nature genetics

Paired exome analysis of Barrett's esophagus and adenocarcinoma

Matthew D Stachler[1,2,12], Amaro Taylor-Weiner[3,12], Shouyong Peng[2], Aaron McKenna[4], Agoston T Agoston[1], Robert D Odze[1], Jon M Davison[5], Katie S Nason[5], Massimo Loda[1,2], Ignaty Leshchiner[3], Chip Stewart[3], Petar Stojanov[3], Sara Seepo[3], Michael S Lawrence[3], Daysha Ferrer-Torres[6], Jules Lin[6], Andrew C Chang[6], Stacey B Gabriel[3], Eric S Lander[3,7], David G Beer[6], Gad Getz[3,8,13], Scott L Carter[3,9–11,13] & Adam J Bass[2,3,13]

Barrett's esophagus is thought to progress to esophageal adenocarcinoma (EAC) through a stepwise progression with loss of *CDKN2A* followed by *TP53* inactivation and aneuploidy. Here we present whole-exome sequencing from 25 pairs of EAC and Barrett's esophagus and from 5 patients whose Barrett's esophagus and tumor were extensively sampled. Our analysis showed that oncogene amplification typically occurred as a late event and that *TP53* mutations often occurred early in Barrett's esophagus progression, including in non-dysplastic epithelium. Reanalysis of additional EAC exome data showed that the majority (62.5%) of EACs emerged following genome doubling and that tumors with genomic doubling had different patterns of genomic alterations, with more frequent oncogenic amplification and less frequent inactivation of tumor suppressors, including *CDKN2A*. These data suggest that many EACs emerge not through the gradual accumulation of tumor-suppressor alterations but rather through a more direct path whereby a *TP53*-mutant cell undergoes genome doubling, followed by the acquisition of oncogenic amplifications.

Barrett's esophagus, the intestinalization of the lower esophagus, develops in response to chronic gastric reflux and is the precursor to EAC[1–3]. Whereas Barrett's esophagus is estimated to exist in at least 1:100 adults[4], relatively few progress to cancer. Those that do develop cancer are typically diagnosed at an advanced, incurable stage. Therefore, there is substantial interest in defining the molecular and genomic features of aggressive Barrett's esophagus to enable means to prevent cancer or identify disease when it is most curable. The transformation of Barrett's esophagus into EAC follows the progressive development of increasing grades of dysplasia, leading to invasive carcinoma. Several studies have characterized differences between Barrett's esophagus, dysplasia and EAC, looking at genomic copy number[5–8] and focused analysis of specific cancer-associated genes[5,9–12]. Through studies of Barrett's esophagus, a linear model of progression has emerged in which early, non-dysplastic Barrett's esophagus represents a clonal or polyclonal expansion, typically following inactivation of the tumor suppressor *CDKN2A*[5,7–10]. Further subclonal expansions are thought to occur, often leading to the emergence of a dysplastic clone with *TP53* inactivation and other somatic alterations, including frequent genome doubling and increasing genomic disruption, resulting in malignant transformation[7,9]. We sought to further clarify the process underlying the transformation of Barrett's esophagus into EAC by performing genomic analysis on Barrett's esophagus and EAC samples derived from the same patient. We then extend this analysis to a cohort of previously sequenced EAC samples.

RESULTS

Paired Barrett's esophagus and esophageal adenocarcinoma analysis

We first performed whole-exome sequencing on 25 patient-matched 'trios', including fresh-frozen EAC, Barrett's esophagus and non-malignant, distant gastric or esophageal squamous tissue as a germline comparator (**Supplementary Table 1**). All samples were obtained by surgical resection from patients without previous chemo- or radiotherapy, with Barrett's esophagus intentionally isolated from a region not immediately adjacent to the tumor (when possible) during processing to avoid contamination of the Barrett's esophagus sample with EAC cells. Upon pathological review, 14 Barrett's esophagus samples contained no dysplasia and 11 of the Barrett's esophagus samples showed evidence of dysplasia. Of the 11 dysplastic samples, 6 contained changes consistent with high-grade dysplasia (HGD). After somatic mutation calling (**Supplementary Data Set 1**), we inferred the degree of shared ancestry for the paired Barrett's esophagus and EAC samples

[1]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [3]Eli and Edythe L. Broad Institute, Cambridge, Massachusetts, USA. [4]University of Washington, Seattle, Washington, USA. [5]University of Pittsburgh Cancer Institute, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [6]Section of Thoracic Surgery, University of Michigan, Ann Arbor, Michigan, USA. [7]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [8]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. [9]Joint Center for Cancer Precision Medicine, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Broad Institute of Harvard and MIT, Harvard Medical School, Boston, Massachusetts, USA. [10]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [11]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [12]These authors contributed equally to this work. [13]These authors jointly supervised this work. Correspondence should be addressed to S.L.C. (scarter@broadinstitute.org), G.G. (gadgetz@broadinstitute.org) or A.J.B. (adam_bass@dfci.harvard.edu).

Received 2 February; accepted 29 May; published online 20 July 2015; doi:10.1038/ng.3343

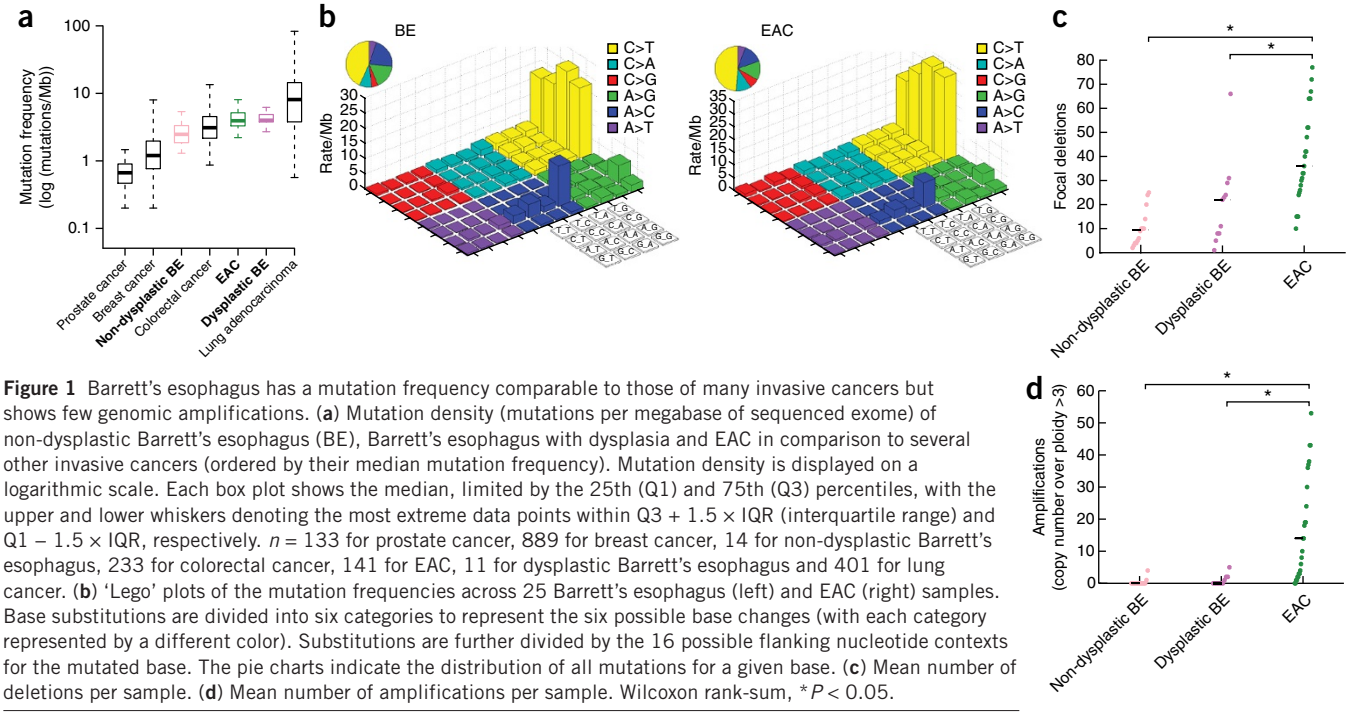

Figure 1 Barrett's esophagus has a mutation frequency comparable to those of many invasive cancers but shows few genomic amplifications. (**a**) Mutation density (mutations per megabase of sequenced exome) of non-dysplastic Barrett's esophagus (BE), Barrett's esophagus with dysplasia and EAC in comparison to several other invasive cancers (ordered by their median mutation frequency). Mutation density is displayed on a logarithmic scale. Each box plot shows the median, limited by the 25th (Q1) and 75th (Q3) percentiles, with the upper and lower whiskers denoting the most extreme data points within Q3 + 1.5 × IQR (interquartile range) and Q1 − 1.5 × IQR, respectively. $n = 133$ for prostate cancer, 889 for breast cancer, 14 for non-dysplastic Barrett's esophagus, 233 for colorectal cancer, 141 for EAC, 11 for dysplastic Barrett's esophagus and 401 for lung cancer. (**b**) 'Lego' plots of the mutation frequencies across 25 Barrett's esophagus (left) and EAC (right) samples. Base substitutions are divided into six categories to represent the six possible base changes (with each category represented by a different color). Substitutions are further divided by the 16 possible flanking nucleotide contexts for the mutated base. The pie charts indicate the distribution of all mutations for a given base. (**c**) Mean number of deletions per sample. (**d**) Mean number of amplifications per sample. Wilcoxon rank-sum, $^*P < 0.05$.

on the basis of the number of shared mutations. In 11 of the 25 trios, the specific region of sequenced Barrett's esophagus appeared to be clonally unrelated to the sampled tumor, as the samples lacked shared somatic mutations (**Supplementary Fig. 1** and **Supplementary Data Set 2**). In addition, hierarchical clustering of the paired samples using somatic copy number alterations (SCNAs) failed to group these unrelated sample pairs together (**Supplementary Fig. 2**). In the remaining 14 trios, the sampled regions of Barrett's esophagus and EAC showed evidence of having emerged from a common neoplastic clone, as they shared 3.4–64% of the coding point mutations with a cancer cell fraction (CCF) of 1 (i.e., mutations present in all neoplastic cells in the tissue sample). Overall, we found no association between the presence of dysplasia and whether the Barrett's esophagus and EAC samples were clonally related (Fisher's exact test, $P = 0.69$). Among the six samples with HGD, five were clonally related to the EAC ($P = 0.18$).

Our Initial analyses compared the degree of genomic disruption for the Barrett's esophagus and EAC samples. The somatic mutation frequencies for Barrett's esophagus ranged from 1.3 to 5.4 mutations/Mb—demonstrating that even non-dysplastic Barrett's esophagus has mutation frequencies higher than those of many common invasive cancers (**Fig. 1a**). Mutational densities increased from non-dysplastic Barrett's esophagus (2.8 mutations/Mb) to dysplastic Barrett's esophagus (4.9 mutations/Mb; Student's *t* test, $P = 0.05$) but were similar for dysplastic Barrett's esophagus and EAC (4.1 mutations/Mb; $P = 0.43$). We also evaluated the base context of mutations in Barrett's esophagus to query the presence of the common pattern of A>C transversions at the 3′ adenine of AA dinucleotides that we described in EAC[13]. The predilection for these transversions was also present in Barrett's esophagus (**Fig. 1b**), suggesting that these mutations occur early during neoplastic progression, possibly as a result of exposure to bile and acid reflux. In the 14 clonally related cases, we separately analyzed the mutation spectra of the early shared mutations in comparison to the spectra for mutations that were private to either the Barrett's esophagus or EAC sample. Although there was a trend for increased A>C mutations private to Barrett's esophagus, this was not statistically significant (**Supplementary Fig. 3**).

Our analysis of structural genomic changes showed that, in contrast to point mutations, structural genomic disruption occurred at higher levels with progression to cancer. The mean number of focal deletions per sample increased steadily between non-dysplastic Barrett's esophagus (10.43), dysplastic Barrett's esophagus (20.73) and the paired EAC (40.20; $P < 0.01$ for both transitions) (**Fig. 1c**). The number of amplifications was strongly associated with progression from dysplastic Barrett's esophagus to EAC (**Fig. 1d**). Non-dysplastic and dysplastic Barrett's esophagus averaged 0.42 and 0.91 amplifications per sample, respectively, whereas EACs averaged 8.44 amplifications per sample ($P < 0.001$ for non-dysplastic BE versus EAC and dysplastic BE versus EAC), suggesting that amplifications might be key mediators of oncogenic transformation. Even between Barrett's esophagus with HGD and EAC we still found a significant increase in the number of amplifications ($P = 0.009$; **Supplementary Fig. 4**). When we lowered the amplification threshold to include lower-level gains, the significant increase in amplification number with progression persisted (**Supplementary Fig. 5**).

We evaluated specific oncogenic alterations within the trios, starting with tumor-suppressor gene (TSG) alterations (**Supplementary Table 2** and **Supplementary Data Set 1**). Of the 11 Barrett's esophagus samples clonally unrelated to the paired EAC sample, only one harbored a *TP53* mutation (**Fig. 2a**). Although four of these Barrett's esophagus cases possessed a homozygous *CDKN2A* deletion, the EACs that emerged in these patients lacked detectable somatic *CDKN2A* alterations. When we next evaluated the clonally related cases, we found that three of the four most distantly related trios (sharing 3.4%, 4.5% and 7.8% of mutations) had *TP53* mutations shared by the Barrett's esophagus and EAC samples (**Fig. 2a**,**b**), indicating that these *TP53* mutations were among the earliest somatic mutations in the development of the tumors. All three of these tumors with early shared *TP53* mutations were determined to have undergone whole-genomic doubling (WGD) using the ABSOLUTE algorithm[14].

Furthermore, in these patients with evidence for an early *TP53* mutation, shared *CDKN2A* somatic alterations were not observed. Seven of the 14 related Barrett's esophagus–EAC cases had shared

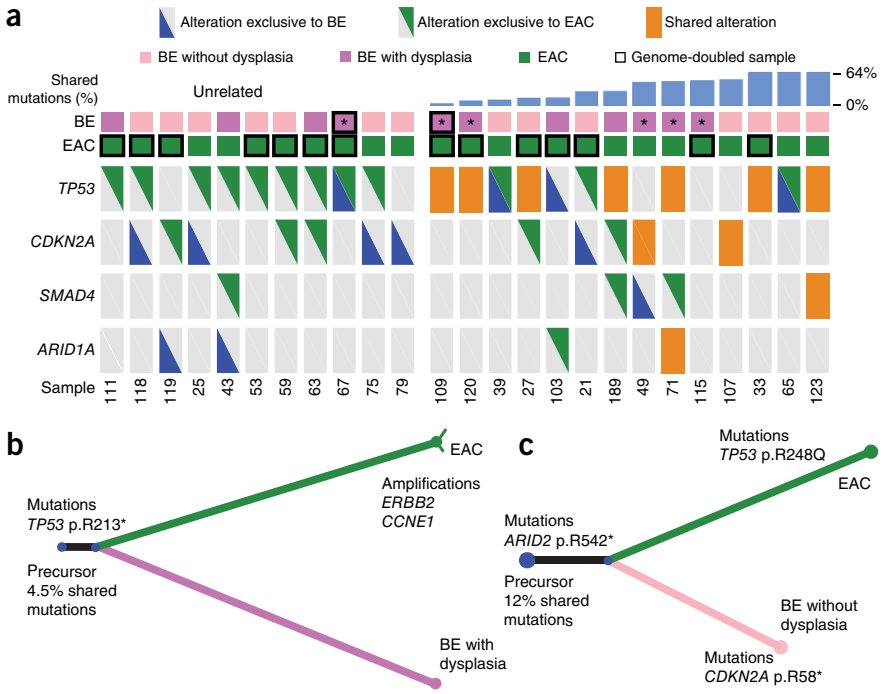

Figure 2 Paired analysis identifies early shared *TP53* alterations. (**a**) Tumor-suppressor plot showing both mutations (heterozygous or homozygous) and homozygous deletions of four commonly altered TSGs in the Barrett's esophagus samples and their paired EACs. Patients are separated into cases where the Barrett's esophagus and EAC samples are clonally unrelated (left) or clonally related (right), with ordering by increasing percentage of shared mutations. An orange box indicates alterations that were shared by the two paired samples, whereas triangles represent alterations private to either the Barrett's esophagus or EAC sample. Sample EAC 71 contained both a shared *ARID1A* alteration and an exclusive *ARID1A* alteration. Black bordering boxes denote samples that have undergone genome doubling, and an asterisk indicates samples suggestive of HGD. (**b**) Example of an evolutionary 'tree' where, despite the paired samples only sharing a small percentage of overall mutations, a *TP53* mutation was found to be one of the early shared events. The lengths of the lines represent the number of mutations in common to the branch according to scale. Thin lines denote alterations with CCF <1. (**c**) Example of an evolutionary tree where a *CDKN2A* mutation occurred late in the Barrett's esophagus sample after the clone that went on to develop into an invasive cancer had already split off.

mutations in *TP53* but not *CDKN2A*, which was either unaltered or had distinct inactivating events in the Barrett's esophagus and EAC samples, suggesting that *TP53* mutation was not preceded by *CDKN2A* inactivation in these cases (**Fig. 2a**,**c**). Only 2 of the 14 related cases appeared to clearly follow the classic model in which the Barrett's esophagus and EAC samples shared a *CDKN2A* alteration but not a *TP53* alteration, indicating that *CDKN2A* inactivation occurred before inactivation of *TP53*. These results suggest that *TP53* mutations may be an earlier event in Barrett's esophagus pathogenesis in relation to other genomic alterations than previously recognized, often preceding (or occurring without) *CDKN2A* inactivation.

In contrast to the prevalent TSG alterations in Barrett's esophagus, oncogenic activation events were far less prevalent in the sampled Barrett's esophagus lesions (**Fig. 3**), even in samples with advanced dysplasia or where the sampled Barrett's esophagus appeared to be closely related to the cancer (**Supplementary Fig. 6**). High-level amplifications of genes encoding oncogenic cell signaling proteins, cell cycle modulators and transcription factors were recurrently present in EACs but more infrequent in Barrett's esophagus (**Fig. 3** and **Supplementary Figs. 7** and **8**). In addition, activating mutations of oncogenes were uncommon in Barrett's esophagus, with only one known activating mutation identified, a *PIK3CA* mutation encoding p.Glu545Gln (**Supplementary Data Set 1**), which was shared with the paired EAC. Oncogenic mutations in the 25 EACs were also uncommon, with 1 *CTNNB1* and 2 *PIK3CA* hotspot mutations identified (**Fig. 3**). Together, oncogenes encoding cell signaling proteins (64% versus 12%; $P = 0.0003$), cell cycle modulators (40% versus 8%; $P = 0.0181$) and transcription factors (64% versus 20%; $P = 0.0037$) were much more likely to have an activating event in the EAC than the paired Barrett's esophagus, respectively. These data suggest that, although Barrett's esophagus harbors common TSG inactivation, oncogene activation is a late step and may mediate transformation to cancer. In the patient sample (63) available for confirmatory testing, FISH analysis for *ERBB2* confirmed our sequencing analysis, showing high-level amplification in the EAC sample and no amplification in the Barrett's esophagus tissue (**Supplementary Fig. 8**).

Expanded multi-sample analysis via laser-capture microdissection

To better understand the progression of Barrett's esophagus to EAC, we identified five additional patients (**Supplementary Table 3**) with a broad field of Barrett's esophagus who underwent surgical esophagectomies (without neoadjuvant therapy), from which we obtained paraffin-embedded samples representing multiple distinct stages of Barrett's esophagus and EAC. We used laser-capture microdissection to isolate normal tissue and pathological samples spanning non-dysplastic Barrett's esophagus, Barrett's esophagus with low-grade dysplasia (LGD), Barrett's esophagus with HGD, EAC and nodal metastatic foci. The number of neoplastic samples characterized per patient ranged from 5 to 11 (**Figs. 4** and **5**). Laser-capture microdissection allowed us to sample regions of Barrett's esophagus in closer proximity to the tumor than was feasible with the earlier trios from fresh tissue. With DNA from these newly dissected samples, we performed whole-exome sequencing and processed the data to identify SCNAs, mutations and genomic doubling events (**Supplementary Fig. 9**). We then constructed phylogenetic trees indicating the evolutionary relationships between the subclones detected in each sample (**Figs. 4** and **5**, and **Supplementary Figs. 10–13**). In three of the five cases, all of the identified Barrett's esophagus regions (both non-dysplastic and dysplastic) were clonally related to the EAC (**Fig. 4**). In the other two cases, all sampled regions of non-dysplastic Barrett's esophagus were clonally unrelated to the EAC, but we identified dysplastic Barrett's esophagus samples related to the sampled tumor (**Fig. 5**). Most individually dissected tissue samples were sufficiently diverged from one another that no sharing of minor subclones occurred (mutations with CCF <1 in two or more samples). However, patients P3 and P7 appeared to contain samples with partially overlapping subclones (**Figs. 4f** and **5b**).

Barrett's esophagus samples from two of the five esophagectomy cases supported our earlier observations that *TP53* mutations might

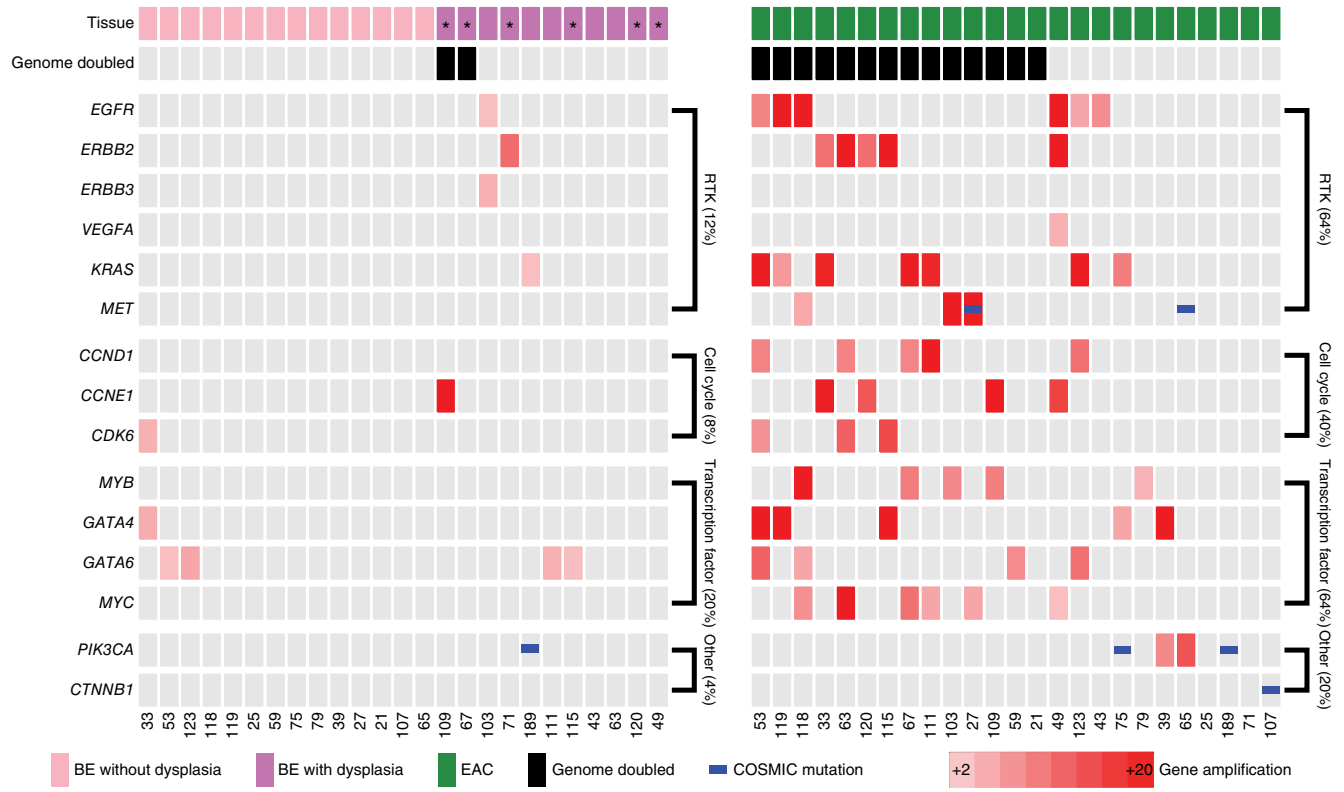

Figure 3 Paired analysis shows a lack of oncogene amplification in Barrett's esophagus samples. The amplification plot shows amplified oncogenes, mutations and pathways in Barrett's esophagus in comparison to EAC, with the genomic doubling status of samples and the presence or absence of dysplasia in the Barrett's esophagus samples marked. An asterisk indicates samples suggestive of HGD. COSMIC, Catalogue of Somatic Mutations in Cancer database.

occur earlier than previously thought in Barrett's esophagus pathogenesis. In patients P1 and P7 (**Fig. 4b**,**f**), *TP53* missense mutations were shared by all sampled tissues, including regions of non-dysplastic Barrett's esophagus (**Supplementary Fig. 14**). Furthermore, these two patients also demonstrated how oncogene amplification can be a late event mediating transformation. In patient P1, we identified a focal *CDK6* amplification present in the two EAC samples but not in the Barrett's esophagus or HGD samples. However, in the *TP53*-mutant regions of non-dysplastic Barrett's esophagus and HGD, we detected a *GATA6* amplification that was absent from the tumor (**Fig. 4b**), demonstrating that amplifications are not exclusive to EAC. In patient P7, we identified a high-level *KRAS* amplification in the tumor and in a focus of HGD immediately adjacent to the cancer (HGD1; **Fig. 4f**). The *KRAS* amplification was notably absent from other regions of LGD and HGD more distant to the tumor (LGD1 and HGD2). FISH analysis for *KRAS* confirmed these findings, with high-level amplification (>25 copies) present in the EAC and amplification absent in HGD2 and LGD1 (**Supplementary Fig. 15**).

Individually dissected samples from patient P7 also demonstrated substantial overlap of distinct subclonal populations. Sample LGD1 contained a major subpopulation (subclone 1a; CCF = 0.80–0.85) that was closely related to the last common ancestor of all neoplastic cells sampled in the patient. In addition, sample LGD1 contained a minor subpopulation (subclone 1b; CCF = 0.15–0.2) defined by 12 additional mutations, as well as single-copy loss of chromosomes 5q, 11p and 13 and copy-neutral loss of heterozygosity (LOH) on chromosomes 7 and 21 (**Supplementary Figs. 9–12**). Subclone 1c was descended from a cell closely resembling subclone 1b, with ten additional mutations; subclone 1c was sampled only in HGD2. Subclone 1c closely resembled the common ancestor of all the subclones sampled in MET1, HGD1, EAC2 and EAC1 (none of which contained shared minor subclones with CCF <1). Sample BE1 contained two major subpopulations (subclone 2a (CCF = 0.4–0.6) and subclone 2b (CCF = 0.5–0.6)). Subclone 2a was descended from a subclone closely resembling subclone 1a, defined by seven additional mutations. Subclone 2b was descended from a cell closely resembling subclone 2a, defined by 23 additional mutations (including an *AKT1* mutation encoding p.Gly10Val) and gains of chromosomes 10 and 18p (**Supplementary Figs. 9–12**). We redrew the phylogenetic tree to represent the relationships between all subclones identified and overlaid the tissue samples onto the tree (**Fig. 4f**).

Patient P4 (**Fig. 4c**,**d**) exemplified a scenario where a *CDKN2A* deletion was present in all sampled regions of Barrett's esophagus and EAC. We therefore inferred that a *CDKN2A*-null progenitor gave rise to five distinct non-dysplastic Barrett's esophagus subclones and one focus of LGD, all lacking *TP53* alteration. However, we identified another subclonal branch in this case, consisting of five HGD and EAC samples all containing a *TP53* mutation (encoding p.Arg175His), focal low-level (two extra copies) amplification of *MET* and genomic doubling, consistent with a cancer in which *TP53* mutation occurred later in clonal evolution, following *CDKN2A* inactivation. Moreover, this case also supports the potential for oncogenic amplification to mediate transformation, as the one EAC sample harbored amplifications at *CCNE1*, *GATA6*, *AKT2* and 3q26 that were absent from all other samples. Samples HGD1, HGD2 and HGD3 all harbored a 7-Mb gain with approximately three extra copies compared to a 4*N* baseline on 17q, including *ERBB2*, which was not present in the EAC.

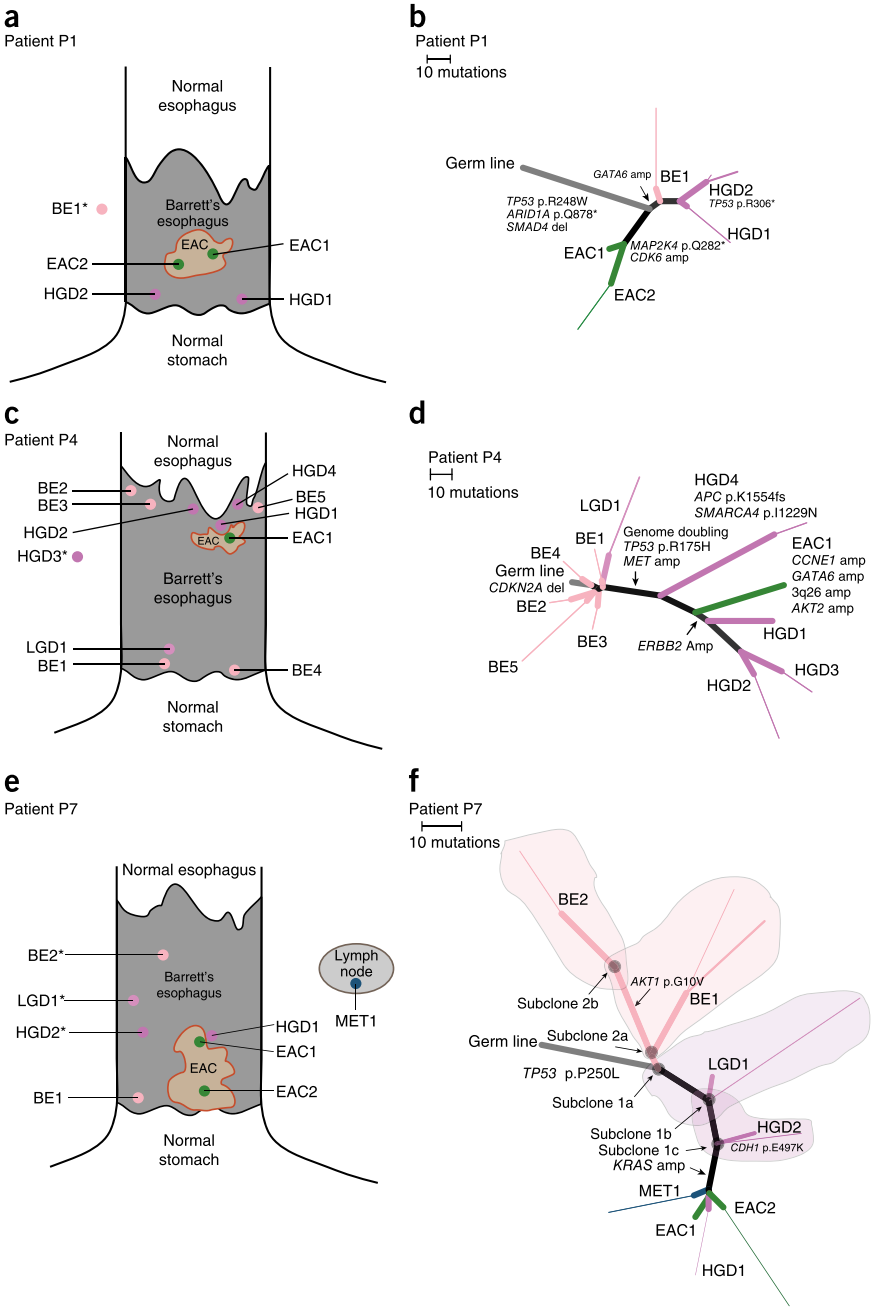

Figure 4 Spatial and phylogenetic relationships of multiple sampled areas, where all samples within a patient share a common set of genomic alterations. (**a**,**c**,**e**) Diagrams showing the relative locations of samples isolated for genomic analysis in patients P1 (**a**), P4 (**c**) and P7 (**e**). Samples immediately adjacent to each other were isolated on the same block. Specimens marked with an asterisk did not contain another sample, other tissue landmark or gross pathologic description to determine where in the field of Barrett's they were located. The sizes of the Barrett's esophagus and EAC samples are roughly proportional to the reported Barrett's esophagus length and tumor size, respectively. (**b**,**d**,**f**) Phylogenetic trees displaying the relationships of the subclones detected in each tissue sample for patients P1 (**b**), P4 (**d**) and P7 (**f**). Branch lengths are proportional to the number of somatic point mutations occurring on that branch. For mutations detected in a single sample, the thickness of the branch is proportional to the CCF of the mutations in that sample. Starting at the germ line, light gray branches represent acquired alterations shared by all samples from a given patient. (**b**) Patient P1 shows a *TP53* mutation shared by all samples as well as a *CDK6* amplification only seen in the EAC samples. (**d**) Patient P4 shows a shared *CDKN2A* deletion in all samples, with *TP53* mutation, WGD and oncogene amplification in the highly related HGD and cancer samples. (**f**) Patient P7 shows a *TP53* mutation shared across all samples and *KRAS* amplification only found in the cancer and a single focus of adjacent HGD. Shaded areas contain the subclones present in the corresponding tissue samples. Amp, amplification; del, deletion.

Analysis of tissue from patients P3 and P6 (**Fig. 5**) demonstrated the potential for heterogeneity within the field of both Barrett's esophagus and EAC. The case for patient P3 contained three separate neoplastic regions, each clonally unrelated to the others. A first region contained an area of non-dysplastic Barrett's esophagus with a focal *CDKN2A* deletion. From this region, a segment of LGD emerged with biallelic loss of *ARID1A*. Separate from this region of Barrett's esophagus were two clonally unrelated primary cancers (**Supplementary Fig. 16**). The sample EAC1 contained a mix of two distinct populations. Subclone 1 harbored 182 unique mutations (including in *SMARCA4*) and additional copy number alterations, including a *CCND1* amplification. EAC1 also contained a subpopulation of cells (subclone 2) that closely resembled the common ancestor of all subclones in the samples of more advanced disease—HGD1, EAC2 and MET1—with 125 mutations unique to this branch and unrelated to subclone 1. HGD1, EAC2 and MET1 also underwent WGD, which was not evident in EAC1–subclone 1. Our focused analysis of this branch of advanced neoplastic tissues demonstrated how oncogene amplification can lead to transformation. The HGD sample contained low-level segmental gains on chromosomes 6, 7, 8 and 17. The EAC and metastasis samples possessed additional amplifications, often occurring on top of the gains in the HGD sample. EAC2 harbored a more focal *MYC* (chromosome 8) amplification and focal amplifications of 3q26 and *MET* (chromosome 7) not identified in the HGD region. In addition, MET1 (from a local lymph node) had increased *ERBB2* (chromosome 17) amplification, demonstrating further heterogeneity.

Our analysis of patient P6 identified one region of clonally distinct non-dysplastic Barrett's esophagus as well as two EAC foci and two HGD regions, which all emerged with WGD and a shared *ATM* nonsense mutation. Interestingly, a focal *CDKN2A* deletion was present in one of the dysplastic samples (HGD1) but not in the HGD2 or EAC sample—another demonstration of late *CDKN2A* inactivation. The two EAC samples shared a *MET* amplification not present in the dysplastic tissue.

Wider analysis of esophageal adenocarcinoma exomes

To validate our findings and better estimate the fraction of EAC cases that do not follow the conventional EAC evolution model, we reanalyzed

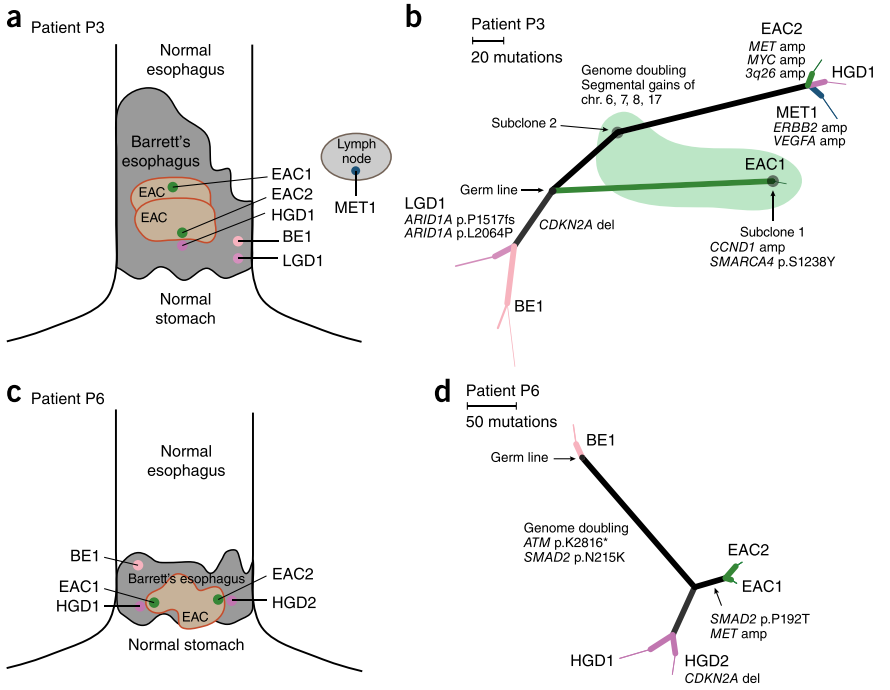

Figure 5 Spatial and phylogenetic relationships of multiple sampled areas in patients showing at least two clonally unrelated populations. (**a**,**c**) Diagrams of sample locations and diagnoses within the patient's field of Barrett's esophagus for patients P3 (**a**) and P6 (**c**). The sizes of the Barrett's esophagus and EAC are roughly proportional to the reported Barrett's esophagus length and tumor size, respectively. Specimens marked with an asterisk did not contain enough information to be properly located. (**b**,**d**) Phylogenetic trees displaying the relationships of the subclones detected in each tissue sample for patients P3 (**b**) and P6 (**d**). Branch lengths are proportional to the number of somatic point mutations occurring on that branch. For mutations detected in a single sample, the thickness of the branch is proportional to the CCF of the mutations in that sample. Black circles mark the starting point (germ line), as there were no somatic alterations common to all samples. (**b**) Patient P3 shows three distinct clonally unrelated branches. The branch with BE2-LGD1 shows a *CDKN2A* deletion in the Barrett's esophagus and LGD samples only. The green shaded region represents tissue sample EAC1, which contains a mixture of distinct cell populations, subclones 1 and 2. (**d**) Patient P6 shows a region of Barrett's esophagus tissue unrelated to the other HGD and cancer samples, all of which share *ATM* and *SMAD2* mutations and WGD.

whole-exome sequencing data from 144 microsatellite-stable EACs that we recently presented[13,15]. Consistent with the conventional progression model, these tumors showed marked aneuploidy[15] and harbored *TP53* mutations in 97 of 144 (67%) cases. However, we identified *CDKN2A* inactivation via mutation or homozygous deletion in only 39 of 144 (27%) tumors[14].

We reanalyzed the EAC whole-exome sequencing profiles using ABSOLUTE[14] to make WGD calls and to determine whether specific mutations likely occurred before or after doubling. The majority (62.5%) of EACs showed evidence of WGD, corroborating previous findings[14]. Although tumors with and without WGD showed different ploidy and genomic disruption, mutation densities were not different between these two groups (**Supplementary Figs. 17** and **18**). *TP53* mutations were distributed evenly across EACs with (67% *TP53* mutant) and without (69% *TP53* mutant) WGD. Within the WGD samples, 54 of 60 (90%) *TP53* mutations were determined to have occurred before doubling. We compared the timing of *TP53* mutations with that inferred for other mutations of TSGs, including *CDKN2A*. *TP53* was the only gene whose mutations were statistically more likely to have occurred before doubling ($P = 8.7 \times 10^{-7}$; **Supplementary Fig. 19** and **Supplementary Table 4**), suggesting a role for *TP53* as an antecedent to genomic doubling[16,17].

We next searched for genomic characteristics that differentiate EACs with and without WGD, finding early evidence for a different spectrum of TSG alterations. *CDKN2A* and *SMAD4* mutations were found in only 8% and 1% of EACs with WGD, respectively, but were present in 19% ($P = 0.065$) and 20% ($P = 0.0001$) of tumors without WGD (**Fig. 6a**). Broadening our TSG analysis, we found that TSG losses involving chromatin modification ($P = 0.005$), the cell cycle ($P = 0.027$) and the transforming growth factor (TGF)-β pathway ($P = 0.0001$) were all more frequent in the EACs without WGD (**Fig. 6b** and **Supplementary Tables 5** and **6**). Although we found that the mean number of deleted segments was not statistically different between tumors with and without WGD (29.13 versus 30.93, respectively; $P = 0.28$), there was a trend for increased numbers of homozygous deletions in the samples without WGD (1.46 for samples without WGD versus 0.867 for samples with WGD; $P = 0.104$; **Supplementary Fig. 20**), as previously observed in serous ovarian cancer[14]. The observation that tumors that emerge from WGD have fewer somatic alterations of TSGs suggested that different alterations might drive transformation in these cases.

Therefore, we next evaluated patterns of oncogene activation, finding that the proportion of alterations predicted to activate oncogenic signaling molecules, including *KRAS*, receptor tyrosine kinases (RTKs) and phosphoinositide 3-kinase (PI3K) signaling, did not differ between the groups with and without WGD (64% versus 54%, $P = 0.22$ for *KRAS* and RTKs; 9% versus 13%, $P = 0.57$ for *PIK3CA*) (**Fig. 7**). However, tumors with WGD showed more frequent amplification of oncogenic transcription factors (43% versus 22%, $P = 0.012$) and a trend toward higher rates of amplification of cell cycle mediators (40% versus 24%, $P = 0.069$). More frequent *CCNE1* amplifications in genome-doubled EACs (16% versus 6%) paralleled recent analyses across tumor types, which identified correlations between *CCNE1* amplification and WGD[18,19]. These findings indicate that the absence or presence of WGD in Barrett's esophagus modifies the most likely pathway available to undergo transformation to cancer.

DISCUSSION

When the results of the EAC reanalysis are interpreted jointly with our earlier data on genomic analysis of paired Barrett's esophagus–EAC cases, the results lead us to refine the previous model of the emergence of EAC from Barrett's esophagus. EAC has been thought to arise from the progressive accumulation of genomic alterations, starting with ones in *CDKN2A* and followed by expansion of a *TP53*-mutant dysplastic clone that is able to develop tetraploidy and genomic instability[7,20]. Consistent with this idea, we confirm that Barrett's esophagus harbors frequent TSG alterations, even in Barrett's esophagus segments clonally unrelated to the cancer, corroborating reports

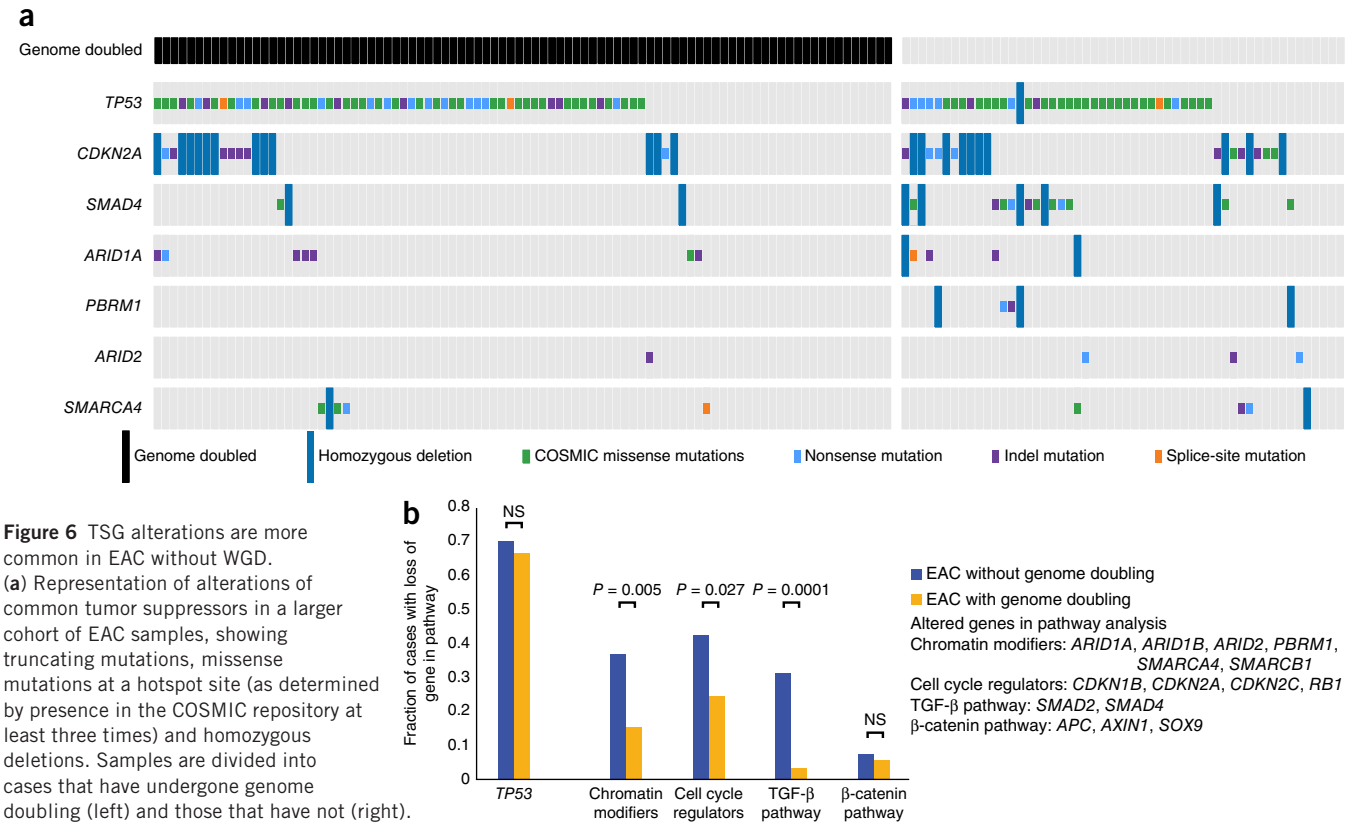

Figure 6 TSG alterations are more common in EAC without WGD. (**a**) Representation of alterations of common tumor suppressors in a larger cohort of EAC samples, showing truncating mutations, missense mutations at a hotspot site (as determined by presence in the COSMIC repository at least three times) and homozygous deletions. Samples are divided into cases that have undergone genome doubling (left) and those that have not (right). The type of mutation identified is represented by the color of the mutation box. (**b**) Expanded analysis of the fraction of cases with and without genome doubling with alterations in the given tumor-suppressor pathways (genes in pathways with multiple identified alterations are listed on the right). Statistically significant differences are highlighted; NS, not significant. Genes in the individual pathways are also shown in **Supplementary Table 5**.

of TSG inactivation in non-dysplastic Barrett's esophagus[21] and clonal diversity within Barrett's esophagus[22]. Our data reinforce models that underscore the importance of *TP53* mutation in the neoplastic progression of Barrett's esophagus[23]. However, our results also suggest that, in many cases, *TP53* mutations occur earlier in the disease process relative to other alterations (including loss of *CDKN2A*) and can be detected in the non-dysplastic Barrett's esophagus of cases that progress to cancer. In addition, our data suggest that oncogene activation via amplification may often be a critical later event in transformation to invasive EAC. This sequence of events is in

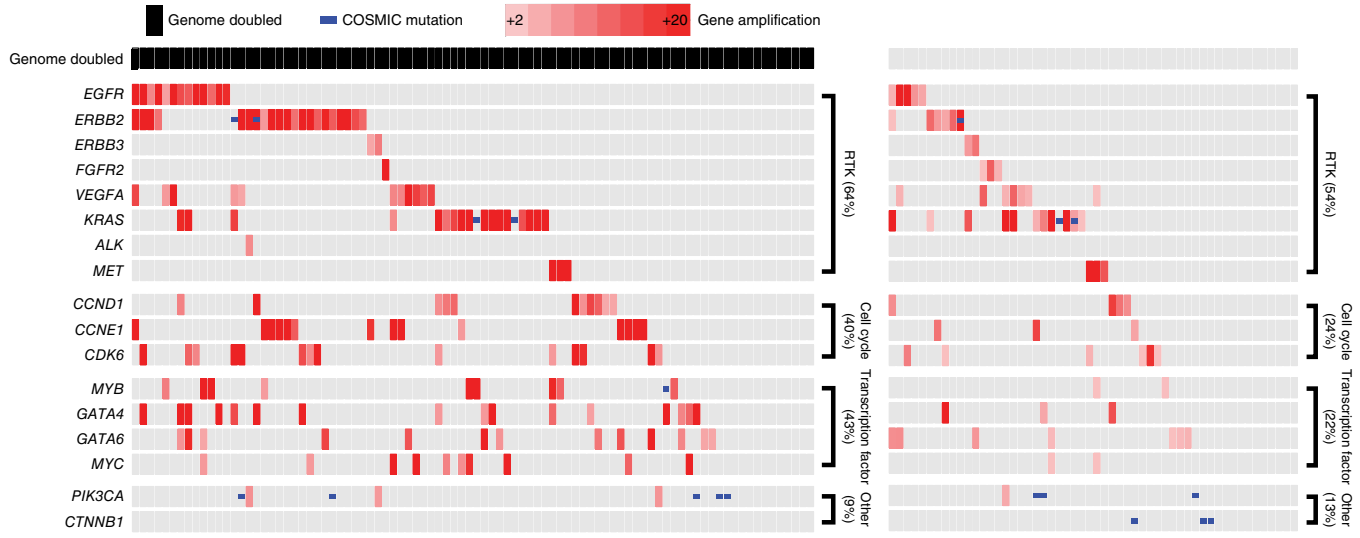

Figure 7 Genome-doubled EAC contains more frequent amplifications in cell cycle regulators and transcription factors. The amplification plot shows amplified oncogenes, mutations and pathways in EAC. Samples are divided into cases that have undergone genome doubling (left) and those that have not (right).

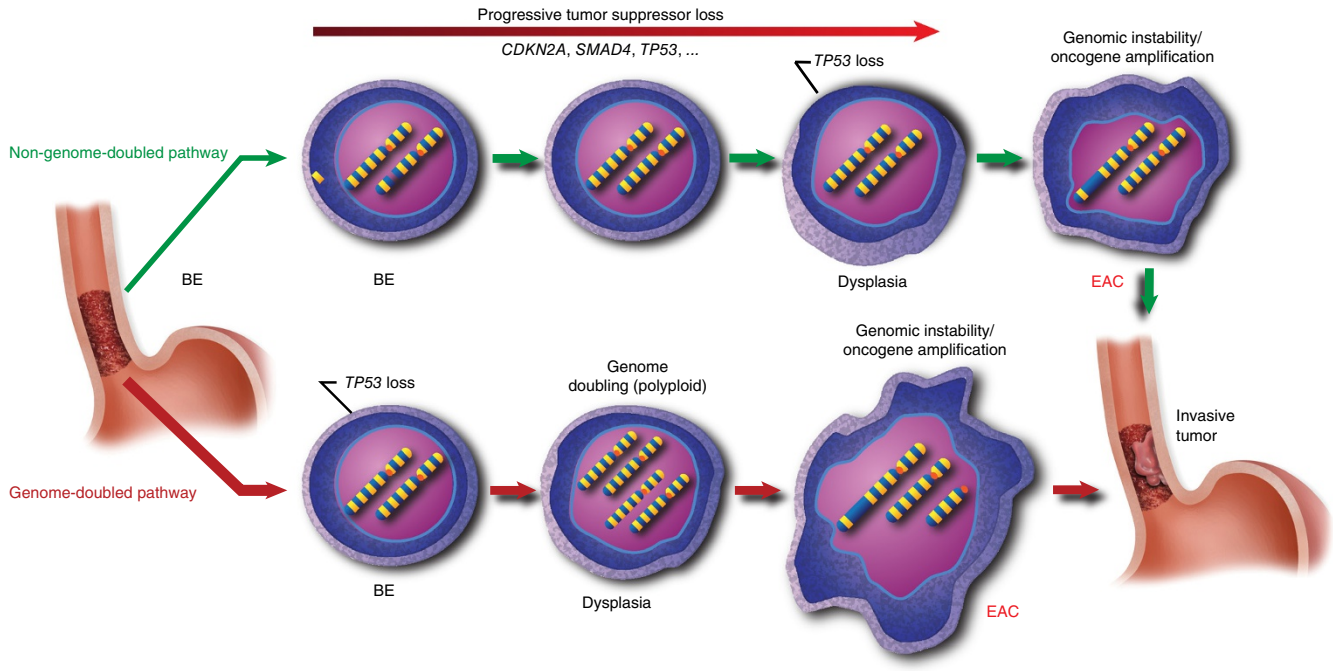

Figure 8 Genome-doubled EAC shows a distinct pathway of development. The schematic shows two general pathways by which Barrett's esophagus can develop into EAC. The top model involves the gradual accumulation of alterations in TSGs followed by the subsequent activation of oncogenes and development of genomic instability. In the bottom model, *TP53* inactivation is acquired as an early event. The sample then undergoes genome doubling, leading to genomic instability, aneuploidy and oncogene amplification.

contrast to that seen in pancreatic or colorectal cancer, where oncogene activation via mutation is believed to occur earlier, typically before *TP53* inactivation[24–27].

Our data therefore suggest that the traditional Barrett's esophagus progression model conflates two general pathways for oncogenic transformation (**Fig. 8**). One pathway, starting similarly to the traditional model, appears to involve the progressive accumulation of TSG losses (commonly *CDKN2A* and *TP53* and also frequently including *SMAD4* and alterations of chromatin-modifying enzymes), leading to genomic instability and oncogenic amplifications without an antecedent WGD event. Despite the prevalence of clonal expansions of Barrett's esophagus tissues with somatic inactivation of TSGs such as *CDKN2A* and *ARID1A*, a minority of EACs appear to emerge following such a path without WGD. Instead, the majority of EACs apparently develop following expansion of a *TP53*-mutant clone that undergoes WGD, with WGD predominately seen in tissues with dysplasia. Genomic doubling has been documented to facilitate the acquisition of genomic instability[16,28,29]. Consistent with this, tumors that emerged following WGD harbored marked genomic disruption and oncogene amplification, with amplifications of known oncogenes observed in 77 of 90 EAC samples with WGD. Following WGD, homozygous inactivation of tumor suppressors becomes more difficult owing to the additional number of events required[14]. Once WGD occurs, the more expedient pathway to transformation is thus likely that of acquisition of oncogene activation via structural genomic instability. The predilection for instability following WGD[16,29–31] likely contributes to catastrophic genomic disruptions resulting in a large number of copy number alterations, as recently identified by whole-genome sequencing of EAC[32]. This alternative pathway to transformation whereby a dysplastic clone with WGD acquires oncogene activation via amplification is supported by our findings that oncogene amplification is typically a later event in the progression to EAC. Such episodes of genomic disruption could lead to positive selection for distinct amplifications that bypass the need for loss of tumor suppressors, as more frequently occurs in EACs without WGD (**Fig. 3**). We note that it is also possible that cells with WGD are better able to survive catastrophic genomic disruption (for example, because such events are unlikely to generate complete gene knockouts) and may be less subject to negative selection. Although our limited data set did not identify differences in clinical stage between EACs with and without WGD (**Supplementary Fig. 21**), future studies will be needed to determine whether tumors following these distinct paths have other distinguishing clinical features.

Both WGD and *TP53* mutations have been recognized for over a decade to be risk factors for the development of EAC in patients with Barrett's esophagus[7,23,33,34]. Here we are able to refine the conventional model of how these tumors emerge, finding that *TP53* mutations may be earlier events than previously recognized. After *TP53* mutation, many EACs may follow a distinct pathway to cancer involving WGD with subsequent transformation to cancer via catastrophic aneuploidy and oncogene amplification. Our model is therefore consistent with a recent complementary report by Li *et al.*[7] that characterized the copy number profiles of serially collected Barrett's esophagus biopsies in which the authors noted that, in the 24 months before cancer diagnosis, a marked increase in DNA content occurred, suggestive of WGD.

Our refined model positing a potentially more rapid path to transformation following the acquisition of a *TP53* mutation and WGD may help explain the failure of endoscopic screening of patients with Barrett's esophagus to prevent cancer diagnoses and deaths. Screening strategies are largely premised upon the concept that Barrett's esophagus is at risk of the gradual accumulation of genomic alterations leading to progression to cancer. This model would predict that, just as aging predisposes to cancer, duration of Barrett's esophagus would also enhance cancer risk. However, contrary to this concept, studies of population cohorts of patients with diagnoses of Barrett's esophagus show that the majority of EACs are detected within the first

2 to 3 years of initial endoscopic diagnosis of Barrett's esophagus, even when the incident endoscopy fails to identify dysplasia or cancer[35–38]. If most EACs emerge from Barrett's esophagus following a more rapid process involving *TP53* mutation and the emergence of an unstable, genomically doubled intermediate, new diagnostic strategies may be required that seek out these *TP53*-mutant precursors and intermediates as a means of detecting and preventing this deadly disease.

URLs. MutSig algorithm, www.broadinstitute.org/cancer/cga/MutSig; CCDS, http://www.ncbi.nlm.nih.gov/CCDS/; Broad Institute Picard sequencing pipeline, http://broadinstitute.github.io/picard/; Broad Institute Firehose Pipeline, http://www.broadinstitute.org/cancer/cga/; Oncotator, http://www.broadinstitute.org/oncotator/.

METHODS

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

Accession codes. Binary sequence alignment/map (BAM) files have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession phs000598.

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

ACKNOWLEDGMENTS

We thank the members of the Broad Institute Genome Sequencing Platform and the molecular laboratories at Brigham and Women's Hospital and Massachusetts General Hospital for their assistance. We are grateful to the patients and families who agreed to contribute their samples to enable this research and to the physicians and hospital staff whose efforts in collecting these samples are essential to this work. This work was supported by US National Institutes of Health grant T32 HL007627 and the Dana-Farber/Harvard Gastrointestinal Cancer Specialized Programs of Research Excellence P50CA127003 (M.D.S.), the National Human Genome Research Institute (NHGRI) Large-Scale Sequencing Program (U54 HG0003067; E.S.L.), National Cancer Institute grant U54 CA163059 (D.G.B.), Broad Institute SPARC funding (A.J.B., S.L.C. and G.G.), a Research Scholar Grant from the American Cancer Society (A.J.B.) and the National Cancer Institute (P01 CA098101 and U54 CA163004; A.J.B.).

AUTHOR CONTRIBUTIONS

M.D.S. performed experiments and interpreted results. A.T.-W., S.P., A.M., P.S., I.L., M.S.L. and S.L.C. performed computational analysis. A.T.A. and R.D.O. performed pathological slide review. J.M.D., K.S.N., D.F.-T., J.L., A.C.C. and D.G.B. contributed samples and clinical annotation. M.L. contributed laser-capture microdissection guidance and manuscript review. C.S., S.S., S.B.G. and E.S.L. organized and supervised sequencing. G.G., S.L.C. and A.J.B. supervised all studies. M.D.S., A.T.-W., S.L.C., G.G. and A.J.B. prepared the manuscript, and all authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Nehra, D., Howell, P., Williams, C.P., Pye, J.K. & Beynon, J. Toxic bile acids in gastro-oesophageal reflux disease: influence of gastric acidity. *Gut* **44**, 598–602 (1999).
2. Wild, C.P. & Hardie, L.J. Reflux, Barrett's oesophagus and adenocarcinoma: burning questions. *Nat. Rev. Cancer* **3**, 676–684 (2003).
3. Lagergren, J., Bergstromeinhold, R., Lingren, A. & Nyren, O. Symptomatic gastoroesophageal reflux as a risk factor for esophageal adenocarcinoma. *N. Engl. J. Med.* **340**, 825–831 (1999).
4. Ormsby, A.H. *et al.* The location and frequency of intestinal metaplasia at the esophagogastric junction in 223 consecutive autopsies: implications for patient treatment and preventive strategies in Barrett's esophagus. *Mod. Pathol.* **13**, 614–620 (2000).
5. Galipeau, P.C., Prevo, L.J., Sanchez, C.a., Longton, G.M. & Reid, B.J. Clonal expansion and loss of heterozygosity at chromosomes 9p and 17p in premalignant esophageal (Barrett's) tissue. *J. Natl. Cancer Inst.* **91**, 2087–2095 (1999).
6. Gu, J. *et al.* Genome-wide catalogue of chromosomal aberrations in Barrett's esophagus and esophageal adenocarcinoma: a high-density single nucleotide polymorphism array analysis. *Cancer Prev. Res. (Phila.)* **3**, 1176–1186 (2010).
7. Li, X. *et al.* Temporal and spatial evolution of somatic chromosomal alterations: a case-cohort study of Barrett's esophagus. *Cancer Prev. Res. (Phila.)* **7**, 114–127 (2014).
8. Li, X. *et al.* Single nucleotide polymorphism–based genome-wide chromosome copy change, loss of heterozygosity, and aneuploidy in Barrett's esophagus neoplastic progression. *Cancer Prev. Res. (Phila.)* **1**, 413–423 (2008).
9. Reid, B.J. *et al.* Barrett's esophagus: ordering the events that lead to cancer. *Eur. J. Cancer Prev.* **5**, 57–65 (1996).
10. Wong, D.J. *et al.* $p16^{INK4a}$ lesions are common, early abnormalities that undergo clonal expansion in Barrett's metaplastic epithelium. *Cancer Res.* **61**, 8284–8289 (2001).
11. Zhang, S. & Wang, X.I. SIRT1 is a useful biomarker for high-grade dysplasia and carcinoma in Barrett's. *Esophagus* **43**, 373–377 (2013).
12. Paulson, T.G. *et al.* p16 mutation spectrum in the premalignant condition Barrett's esophagus. *PLoS ONE* **3**, e3809 (2008).
13. Dulak, A.M. *et al.* Exome and whole genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478–486 (2013).
14. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
15. Dulak, A.M. *et al.* Gastrointestinal adenocarcinomas of the esophagus, stomach, and colon exhibit distinct patterns of genome instability and oncogenesis. *Cancer Res.* **72**, 4383–4393 (2012).
16. Fujiwara, T. *et al.* Cytokinesis failure generating tetraploids promotes tumorigenesis in *p53*-null cells. *Nature* **437**, 1043–1047 (2005).
17. Davoli, T. & de Lange, T. Telomere-driven tetraploidization occurs in human cells undergoing crisis and promotes transformation of mouse cells. *Cancer Cell* **21**, 765–776 (2012).
18. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
19. Etemadmoghadam, D. *et al.* Resistance to CDK2 inhibitors is associated with selection of polyploid cells in *CCNE1*-amplified ovarian cancer. *Clin. Cancer Res.* **19**, 5960–5971 (2013).
20. Maley, C.C. *et al.* Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. *Cancer Res.* **64**, 3414–3427 (2004).
21. Weaver, J.M.J. *et al.* Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat. Genet.* **46**, 837–843 (2014).
22. Maley, C.C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38**, 468–473 (2006).
23. Reid, B.J. *et al.* Predictors of progression in Barrett's esophagus II: baseline 17p (*p53*) loss of heterozygosity identifies a patient subset at increased risk for neoplastic progression. *Am. J. Gastroenterol.* **96**, 2839–2848 (2001).
24. van Wyk, R. *et al.* Somatic mutations of the *APC*, *KRAS*, and *TP53* genes in nonpolypoid colorectal adenomas. *Genes Chromosom. Cancer* **27**, 202–208 (2000).
25. Prestlow, T.P. & Prestlow, T.G. No mutant *KRAS* in aberrant crypt foci (ACF): initiation of colorectal cancer? *Biochim. Biophys. Acta* **1756**, 83–96 (2005).
26. Lüttges, J. *et al.* The K-*ras* mutation pattern in pancreatic ductal adenocarcinoma usually is identical to that in associated normal, hyperplastic, and metaplastic ductal epithelium. *Cancer* **85**, 1703–1710 (1999).
27. Deramaudt, T. & Rustgi, A. Mutant *KRAS* in the initiation of pancreatic cancer. *Biochim. Biophys. Acta* **1756**, 97–101 (2005).
28. Gordon, D.J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nat. Rev. Genet.* **13**, 189–203 (2012).
29. Dewhurst, S.M. *et al.* Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014).
30. Davoli, T. & de Lange, T. The causes and consequences of polyploidy in normal development and cancer. *Annu. Rev. Cell Dev. Biol.* **27**, 585–610 (2011).
31. Ganem, N.J. *et al.* Cytokinesis failure triggers Hippo tumor suppressor pathway activation. *Cell* **158**, 833–848 (2014).
32. Nones, K. *et al.* Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nat. Commun.* **5**, 5224 (2014).
33. Rabinovitch, P.S., Reid, B.J., Haggit, R.C., Norwood, T.H. & Rubin, C.E. Progression to cancer in Barrett's esophagus is associated with genomic instability. *Lab. Invest.* **60**, 65–71 (1989).
34. Davelaar, A.L. *et al.* Aberrant *TP53* detected by combining immunohistochemistry and DNA-FISH improves Barrett's esophagus progression prediction: a prospective follow-up study. *Genes Chromosom. Cancer* **54**, 82–90 (2015).
35. Bytzer, P., Christensen, P.B., Damkier, P., Vinding, K. & Seersholm, N. Adenocarcinoma of the esophagus and Barrett's esophagus: a population-based study. *Am. J. Gastroenterol.* **94**, 86–91 (1999).
36. Weston, A.P. *et al.* Long-term follow-up of Barrett's high-grade dysplasia. *Am. J. Gastroenterol.* **95**, 1888–1893 (2000).
37. Corley, D.A., Levin, T.R., Habel, L.A., Weiss, N.S. & Buffler, P.A. Surveillance and survival in Barrett's adenocarcinomas: a population-based study. *Gastroenterology* **122**, 633–640 (2002).
38. Hvid-Jensen, F., Pedersen, L., Drewes, A., Sorensen, H. & Funch-Jensen, P. Incidence of adenocarcinoma among patients with Barrett's esophagus. *N. Engl. J. Med.* **365**, 1375–1383 (2011).

ONLINE METHODS

Sample selection and DNA extraction (fresh frozen). Samples were obtained with documented informed consent and institutional review board (IRB) approval. Fresh-frozen samples were obtained at the time of surgical esophagectomy from patients with a diagnosis of EAC and without neoadjuvant therapy from the University of Michigan. Samples of Barrett's esophagus were selected not immediately adjacent to the tumor (when possible) to minimize tumor contamination within the Barrett's esophagus sample. Slides stained with hematoxylin and eosin from 27 cases were examined by a pathologist with subspecialty training in gastrointestinal pathology. Slides from Barrett's esophagus tissue were classified as either non-dysplastic or dysplastic. Samples with HGD and LGD were combined for primary analysis because of the overall low numbers and the diagnostic challenges and controversies associated with distinguishing between these entities (especially on frozen section slides). Additional subanalyses looked separately at these diagnoses. Two cases were excluded from further analysis as one case contained too low of a percentage of Barrett's esophagus tissue and one Barrett's esophagus tissue sample was contaminated with invasive tumor. DNA was extracted using phenol-chloroform and ethanol precipitation and quantified using PicoGreen dsDNA Quantification reagent (Invitrogen).

Sample selection and DNA extraction (formalin fixed and paraffin embedded). With documented informed consent and IRB approval, formalin-fixed, paraffin-embedded (FFPE) esophagectomy samples without neoadjuvant therapy were identified in the pathology archives of the University of Pittsburgh Medical Center and Brigham and Women's Hospital. Slides stained with hematoxylin and eosin were reviewed by two gastrointestinal pathologists to determine consensus areas of Barrett's esophagus, Barrett's esophagus with LGD, Barrett's esophagus with HGD and EAC. If uncertainty for a diagnosis was present, a third pathologist reviewed the sample. Any sample without a consensus diagnosis was eliminated from analysis. Ten 8-micron sections were cut onto PEN membrane frame slides (Life Technologies) bracketed by standard slides for hematoxylin and eosin staining. The frame slides were stained using Arcturus paradise plus stain (Life Technologies) following the manufacturer's recommendations. Areas of interest were microdissected using the ArcturusXT laser-capture microdissection instrument (Life Technologies). DNA was isolated using the Qiagen FFPE DNA isolation kit following the manufacturer's protocol with the exception that the tissue was digested with proteinase K overnight. DNA was quantified using PicoGreen dsDNA Quantification reagent.

Whole-exome sequencing. Whole-exome capture libraries were constructed from 100 ng of DNA following shearing, end repair, phosphorylation and ligation to barcoded sequencing adaptors[39]. DNA was size selected for lengths between 200 and 350 bp and subjected to exonic hybrid capture using SureSelect v2 Exome bait (Agilent). Samples were multiplexed and sequenced on multiple Illumina HiSeq flow cells. Mean target exome coverage of 95× was achieved in the DNA from Barrett's esophagus samples, 85× was achieved for DNA from the neoplastic samples and 87× was achieved for the DNA from normal tissue.

Sequencing data processing. Exome sequence data processing and analysis were performed using Broad Institute pipelines as previously described[40–43]. A BAM file aligned to the hg19 human genome build was generated from sequencing reads for each sample by the Picard pipeline.

Mutation calling. The MuTect algorithm was used to identify somatic mutations[41,42,44]. We required a minimum of 14 reads covering a site in the tumor and 8 reads in the normal sample for declaring a site to be adequately covered for mutation calling. We determined the lowest allelic fraction at which somatic mutations could be detected on a per-sample basis, using estimates of cross-contamination from the ContEst pipeline[45]. Small somatic insertions and deletions were detected using the Indelocator algorithm after local realignment of tumor and normal sequences[44]. All somatic mutations detected by whole-exome sequencing were analyzed for potential false positive calls by performing a comparison to mutation calls from a panel of 2,500 germline DNA samples. Mutations found in 2% of the germline samples or 2% of sequencing reads were removed from analysis.

Because of our goal of comparing the presence and absence of mutations between distinct samples taken from the same patient, we used a tool designed to specifically query evidence for mutations or indels called in one sample for evidence of their presence, even at low allelic fraction, in other samples from the same patient. The strong prior of having been called *de novo* in one sample allowed for more sensitive detection in other related samples. This method, termed 'force calling', uses outputs from MuTect and Indelocator to generate an aggregate set of somatic events for each patient. It then adopts SAMtools to count the number of reads supporting the reference or alternate alleles at those sites in the other matched samples. Reads are considered if they are from unique pairs, have a base quality at the site of interest of greater than or equal to 20 and have a read quality of greater than or equal to 5.

Mutation annotation. Somatic single-nucleotide variants, insertions and deletions were annotated using Oncotator, which uses information from publicly available databases[46–50].

Calculation of total and allelic copy numbers from whole-exome sequencing data. Genome-wide copy-ratio profiles were inferred using CAPSEG. Read depth at informative capture targets in tumor samples was calibrated to estimate the copy ratio using depths observed in a panel of normal (non-cancer) diploid genomes. The resulting copy-ratio profiles were then segmented using the circular binary segmentation (CBS) algorithm[51]. Allelic copy number analysis was then performed by examination of alternate and reference read counts at heterozygous SNP positions (as determined by analysis of the matched normal sample). These counts were used to infer the contribution of the two homologous chromosomes to the observed copy ratio in each segment. Further analysis of change points in these allelic ratios was performed using PSCBS[52], refining the segmentation. Finally, for each segment, we combined the copy-ratio and allelic data to derive allelic copy ratios, which were input for analysis with ABSOLUTE[14].

The ABSOLUTE[14] computational tool (v1.2) was used to provide computational estimates of several parameters for each neoplastic sample in this study. These estimates include (i) the purity of each sample (fraction of nuclei in the sample originating from tumor or Barrett's esophagus); (ii) the average ploidy of the cancer or Barrett's esophagus genome; (iii) the presence of antecedent genomic doubling for each genome; and (iv) the absolute allelic copy number across the genome. ABSOLUTE takes as input the segmented allelic copy number ratio data (as described above) as well as the allele fractions of somatic point mutations (aberrant reads as a ratio of total reads covering the locus) and then determines possible combinations of tumor purity, ploidy and antecedent genomic doubling, which fit the allelic copy number ratio data and point mutation variant allele fraction (VAF; **Supplementary Fig. 9**). The ABSOLUTE solutions were reviewed manually to maximize concordance with the data (A.T.-W.).

Calculation of point mutation CCF distributions. For each somatic mutation, we computationally estimated the fraction of neoplastic cells within a specific DNA sample that harbored the mutation, i.e., the CCF. This fraction is represented as a distribution between 0 and 1 (refs. 14,53,54). A CCF value of 1 corresponds to a mutation present in 100% of the neoplastic cells in a sample. A CCF value of <1 indicates that the mutation is present in a subset of the neoplastic cells in a sample and thus is subclonal. Probability distributions for CCF values were computed by correcting mutant and reference read counts from Illumina sequencing for the estimated sample purity and local copy number[14] (**Supplementary Fig. 9**), as previously described[53,54].

After the initial determination of a CCF distribution for each mutation, we then performed an additional analysis to refine our CCF estimates, on the basis of the assumption that each neoplastic sample contained a small (but unknown) number of distinct populations defined by mutations that share the same CCF. We clustered mutations with similar CCF values by sampling from a mixture of Dirichlet processes using a Markov chain Monte Carlo (MCMC) technique, as previously described (**Supplementary Fig. 11a**,**b**)[53,54]. We used 250 MCMC iterations where the 125 initial ones were discarded as burn-in. A prior over the number of mutation clusters in a given sample was specified using a negative binomial distribution ($r = 10$, $\mu = 3$), which favored 1–5 clusters (**Supplementary Fig. 11d**). A partition of mutations was obtained

doi:10.1038/ng.3343

on the basis of how often each pair of mutations was assigned to the same cluster during the MCMC simulation (after convergence). A distance metric was generated from the inverse of each pair count, and hierarchical clustering was performed using complete linkage. The resulting tree was divided into k clusters, with k chosen as the lowest number of sampled clusters in the MCMC (**Supplementary Fig. 11c**,**d**).

Calculation of statistical power for detection of shared somatic mutations. To calculate power, we considered the expected VAF of a point mutation with CCF = 1 and multiplicity = 1, given the sample purity and local copy number. For mutations detected in only a subset of the tissue samples comprising a given case, we calculated the paired detection power. Because shared mutations with a single supporting read matching the called allele were called by the forced calling procedure (described above), we calculated power as the probability of observing one or more such reads, given expected VAF and sequence coverage.

Relatedness calculations. Relatedness was calculated by taking the total number of shared mutations that were present at CCF = 1 in both samples divided by the sum of the total number of mutations with CCF = 1 in either sample (**Supplementary Fig. 1**). The relatedness using all mutations (CCF = 1 and CCF < 1) was also calculated and is reported in **Supplementary Table 1**. We also searched for evidence of shared copy number aberrations, which would provide independent support for a potential common origin in an EAC–Barrett's esophagus pair beyond shared point mutations. For each segment in the EAC samples with copy number greater than 2.5 or less than 1.5, we looked at the matched Barrett's esophagus sample to see whether there was also a segment greater than 2.5 or less than 1.5 that was 50% mutually overlapping with the segment in the EAC sample. If so, the segment was called shared. Using this analysis, we did not find any additional shared EAC–Barrett's esophagus pairs not detected on the basis of shared somatic mutations.

Phylogenetic inference. Tumors can exhibit genetic heterogeneity, both across different regions[55–58] and within single cancer tissue samples[14,53,54,59–61]. Heterogeneity within individual tissue samples presents a difficulty to simple phylogenetic inference algorithms, which typically distinguish only between the presence or absence of mutations in each sample. These algorithms attempt to construct phylogenetic trees relating each tissue sample, which may not accurately reflect the evolutionary relationships between the neoplastic cell populations represented in the tissue samples. For example, shared mutations (present in multiple samples) due to overlapping subclonal populations would be mistaken as evidence for shared ancestry between the samples.

To address this difficulty, we used quantitative information about each mutation's prevalence in each neoplastic tissue sample (CCF) to determine whether the tissue samples were sufficiently diverged from one another such that no detectable overlap of minor subclones occurred (CCF < 1), a scenario we term the 'branched sibling' model. In this scenario, it is valid to construct standard phylogenetic trees relating each tissue sample, with minor subclones (CCF < 1) private to each tissue sample represented as subtrees (microphylogenies) grafted on to each sample tip. The branched sibling scenario implies that such trees accurately represent the evolutionary relationship of all subclonal populations detected with CCF = 1 in the sampled neoplastic tissues. A corollary of the branched sibling model is that all mutations shared in two or more samples must have CCF = 1 wherever they are present (**Supplementary Fig. 12a**,**b**). Thus, the appearance of mutations shared in two or more samples with CCF <1 in any of them either represents a technical artifact or constitutes evidence that the branched sibling approximation is not an accurate description of those samples.

We constructed phylogenetic trees representing the evolutionary relationship between the neoplastic tissue samples sequenced from each patient using a semiautomated four-stage process, described below. First, we searched for the optimal tree that would explain the observed matrix of binary point mutation presence or absence data in each sample, given the standard phylogenetic assumptions that specific mutations arise uniquely in each patient and that there were negligible rates of mutation loss (for example, due to chromosomal deletion of a mutant allele). We searched for the phylogenetic tree with maximum parsimony using the standard parsimony ratchet method[62].

Second, we applied the Bayesian CCF clustering procedure described above to each sample individually, retaining all mutations provisionally called with >0 supporting reads in that sample. A single pseudocount observation was added corresponding to a cluster at CCF = 1. We then identified all provisional mutation calls (>0 supporting reads) made in at least two samples of the case that were assigned to a CCF cluster with a posterior mode <1.0 (**Supplementary Fig. 12**). These mutation calls represent either sequencing artifacts or evidence for overlapping minor subclones in the sampled neoplastic tissues (violating the branched sibling approximation). We rejected such sites if the number of supporting reads was <3, and this modified matrix of mutation calls was then used to assign each mutation to a branch of the phylogenetic tree (**Supplementary Fig. 12b**).

We also distinguished mutations that were underpowered for detection in some samples (as described above). For each sample, the number of mutations in each category is shown in **Supplementary Figure 12c**. Assignment of gene-level SCNAs to branches was performed in a similar manner (**Supplementary Fig. 12d**).

Third, we refined the tips of each phylogenetic tree by distinguishing between private mutations that occurred in all neoplastic cells of a given sample (CCF = 1) compared to those that occurred in only a minor subclone (CCF < 1) specific to that sample. For this distinction, we applied Bayesian clustering techniques (described above) to the mutations identified only in that sample. We added n pseudocount observations of CCF = 1, with n representing the number of mutations called in >1 sample of the case that were also called in the sample being considered. This process partitioned the private mutations in each sample into putative subclones with common CCF values (**Supplementary Fig. 10**). We modified the phylogenetic trees by replacing each (non-germline) tip with the subtree consistent with the maximally branching microphylogeny respecting the rule that the sum of sibling subclone CCF values could not exceed that of their most recent common ancestor (**Figs. 4** and **5**, and **Supplementary Fig. 13**).

Fourth, we examined whether evidence that the branched sibling model was not an adequate approximation of the sampled neoplastic tissues could be discerned. For each case, we analyzed the two-dimensional CCF distributions of point mutations for all unique tissue sample pairs (**Supplementary Fig. 11**) using a two-dimensional version of the Bayesian clustering algorithm described above[63]. Examination of these data showed robust clusters of mutations with CCF = 1 in both samples of clonally related pairs. In addition, many pairs harbored mutations with CCF = 1 in one sample that were undetected in the paired sample (and vice versa). Most samples also harbored mutations with CCF < 1 that were undetected in the paired sample. Furthermore, for most patients, only a small fraction of mutations with CCF < 1 appeared to be detected in both sample pairs, and these mutations did not tend to form strong clusters (consistent with sequencing artifacts; **Supplementary Fig. 11**). Taken together, these observations implied that most tissue samples were well approximated by the branched sibling model (**Supplementary Fig. 11**), as true overlap of minor populations from distinct subclonal branches would tend to displace the CCF values of mutations private to each sample, such that none would have CCF = 1.

In the two cases where evidence contradicting the branched sibling model was observed, phylogenetic trees were manually adjusted (as described below) to accurately reflect the evolutionary relationship between the different clonal lineages (shown in **Figs. 4** and **5**). This was done in a manner analogous to that described in a recent report[61]; here we extended similar logic to the scenario where the same subclone was present in multiple tissue samples. Detailed analysis of mutation CCFs for each patient, including the automatically generated phylogenetic trees (before manual adjustment), are available in **Supplementary Data Sets 3** and **4**.

For patients 4 and 6, nearly every shared mutation had CCF = 1. For patient P1, the shared mutations with CCF < 1 did not appear to form a strong cluster or to displace other clusters away from CCF = 1. We therefore assumed that these mutations did not constitute strong evidence of a branched sibling violation and accepted the automatically produced phylogenies for these cases.

For patient P3, we detected a minor population in sample EAC1 (subclone 1; CCF = 0.2) defined by 45–200 mutations (which were present at CCF = 1 in the shared MET1-HGD1-EAC2 branch. We adjusted the phylogenetic tree

(**Fig. 5b**) to move EAC1 onto a distinct branch from these samples to represent the dominant subclone (subclone 2; CCF = 0.8) in EAC1, which had an evolutionary origin distinct from that of subclone 1. In addition, samples HGD1 and MET1 appeared to share a small subclone (CCF = 0.05) defined by five mutations that did not appear elsewhere in the case. We removed the shared branch defined by these mutations from the phylogeny as these mutations did not reflect shared recent ancestry of the dominant subclones in HGD1 and MET1 (**Fig. 5b**). For patient P7, details regarding phylogenetic tree generation are provided in the main text.

FISH analysis. In patient P3, we wanted to confirm the genome-doubling status in two apparent genomically unrelated tumors. ABSOLUTE predicted that one of the samples had undergone a genomic-doubling event, whereas the other did not. We performed FISH analysis co-labeling a 4-μm FFPE section with two differently labeled centromeric markers for chromosomes that are relatively stable in EAC (CEP 2 and CEP 4). After review and identification of the areas of interest, the probes were enumerated by two senior research technologists within the Brigham and Women's cytogenomics core facility. Each technologist counted 50 cells per sample, and the number of CEP 2 and CEP 4 probes per cell were averaged. In patient P7, we performed FISH for *KRAS* and CEP 12 from samples LGD1, HGD2 and EAC1 to confirm the finding that the amplification was present in EAC1 but not the premalignant lesions. FISH was performed on the paired samples BE 63 and EAC 63 for *ERBB2* and CEP 17. All FISH studies were performed with standard techniques. The gene to CEP ratio was calculated, and a ratio of greater than two was considered positive.

39. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
40. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
41. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
42. Barbieri, C.E. *et al.* Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
43. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
44. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
45. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601–2602 (2011).
46. Fujita, P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876–D882 (2011).
47. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
48. Griffith, O.L. *et al.* ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* **36**, D107–D113 (2008).
49. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39**, D214–D219 (2011).
50. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
51. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
52. Olshen, A.B. *et al.* Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics* **27**, 2038–2046 (2011).
53. Landau, D.A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).
54. Lohr, J.G., Stojanov, P., Carter, S.L., Cruz-gordillo, P. & Lawrence, M.S. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell* **25**, 91–101 (2014).
55. Campbell, P.J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
56. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
57. Liu, W. *et al.* Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat. Med.* **15**, 559–565 (2009).
58. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
59. Shah, S.P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
60. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
61. McFadden, D.G. *et al.* Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell* **156**, 1298–1311 (2014).
62. Nixon, K.C. The Parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics* **15**, 407–414 (1999).
63. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).