Whole-genome sequencing provides new insights into the clonal architecture of Barrett’s esophagus and esophageal adenocarcinoma

Caryn S Ross-Innes1,6, Jennifer Becq2,6, Andrew Warren2, R Keira Cheetham2, Helen Northen2, Maria O’Donovan3, Shalini Malhotra3, Massimiliano di Pietro1, Sergii Ivakhno2, Miao He2, Jamie M J Weaver1, Andy G Lynch4, Zoya Kingsbury2, Mark Ross2, Sean Humphray2, David Bentley2 & Rebecca C Fitzgerald1 for the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Study Group5

The molecular genetic relationship between esophageal adenocarcinoma (EAC) and its precursor lesion, Barrett’s esophagus, is poorly understood. Using whole-genome sequencing on 23 paired Barrett’s esophagus and EAC samples, together with one in-depth Barrett’s esophagus case study sampled over time and space, we have provided the following new insights: (i) Barrett’s esophagus is polyclonal and highly mutated even in the absence of dysplasia; (ii) when cancer develops, copy number increases and heterogeneity persists such that the spectrum of mutations often shows surprisingly little overlap between EAC and adjacent Barrett’s esophagus; and (iii) despite differences in specific coding mutations, the mutational context suggests a common causative insult underlying these two conditions. From a clinical perspective, the histopathological assessment of dysplasia appears to be a poor reflection of the molecular disarray within the Barrett’s epithelium, and a molecular Cytosponge technique overcomes sampling bias and has the capacity to reflect the entire clonal architecture.

Most epithelial cancers present de novo but have progressed from a clinically silent preinvasive state or so-called intraepithelial neoplasia. There is growing interest in understanding the life history of cancers at a molecular level so that more focused cancer prevention strategies can be implemented that circumvent the current problems of over-diagnosis inherent in mass screening programs. Barrett’s esophagus is the precursor lesion to the aggressive cancer EAC. In a minority of patients, estimated at 0.33% per year, Barrett’s can progress from non-dysplastic Barrett’s esophagus, through intermediate stages of low-grade dysplasia (LGD) and high-grade dysplasia (HGD), to adenocarcinoma. This condition is a classic example of a disease in which clinical strategies characterized by endoscopic random biopsy sampling and pathological diagnosis have arguably failed to improve outcomes for patients.5,6

Historically, studies have focused on mutation, methylation changes and/or loss of heterozygosity (LOH) of specific target genes, most commonly CDKN2A (p16) and TP53, to try to understand the natural history of Barrett’s esophagus and to predict patients at high risk of cancer.6-10 However, there are conflicting interpretations of these data concerning the clonal evolution of this disease. One model formulated by Maley et al.7,11 proposed that a mutation (most commonly inactivation of CDKN2A) that confers a selective advantage to a cell will sweep across the Barrett’s segment, resulting in this mutation being present in the majority of the cells in that Barrett’s segment, in a so-called selective sweep. As additional advantageous mutations arise (commonly TP53 loss), the resulting cell clones can then also expand across the Barrett’s segment. This scenario results in a cancer with serial accumulation of mutations, including drivers and hitchhikers, a proportion of which would be present in all Barrett’s epithelium. Leedham et al.6 have subsequently described a more heterogeneous model whereby multiple independent clones arise, some of which die out and some of which are maintained. Hence, genetic aberrations present within these clones would not necessarily sweep across the whole Barrett’s segment, but such expansion would depend on the competitive advantage of the individual clones. This scenario could lead to far more heterogeneous Barrett’s esophagus and cancer tissue. Although the two different models are not mutually exclusive, the limited resolution of molecular genetic alterations, stemming from analysis of a small number of candidate genes, has made it difficult to resolve the issue.

Recent DNA sequencing studies have demonstrated a high mutational burden in EAC13–16 and have described distinct mutational signatures within this cancer type13,16,17. TP53 is by far the most commonly mutated gene in EAC, followed by a plethora of genes that are mutated in a smaller proportion of cases (<25%), such as ARID1A, SMARCA4, SMAD4 and SYNE1 (refs. 15,16). In a study

1Medical Research Council Cancer Unit, Hutchison/Medical Research Council Research Centre, University of Cambridge, Cambridge, UK. 2Illumina, Chesterford Research Park, Little Chesterford, UK. 3Department of Histopathology, Addenbrooke’s Hospital, Cambridge, Cambridge, UK. 4Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. 5A full list of members and affiliations appears at the end of the paper. 6These authors contributed equally to this work. Correspondence should be addressed to R.C.F. (rcf29@mrc-cu.cam.ac.uk).

Received 16 January; accepted 23 June; published online 20 July 2015; doi:10.1038/ng.3357
using exome sequencing data from two matched Barrett’s esophagus and EAC samples, Agrawal et al.\textsuperscript{14} found that approximately 80% of the mutations in the cancer samples were already present in DNA from the adjacent Barrett’s epithelium. However, the grade of dysplasia and the spatial relationship of these samples were not reported. We recently demonstrated that putative esophageal driver genes, such as \textit{ARID1A} and \textit{SMARCA4}, were also recurrently mutated in patients with a very stable phenotype who had never shown any evidence of dysplasia in their Barrett’s segment over multiple years of follow-up (median of 58 months)\textsuperscript{13}. This finding would argue against these genes having a causal role in cancer progression. In contrast, \textit{TP53} and \textit{SMAD4} mutations were highly specific to patients with HGD and EAC, respectively. However, although informative, these studies were not designed to deduce clonal architecture over time and space for Barrett’s esophagus carcinogenesis.

The aims of this study were therefore to examine the clonal ordering and heterogeneity of Barrett’s esophagus (including degrees of dysplasia) that had progressed to EAC. We were able to examine these aspects to a higher level of detail than had been previously possible by virtue of some highly characterized sample sets and recent technological developments that permit genome-wide sequencing from minute pieces of paraffin-embedded archival material, as well as fresh-frozen tissues. Thus, we interrogated the mutational landscape, on a genome-wide scale, of 23 paired Barrett’s esophagus and EAC samples as well as 73 samples taken over a 3-year period from one patient’s Barrett’s segment (displaying all the different stages of progression, from non-dysplastic Barrett’s esophagus to intramucosal adenocarcinoma). From these data, we were able to determine the mutational load and mutational context as well as the clonal composition and heterogeneity in pathologically defined steps of Barrett’s esophagus carcinogenesis.

**RESULTS**

**Barrett’s esophagus is highly mutated and polyclonal**

Paired Barrett’s esophagus and cancer samples taken at the same time point from 23 patients with EAC (see Supplementary Tables 1 and 2 for demographic and tumor information) with macroscopically visible Barrett’s esophagus were submitted to whole-genome sequencing. All Barrett’s and EAC samples were sequenced to a minimum of 54-fold coverage, and the germline, normal comparison sample (blood or normal esophageal squamous) was sequenced to a minimum of 31-fold coverage (Supplementary Table 3). The median numbers of single-nucleotide variants (SNVs) present in the EAC and Barrett’s esophagus samples were 18,786 (interquartile range (IQR) of 15,007–32,034) and 12,714 (IQR of 6,604–21,559), respectively. The average somatic mutation rate for the Barrett’s samples was 6.76 SNVs/Mb, which is higher than the rates for multiple myeloma (2.9 SNVs/Mb)\textsuperscript{18}, luminal breast cancer (1.1 SNVs/Mb)\textsuperscript{19}, hepatocellular carcinoma (3.69 SNVs/Mb)\textsuperscript{20} and colorectal adenocarcinoma (5.9 SNVs/Mb)\textsuperscript{21}. Although Barrett’s esophagus adjacent to EAC was found to be highly mutated and contained thousands of SNVs, even samples with no histological dysplasia, there were significantly more
SNVs called in the tumor samples (average somatic mutation rate of 10.02 SNVs/Mb) than in the Barrett’s esophagus samples (Wilcoxon signed-rank test, \( P < 0.001 \)) (Fig. 1a). Furthermore, this was apparently not driven by a difference in the purity of the Barrett’s esophagus and tumor samples, as there was no significant correlation between the number of SNVs called and the purity of the samples (Spearman’s rho, \( r = 0.108 \); \( P = 0.475 \)) (Supplementary Table 3). Of note, the Barrett’s esophagus samples with dysplasia (either LGD, HGD or indefinite for dysplasia; called by two independent expert pathologists) did not have significantly more mutations than the Barrett’s esophagus samples with no dysplasia (Mann-Whitney test, \( P = 0.271 \)).

Surprisingly, when we looked at the percentage of SNVs that were common to both the paired Barrett’s esophagus and EAC samples, after performing additional filtering and only considering SNVs with good coverage in both paired samples (Online Methods), we found a lower degree of overlap than we would have first expected (<20% overlap between the SNVs in the paired Barrett’s esophagus and EAC samples in 13/23 (57%) of the samples), and this low degree of overlap was not primarily due to LOH in the paired sample (Fig. 1b) or differences in purity or sequencing depth between the paired samples. The Barrett’s esophagus samples that had better overlap with their paired EAC samples were more likely to be dysplastic (Mann–Whitney test, \( P = 0.019 \)). We did not find any other association between samples with a poor overlap and any clinical features, such as stage or differentiation status of the tumor, age or sex of the patient, and the length of the Barrett’s segment or the tumor (\( P > 0.5 \) for all comparisons). As expected, TP53 was the most recurrently mutated gene, with mutations occurring in 19 of the 23 (82.6%) EAC samples. Barrett’s esophagus samples harbored TP53 mutations less commonly (9/23 (39.1%), of which 5 were dysplastic). Of note, the TP53 mutations present in Barrett’s esophagus adjacent to EAC were not always present in the paired EAC sample (4/9 cases); however, of these 4 cases, 3 had a different TP53 mutation in the paired EAC sample (Barrett’s esophagus private mutations in Fig. 1c). Other previously reported putative EAC driver genes, such as EYS, ARID1A and ABCB1, were mutated less commonly (<30%) and, if mutated in a patient’s tissue (either Barrett’s esophagus or EAC), were seldom shared (21/73; 28.8%) by the paired Barrett’s esophagus and EAC samples (Fig. 1c).

To be sure that the poor overlap was not due to sampling bias, we performed whole-genome sequencing on additional samples taken from 5 of the 23 cases. These data showed a very similar degree of overlap to the paired Barrett’s esophagus and EAC data (Fig. 1), regardless of whether we considered all samples together or any two samples from the Barrett’s esophagus in comparison to the EAC from the same patient (Supplementary Fig. 1). Taken together, these data show that there is considerable heterogeneity in the spectrum of mutations, with surprisingly little overlap in the molecular genetics of EAC and adjacent Barrett’s esophagus.

**Copy number increases as EAC develops**

Although Barrett’s esophagus adjacent to EAC was found to be highly mutated, we observed a stark contrast between the copy number aberrations in the Barrett’s esophagus and EAC samples (Fig. 2a,b and Supplementary Fig. 2). With the exception of two Barrett’s esophagus samples (from patients P3 and P21), both of which showed features in keeping with LGD, the Barrett’s esophagus samples, even those with HGD (patients P17 and P22), contained very few copy number changes, with the vast majority of their genomes being diploid (median number of at least 5. (b) Stacked mountain plots summarizing the copy number variation in all 23 Barrett’s esophagus and EAC samples. Gains of at least two copies on top of the normal copy number in a region are denoted by yellow-orange-red mountains, and deletions are represented by green-blue valleys. The height or depth of the corresponding mountain or valley indicates the summed copy number status across all patients for that region. The colors represent different samples such that the greater the number of different colors in a region, the greater the number of samples that display that copy number change.

Figure 2  EAC samples display multiple copy number changes in comparison to paired Barrett’s esophagus samples. (a) Graphs showing the percentage of the genome at different copy number states for each patient (P) in turn (P1–P23) for the paired Barrett’s esophagus (left graph) and EAC (right graph) samples. CN0, copy number 0; CN1, copy number 1; CN2, copy number 2; CN3, copy number 3; CN4, copy number 4; CN5+, copy number of at least 5. (b) Stacked mountain plots summarizing the copy number variation in all 23 Barrett’s esophagus and EAC samples. Gains of at least two copies on top of the normal copy number in a region are denoted by yellow-orange-red mountains, and deletions are represented by green-blue valleys. The height or depth of the corresponding mountain or valley indicates the summed copy number status across all patients for that region. The colors represent different samples such that the greater the number of different colors in a region, the greater the number of samples that display that copy number change.
The mutational context is similar for early and late SNVs. (a) Heat maps showing log-transformed values representing the fraction of each mutation type in each trinucleotide context corrected for the frequency of each trinucleotide in the reference genome, as described in Nik-Zainal et al.\textsuperscript{27}. The values for the mutational contexts were calculated separately for the three subsets of SNVs for each patient: (i) Barrett’s-unique SNVs not found in a region of LOH in the EAC sample, (ii) SNVs common to the Barrett’s and EAC samples, and (iii) EAC-unique SNVs not found in a region of LOH in the Barrett’s sample. Teal, SNVs common to both EAC and Barrett’s; pink, SNVs unique to Barrett’s; gray, SNVs unique to EAC. (b) Mutational context plotted as a dot plot showing enrichment of the trinucleotide mutational contexts for every possible option for all 23 paired Barrett’s and EAC samples.
percentage of the genome with copy number 2 = 99.7%; range of 62.8–100.0%). This scenario was significantly different from that for the EAC samples (Wilcoxon signed-rank test, P < 0.001), which showed a range of copy numbers, including some highly amplified regions (15–20 copies), with the median percentage of the genome that was not diploid equal to 37.6% (range of 2.1–87.8%).

The only common copy number change in the Barrett’s samples was 9p LOH, which was present in 11 of 23 (48%) samples. The majority of EAC samples (16/23; 70%) also had 9p LOH. However, of the 11 Barrett’s samples with 9p LOH, only 7 of the corresponding cancer samples also had the same alteration. In the EAC samples, when considering more focal copy number changes (those involving less than half of the particular chromosome arm), we found 17 commonly amplified regions (with at least 4 copies) and 18 deleted regions that were present in at least 3 of the 23 (13%) samples. The commonly amplified regions included previously reported amplified regions in EAC, namely GATA4, KLF5, MYB, PRKCI, CCND1, FGFR3, FGFR4, FGFR19 and VEGFA, some of which are potential therapeutic targets. The deleted regions included some known fragile sites (PHIT and WWOX) as well as some previously reported22 and potentially interesting targets, namely RBFOX1, CDKN2A, PDE4D, PTPRD and PARK2.

Mutational context is unchanged between Barrett’s and EAC.

From the SNV overlap data, we could infer early SNVs (those present in both the paired Barrett’s and EAC samples) as well as later events (SNVs unique to either the Barrett’s or tumor sample). From these categorizations, we wanted to know whether there was a common mutagenic stimulus for the early and late events. The previously identified EAC mutational signature (A>T>G specifically at AAG trinucleotides)13,16 was similarly enriched for both the early and late SNVs (both Barrett’s and EAC unique) (Fig. 3a,b). There was good correlation when comparing the mutational contexts of all the Barrett’s-unique, EAC-unique and common SNVs (average correlation of 0.83, 0.86 and 0.86, respectively); however, there appeared to be some difference between the early and late SNVs, albeit small (6.9% variance in the second component), when using principal-component analysis (Supplementary Fig. 3a). This difference seemed to be driven mainly by four different mutational patterns: A→C in a CAC context, C→G in a CCG context, A→T in a CAC context and C→G in a TCA context (Supplementary Fig. 3b).

Spatial and temporal characterization of a Barrett’s segment

As we observed such heterogeneity in the genetic landscape between the paired Barrett’s and EAC samples, we sought to characterize, in precise detail, the Barrett’s segment from a single patient (AHM1051) to better understand clonal evolution and how this relates to cancer development. For this analysis, we studied multiple samples from a 58-year-old male patient who displayed all the stages of the Barrett’s esophagus progression series: gastric metaplasia, intestinal metaplasia, LGD, HGD and intramucosal adenocarcinoma. Following a tertiary referral, this patient underwent five endoscopies between May 2009 and March 2012, and samples were available across the full length of the patient’s 10-cm Barrett’s segment for this time period (outlined in Fig. 4). This patient was also selected as he had previously swallowed a Cytosponge, which is a non-endoscopic cell sampling device that collects cells from the gastroesophageal junction and the entire length of the esophagus, as well as the oropharynx, providing an opportunity to assess clonal architecture from a single sample, which may have clinical application.

Somatic SNVs (both coding (variant allele fraction (VAF) > 0.08) and noncoding (present at the highest VAFs in the different Barrett’s esophagus whole-genome sequencing libraries)) were selected for targeted resequencing, with the expectation that these SNVs would define the clones present in this patient’s Barrett’s segment and could be assessed across a larger number of samples. Using a panel of 1,443 targets, we performed sequencing on 73 individual samples (Fig. 4). We used a custom-made data browser to investigate this complex data set (available with the Supplementary Data Set). All but 2 of the 1,443 SNVs gave useful data with sufficient sequencing depth for the 73 samples with a median coverage of 7,760× (IQR of 6,015–10,100×) for the samples and a median coverage of 6,504× (IQR of 3,285–10,940×) for the 1,441 targets. Of the selected SNVs, 99.7% (1,437/1,441) were verified to be real and somatic (VAF > 0.01).

Clonal heterogeneity within non-dysplastic Barrett’s

Of the patient’s sequenced Barrett’s samples, 22 biopsies showed no evidence of dysplasia. These samples were particularly interesting as we wanted to determine the clonal ordering of low-risk Barrett’s epithelium that looked cytologically normal. Of these biopsies, one had gastric metaplasia and all the rest contained intestinal metaplasia. We performed an initial analysis using Pearson correlation of VAF values from the targeted amplicon sequencing. This method allowed us to identify clones on the basis of their overall mutation patterns, taking advantage of the large proportion of VAFs that were affected by the extensive copy number changes present in a number of the patient’s samples. Because this method assigns groups on the basis of VAF patterns, it is more resistant to variable clonality and purity. The analysis of the 22 Barrett’s samples identified six different groups, and by pairwise comparison of the samples we derived the clonal hierarchy (Figs. 5 and 6a).
Three SNVs, all occurring in noncoding regions and representing the most recent common ancestor (chromosome 4: g.33658353T>C; chromosome 13: g.88285478A>C; and chromosome 18: g.11483749A>C), were found in all six clones, providing evidence for an initial clonal sweep of the patient’s Barrett’s segment. After this initial clonal sweep, two very different clones arose, clone 1 and clone 3 (Fig. 5b, c), displaying many more genome-wide SNVs (clone 1 had 819 of the 1,437 SNVs assessed, and clone 3 had 157 of the 1,437 SNVs assessed). Clone 3 had 9p LOH, a common event observed in non-dysplastic Barrett’s esophagus and one that has previously been shown to impart selective advantage to Barrett’s esophagus cells; however, this clone did not appear to seed further clones in this patient’s Barrett’s segment (Fig. 5c). Unlike clone 3, clone 1 was able to seed a daughter clone, which we have called clone 2. Clone 2 contained 234 additional SNVs in comparison to clone 1 (colored orange in Fig. 5b) and subsequently gave rise to two different clones, namely clone 4 and clone 6. Clone 4, containing 1,178 of the 1,437 SNVs assessed, subsequently gave rise to an additional clone, clone 5 (having 1,184 of the 1,437 SNVs assessed as well as a large-scale deletion-amplification on chromosome 15). Clone 6 (with 1,146 of the 1,437 SNVs assessed), in addition to the 93 additional SNVs in comparison to clone 2, displayed a large number of copy number changes, specifically large-scale deletions on chromosomes 5, 11, 13 and 18, as evidenced by the jagged appearance of the VAF plots in Figure 5a and supported by the whole-genome sequencing data (Supplementary Fig. 4). This finding is especially interesting as all three samples that were classified as clone 6 were assigned a histopathological diagnosis of non-dysplastic Barrett’s esophagus although they contained a large number of somatic mutations and copy number changes.

More information on how the clonal hierarchy was derived can be found in the Online Methods. It should be noted that whether the sample sequenced was fresh frozen or paraffin embedded did not lead to any systematic bias that affected the clonal assignments.

**Dysplasia can develop from multiple different clones**

In addition to multiple regions of Barrett’s esophagus with no dysplasia, patient AHM1051 had multiple areas with cellular abnormalities that were graded according to the degree of dysplasia by consensus review of two expert gastrointestinal pathologists. These samples allowed us to assess the clonal architecture in the Barrett’s esophagus progression sequence. Clone 6 (Fig. 5a and Supplementary Fig. 3) displayed the highest number of copy number changes and would be the obvious culprit for seeding HGD. However, interestingly, clones 2, 3 and 6 were all found to give rise to HGD (Fig. 6a–c). Similarly, clones 2, 3, 4 and 6 all appeared to give rise to LGD. Clones 1 and 5 were the exceptions for LGD; however, as there were few samples available from these two clones, this may be a result of sampling bias.
Stability of the Barrett’s esophagus segment over time

Taken together, these data demonstrate a genetically stable clonal pattern in this patient’s 10-cm Barrett’s segment. All six identified clones were present in 2009, the year that this patient was diagnosed as having Barrett’s esophagus with dysplasia. Almost 3 years later, the same six clones could be identified (Fig. 6c). The only alteration to the Barrett’s esophagus segment appeared to have occurred as a result of clinically ineffective endoscopic treatment of this patient, which resulted in the shrinking of clone 3 (only partially visible in one biopsy after treatment). This decrease in clone 3 was further evident using a non-endoscopic cell sampling device, the Cytosponge. Amplicon sequencing of cells collected using the Cytosponge taken after endoscopic treatment (March 2012) showed no evidence of clone 3 (Fig. 6d), suggesting that this clone had substantially shrunk in response to the endoscopic therapy. However, SNVs defining clones 1, 2, 4, 5 and 6 were all present in the Cytosponge sample, and clone 6 was the most abundant, in keeping with the large size of this clone as ascertained from the biopsy data (Fig. 6d). These findings suggest that the Cytosponge could simultaneously sample all five clones present in the patient’s Barrett’s segment and provide some indication of relative clone size.

Given the pathogenic importance of TP53, in a more targeted approach, we also mapped mutations at this locus in the Barrett’s segment from this patient. TP53 sequencing on the Cytosponge sample13 identified a dominant TP53 mutation (chromosome 17: g.7577538C>T) that was also identified by whole-genome sequencing. This particular TP53 mutation established itself in clone 2 (VAF = 0.28–0.56) and was therefore also present in all the daughter clones (that is, clones 4, 5 and 6) (shown by the slanted white lines in Fig. 6c), however, with a much higher VAF in clone 6 (41–73%), probably owing to LOH. The presence of a TP53 mutation in non-dysplastic Barrett’s esophagus from a patient with early-onset cancer is similar to findings from the whole-genome sequencing data that showed that 4 of 17 (23.5%) non-dysplastic Barrett’s samples taken adjacent to EAC contained TP53 mutations. These data from patients who have progressed to cancer should not be confused with studies that look at the TP53 mutation prevalence in non-dysplastic Barrett’s esophagus from patients who never progress to dysplasia and/or EAC13,23,24.

DISCUSSION

In summary, the application of powerful sequencing technology to cases of Barrett’s-associated carcinogenesis has led to new insights about the life history of the disease. First, there are numerous somatic point mutations as well as small insertions and deletions that occur...
in all pathological stages of disease progression. The specific SNVs generally overlap poorly between paired Barrett’s and EAC samples; however, the mutational context of these SNVs is mostly common throughout the course of the disease (from non-dysplastic Barrett’s esophagus to EAC), suggesting exposure to common mutagens throughout the progression sequence. Although the mutational context is consistent along the progression sequence, there is a marked increase in the number of copy number changes in EAC, which is rarely seen in Barrett’s epithelium. From a clinical perspective, the histopathological assessment of dysplasia appears to be a poor reflection of the molecular disarray in Barrett’s epithelium, as the same aberrant genetic profile was seen in both dysplastic and non-dysplastic Barrett’s tissue.

The variation in the mutational overlap between the paired Barrett’s and EAC samples in this study, especially the high proportion of patient samples that showed such heterogeneity (13/23 Barrett’s-EAC pairs with <20% SNV overlap), was surprising. However, these data do demonstrate how much the Barrett’s segment has evolved since seeding the tumor, giving an indication of the long natural history of the condition. One way to capture this genetic heterogeneity in patients with a patent lumen is to use the non-biased Cysponge, which we have shown could sample all five clones remaining after endoscopic therapy in the Barrett’s segment from patient AHM1051. This is a proof-of-principle experiment, and further work is required to demonstrate the clinical usefulness of this sampling approach for detecting mutations representative of the entire Barrett’s segment.

The in-depth study of the Barrett’s segment of patient AHM1051 led to a detailed map of clonal ordering and heterogeneity in a 10-cm Barrett’s segment. All six clones identified in this Barrett’s segment contain multiple SNVs and have varying abilities to seed further daughter clones. Furthermore, not all of the six clones span the whole length of the Barrett’s segment, but some appear to remain more localized. The clonal pattern present in this patient has aspects of both models that have previously been proposed for the clonal evolution of Barrett’s esophagus6,7. There is evidence of a common ancestor before branched evolution in the patient’s Barrett’s segment, indicated by the presence of three noncoding mutations in all Barrett’s esophagus samples. This model is also supported by the common SNVs identified for the 23 paired Barrett’s esophagus and EAC samples. However, there is also evidence for the emergence of distinct clones, such as clones 1 and 3, neither of which were able to sweep across the whole Barrett’s segment. Furthermore, deep sequencing of individual Barrett’s esophagus biopsies has allowed the identification of multiple clones in the same biopsy, demonstrating that this approach is sufficient to identify the clonal heterogeneity in Barrett’s esophagus without requiring further microdissection.

This study is only one step in further understanding clonal evolution and heterogeneity in Barrett’s esophagus. The ultimate goal is still to be able to predict which patients will develop dysplasia and, ultimately, cancer. Approaches such as the one used here have the ability to further improve understanding of the clonal structure in Barrett’s esophagus before and after the development of dysplasia, with the hope of being able to predict which patients will progress to cancer. Li et al.25 recently reported that patients with Barrett’s esophagus who do not progress to cancer have stable genomes, whereas patients who progress to cancer have unstable and diverse genomes that evolve at least 4 years before the development of cancer. However, our data suggest a more complex situation, as HGID can arise from multiple different clones, some of which appear to have very few copy number aberrations. The clinical implication of these findings is that, when endoscopic therapies, such as mucosal resection or ablation, are undertaken, there is a need to treat the whole Barrett’s esophagus segment to ensure that no residual Barrett’s epithelium remains. This is important, as meta-analysis of 18 studies demonstrated that Barrett’s esophagus with intestinal metaplasia was still present in 22% of individuals (95% confidence interval of 14–30%) who were treated with radiofrequency ablation26.

In conclusion, this study has identified genetic aberrations that may contribute to the progression from Barrett’s esophagus to EAC. In the future, it will be important to integrate transcriptomic and epigenetic data with genome-wide DNA sequence data for patients who span the disease spectrum. This approach, which incorporates state-of-the-art technology on fresh and archival specimens, paves the way for further studies in other tissue types aiming to chart the progress from intraepithelial neoplasia to invasive cancer.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The whole-genome sequencing data can be found at the European Genome-phenome Archive (EGA) under accession EGAD00001001394. Information to match the sample identifiers to the patients presented in this manuscript is provided in Supplementary Table 4.

Note: Any Supplementary Information and Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the Human Research Tissue Bank, which is supported by the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. This study was partly funded by a project grant from Cancer Research UK. R.C.F. has programmatic funding from the Medical Research Council and infrastructure support from the Biomedical Research Centre and the Experimental Medicine Centre. We would like to thank all the patients who took part in the study. We thank M. Dunning for his bioinformatics assistance. We thank the Edinburgh Experimental Cancer Medicine Centre.

AUTHOR CONTRIBUTIONS

R.C.F. conceived the overall study and takes responsibility for the data integrity. C.S.R.-I., J.B. and R.K.C. analyzed the data. C.S.R.-I. extracted the samples for patient AHM1051. A.W. developed the targeted sequencing data visualization tool. C.S.R.-I., J.B., R.K.C., H.N., M.J., W., M.R., S.H., D.B. and R.C.F. designed various aspects of the study. H.N. performed the TruSeq Custom Amplicon (TSCA) assay. C.S.R.-I., J.B. and A.G.L. performed the statistical analysis. M.P.D. collected endoscopic samples for patient AHM1051. M.O.D. and S.M. performed the histopathological diagnosis. S.I. developed the copy number pipeline, and R.K.C. and M.H. performed the copy number analysis. Z.K. ran the whole-genome sequencing of patient AHM1051. R.C.F., S.H., D.B. and M.R. supervised the study. C.S.R.-I., J.B., R.K.C. and R.C.F. wrote the manuscript. All authors approved the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Nik-Zainal, S. et al. The life history of 21 breast cancers. Cell 149, 994–1007 (2012).
2. Esserman, L.J. et al. Addressing overdiagnosis and overtreatment in cancer: a prescription for change. Lancet Oncol. 15, e234–e242 (2014).
3. Desai, T.K. et al. The incidence of oesophageal adenocarcinoma in non-dysplastic Barrett’s oesophagus: a meta-analysis. Gut 61, 970–976 (2012).
4. Corley, D.A. et al. Impact of endoscopic surveillance on mortality from Barrett’s esophagus-associated esophageal adenocarcinomas. Gastroenterology 145, 312–319 (2013).
5. Shabazz, N.J. & Hur, C. Garlic, silver bullets, and surveillance upper endoscopy for Barrett’s esophagus. Gastroenterology 145, 273–276 (2013).
The OCCAMS consortium members are: Stephen J Hayes\textsuperscript{7,8}, Yeng Ang\textsuperscript{7,8}, Ian Welch\textsuperscript{9}, Shaun Preston\textsuperscript{10}, Sarah Oakes\textsuperscript{10}, Vicki Save\textsuperscript{11}, Richard Skipworth\textsuperscript{11}, Olga Tucker\textsuperscript{12}, Jim Davies\textsuperscript{13}, Charles Crichton\textsuperscript{13}, Christian Schusterreiter\textsuperscript{13}, Tim Underwood\textsuperscript{14}, Fergus Noble\textsuperscript{14}, Bernard Stacey\textsuperscript{14}, Jamie Kelly\textsuperscript{14}, James Byrne\textsuperscript{14}, Annette Haydon\textsuperscript{14}, Donna Sharland\textsuperscript{14}, Jack Owlsley\textsuperscript{14}, Hugh Barr\textsuperscript{15}, Jesper Lagergren\textsuperscript{16–18}, James Gossage\textsuperscript{16,17}, Andrew Davies\textsuperscript{16,17}, Robert Mason\textsuperscript{16,17}, Fuju Chang\textsuperscript{16,17}, Janine Zylstra\textsuperscript{16,17}, Grant Sanders\textsuperscript{19}, John Saunders\textsuperscript{21}, Vincent Pang\textsuperscript{21}.

21. Dolan, K., Walker, S.J., Gosney, J., Field, J.K. & Sutton, R. TP53 mutations in malignant and premalignant Barrett’s oesophagus. Dis. Esophagus 16, 83–89 (2003).
22. Li, X. et al. Temporal and spatial evolution of somatic chromosomal alterations: a case-cohort study of Barrett’s oesophagus. Cancer Prev. Res. (Phila.) 7, 114–127 (2014).
23. Orman, E.S., Li, N. & Shaheen, N.J. Efficacy and durability of radiofrequency ablation for Barrett’s Esophagus: systematic review and meta-analysis. Clin. Gastroenterol. Hepatol. 11, 1245–1255 (2013).
24. 27. Nik-Zainal, S. et al. Mutational processes modeling the genomes of 21 breast cancers. Cell 149, 979–993 (2012).
ONLINE MATERIALS

Patients, clinical material and consent. This study was approved by the Institutional Ethics Committees (REC numbers 07/H0305/52, 10/H0305/1 and 10/H0308/71), and all patients gave individual informed consent. Patients with EAC were recruited prospectively at six different centers (Addenbrooke’s Hospital, Cambridge; Royal Surrey County Hospital, Guildford; St Thomas’s Hospital, London; Gloucester Royal Hospital, Gloucester; Edinburgh Royal Infirmary, Edinburgh; and Salford Royal Infirmary, Manchester), and samples (both cancer and Barrett’s esophagus) were obtained at the same time from surgical resection, endoscopic ultrasound or endoscopic mucosal resection. Blood (for 22/23 patients with EAC) or frozen normal squamous esophageal samples at least 5 cm distant from the tumor (for 1/23 patients) were used as a germline reference. For the patient with Barrett’s esophagus who was studied in detail (AHM1051), a frozen duodenum sample was used as the germline control. Tissue samples were either snap frozen in liquid nitrogen immediately after collection and stored at –80 °C or fixed in formalin and embedded in paraffin according to the usual clinical protocol. Before DNA extraction, one section was cut from each esophageal tissue sample and hematoxylin and eosin staining was performed to assess the exact histopathology for every individual sample. EAC samples were deemed suitable for DNA extraction only after consensus review by two expert gastrointestinal pathologists confirmed the tumor cellularity to be ≥70%. Where blood was not available for the germline reference, the same review process was applied to the normal esophageal samples to ensure that only squamous epithelium was present. Barrett’s esophagus samples, both frozen and FFPE, were reviewed by two expert gastrointestinal pathologist (M.O. and S.M.) to identify any dysplasia or cancer cells. For the whole-genome sequencing, Barrett’s esophagus samples with any evidence of cancer cells or cancer samples with any evidence of Barrett’s esophagus were excluded.

DNA extraction from clinical material. For the frozen normal, Barrett’s esophagus and EAC samples, DNA was extracted from the tissue using either the DNeasy kit (Qiagen) or the AllPrep DNA/RNA Mini kit (Qiagen), according to the manufacturer’s instructions. DNA was extracted from blood samples using the Nucleon Genomic Extraction kit (Gen-Probe,) according to the manufacturer’s instructions. For the Barrett’s esophagus FFPE biopsy samples, genomic DNA was extracted from the rest of the diagnostic biopsy, and each diagnostic biopsy was extracted separately. For the Cytosponge sample, genomic DNA was extracted from blood samples using the Nucleon Genomic Extraction kit (Gen-Probe,) according to the manufacturer’s instructions. DNA was extracted from blood samples using the DNeasy kit (Qiagen) or the AllPrep DNA/RNA Mini kit (Qiagen). The protocol was followed as described by the manufacturer, with the exceptions that samples were incubated at 56 °C for 24 h instead of the described 1 h and 10 µl of extra proteinase K was added to the samples roughly halfway through the 24-h incubation. For all sample types, the DNA was quantified using the Qubit dsDNA BR or HS Assay kit (Life Technologies).

Whole-genome sequencing analysis. For the patients with EAC, a normal sample, a Barrett’s esophagus sample and a cancer sample were sequenced for each of the 23 patients (see Supplementary Tables 1–4). For 5 of 23 patients, additional Barrett’s esophagus and EAC samples were also sequenced. For patient, AHM1051 with Barrett’s esophagus, 13 samples were sequenced, of which 10 were from fresh-frozen biopsies and 3 were from FFPE diagnostic biopsies. The 13 samples included 2 normal esophageal squamous samples, 1 duodenum sample, 3 Barrett’s esophagus samples with intestinal metaplasia, 2 Barrett’s esophagus samples with gastric metaplasia, 2 Barrett’s esophagus samples with LGD, 2 Barrett’s esophagus samples with HGD and 1 sample with intramucosal adenocarcinoma. The 13 samples were collected between May 2009 and March 2012. For all fresh-frozen samples, libraries were constructed with an insert length of ~300 bp. For all FFPE samples, library preparation followed a modified TruSeq PCR-Free protocol designed to retain more fragments of low molecular weight, an abundance of which are often present in DNA derived from FFPE samples as a result of degradation of the sample. Whole-genome sequencing was performed on the Illumina HiSeq 2010 instrument. We generated 100-bp paired-end sequence reads using v3 clustering and sequencing chemistry. Alignment to human GRCh37.1 and quality control were performed using the Illumina CASAVA v1.8 pipeline for the whole-genome sequencing data from patient AHM1051 and using ISAAC28 for all samples of the patients with EAC. Identification of somatic SNVs and small somatic indels (<50 bp) was performed by Strelka28 in the 12 Barrett’s esophagus samples from patient AHM1051 using the duodenum sample as the matching normal sample and in both cancer and Barrett’s esophagus samples using the matched normal for the selection of 1,801 SNVs that were deemed to have the potential to define clones within this patient’s Barrett’s esophagus segment. After running Illumina’s Design Studio software for the TruSeq Custom Amplicon (TSCA) Protocol, 1,443 loci were deemed fit for the TSCA Protocol (Illumina).

Targeted sequencing, using the TSCA protocol followed by high-throughput sequencing on the HiSeq 2000, was performed on 73 individual samples (70 Barrett’s esophagus samples and 3 normal samples), including 10 of the original 13 samples used for whole-genome sequencing. Of the 73 samples,
Visualization of the targeted amplicon sequencing data. A custom bioinformatics tool was designed and constructed using the JavaScript library d3.js, to allow the interrogation and exploration of the TSCA data. The tool, available in the Supplementary Data Set, loads three data files: the sample metadata (date of sampling, location down the esophagus, histology report, DNA concentration and TSCA average depth), the sequencing metadata (genomic location, base change and annotation, including gene and consequence when relevant) and the VAF values for all targets in all samples (null values were converted to 0). The tool allows three main interactive explorations of the data: (i) visualization of the VAF values of the targets within each given sample, ordered on the x axis according to genomic coordinates; (ii) a pairwise scatterplot of VAF values between two selected samples; and (iii) a hierarchical tree of the samples. The tree display is dynamic and allows the user to remove samples and/or targets and to choose between five distance metrics between samples.

The five distance metrics (binary, Euclidean, Manhattan, Max and Pearson) are described in the help page of the tool. All analysis results were given using the Pearson metric:

\[
D(A,B) = 1 - \frac{n\sum VAF_i^A \times VAF_i^B - \sum VAF_i^A \times \sum VAF_i^B}{\sqrt{n\sum (VAF_i^A)^2 - (\sum VAF_i^A)^2} \times \sqrt{n\sum (VAF_i^B)^2 - (\sum VAF_i^B)^2}}
\]

where \(D(A,B)\) is the Pearson distance between samples \(A\) and \(B\) and \(VAF_i^A\) is the variant read fraction of mutation \(i\) in sample \(A\). A distance matrix is generated by calculating all pairwise distances between samples. This distance matrix is then used to generate the tree using hierarchical clustering with the complete methodology for linkage of the nodes.

Determining the clonal hierarchy within the Barrett’s segment of patient AHM1051. Clone 3 was distinctly different from the five other identified clones and only had three SNVs in common with the other five clones. Clone 1 was the oldest clone of the other five clones and contained 819 of the 1,437 assessed SNVs. Clone 2 then arose from clone 1, as clones 1 and 2 shared the same 819 SNVs and clone 2 in addition contained 234 SNVs (clone 2 had a total of 1,053 of the 1,437 SNVs assessed). Clone 4 must have arisen from clone 2, as it contained the same 1,053 SNVs as clones 2 as well as 125 additional SNVs that were not present in either clone 1 or 2 (clone 4 had a total of 1,178 of the 1,437 SNVs assessed). Clone 5 then arose from clone 4, as it contained the same 1,178 SNVs as clone 4 as well as 6 additional SNVs and a large-scale deletion on chromosome 15 (as evidenced by the increased VAF in the VAF plot in Fig. 5a). Clone 6 must have arisen from clone 2, as clones 4 and 5 did not share all their SNVs with clone 6. Clones 1 and 2 did share the majority of their SNVs with clone 6 (with the exception of some SNVs that were probably lost through copy number changes present in clone 6). Clone 6 contained an additional 93 SNVs in comparison to clone 2, as well as multiple copy number changes on chromosomes 5, 11, 13 and 18.

Statistical analysis. The Mann–Whitney test was used to compare continuous variables between groups, and a Fisher’s exact test was used to compare counts between categorical variables. All reported \(P\) values were two-sided.

28. Raczy, C. et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics 29, 2041–2043 (2013).
29. Saunders, C.T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
30. Gusnanto, A., Wood, H.M., Pawitan, Y., Rabbitts, P. & Berri, S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics 28, 40–47 (2012).