nature genetics

Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis

Jamie M J Weaver[1,11], Caryn S Ross-Innes[1,11], Nicholas Shannon[2,11], Andy G Lynch[2,11], Tim Forshew[2], Mariagnese Barbera[1], Muhammed Murtaza[2], Chin-Ann J Ong[1], Pierre Lao-Sirieix[1], Mark J Dunning[2], Laura Smith[1], Mike L Smith[2], Charlotte L Anderson[2], Benilton Carvalho[2], Maria O'Donovan[3], Timothy J Underwood[4], Andrew P May[5], Nicola Grehan[1], Richard Hardwick[6], Jim Davies[7], Arusha Oloumi[8], Sam Aparicio[8], Carlos Caldas[2], Matthew D Eldridge[2], Paul A W Edwards[9], Nitzan Rosenfeld[2], Simon Tavaré[2], Rebecca C Fitzgerald[1] & the OCCAMS Consortium[10]

Cancer genome sequencing studies have identified numerous driver genes, but the relative timing of mutations in carcinogenesis remains unclear. The gradual progression from premalignant Barrett's esophagus to esophageal adenocarcinoma (EAC) provides an ideal model to study the ordering of somatic mutations. We identified recurrently mutated genes and assessed clonal structure using whole-genome sequencing and amplicon resequencing of 112 EACs. We next screened a cohort of 109 biopsies from 2 key transition points in the development of malignancy: benign metaplastic never-dysplastic Barrett's esophagus (NDBE; *n* = 66) and high-grade dysplasia (HGD; *n* = 43). Unexpectedly, the majority of recurrently mutated genes in EAC were also mutated in NDBE. Only *TP53* and *SMAD4* mutations occurred in a stage-specific manner, confined to HGD and EAC, respectively. Finally, we applied this knowledge to identify high-risk Barrett's esophagus in a new non-endoscopic test. In conclusion, mutations in EAC driver genes generally occur exceptionally early in disease development with profound implications for diagnostic and therapeutic strategies.

© 2014 Nature America, Inc. All rights reserved.

Most epithelial cancers develop gradually from preinvasive lesions, in some instances after an initial metaplastic conversion. Research to characterize the genomic landscape of cancer has focused on established invasive disease with the goal of developing biomarkers for personalized therapy[1]. However, it is becoming increasingly clear that extensive genomic heterogeneity is present in the majority of advanced cancers[2]. The most appropriate therapeutic targets are therefore those mutations that occur early in the development of disease and are thus clonal in the resulting malignancy. The identification of causative mutations occurring early in pathogenesis is also pivotal to developing clinically useful biomarkers. In this context, mutations occurring at disease stage boundaries, for example, the transition from non-dysplastic epithelium to dysplasia and then to cancer, would be most informative. The evidence thus far on the genetic evolution of cancer from premalignant lesions suggests that the accumulation of mutations is stepwise[3–5]. In the most well-studied example, the adenoma-dysplasia–colorectal adenocarcinoma progression sequence, it has been possible to assign timings for mutations in a limited number of candidate genes by comparative lesion sequencing[3]. More recent studies have sought to use statistical algorithms to infer the life history[4,5] of a tumor from single samples.

EAC arises from metaplastic Barrett's esophagus in the context of chronic inflammation secondary to exposure to acid and bile[6,7]. Barrett's esophagus lends itself well to studies of genetic evolution owing to the repeated sampling of mucosa during clinical surveillance before therapeutic intervention[8]. Previous studies of EAC and Barrett's esophagus have generally used candidate gene approaches with the goal of identifying clinical biomarkers to complement histological examination, which is an approach fraught with difficulties[8,9]. Data from high-density SNP arrays and exome sequencing studies are now accumulating, with a plethora of mutations identified in many different genes[10,11]. However, little work has yet focused on the precise ordering of these alterations in large cohorts of individuals with premalignant disease and associated clinical follow-up data.

Recently, Agrawal *et al.* performed exome sequencing on 11 EAC samples and 2 samples of Barrett's esophagus adjacent to the cancer. Intriguingly, the majority of mutations were found to be present even in apparently normal Barrett's esophagus[12], similar to the observation in colorectal adenocarcinoma. This finding raises the possibility that, before progression to malignancy, mutations that predict risk of progression might be detectable in cytologically benign tissue.

[1]Medical Research Council (MRC) Cancer Unit, University of Cambridge, Cambridge, UK. [2]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [3]Department of Histopathology, Addenbrooke's Hospital, Cambridge, UK. [4]Cancer Sciences Division, University of Southampton, Southampton, UK. [5]Fluidigm Corporation, South San Francisco, California, USA. [6]Oesophago-Gastric Unit, Addenbrooke's Hospital, Cambridge, UK. [7]Oxford Computing Laboratory, University of Oxford, Oxford, UK. [8]British Columbia Cancer Research Centre, Cancer Agency Research Centre, Vancouver, British Columbia, Canada. [9]Department of Pathology, University of Cambridge, Cambridge, UK. [10]A full list of members and affiliations appear at the end of the paper. [11]These authors contributed equally to this work. Correspondence should be addressed to R.C.F. (rcf29@mrc-cu.cam.ac.uk).

Received 22 November 2013; accepted 28 May 2014; published online 22 June 2014; doi:10.1038/ng.3013

However, it is unclear to what extent the same mutations might be present in Barrett's esophagus tissue from individuals who have not progressed to cancer. This question is noteworthy as the majority of individuals with Barrett's esophagus will not progress to cancer, and somatic alterations occurring early, before dysplasia, are unlikely to provide clinically discriminatory biomarkers. Biomarker research in this area is critical because current endoscopic surveillance strategies are increasingly recognized to be ineffective[13], and novel approaches are therefore required[14,15].

The aims of this study were (i) to identify a list of candidate recurrently mutated genes in EAC; (ii) to accurately resolve the stage of disease at which mutation occurs, thereby providing insight on the role of these recurrent mutations in cancer progression; and (iii) to test the usefulness of these mutations in clinical applications, that is, using the non-invasive, non-endoscopic cell sampling device the Cytosponge.

RESULTS

High mutation burden and unusual mutational signature in EAC

The discovery cohort (22 EACs subjected to whole-genome sequencing; **Fig. 1**) reflected the known clinicodemographic features of the disease, including male predominance (male:female ratio of 4.5:1), a mean age of 68 years (range of 53 to 82 years) and a majority with advanced disease (81.8% (18/22) above stage I). Of the 22 cases, 17 (77.3%) had evidence of Barrett's esophagus in the resection specimen (**Table 1** and **Supplementary Table 1**). Case samples were sequenced to mean coverage of 63-fold and 67-fold for tumor and normal samples, respectively (**Supplementary Table 2**; normal squamous tissue or blood was used as outlined in **Supplementary Table 1**).

We identified a median of 16,994 somatic single-nucleotide variants (SNVs; range of 4,518–56,528) and 994 small indels (range of 262–3,573) per sample. In this final data set, a total of 1,086 coding-region mutations were subjected to verification as part of a larger pipeline benchmarking study (Online Methods). We used ultra-deep targeted resequencing, achieving a median coverage of >13,000-fold, and confirmed 1,081 mutations (99.5%) as somatic. Using Sanger sequencing, 23 of 25 indels (92%) were verified as real and somatic. As observed by Dulak *et al.* in the intervening time since our study commenced[11], the most frequent mutation type across the discovery cohort was T:A>G:C transversions, with a striking enrichment at CTT trinucleotides (**Supplementary Fig. 1**). This enrichment for T:A>G:C transversions differentiates EAC from other cancers that have been studied by whole-genome sequencing, including breast, colorectal and hepatocellular cancers[16–18].

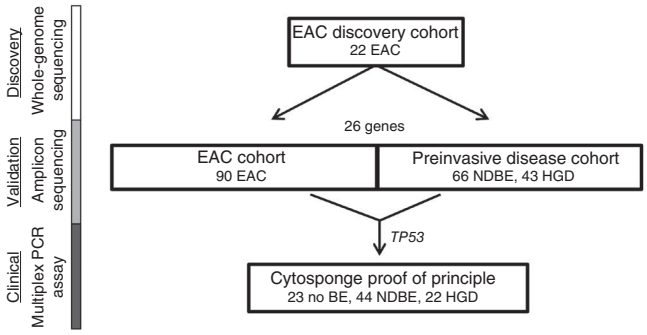

Figure 1 Flow chart showing the study outline. The number of samples used at each stage is given. The methodology used for each study phase is shown on the left side. EAC, esophageal adenocarcinoma; BE, Barrett's esophagus; HGD, high-grade dysplasia; NDBE, never-dysplastic Barrett's esophagus.

Targeted amplicon resequencing in a validation cohort of EACs

To highlight the genes most likely to be relevant in the development of EAC in individuals with Barrett's esophagus, we sought to determine the degree to which the mutated genes identified in our discovery cohort (n = 22 cases) were representative of the spectrum of mutations in an expanded cohort. Hence, a final list of 26 genes that were either mutated above the background rate or in pathways of interest was selected (**Supplementary Note**) and tested in a larger cohort (90 additional EACs; **Table 1** and **Supplementary Table 3**) using targeted amplicon resequencing. The findings confirmed and extended those of our discovery cohort and previous work from others[11,12,19], including the identification of recurrent mutations in SWI/SNF complex genes, such as *ARID1A* (**Supplementary Fig. 2**). Analysis of loss of ARID1A protein expression by immunohistochemistry in a cohort of 298 additional EACs found absent or decreased expression in 41% of samples (122/298). This finding suggests that alternative mechanisms of ARID1A downregulation might be present, although we did not identify any large-scale structural variants in the whole-genome sequencing data from our discovery cohort (data not shown).

We next combined the data from both the discovery and validation cohorts and identified 15 genes that were mutated in 4 or more samples (**Fig. 2**). These included genes previously identified as EAC candidate genes and several new candidates: *MYO18B*, *SEMA5A* and *ABCB1*. Comparison with recent EAC exome sequencing from Dulak *et al.* confirmed that these genes were recurrently mutated in an external data set (**Supplementary Table 4**). *TP53* was mutated in the majority of cases; however, 31% of cases had wild-type *TP53*. Although we did not have enough power to detect mutually exclusive mutations in our cohort, we could detect significantly co-occurring mutations. *SEMA5A* and *ABCB1* mutations occurred more often in the same tumor than would be expected by chance (Benjamini-Hochberg adjusted P value = 0.0021), although the reason for this association remains unclear.

Similar mutation frequency across disease stages

The stage specificity of mutations can be determined by examining cases at discrete stages of Barrett's esophagus carcinogenesis. Mutations occurring at disease stage boundaries would be candidate biomarkers of malignant progression. In addition, mutations occurring early in the development of disease should represent ideal targets for new therapeutic interventions because of their presence in the majority of cells in more advanced lesions owing to clonal expansion early in the natural history of tumors. We therefore sought to identify the mutation status of the 26 genes in our panel in Barrett's esophagus samples obtained from a prospective cohort of individuals undergoing endoscopic surveillance. This cohort included 109 Barrett's esophagus biopsies from 79 individuals (**Fig. 1**). We selected 66 NDBE samples from 40 individuals with Barrett's esophagus for whom there was no evidence for progression to dysplasia or malignancy (median follow-up time of 58 months, range of 4–132 months) and 43 Barrett's esophagus biopsy samples from 39 individuals of histopathologically confirmed HGD, the stage just before the development of invasive EAC (**Table 1**). We did not include low-grade dysplasia because of the poor agreement on histopathological grading of this lesion[20].

© 2014 Nature America, Inc. All rights reserved.

Table 1 Demographics of the case cohorts

		EAC cohorts		Barrett's esophagus cohorts		*TP53* analysis on Cytosponge		
		Discovery	Validation	NDBE	Barrett's esophagus with HGD	No Barrett's esophagus controls	NDBE	Barrett's esophagus with HGD
Number		22	90	40	39	23	44	22
Age (years)		68 (53–82)	66 (32–83)	63 (32–81)	71 (50–87)	53 (28–74)	61 (41–85)	66 (41–82)
Sex (M:F)		5:1	5:1	2:1	12:1	1:2	4:1	10:1
Stage (%)	I	4 (18.2)	14 (15.6)					
	II	6 (27.3)	14 (15.6)					
	III	11 (50.0)	49 (54.4)					
	IV	1 (4.5)	4 (4.4)					
	NA	0 (0.0)	9 (10.0)					
Barrett's esophagus length (cm)				4.8 (1–9)	8.6 (2–16)		5.8 (1–12)	8.5 (4–16)
Follow-up from EAC diagnosis (months)		28.5 (5–63)	18 (1–134)					
Total Barrett's esophagus surveillance (months)				58 (4–132)	1 (0–45)		56 (0–175)	24 (0–180)

Data shown reflect mean (range) for age and Barrett's esophagus length, number (percentage) for stage and median (range) for follow-up from EAC diagnosis and total Barrett's esophagus surveillance. Sex ratios (male:female) are rounded to the nearest whole number. NA, not available.

The findings were striking and unexpected. In the NDBE cohort, 21 of 40 individuals (53%) were found to have mutations in their Barrett's esophagus segment (**Fig. 3a**), with several biopsies containing multiple mutations (**Supplementary Table 5**). In total, we identified 29 SNVs and 7 indels in this cohort. Notably, the mutations identified in NDBE occurred in several genes previously identified as drivers in EAC[11,19] and other cancers[21,22], including in *SMARCA4*, *ARID1A* and *CNTNAP5* (**Fig. 3b**). Of interest, 7 of the 29 SNVs were mutations at T:A base pairs. Of these, 5 of 7 (71%) occurred at TT dinucleotide sequences, the mutational context identified as highly enriched in the EAC whole-genome sequencing data. Thus, this mutational process might well be active at the earliest stages of disease. Of the 43 HGD biopsy samples, 39 (91%) were found to have mutations in at least 1 of the genes in our panel, with a total of 67 SNVs and 7 indels. Hence, rather than the frequency of mutation in a given gene increasing across disease stages, we observed that, for the vast majority of genes, the mutational frequency was not significantly different between NDBE, HGD and EAC (Fisher's exact test with Benjamini-Hochberg correction for multiple testing; **Fig. 3b** and **Supplementary Table 6**). For 2 genes, *MYO18B* and *ARID1A*, we performed amplicon sequencing in an additional 25 NDBE and 11 HGD samples, increasing the cohort to a total of 91 NDBE and 54 HGD samples, but we did not identify any significant difference in the frequency of mutation between disease stages (**Supplementary Table 7**). Only *TP53* ($P < 0.0001$) and *SMAD4* ($P = 0.0061$) (**Fig. 3b**,**c**) exhibited mutational frequencies that would distinguish between disease stages and thus identify progression toward malignancy. *TP53* was found to be recurrently mutated in both HGD (72%) and EAC (69%) samples but was mutated in only a single NDBE sample (2.5%). *SMAD4* was mutated at a lower frequency (13%) and, intriguingly, was only found in EAC, the invasive stage of disease.

Clonal analysis of recurrent mutations

Having identified the occurrence of mutations in the earliest stages of disease development, we next sought to determine whether these mutations were fully clonal or subclonal in our original discovery cohort of 22 EACs. For each of the 15 genes mutated in ≥4 samples from our expanded cohort, we combined our high-depth resequencing of SNVs, copy number variant data and loss-of-heterozygosity (LOH) analysis to determine the fraction of tumor cells containing the mutation (**Supplementary Note**). If a mutation occurs at the earliest stage of disease development, before clonal expansion of the malignancy, we would expect that the mutation would be present in all cells of the tumor. For 7 of 15 genes—*SMAD4*, *TP53*, *ARID1A*, *SMARCA4*, *TLR4*, *CDKN2A* and *PNLIPRP3*—this was the case. Mutation in the other eight genes (*MYO18B*, *TRIM58*, *CNTNAP5*, *ABCB1*, *PCDH9*, *UNC13C*, *SEMA5A* and *CCDC102B*) was not always present in the major clone (**Supplementary Fig. 3**), suggesting that mutation of these genes might be selected for at multiple stages of tumorigenesis.

Application of mutational knowledge to a diagnostic test

The current clinical strategy for patients with Barrett's esophagus involves regular endoscopic examinations to try to identify patients

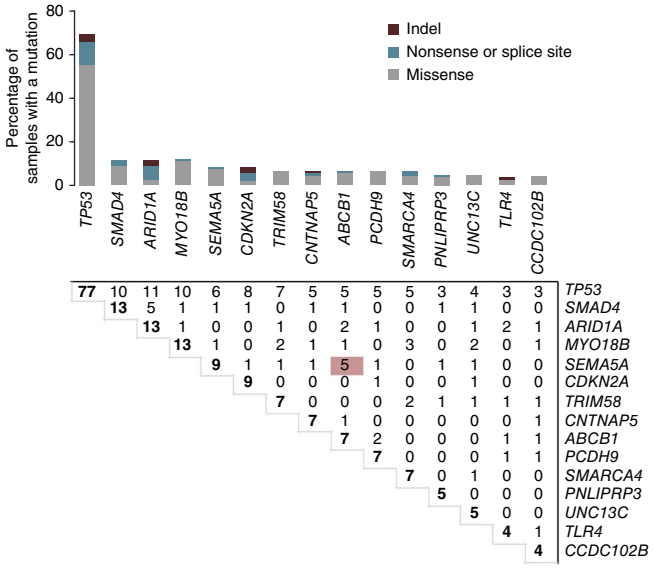

Figure 2 Mutation in esophageal adenocarcinoma. The bar graph on the top indicates the percentage of samples with aberrations for a given gene. The number in bold denotes the total number of mutations for each gene. Genes with 4 or more mutations in our EAC discovery and validation cohorts (combined total of 112 cases) were included. The proportions of missense, nonsense or splice-site, and indel mutations are shown. The matrix below shows the number of samples with mutations in both genes for each possible pairing of genes. The red highlighted box indicates significantly co-occurring mutations (significance was assessed empirically from 100,000 permutations, and the false discovery rate was nominally controlled using the Benjamini-Hochberg procedure to be less than 0.05; the co-occurrence of mutations in *ABCB1* and *SEMA5A* has an adjusted *P* value of 0.0021).

© 2014 Nature America, Inc. All rights reserved.

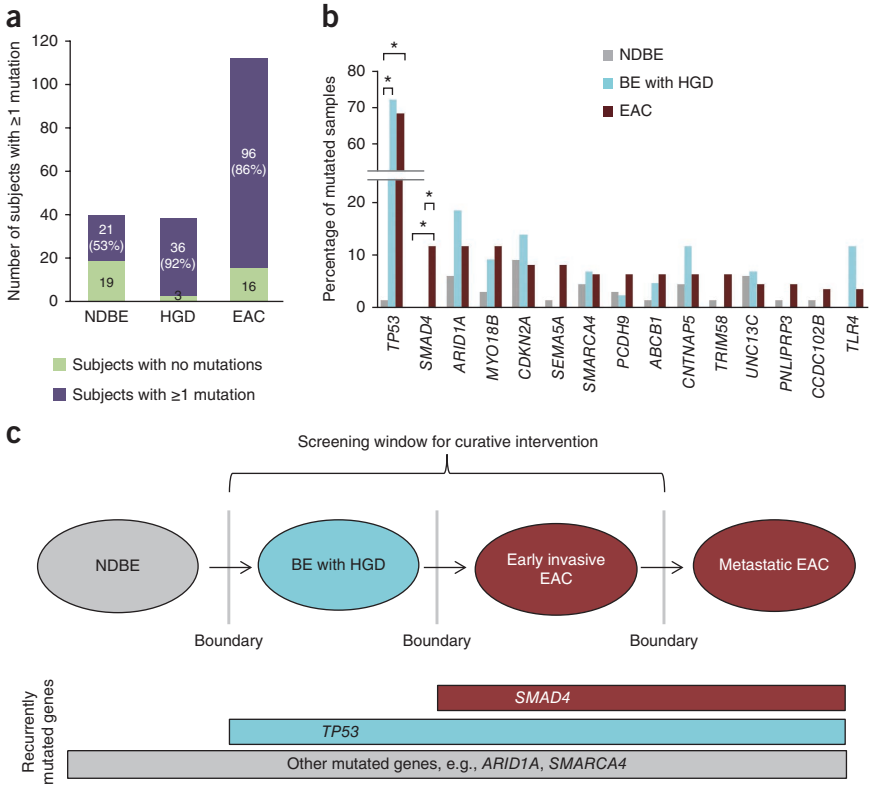

Figure 3 *TP53* and *SMAD4* mutations accurately define stage boundaries in the progression toward cancer, whereas other mutations appear to occur independent of disease stage. (**a**) Bar graph showing the number of subjects with NDBE ($n = 40$), Barrett's esophagus with HGD ($n = 39$) and EAC ($n = 112$) with at least 1 mutation in our panel of 26 genes. (**b**) Percentage of NDBE, Barrett's esophagus with HGD and EAC samples with mutations in recurrently mutated genes (mutated in ≥4 samples) identified in the EAC discovery cohort and EAC validation cohort. *TP53* and *SMAD4* are the only genes for which mutations separate the boundaries between never-dysplastic and dysplastic Barrett's esophagus (*TP53*) or cancer (*SMAD4*) (two-tailed Fisher's exact test with Benjamini-Hochberg correction for multiple testing, $*P < 0.05$). (**c**) Proposed model for the boundary-defining mutations in Barrett's esophagus carcinogenesis. The gray box depicts multiple other mutations that might occur and provide selective advantage at any stage of disease.

with dysplasia who are at high risk of progression to adenocarcinoma. This approach is highly controversial because of inherent difficulties in the accurate identification of dysplastic lesions, and recent data suggest that endoscopic surveillance of Barrett's esophagus is not effective[13,23]. The difficulties involved in endoscopic surveillance for Barrett's esophagus include sampling bias inherent in random-biopsies protocols and the subjective and time-consuming histopathological diagnosis of dysplasia. We therefore developed a new approach that has the potential to overcome these limitations of Barrett's esophagus surveillance. The strategy comprises a non-endoscopic device called the Cytosponge, which can be provided to patients in the primary care setting. This device collects cells from the entire esophageal mucosa, thus avoiding sampling bias, and can be combined with objective analysis of biomarkers for diagnosis[24,25]. Thus far, our focus has been on a biomarker for diagnosing Barrett's esophagus; however, because most patients with Barrett's esophagus will not progress to EAC, this biomarker for Barrett's esophagus needs to be combined with a biomarker (or a panel of biomarkers) to identify high-risk dysplastic patients. From the aforementioned sequencing data, *TP53* mutations fit the criteria of a good candidate marker for risk stratification, as these mutations discriminate between individuals with and without HGD, the key point of therapeutic intervention. Although the device samples abnormal tissue, the majority of cells collected are from normal gastric glandular tissue at the top of the stomach, as well as normal squamous areas of the esophagus, and any mutant DNA would therefore theoretically be in the minority, requiring a very sensitive assay (**Supplementary Fig. 4**). This situation is analogous to the detection of tumor cell–free DNA in blood as a biomarker in advanced malignant disease: sensitive assays have been developed to detect extremely low levels of mutant DNA against normal background[26,27]. We therefore took an analogous approach to detect mutations in Cytosponge material.

To determine whether mutations in Barrett's esophagus lesions could be detected in material collected from the Cytosponge, we first tested mutations previously identified in endoscopic Barrett's esophagus biopsies. Four individuals with HGD had *TP53* mutations and had also swallowed the Cytosponge (twice in the case of subject 4). For all four individuals, the specific *TP53* mutations were detected at an allele fraction (proportion of variant reads) of between 0.04 and 0.24 (**Table 2**).

We then tested whether we could detect unknown *TP53* mutations in material collected using the Cytosponge, as this ability would be required for a clinical test. We amplified the majority of the coding region of *TP53* (1,019/1,182 bp; 86%) by multiplexed PCR and sequenced the amplified DNA by massively parallel sequencing. *TP53* mutations were called *de novo* using TAm-seq (tagged-amplicon deep sequencing)[26] on samples from controls (no Barrett's esophagus) and individuals with Barrett's esophagus with no dysplasia, as well as individuals with Barrett's esophagus with HGD. As we expected, no *TP53* mutations were identified in samples from controls or individuals with Barrett's esophagus with no dysplasia (**Fig. 4a**), demonstrating 100% specificity for these mutations in differentiating between individuals with HGD and no dysplasia. In contrast, we identified *TP53* mutations in 19 of 22 individuals (86%) with HGD (details for

Table 2 Allele fractions for known *TP53* mutations previously identified by sequencing *TP53* on diagnostic biopsies

		AF on biopsy		AF on Cytosponge	
Case	Mutation	1	2	1	2
HGD_01	Chr. 17: c.7574003G>A	0.35	NA	0.04	NA
HGD_40	Chr. 17: c.7577538C>T	0.23	0.52	0.10	NA
HGD_03	Chr. 17: c.7578406C>T	0.51	0.72	0.06	NA
HGD_04	Chr. 17: c.7577551C>T	0.19	NA	0.14	0.24

For these four cases, the mutation can also be detected in material collected using the Cytosponge. Subject 4 swallowed the Cytosponge on two different occasions, 8 months apart, and the data for both Cytosponge samples are shown. NA, not applicable as no sample was taken; AF, allele fraction.

© 2014 Nature America, Inc. All rights reserved.

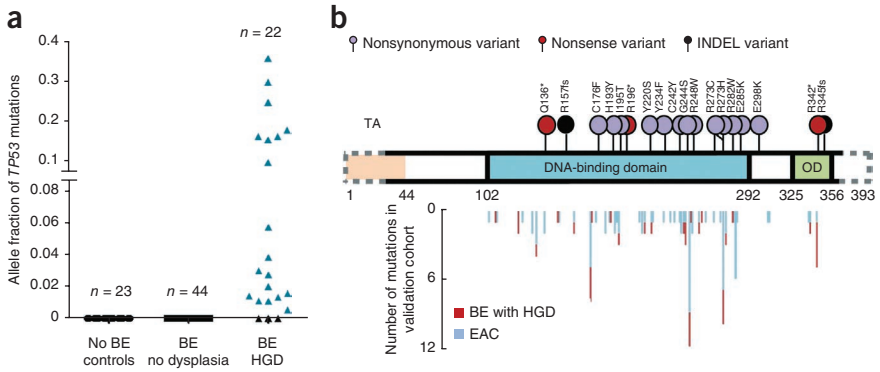

Figure 4 *TP53* mutations can be used to diagnose Barrett's esophagus with prevalent HGD on the Cytosponge. (**a**) The allele fraction of *TP53* mutations identified in Cytosponge samples is shown for the three case groups: no Barrett's esophagus ($n = 23$), Barrett's esophagus with no dysplasia ($n = 44$) and Barrett's esophagus with HGD ($n = 22$). (**b**) The positions of the *TP53* mutations identified for the Cytosponge samples are shown above the gene diagram compared with those found in the EAC and Barrett's esophagus HGD biopsy cohorts. The dashed line on the gene outline denotes the two small areas not covered by the multiplex PCR assay (amino acids 1–27 and 361–393). TA, transcription activation domain; OD, oligomerization domain.

individual mutations can be found in **Supplementary Table 8**). The allele fractions of the *TP53* mutations varied widely (between 0.006 and 0.357), but anything in this range could be called successfully, and mutations were mostly clustered in the sequence encoding the DNA-binding domain, as expected (**Fig. 4b**).

DISCUSSION

Barrett's esophagus is the only known precursor lesion of EAC, co-occurring in >80% of cases presenting *de novo*[28]; however, the majority of individuals with Barrett's esophagus will never progress to invasive disease[29]. There is therefore a need for sensitive and specific biomarkers that can identify those who are at risk of progression. As long ago as the Nowell hypothesis, a stepwise selection of genomic mutations has been assumed to be necessary for cancer development[30]. Somatic genomic variants should therefore be highly sensitive and specific markers of disease stage. By screening for our panel of recurrently mutated genes in a cohort of individuals with Barrett's esophagus who had never developed dysplasia and a cohort of individuals with Barrett's esophagus with HGD, we hoped to identify a stepwise accumulation of mutations across these disease stages. Surprisingly, we identified numerous mutations occurring in never-dysplastic Barrett's esophagus at detectable allele fractions (>10%). Intriguingly, the most prevalent gene mutations in EAC were also present at similar frequencies in NDBE and HGD samples including mutations in cancer-associated genes, for example *ARID1A* and *SMARCA4*, which encode members of the SWI/SNF chromatin-remodeling complex. These data demonstrate the complex mutational landscape that may be present even within tissue with a very low risk of malignant progression that has an entirely benign histopathological appearance. The exact role of these mutations at such an early stage of disease development remains unclear. However, it is known that clonal expansions occur frequently in Barrett's esophagus, and it is possible that these mutations provide an increase in fitness for a clone without leading to disruption of the epithelial architecture or providing the necessary cellular characteristics for invasion. A similar observation has been reported in endometrial cancer. In the normal population, ~35% of women harbor *PTEN*-mutant glands in their endometrial tissue, yet the lifetime risk of endometrial cancer is ~2.5% (ref. 31).

Our result has substantial implications for the specificity of tests aiming to use highly sensitive detection of mutations for the early diagnosis of malignancy[32]. Biomarkers predicting individuals at risk for cancer need to have substantial predictive power to distinguish between those who will and will not develop cancer. In our study, almost all recurrently mutated genes in EAC, including *ABCB1*, *CNTNAP5* and *MYO18B*, among others, are ruled out for use as surveillance tools for progression risk. Only mutation in *TP53* and *SMAD4* accurately defined the boundaries of disease states. The fact that mutation of *SMAD4* was only found in EAC provides a clear genetic distinction between EAC and HGD. However, the low frequency of *SMAD4* mutation (13%) makes it a suboptimal candidate for biomarker development. Furthermore, HGD rather than EAC is now the ideal point of clinical intervention owing to the advent of endoscopic therapy. We therefore focused on *TP53* for the proof-of-principle Cytosponge study. Sequencing technologies are now being introduced to routine clinical use, and genes of interest can be sequenced rapidly and with exquisite sensitivity, providing a quantitative readout[26].

We detected mutations in 86% of HGD Cytosponge samples using a simple, clinically applicable test. To improve the sensitivity of any early detection program, it will also be important to identify the genetic or epigenetic changes that drive HGD and EAC in the minority of patients without a detectable *TP53* mutation. In addition, as genetic diversity has been shown to predict progression to Barrett's esophagus, it may be possible to perform somatic mutation testing looking at both the presence and relative proportions of mutations in a panel of genes to identify patients with high-risk disease[33].

In conclusion, NDBE harbors frequent mutations affecting recurrently mutated genes in EAC. Given the low rate of progression to malignant disease in NDBE, the role of these mutations on the road to malignancy is unclear. It is generally accepted that the mutations observed in a tumor are accrued in a linear progression, with each step bringing the clone closer to the invasive endpoint. Our observation of mutations in almost all of the recurrently mutated genes in the tissue of individuals who have not gone on to develop malignancy argues against a major role of these mutations in the progression toward cancer. Although their recurrent nature suggests a role in clonal expansion at the premalignant stage, these mutations do not seem to provide any long-term increase in the likelihood of malignant progression. It is likely that sequencing of additional cohorts with greater sample numbers and differing demographics will identify further recurrently mutated genes in EAC, and these too will need careful analysis to determine the disease stage at which they occur.

From a clinical perspective, because the vast majority of recurrently mutated genes in EAC do not differentiate between the premalignant and malignant stages of disease, they therefore cannot be applied in a simple binary test—characterizing samples as mutant or non-mutant—as biomarkers of malignant progression. The Cytosponge provides a representative sample of the entire esophageal mucosa and, coupled with high-throughput sequencing,

© 2014 Nature America, Inc. All rights reserved.

is capable of sensitive and objective detection of HGD. This approach could be readily adapted as understanding of the genetic basis for this disease evolves. Furthermore, our systematic molecular approach to identify key mutations involved in the steps distinguishing preinvasive from invasive disease has applicability to other epithelial cancers amenable to early detection.

URLs. International Cancer Genome Consortium (ICGC), http://dcc.icgc.org/; Picard, http://picard.sourceforge.net/; FastQC, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

METHODS

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

Accession codes. Whole-genome sequencing data are available at the European Genome-phenome Archive (EGA; accession EGAD00001000704); however, data access approval is required via the ICGC data portal.

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

ACKNOWLEDGMENTS

Whole-genome sequencing of EAC is part of the International Cancer Genome Consortium (ICGC) through the Esophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium and is funded by Cancer Research UK. We thank the ICGC members for their input on verification standards as part of the benchmarking exercise. Cytosponge samples were collected as part of the Cancer Research UK–funded BEST2 trial. We thank M. Griffin, L. Lovat and K. Ragunath for their contribution to Cytosponge collection. The MRC developed the Cytosponge and also funded laboratory work through a program grant to R.C.F. J.M.J.W. was funded by a Wellcome Trust Translational Medicine and Therapeutics grant. R.C.F. and C.C. are supported by additional clinical research infrastructure funding from the NHS National Institute for Health Research (NIHR), the Experimental Cancer Medicine Centre Network and the NIHR Cambridge Biomedical Research Centre. Bioinformatics work was also supported by a Cancer Research UK program grant to S.T.

We thank the Genomics Core at the Cancer Research UK Cambridge Institute for their help with processing some of the Access Array experiments as well as for running the targeted resequencing experiments. We thank the IT department at the Cancer Research UK Cambridge Institute for their support. We thank F. Marass for assistance with data analysis. We thank the Human Research Tissue Bank, supported by the NIHR Cambridge Biomedical Research Centre, from Addenbrooke's Hospital as well as the University Hospital of Southampton Trust and the Southampton Experimental Cancer Medicine Centre. We are grateful to all patients who provided written consent for participation in this study, and the staff at Addenbrooke's and the University of Southampton Tissue Bank.

AUTHOR CONTRIBUTIONS

R.C.F. obtained funding and conceived and supervised the study. M.D.E. undertook and supervised the development of the whole-genome analysis pipeline. M.J.D., N.S., A.G.L., M.L.S., B.C. and C.L.A. undertook development of the whole-genome analysis pipeline. S.T., P.A.W.E., N.R., M.D.E., J.M.J.W., C.S.R.-I., N.S. and A.G.L. designed various aspects of the study. J.M.J.W., C.S.R.-I., T.F., M.B. and P.L.-S. extracted the samples and performed the molecular analyses. M.O. performed histopathological diagnosis. T.J.U., N.G., R.H., C.-A.J.O. and L.S. identified and collected samples. J.M.J.W., C.S.R.-I., N.S., A.G.L., M.M., M.J.D., M.L.S., C.L.A., B.C. and M.D.E. analyzed the data. J.M.J.W. performed the analysis of mutational context. A.P.M. designed the Fluidigm primers. J.D. developed the clinical database. C.C., A.O. and S.A. developed the strategy for and performed the verification experiments. J.M.J.W., C.S.R.-I., N.S., A.G.L. 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. Chin, L., Andersen, J.N. & Futreal, P.A. Cancer genomics: from discovery science to personalized medicine. *Nat. Med.* **17**, 297–303 (2011).
2. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
3. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. USA* **105**, 4283–4288 (2008).
4. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
5. Vogelstein, B. *et al.* Genetic alterations during colorectal-tumor development. *N. Engl. J. Med.* **319**, 525–532 (1988).
6. Goh, X.Y. *et al.* Integrative analysis of array–comparative genomic hybridisation and matched gene expression profiling data reveals novel genes with prognostic significance in oesophageal adenocarcinoma. *Gut* **60**, 1317–1326 (2011).
7. Quante, M. *et al.* Bile acid and inflammation activate gastric cardia stem cells in a mouse model of Barrett-like metaplasia. *Cancer Cell* **21**, 36–51 (2012).
8. Greaves, M. & Maley, C.C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
9. Varghese, S., Lao-Sirieix, P. & Fitzgerald, R.C. Identification and clinical implementation of biomarkers for Barrett's esophagus. *Gastroenterology* **142**, 435–441 (2012).
10. 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).
11. 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).
12. Agrawal, N. *et al.* Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discov.* **2**, 899–905 (2012).
13. Corley, D.A. *et al.* Impact of endoscopic surveillance on mortality from Barrett's esophagus–associated esophageal adenocarcinomas. *Gastroenterology* **145**, 312–319 (2013).
14. Shaheen, N.J. & Hur, C. Garlic, silver bullets, and surveillance upper endoscopy for Barrett's esophagus. *Gastroenterology* **145**, 273–276 (2013).
15. Hayes, D.F. *et al.* Breaking a vicious cycle. *Sci. Transl. Med.* **5**, 196cm6 (2013).
16. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
17. Fujimoto, A. *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44**, 760–764 (2012).
18. Bass, A.J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion. *Nat. Genet.* **43**, 964–968 (2011).
19. Streppel, M.M. *et al.* Next-generation sequencing of endoscopic biopsies identifies *ARID1A* as a tumor-suppressor gene in Barrett's esophagus. *Oncogene* **33**, 347–357 (2014).
20. Curvers, W.L. *et al.* Low-grade dysplasia in Barrett's esophagus: overdiagnosed and underestimated. *Am. J. Gastroenterol.* **105**, 1523–1530 (2010).
21. Wang, K. *et al.* Exome sequencing identifies frequent mutation of *ARID1A* in molecular subtypes of gastric cancer. *Nat. Genet.* **43**, 1219–1223 (2011).
22. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene *ARID1A* in ovarian clear cell carcinoma. *Science* **330**, 228–231 (2010).
23. Reid, B.J., Li, X., Galipeau, P.C. & Vaughan, T.L. Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis. *Nat. Rev. Cancer* **10**, 87–101 (2010).
24. Kadri, S.R. *et al.* Acceptability and accuracy of a non-endoscopic screening test for Barrett's oesophagus in primary care: cohort study. *Br. Med. J.* **341**, c4372 (2010).
25. Lao-Sirieix, P. *et al.* Non-endoscopic screening biomarkers for Barrett's oesophagus: from microarray analysis to the clinic. *Gut* **58**, 1451–1459 (2009).
26. Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra68 (2012).
27. Dawson, S.J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368**, 1199–1209 (2013).
28. Theisen, J. *et al.* Preoperative chemotherapy unmasks underlying Barrett's mucosa in patients with adenocarcinoma of the distal esophagus. *Surg. Endosc.* **16**, 671–673 (2002).
29. Bhat, S. *et al.* Risk of malignant progression in Barrett's esophagus patients: results from a large population-based study. *J. Natl. Cancer Inst.* **103**, 1049–1057 (2011).
30. Nowell, P.C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
31. Mutter, G.L. *et al.* Molecular identification of latent precancers in histologically normal endometrium. *Cancer Res.* **61**, 4311–4314 (2001).
32. Kinde, I. *et al.* Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. *Sci. Transl. Med.* **5**, 167ra4 (2013).
33. Maley, C.C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38**, 468–473 (2006).

© 2014 Nature America, Inc. All rights reserved.

OCCAMS Consortium: Stephen J Hayes[12,13], Ang Yeng[12], Anne-Marie Lydon[12], Soney Dharmaprasad[12], Sandra Greer[14], Shaun Preston[15], Sarah Oakes[15], Vicki Save[16], Simon Paterson-Brown[16], Olga Tucker[17,18], Derek Alderson[17], Philippe Taniere[17], Jamie Kelly[19], James Byrne[19], Donna Sharland[19], Nina Holling[19], Lisa Boulter[19], Fergus Noble[19], Bernard Stacey[19], Charles Crichton[18], Hugh Barr[20], Neil Shepherd[20], L Max Almond[20], Oliver Old[20], Jesper Lagergren[21–23], James Gossage[21–23], Andrew Davies[21–23], Robert Mason[21–23], Fuju Chang[21,22], Janine Zylstra[21,22], Grant Sanders[24], Tim Wheatley[24], Richard Berrisford[24], Tim Bracey[24], Catherine Harden[24], David Bunting[24], Tom Roques[25], Jenny Nobes[25], Suat Loo[25], Mike Lewis[25], Ed Cheong[25], Oliver Priest[25], Simon L Parsons[26], Irshad Soomro[26], Philip Kaye[26], John Saunders[26], Vincent Pang[26], Neil T Welch[26], James A Catton[26], John P Duffy[26], Krish Ragunath[26], Laurence Lovat[27], Rehan Haidry[27], Haroon Miah[27], Sarah Kerr[27], Victor Eneh[27], Rommel Butawan[27], Laszlo Igali[28], Hugo Ford[29], David Gilligan[29], Peter Safranek[29], Andy Hindmarsh[29], Vijayendran Sudjendran[29], Andy Metz[29], Nick Carroll[29], Michael Scott[30], Alison Cluroe[3], Ahmad Miremadi[3], Betania Mahler-Araujo[3], Olga Knight[1], Barbara Nutzinger[1], Chris Peters[21], Zarah Abdullahi[1], Irene Debriram-Beecham[1], Shalini Malhotra[3], Jason Crawte[1], Shona MacRae[1], Ayesha Noorani[1], Rachael Fels Elliott[1], Xiaodun Li[1], Lawrence Bower[2], Achilleas Achilleos[2], Jan Bornschein[1], Sebastian Zeki[1], Hamza Chettouh[1], Maria Secrier[2], Nadeera de Silva[1], Eleanor Gregson[1], Tsun-Po Yang[1] & J Robert O'Neil[31]

[12]Salford Royal National Health Service (NHS) Foundation Trust, Salford, UK. [13]Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK. [14]Wigan and Leigh NHS Foundation Trust, Manchester, UK. [15]Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK. [16]Edinburgh Royal Infirmary, Edinburgh, UK. [17]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [18]Department of Computer Science, University of Oxford, Oxford, UK. [19]Southampton General Hospital, Southampton, UK. [20]Gloucester Royal Hospital, Gloucester, UK. [21]St Thomas's Hospital, London, UK. [22]King's College London, London, UK. [23]Karolinska Institutet, Stockholm, Sweden. [24]Plymouth Hospitals NHS Trust, Plymouth, UK. [25]Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, UK. [26]Nottingham University Hospitals NHS Trust, Nottingham, UK. [27]University College London, London, UK. [28]Norfolk and Waveney Cellular Pathology Network, Norwich, UK. [29]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [30]Department of Pathology, Wythenshawe Hospital, Manchester, UK. [31]Edinburgh Cancer Research Centre, Edinburgh University, Edinburgh, UK.

© 2014 Nature America, Inc. All rights reserved.

ONLINE METHODS

Sample collection, pathology review and extraction. The study was approved by the Institutional Ethics Committees (REC 07/H0305/52 and 10/H0305/1), and all subjects gave individual informed consent. For the discovery cohort, patients with EAC were recruited prospectively, and samples were obtained either from surgical resection or by endoscopic ultrasound (EUS). This was an exploratory pilot with no prespecified effect size. Blood or normal squamous esophageal samples, distant by at least 5 cm from the tumor, were used as germline reference. All tissue samples were snap frozen in liquid nitrogen immediately after collection and stored at −80 °C. Before DNA extraction, one section was cut from each esophageal tissue sample, and staining with hematoxylin and eosin was performed. Cancer samples were deemed suitable for DNA extraction only after consensus review by two expert pathologists had confirmed tumor cellularity of ≥70%. Where blood was not available, the same review process was applied to normal esophageal samples to ensure that only squamous epithelium was present. For the discovery cohort, 127 cases were screened from 2 centers (Cambridge and Southampton). Sixty-three cases had the 70% cellularity required to meet ICGC criteria, and, of these, 22 tumor-normal pairs had sufficient quality and quantity of DNA extracted (total yield of ≥5 μg) and were submitted for whole-genome sequencing. Of the remaining 105 cases available, 90 had >50% cellularity, and all of these had sufficient DNA for amplicon sequencing. For all cases in the discovery and validation cohorts, there was a 260/280 ratio of 1.8–2.1. For the preinvasive disease cohort, we screened our entire 10-year prospective Barrett's cohort of >500 patients and selected cases for which there was frozen material available and for which review of the frozen section revealed a homogeneous grade of dysplasia in expert histopathological review. Cytosponge samples were all those available as part of an interim analysis from an ongoing prospective case-control study (BEST2).

DNA was extracted from frozen esophageal tissue using the DNeasy kit (Qiagen) and from blood samples using the Nucleon Genomic Extraction kit (Gen-Probe) according to the manufacturers' instructions. For validation, DNA was extracted using the AllPrepDNA/RNA Mini kit (Qiagen) according to the manufacturer's instructions.

Whole-genome sequencing. A single library was created for each sample, and 100-bp paired-end sequencing was performed under contract by Illumina to a typical depth of at least 50×, with 94% of the known genome being sequenced to at least 8× coverage and achieving a Phred quality of at least 30 for at least 80% of mapping bases. Typically, five lanes of a HiSeq 2000 (Illumina) flow cell were sufficient to achieve this, but samples were not multiplexed, so some exceeded these minimum standards by a large margin. Filtered read sequences were mapped to the human reference genome (GRCh37) using Burrows-Wheeler Alignment (BWA)[34], and duplicates were marked using Picard. As part of an extensive quality assurance process, quality control metrics and alignment statistics were computed on a per-lane basis. Aggregated quality control for each discovery cohort sample is given in **Supplementary Table 9**. Details of any tiles within flow cells that were removed after quality control are shown in **Supplementary Table 10**.

The FastQC package was used to assess the quality score distribution of the sequencing reads and enabled the identification of three lanes of sequencing that required trimming owing to a drop in quality in the later cycles of sequencing (see details in **Supplementary Table 11**).

Whole-genome sequencing mutation calling. Somatic SNVs were predicted using SomaticSniper V1.0.2 (ref. 35) run with the following command: somaticsniper -q 1 -Q 15 -F vcf -J -r 0.001000 -T 0.850000 -N 2 -s 0.01 –f. Output from SomaticSniper was then filtered using the following criteria derived from comparison of heuristic filters applied to SomaticSniper and VarScan 2 (ref. 36) and implemented using scripts provided in Koboldt *et al.*[36] and custom scripts (homopolymer filter). The filtering criteria were (i) germline and tumor sample coverage of ≥10, (ii) average variant position in the read between positions 10 and 90, (iii) percentage of supporting reads from each strand of ≥1% and ≤99%, (iv) total number of supporting reads of ≥4, (v) average distance of variant base from effective 3′ end of supporting reads of ≥20 bp, (vi) average mapping quality difference between reference and variant-supporting reads of <30, (vii) average difference in length of trimmed sequences between reference and variant reads of <25 bp, (viii) mismatch quality sum difference of <100 between reference and variant reads, (ix) adjacent homopolymer <5 bp away, and (x) nearest indel ≥40 bp away. In addition, all variants were compared to dbSNP129 and removed if overlapping with predicted germline SNPs. A median of 99.7% of the mappable genome was covered by at least tenfold in the tumor and matched germline sample and was thus defined as callable (**Supplementary Table 12**).

Candidate somatic indels were taken as the consensus between SAMtools[37] and Pindel[38], filtered to exclude indels present in the matched normal genome of any of the 22 samples (including non-consensus indel calls). Indels falling within coding regions and splice sites were manually inspected to generate a final list of calls. Variants were annotated with sequence ontology terms to describe consequence and position relative to Ensembl gene annotations. SNVs and indels were also annotated with matching or nearest features in UniSNP.

Verification of indel variants by PCR. A total of 25 coding indels, confirmed by manual review, were randomly selected for verification. Primers (sequences available upon request) were designed to amplify the predicted variant location. PCR was performed on both the tumor and normal DNA, and resulting products underwent Sanger sequencing. All traces were visualized using Chromas lite 2.01 and were manually reviewed for presence of the variant. An indel was considered somatic if it was present only in the tumor trace.

Verification of single-nucleotide variants by targeted resequencing. As part of a larger benchmarking exercise of our SNV calling pipeline, we selected 2,007 SNVs to be verified. These SNVs included those that had failed filters and those that had been predicted using the Illumina pipeline ELAND alignment plus STRELKA. The complete analysis of these data is ongoing, with the overall aim of optimizing the sensitivity of our SNV calling pipeline. After a preliminary analysis and comparison to the ICGC benchmarking exercise, we chose to increase the stringency of our filters for this pilot data set. The verification data in this manuscript are for only those SNVs passing these additional filters. Putative nonsynonymous SNVs (1,330 in total) underwent ultra-high-depth targeted sequencing. For eight samples, all nonsynonymous variants were sent for verification. In the remaining 14 cases, the selected SNVs were restricted to nonsynonymous variants in genes mutated in more than 1 sample. Amplicons were generated, indexed and pooled, and libraries were constructed as described by Shah *et al.*[39]. Samples were pooled separately, and a single lane of HiSeq 2000 data was generated for each, leading to a typical depth of coverage of 13,855 (interquartile range (IQR) of 3,408 to 39,059 for the amplicons). For 1,086 of these, ≥50-fold coverage was generated for both tumor and normal samples. An SNV was confirmed as somatic if the variant allele frequency was ≤1% in the matched normal sample and ≥2% in the tumor sample, and 1,081 SNVs met these criteria, giving a verification rate of 1,081/1,086 (99.5%).

Mutation validation in independent samples. Mutation validation was performed in a cohort of 90 additional EACs and 109 Barrett's esophagus biopsies, including 43 Barrett's esophagus biopsies with histopathologically confirmed HGD and 66 with no dysplasia. The Access Array microfluidics PCR platform (Fluidigm) together with high-throughput sequencing (Illumina) was used for targeted resequencing.

Amplicons with a median size of 180 bp (range of 100–200 bp) were designed using Fluidigm in-house software (primers available upon request)[26]. After two iterations of primer design, one gene remained uncovered by suitable amplicons (*DIRC3*), and this gene was removed from further analysis. Hence, in total, 26 genes were selected (**Supplementary Tables 13** and **14**). All primers were synthesized with universal sequences (termed CS1 and CS2) appended at the 5′ end.

Target amplification and sample barcoding were performed using the manufacturer's standard multiplex protocol (Fluidigm, Access Array User Guide). Primers were combined into multiplex pools ranging from 1 to 12 primer pairs. The Access Array system was used to combine PCR reagents (FastStart High-Fidelity PCR System, Roche) with 47 DNA samples (50 ng) plus a single

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/ng.3013

negative control and 48 sets of multiplexed primers in 2,304 unique 35-nl PCR reactions. Thermal cycling was then applied to amplify all selected targets by PCR. After PCR, a harvesting reagent was used to collect the amplified products of the 48-plex reactions, for each sample, through the sample inlets, for subsequent sequencing. Illumina sequencing adaptors and a 10-bp sample-specific barcode were attached through an additional 15 cycles of PCR. After the PCR products were barcoded, the PCR products from a small number of samples, as well as the water controls, were analyzed using the Agilent 2100 Bioanalyzer to ensure that the expected amplicon size was obtained and that there was no contamination across the PCR reactions. Products were then pooled and purified using AMPure XP beads with a bead to amplicon ratio of 1.8:1.0. The library was quantified using the Agilent Bioanalyzer and subjected to Illumina cluster generation. Paired-end sequencing (100 to 150 bp) was performed on a HiSeq 2000 or MiSeq instrument with a 10-base indexing (barcode) read, using custom sequencing primers targeted to the CS1 and CS2 tags for read 1, read 2 (index read) and read 3, according to the manufacturer's recommendations.

The methods used for the analysis of targeted sequencing data generated using TAm-seq have been reported previously[26]. Reads were demultiplexed using a known list of barcodes, allowing zero mismatches. Each set of reads was aligned independently to the hg19 reference genome using BWA in the paired-end mode[34]. Using expected genomic positions, each set of aligned reads was separated further into its constituent amplicons. A pileup was generated for each amplicon using SAMtools v1.17 (ref. 37). Using base quality and mapping quality cutoffs of 30, observed frequencies of non-reference alleles for every sequenced locus across all amplicons and barcodes were calculated. For each locus and base, the distribution of non-reference background allele frequencies/reads was modeled, and the probability of obtaining the observed frequency/number of reads (or greater) was calculated. Putative substitutions were identified on the basis of a probability cutoff (confidence margin) of 0.9995. Known SNPs obtained from the 1000 Genomes Project, dbSNP version 135 and regions covering amplification primers were discarded. Any substitutions observed at >5% allele frequency in more than half of the sequenced samples were discarded. Small insertions and deletions of sequence were predicted using the Genome Analysis Toolkit (GATK; this tool was preferred to SAMtools and Pindel for the higher depth of sequencing). All remaining putative mutations were annotated with sequence ontology terms to describe consequence and position relative to Ensembl gene annotations. In the final list, all nonsense or missense exonic mutations and splicing mutations with an allele frequency of 10% or greater at loci covered by at least 100-fold were retained. Three genes were removed at this stage owing to poor sequence coverage in all samples, *TLR1*, *TLR7* and *TLR9*, leaving a total of 23 genes for further analysis (**Supplementary Table 14**).

To verify the called mutations, all nonsynonymous mutations identified through Fluidigm Access Array sequencing were reamplified using the CS1- and CS2-tagged primer pair targeting the region and DNA from the original sample. Where available, DNA from a matched normal sample (blood, duodenum or normal squamous epithelium) was also amplified using the same tagged primer pair. Amplification was performed in 5-µl reactions (0.1 Phusion High-Fidelity DNA Polymerase (New England BioLabs), 1× Phusion Buffer, 4.5 mM $MgCl_2$, 5% DMSO, 0.2 mM dNTPs, 1 µM forward and reverse primer, 25 ng of genomic DNA). PCR cycling conditions were as follows; 50 °C for 2 min, 70 °C for 20 min and 95 °C for 10 min followed by 10 cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 1 min followed by 2 cycles of 95 °C for 15 s, 80 °C for 30 s, 60 °C for 30 s and 72 °C for 1 min followed by 8 cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 1 min followed by 2 cycles of 95 °C for 15 s, 80 °C for 30 s, 60 °C for 30 s and 72 °C for 1 min and 8 cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 1 min followed by 5 cycles of 95 °C for 15 s, 80 °C for 30 s, 60 °C for 30 s and 72 °C for 1 min. After amplification, 2 µl of each PCR reaction was collected and samples were pooled in batches of 12 reactions, such that only unique amplicons were contained in each pool. Thereafter, 5 µl of the pooled reaction mix was added to 2 µl of ExoSAP-IT (Affymetrix). Samples were incubated at 37 °C for 15 min followed by 80 °C for 15 min. The resulting product was diluted 1:100 in sterile water, and Illumina sequencing adaptors and a 10-bp barcode were attached to each pool using an additional 15 cycles of PCR (0.1 U of Phusion High-Fidelity DNA Polymerase (New England BioLabs), 1× Phusion Buffer, 4.5 mM $MgCl_2$, 5% DMSO, 0.2 mM dNTPs, 1 µM forward and reverse barcoding primers, 1 µl of ExoSAP-IT-treated PCR product (1:100 dilution)). Cycling conditions were as follows: heat activation at 95 °C for 2 min followed by 15 cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 1 min followed by a final elongation step of 72 °C for 3 min.

As previously, PCR products after barcoding were first analyzed using an Agilent 2100 Bioanalyzer to ensure that the expected amplicon size was obtained. They were then pooled together and purified using AMPure XP beads with a bead to amplicon ratio of 1.8:1.0. The library was quantified using the KAPA-Library Quantification kit (KAPA Biosystems) on a LightCycler 480 (Roche), diluted to 2 nM and subjected to Illumina cluster generation and sequencing on an Illumina MiSeq instrument (150-bp paired-end sequencing). Reads were demultiplexed using a known list of barcodes, allowing zero mismatches. Each set of reads was aligned independently to the hg19 reference genome using BWA in the paired-end mode[34]. SAMtools mpileup v1.17 (ref. 37) was used to generate counts for each nucleotide at the position of the putative somatic mutation. Samples with a mutant allele frequency of ≥3% and a depth of coverage of ≥50-fold were considered to be verified mutations. In addition, mutant allele frequency in the matched normal sample was required to be <1%. We additionally removed all mutations from the samples without a matched normal sample that were confirmed to be germline in the cohort of samples with sequenced matched normal samples.

Processing of capsule sponge specimens. Cytosponge capsules were swallowed by patients and then placed directly into preservative solution at 4 °C until processed further. Samples were vortexed extensively and shaken vigorously to remove any cells from the sponge material. Preservative liquid containing the cells was centrifuged at 1,350*g* for 5 min to pellet the cells. The resulting pellet was resuspended in 500 µl of plasma, and thrombin (Diagnostic Reagents) was then added in 10-µl increments until a clot formed. The clot was placed in formalin for 24 h before processing into a paraffin block. Sections (8 × 10 µm) were cut and placed in a tube for DNA extraction.

DNA extraction from Cytosponge samples. Genomic DNA was extracted from sections (8 × 10 µm) of the processed Cytosponge formalin-fixed, paraffin-embedded clot using Deparaffinization Buffer (Qiagen) and the QIAamp FFPE DNA Tissue 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 additional proteinase K was added to the samples roughly halfway through the 24-h incubation. After extraction, DNA was quantified using the Qubit dsDNA HS Assay kits (Invitrogen).

Sequencing for *TP53* mutations. A multiplex *TP53* PCR assay was used to sequence the coding region of the *TP53* gene. The multiplex consisted of 14 primer pairs[26], and these 14 primer pairs were divided into 2 different pools. The sequence of each of the primers, the genomic region that it amplified (coordinates are correct for the hg19 version of the human genome) and which pool it was part of are described in **Supplementary Tables 15** and **16**.

All *TP53* multiplex PCR reactions were performed in duplicate using Q5 Hot-Start High-Fidelity 2× Master Mix (New England BioLabs). The coding region of the *TP53* gene was first amplified using a PCR mix consisting of 1× Q5 master mix, 5% DMSO, a final concentration of 50 nM of each primer pair and up to 70 ng of FFPE DNA extracted from Cytosponge samples. Cycling conditions for PCR were as follows: initial denaturation at 95 °C for 30 s followed by 30 cycles of 95 °C for 10 s, 60 °C for 10 s and 72 °C for 15 s. A final extension at 72 °C for 2 min was also included to ensure elongation of all PCR products.

After the first round of PCR, 2.5 µl of pool 1 and 2.5 µl of pool 2 were combined. We added 2 µl of IllustraExostar, 1-step (GE Healthcare UK) to the 5 µl of pooled PCR products, and the Exostar reaction was performed (15 min at 37 °C followed by 15 min at 80 °C) to degrade the primers from the first reaction. Then, 1 µl of the pooled, Exostar-treated products was added to the barcode PCR in order to add a unique barcode as well as to add the sequencing primers onto the PCR products. The barcodes used for this second PCR were taken from Forshew *et al.*[26], and the core sequence for the barcode primers can be

© 2014 Nature America, Inc. All rights reserved.

found in **Supplementary Table 17**. Fluidigm barcode primers were used as they contain a sequence that binds to the CS1 and CS2 sequences present in the first *TP53* primers as well as the Illumina adaptors. The barcode PCR mix consisted of 1× Q5 master mix, 5% DMSO, a final concentration of 400 nM of each barcode primer pair and 1 µl of undiluted Exostar-treated DNA. Cycling conditions for PCR were as follows: initial denaturation at 98 °C for 30 s followed by 14 cycles of 98 °C for 10 s, 60 °C for 10 s and 72 °C for 30 s. A final extension at 72 °C for 2 min was also included to ensure elongation of all PCR products.

TAm-seq single-nucleotide variant and indel calling for detecting *TP53* mutations in Cytosponge samples. Indels were called by selecting outliers from locus-specific distributions of background mutation rates. Candidate insertions and deletions in each sample were compared with insertion and deletion rates at the same locus in samples from every other patient and were scored by means of *z* scores. Indels with a *z* score greater than or equal to 10, at least 200× coverage and at least five supporting reads were retained.

34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
35. Larson, D.E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
36. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
39. Shah, S.P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/ng.3013