with germline coverage between 16 and 90, with at least 4 copies of each allele present, sites where the strand bias lies between 0.1 and 0.9 and sites that are not in obvious regions of germline copy number variation, identified with fastseg (Klambauer et al. 2012). This results in approximately 2,000,000 such loci per trio. The depths of coverage and allele fractions for these loci were recorded for all samples in the trio.

To aid segmentation, a running median was applied to the depth and allele fraction data. A single segmentation of these values was created for each patient by combining, for each tumor sample, a sliding analysis-of-variance procedure and careful manual review of the genome. We erred on the side of over-segmentation as there is little to no penalty for this in the analyses that follow. The cellularity and baseline copy number for each sample was identified using the Crambled tool (Lynch 2015) and depth and allele-fraction values for clonal copy number states were predicted. Segments were assigned to these copy number states, or sub-clonal combinations of those states, based on the mean values for the segments. Where solutions for a segment appeared to be sub-clonal, or differed between the multiple samples for a patient they were reviewed for possible technical explanations such as mis-segmentation. Neighbouring 'segments' assigned the same copy number state in both samples were merged. Segments were compared across samples to confirm the consistency of allele assignment (e.g. if both samples show two copies of one allele and one copy of the other, is the same allele duplicated in both cases) and corrected if not.

SNVs were called with Strelka and annotated with VEP as described elsewhere. SNVs were mapped to a copy number state pre- and post-chemotherapy. SNVs with the same copy-number combination pre- and post- chemotherapy were partitioned into early (coming before a copy number change) and late mutations where copy number states and power allowed. Vectors of trinucleotide mutation counts were deconstructed into the 30 cosmic signatures (http://cancer.sanger.ac.uk/cosmic/signatures) using a quadratic programming approach (Lynch et al. 2016).

DATA ACCESS

The whole-genome sequencing data have been submitted to the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) under accession number EGAD00001002241. Mutation calls can be found within the ICGC data portal (https://dcc.icgc.org/) under project ID ESAD-UK and library IDs listed in Supplemental Table S2.

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

Rachael Fels Elliott[1], Paul A.W. Edwards[1], Xiaodun Li[1], Hamza Chettouh[1], Gianmarco Contini[1], Eleanor Gregson[1], Sebastian Zeki[1], Laura Smith[1], Zarah Abdullahi[1], Rachel de la Rue[1], Ahmad Miremadi[1,3], Shalini Malhotra[1,3], Mike L. Smith[2], Jim Davies[6], Charles Crichton[7], Nick Carroll[8], Peter Safranek[8], Andrew Hindmarsh[8], Vijayendran Sujendran[8], Richard Turkington[9], Stephen J. Hayes[10,17], Yeng Ang[10,11, 32], Shaun R. Preston[12], Sarah Oakes[12], Izhar Bagwan[12], Vicki Save[13], Richard J.E. Skipworth[13], Ted R. Hupp[13], J. Robert O'Neill[13,26], Olga Tucker[14,31], Philippe Taniere[14], Timothy J. Underwood[15,16], Fergus Noble[15], Jack Owsley[15], Hugh Barr[18], Neil Shepherd[18], Oliver Old[18], Jesper Lagergren[19, 28], James Gossage[19,27], Andrew Davies[19,27,] Fuju Chang[19,27], Janine Zylstra[19,27],Grant Sanders[20], Richard Berrisford[20], Catherine Harden[20], David Bunting[20], Mike Lewis[21], Ed Cheong[21], Bhaskar Kumar[21],Simon L Parsons[22], Irshad Soomro[22], Philip Kaye[22], Laurence Lovat[23], Rehan Haidry[23], Victor Eneh[23], Laszlo Igali[24], Michael Scott[25], Shamila Sothi[29], Sari Suortamo[29], Suzy Lishman[30.]

[6]Oxford ComLab, University of Oxford, UK
[7]Department of Computer Science, University of Oxford, UK
[8]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[9]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Northern Ireland, UK
[10]Salford Royal NHS Foundation Trust, Salford, UK
[11]Wigan and Leigh NHS Foundation Trust, Wigan, Manchester, UK
[12]Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK
[13]Edinburgh Royal Infirmary, Edinburgh, UK
[14]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[15]University Hospital Southampton NHS Foundation Trust, Southampton, UK
[16]Cancer Sciences Division, University of Southampton, Southampton, UK
[17]Faculty of Medical and Human Sciences, University of Manchester, UK
[18]Gloucester Royal Hospital, Gloucester, UK
[19]St Thomas's Hospital, London, UK
[20]Plymouth Hospitals NHS Trust, Plymouth, UK
[21]Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, UK
[22]Nottingham University Hospitals NHS Trust, Nottingham, UK
[23]University College London, London, UK
[24]Norfolk and Waveney Cellular Pathology Network, Norwich, UK
[25]Wythenshawe Hospital, Manchester, UK
[26]Edinburgh University, Edinburgh, UK
[27]King's College London, London, UK
[28]Karolinska Institutet, Stockholm, Sweden

[29]University Hospitals Coventry and Warwickshire NHS, Trust, Coventry, UK
[30]Peterborough Hospitals NHS Trust, Peterborough City Hospital, Peterborough, UK
[31]Institute of cancer and genomic sciences, University of Birmingham
[32]GI science centre, University of Manchester, UK

ACKNOWLEDGEMENTS

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. 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. This study was partly funded by a project grant from Cancer Research UK. R.C.F. is funded by an NIHR Professorship and receives core funding from the Medical Research Council and infrastructure support from the Biomedical Research Centre and the Experimental Cancer Medicine Centre. We acknowledge the support of The University of Cambridge, Cancer Research UK (C14303/A17197) and Hutchison Whampoa Limited. We are grateful to all the patients who provided written consent for participation in this study and the staff at all participating centres.

DISCLOSURE DECLARATION

The authors declare no conflict of interest

References

Alderson D, Langley RE, Nankivell MG, Blazeby JM, Griffin M, Crellin A, Grabsch HI, Okines AFC, Goldstein C, Falk S et al. 2015. Neoadjuvant chemotherapy for resectable oesophageal and junctional adenocarcinoma: Results from the UK Medical Research Council randomised OEO5 trial (ISRCTN 01852072). *J Clin Oncol* **33**(15).

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500**(7463): 415-421.

Allum WH, Blazeby JM, Griffin SM, Cunningham D, Jankowski JA, Wong R, Association of Upper Gastrointestinal Surgeons of Great B, Ireland tBSoG, the British Association of Surgical O. 2011. Guidelines for the management of oesophageal and gastric cancer. *Gut* **60**(11): 1449-1472.

Almendro V, Cheng YK, Randles A, Itzkovitz S, Marusyk A, Ametller E, Gonzalez-Farre X, Munoz M, Russnes HG, Helland A et al. 2014. Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell reports* **6**(3): 514-527.

Cunningham D, Allum WH, Stenning SP, Thompson JN, Van de Velde CJ, Nicolson M, Scarffe JH, Lofts FJ, Falk SJ, Iveson TJ et al. 2006. Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. *The New England journal of medicine* **355**(1): 11-20.

Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C, Bandla S, Imamura Y, Schumacher SE, Shefler E et al. 2013. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature genetics* **45**(5): 478-486.

Findlay JM, Castro-Giner F, Makino S, Rayner E, Kartsonaki C, Cross W, Kovac M, Ulahannan D, Palles C, Gillies RS et al. 2016. Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. *Nature communications* **7**: 11111.

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. 2011. Global cancer statistics. *CA: a cancer journal for clinicians* **61**(2): 69-90.

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids research* **40**(9): e69.

Lepage C, Drouillard A, Jouve JL, Faivre J. 2013. Epidemiology and risk factors for oesophageal adenocarcinoma. *Digestive and liver disease : official journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver* **45**(8): 625-629.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Lynch A. 2015. Crambled: A Shiny application to enable intuitive resolution of conflicting cellularity estimates. *F1000Research* **4**: 1407.

Lynch MJ, Mulvaney MJ, Hodges SC, Thompson TL, Thomason WE. 2016. Decomposition, nitrogen and carbon mineralization from food and cover crop residues in the central plateau of Haiti. *SpringerPlus* **5**(1): 973.

Masclee GM, Coloma PM, de Wilde M, Kuipers EJ, Sturkenboom MC. 2014. The incidence of Barrett's oesophagus and oesophageal adenocarcinoma in the United Kingdom and The Netherlands is levelling off. *Alimentary pharmacology & therapeutics* **39**(11): 1321-1330.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**(9): 1297-1303.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome biology* **17**(1): 122.

Medical Research Council Oesophageal Cancer Working G. 2002. Surgical resection with or without preoperative chemotherapy in oesophageal cancer: a randomised controlled trial. *Lancet* **359**(9319): 1727-1733.

Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, Raine K, Maddison M, Anderson E, Stratton MR et al. 2014. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome research* **24**(10): 1624-1636.

Murugaesu N, Wilson GA, Birkbak NJ, Watkins TB, McGranahan N, Kumar S, Abbassi-Ghadi N, Salm M, Mitter R, Horswell S et al. 2015. Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer discovery* **5**(8): 821-831.

Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M et al. 2012. The life history of 21 breast cancers. *Cell* **149**(5): 994-1007.

Nones K, Waddell N, Wayte N, Patch AM, Bailey P, Newell F, Holmes O, Fink JL, Quinn MC, Tang YH et al. 2014. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nature communications* **5**: 5224.

Orditura M, Galizia G, Di Martino N, Ancona E, Castoro C, Pacelli R, Morgillo F, Rossetti S, Gambardella V, Farella A et al. 2014. Effect of preoperative chemoradiotherapy on outcome of patients with locally advanced esophagogastric junction adenocarcinoma-a pilot study. *Current oncology* **21**(3): 125-133.

Rebucci M, Michiels C. 2013. Molecular aspects of cancer cell resistance to chemotherapy. *Biochemical pharmacology* **85**(9): 1219-1226.

Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**(14): 1811-1817.

Sjoquist KM, Burmeister BH, Smithers BM, Zalcberg JR, Simes RJ, Barbour A, Gebski V, Australasian Gastro-Intestinal Trials G. 2011. Survival after neoadjuvant chemotherapy or chemoradiotherapy for resectable oesophageal carcinoma: an updated meta-analysis. *The Lancet Oncology* **12**(7): 681-692.

The 1000 Genomes Project C. 2015. A global reference for human genetic variation. *Nature* **526**(7571): 68-74.

Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B et al. 2010. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America* **107**(39): 16910-16915.

Weaver JM, Ross-Innes CS, Shannon N, Lynch AG, Forshew T, Barbera M, Murtaza M, Ong CA, Lao-Sirieix P, Dunning MJ et al. 2014. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nature genetics* **46**(8): 837-843.

Woods D, Turchi JJ. 2013. Chemotherapy induced DNA damage response: convergence of drugs and pathways. *Cancer biology & therapy* **14**(5): 379-389.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L et al. 2016. Ensembl 2016. *Nucleic acids research* **44**(D1): D710-716.

Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, Aas T, Alexandrov LB, Larsimont D, Davies H et al. 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature medicine* **21**(7): 751-759.