## SUPPLEMENTARY METHODS

### Patient Selection

After IRB approval (Partners Healthcare 2012P002411; University of Pittsburgh PRO15080138), patients were identified from a retrospective cohort with endoscopic biopsies for surveillance of BE performed at one of 4 high-volume endoscopy centers within the University of Pittsburgh Medical Center (UPMC) clinical system between 1999 and 2016. All patients had histologically confirmed intestinal metaplasia.

We identified 143 potential cases, patients diagnosed with HGD or EAC more than 1 year after their index BE biopsy (originally diagnosed as either NDBE, IFD, or LGD) while under surveillance (also referred to as "progressors"). The original pathology slides and corresponding paraffin embedded biopsy samples could be obtained for 50 of these. Of these, 24 had sufficient archival tissue available for DNA sequencing analysis using tissue obtained > 1 year prior to first diagnosis of HGD or EAC. All HGD/EAC diagnoses were independently confirmed and/or diagnosed in multiple successive diagnostic/therapeutic procedures. We also identified 1941 potential controls with more than 5 years of total surveillance for BE and no evidence of progression to HGD/EAC. We selected 143 controls for detailed histopathologic review, matched in a ratio of approximately 3:1 with cases. As over 50% of progressors had a history of IFD/LGD during surveillance, we attempted to match the proportion of patients with a history of IFD/LGD during surveillance among controls. Among the 143 control samples initially selected, 73 had sufficient archival tissue for DNA analysis using a tissue sample taken at least 2 years prior to the end of follow up (last endoscopic exam or first ablation). Patient selection is described in a flow diagram in Supplementary Figure 1. The clinical characteristics of the cases and controls are compared to their respective UPMC reference populations in Supplementary Table 1. Eight controls had a radio frequency ablation (RFA) procedure to treat IFD/LGD with no evidence of progression to HGD or EAC in the electronic record (Supplementary Table 2). In these cases, the RFA procedure was used as the date of last follow up and to calculate the total surveillance time. With at least 5 years of pre-RFA surveillance, none of these patients had a diagnosis of HGD or greater. The clinical characteristics of the 73 controls are compared to the reference population in Supplementary Table 1.

### Sample Selection

The original pathology slides were reviewed by a single experienced gastrointestinal pathologist (JD) to confirm the absence of HGD or EAC and to select a representative formalin-fixed, paraffin-embedded (FFPE) block for analysis. A block was chosen that represented the highest grade of neoplasia (NDBE, IFD, or LGD) originally diagnosed up to that point in surveillance for 83/97 patients. In the remaining 14 patients, a secondary block had to be chosen due to lack of availability of the initially selected block. For 11 patients (3 cases and 8 controls) a sample with an original diagnosis of NDBE was chosen when there was a prior diagnosis of IFD or LGD. For 3 controls a sample with a diagnosis of IFD was used when the patient had a preceding diagnosis of LGD (sample selection is detailed in Supplementary Table 2 and Supplementary Figure 2). In all cases and controls, the selected block represented the highest grade of dysplasia originally diagnosed at that pre-progression surveillance endoscopy.

### P53 Immunohistochemistry and interpretation

If adequate tissue remained in the paraffin block, additional sections were prepared for p53 immunohistochemistry, which was performed using clone DO-7 (Ventana, prediluted). Antigen retrieval was performed with solution CC1 (Ventana) for 64 minutes; detection with Ventana Ultraview DAB kit on Ventana Benchmark Ultra. A single pathologist (JD) reviewed all p53 immunostains in conjunction with a hematoxylin and eosin stained slide on coded slides blinded to mutation and outcome data. Aberrant expression was defined as strong nuclear overexpression in at least 1 glandular complex (pit or crypt and associated glands) with or without surface expression. Definite absence of expression was not seen in any case (defined as complete absence of expression in at least a single glandular complex with adjacent wild

type glandular expression to serve as an internal control). The basal zone of the squamous epithelium, when present, also served as an internal control for positive expression.

**DNA isolation**

For each sample, 9-14 5 micron slides were macrodissected using a scalpel after expert pathologic review identified the areas on the slide with the highest percent of Barrett's epithelium (MS). DNA was then isolated using the Qiagen QIAamp FFPE DNA isolation kit (Germantown, MD). DNA was quantified using picogreen assay (Thermo Fisher, Waltham MA).

**Sequencing**

Prior to library preparation, DNA was fragmented (Covaris sonication) to 250 bp and further purified using Agentcourt AMPure XP beads (Beckman Coulter, Danvers MA). Size-selected DNA was ligated to sequencing adaptors (IDT) with sample-specific barcodes during manual library preparation (Kapa HTP KK8234, Kapa Biosciences/Roche).

Libraries were pooled in equal volume (1 ul each) and sequenced on an Illumina Miseq (Illumina, San Diego CA) to estimate the concentration based on the number of observed reads per sample. Normalized libraries underwent hybrid capture to enrich for the target genes (Supplementary Table 1) using a Agilent Sureselect Hybrid Capture (SureSelectXT, G9611A, Agilent, Santa Clara CA) kit designed for our custom set of genes and sequenced on an Illumina Hiseq 2500 and 3000. Pooled sample reads were de-convoluted (de-multiplexed) and sorted using Picard tools (see http://broadinstitute.github.io/picard/command-line-overview.html for details).

Reads were aligned to the reference sequence b37 edition from the Human Genome Reference Consortium using bwa aln (http://bio-bwa.sourceforge.net/bwa.shtml) using the following parameters "-q 5 -l 32 -k 2 -o 1" and duplicate reads were identified and removed using Picard tools. The alignments were further refined using the GATK for localized realignment around indel sites (http://gatkforums.broadinstitute.org/discussion/38/local-realignment-around-indels). Recalibration of the quality scores was also performed using GATK tools (http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr).

Samples were sequenced to an average mean target coverage of ~160x. Coverage of 30x or more for at least 80% of the targeted bases was required for samples to be considered for analysis. This was achieved for all but five samples, which were discarded from further analysis (not included in above sample totals).

Mutation analysis for single nucleotide variants (SNV) was performed using MuTect v1.1.4 (http://archive.broadinstitute.org/cancer/cga/mutect) and annotated by Variant Effect Predictor (VEP). The SomaticIndelDetector tool that is part of the GATK for indel calling was used. Consecutive variants in the same codon were reannotated to maximize the effect on the codon and marked as "Phased" variants. MuTect was run in paired mode using normal cell line DNA (CEPH1328) (Coriell Institute for Medical Research/NIGMS Human Genetic Cell Repository) that was processed along with the biopsy specimens. To identify any possible *TP53* mutations missed by MuTect, the BAM files for each sample (blinded to outcome status) were manually reviewed around the region of *TP53* in the Integrated Genome Viewer (IGV) (http://software.broadinstitute.org/software/igv/). To remove known germline polymorphisms and likely passenger events, all called variants were filtered to retain only likely pathogenic events. All variants that are listed in the exome sequencing project (ESP, http://evs.gs.washington.edu/EVS/) at an allele fraction greater than 0.005 in either the European American or African American normal population was removed. For tumor suppressor genes, only alterations that may lead to loss of function were retained. This included nonsense, frame shift, splice site, in-frame insertion and deletions and select missense mutations. Missense mutations were included only if they were listed in the COSMIC database (version 63, http://cancer.sanger.ac.uk/cosmic). For oncogenes, only COSMIC missense mutations or other known activating mutations, such as *EGFR* exon 19 in-frame deletions, were

retained for further analysis. To compare the *TP53* mutations found in this data set to *TP53* mutations commonly found in gastroesophageal adenocarcinoma (GEAC), sequencing results from the 705 GEACs in cBioPortal (cBioPortal.org) were used. While this procedure has been used multiple times in both a clinical and research setting to identify and analyze only likely pathogenic variants, it is a possibility that a few rare germline events escaped filtering.

**Calculation of total and allelic copy numbers from sequencing data.**

Copy number variants were identified using RobustCNV, an algorithm developed by the Center for Cancer Genome Discovery at the Dana-Farber Cancer Institute and validated for clinical use within the Center for Advanced Molecular Diagnostics at Brigham and Women's Hospital. RobustCNV relies on localized changes in the mapping depth of sequenced reads in order to identify changes in copy number at the loci sampled during targeted capture. This strategy includes a normalization step in which systematic bias in mapping depth is reduced or removed using robust regression to fit the observed tumor mapping depth against a panel of normals (PON) sampled using the same capture bait set. Observed values are then normalized against predicted values and expressed as log2ratios. A second normalization step is then done to remove GC bias using a loess fit. Finally, log2ratios are centered on segments determined to be diploid based on the allele fraction of heterozygous SNPs in the targeted panel.

The resulting copy-ratio profiles were then segmented using the circular binary segmentation (CBS) algorithm[1]. 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[2], refining the segmentation. For each segment, we combined the copy-ratio and allelic data to derive allelic copy-ratios.

The ABSOLUTE[3] computational tool (v1.4) was applied to the above allelic copy-ratios estimate several parameters for each sample in this study. These estimates include (i) the clonal fraction(s) of somatic copy-number alterations; (ii) the average ploidy of the aberrant subclones; (iii) the presence of antecedent genomic doubling in the dominant clonal lineage; and (iv) genome-wide absolute allelic copy numbers. 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 File 1). The ABSOLUTE solutions were reviewed manually to maximize concordance with the data (MDS, NC, and SLC), Supplementary file 1.

Counts of chromosome arm-level alterations were computed by taking the median of absolute major and minor allelic copy-numbers across each chromosome arm and reducing concordant events on each arm into single events. In cases where subclonal alterations were detected at chromosome-arm level, these were counted as additional events. Notably, several samples had no detectable SCNAs. We estimate that our data allowed detection of chromosome arm-level aneuploidies present in at least 5% of cells. Given this percentage of cells is lower than the estimated percent of Barrett's epithelium in all of the samples tested, we concluded that these samples were negative for such clones.

**Validation of ABSOLUTE for targeted sequencing panel**

*Cancer cell-line DNA mixing experiment.*

DNA extracted from a cancer cell line (HCC1143) was mixed with DNA from matched B-lymphocyte cell line (HCC1143BL) in various proportions (by mass). Each of these fractions

underwent whole-exome sequencing (WES) at the Broad Institute via the Illumina ICE hybrid-capture bait set.

*Simulating targeted panel data.*

The number of original targets in the WES data was downsampled to approximate the genomic and allelic coverage obtained using our targeted-exon + SNPs platform at DFCI. For each chromosome, we randomly downsampled from the set of exons containing the most heterozygous single nucleotide polymorphisms (SNPs). A total of ~5000 exons and ~1500 SNP regions were selected on average across samples, which was comparable to the number of baited regions for the targeted-panel study (4880 exons, 2088 SNP regions). The median and mean number of heterozygous SNPs was nearly identical when comparing the study sequencing to the down sampled validation data (Study panel; Median:2086 Mean:1979 vs Downsampled panel; Median: 1957 Mean:1849). Allelic copy-ratios were then computed using this downsampled set of targets and input into ABSOLUTE. The WES and downsampled allelic copy-ratios were plotted and compared. These were highly concordant for both the minor and major alleles across the different mixing fractions (RSME range = 0.072-0.1604, Supplementary Figure 3A). The estimated purity at each DNA mixing fraction (corrected for estimated-ploidy) for both the downsampled and WES ABSOLUTE results were plotted and have very high concordance with the true aliquot purities (RSME=0.06048538 and RSME=0.05688462 for the downsampled and WES data respectively), Supplementary Figure 3B. The size of the copy number calls for the downsampled data were analyzed and ranged from whole chromosomes to smaller than 50KB.

Method References:


1. Venketraman ES & Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 2007;**23**:657-63.

2. Olshen AB, Bengtsson H, Neuvial P, Spellman PT, et al. Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. Bioinformatics 2011;**27**:2038-46.

3. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. 2012;**30:**413–21.