**Sample description**

Briefly, the average number of reads across the 51 sequenced tissues was approximately 24.3 million (NE 33.6m, NDBE 24.7m, LGD 16.6m, EAC 25.0m) with ~87% uniquely mapped reads (NE 88.2%, NDBE 88.6%, LGD 86.7%, EAC 88.6%). The cohort consisted of 44 patients (43 males and one female) with a median age of 62 years (range 22 – 84 years). At the time of study completion, median length of follow-up for patients with NDBE and LGD was 53 months (range 33 – 74 months) and 63 months (range 21 – 76 months) respectively. During this time, one patient with BE progressed to LGD and one patient with LGD progressed to form invasive cancer. Both of these patients provided samples for sequencing of both lower and higher disease stages. However, due to the small sample size, meaningful longitudinal transcriptomic analyses for progression prediction were not possible. A summary of RNA-seq run metrics and patient demographics is provided in Supplementary Table 1.

**RNA Sequencing**

*RNA extraction and quality control prior to cDNA library construction*

After tissue-homogenization using a Polytron Homogenizer PT 3100 (Thomas Scientific, Swedesboro, NJ) RNA extraction was performed using the QIAGEN RNeasy Mini Kit, (Cat #74104, QIAGEN, Valencia, CA) following the manufacturer’s protocol. RNA yield and quality was measured using a BioSpec nano spectrophotometer (Shimadzu Scientific Instruments, Sydney, Australia) and integrity was verified on a Bioanalyzer separation chip (Agilent Technologies, Forest Hill, Victoria, Australia). 260/280 ratios were > 2.0 for all cases and a minimum RNA Integrity Number (RIN) of 6 was required for samples to be taken forward to library preparation for RNA sequencing. Median RNA yield per sample was 5.9 µg (IQR 3.3 – 10.5µg).

*cDNA library construction*

500ng of total RNA was used as input material for library preparation using the TruSeq Stranded Total RNA Sample Prep Kit (Illumina, USA) according to manufacturer’s instructions. Individual libraries were indexed as recommended by Illumina.

*Quantification and quality control of cDNA libraries*

Indexed DNA libraries were analyzed individually using an Agilent Technologies 2100 Bioanalyzer with the DNA 1000 kit according to the manufacturer’s instructions (Agilent Technologies, USA). Libraries were diluted and pooled to a final concentration of 10 nM each in nuclease-free H2O (Ambion, USA). Pooled libraries were quantitated using a Life Technologies Qubit 2.0 Fluorometer with the Qubit dsDNA HS Assay Kit (Life Technologies, USA) and further diluted to 2 nM. Final DNA library concentration was confirmed using a Qubit dsDNA HS Assay Kit. PCR-competent library DNA concentration was verified using the universal KAPA Library Quantification Kit for Illumina Sequencing Platforms according to manufacturer’s instructions (KAPA Biosystems, USA). An Applied Biosystems 7900HT Fast Real-Time PCR machine (Life Technologies, USA) was used for quantitative real-time PCR.

*Sequencing*

Total RNA sequencing was performed using the Illumina HiSeq2500 platform and version 3 chemistry for cluster generation and sequencing with 100 bp paired-end sequencing. Average library size was 290 bp. Twelve samples (NE 2, NDBE 5, EAC 5) were sequenced twice to allow for across-batch normalization as described below.

**Trimming, mapping and counting**

Fastq files were trimmed using trimgalore 0.33 (http://www.bioinformatics.babraham.ac.uk/projects/trim\_galore/) to ensure the quality of the reads. Remaining reads were mapped to Hg19 with STAR1 using Gencode v.18 as reference. Only unique reads were used in the downstream analysis. Counting of reads to features was conducted using HTSeq2 in union mode to Gencode v.18 and to the Hg19 RepeatMasker track downloaded from UCSC.

**Tissue purity**

Despite having microscopically estimated tissue purity, immune cell infiltration and stromal contamination was measured posthoc using ESTIMATE3 v 0.9. The combined score gives the tissue purity score where 1 is 100% pure tissue.

**Differential gene expression analysis**

The resulting count matrix from HTSeq was imported into R, and to filter out lowly expressed genes, only features with at least 1 count per million reads in minimum 8 samples were used for further analysis. Batches were normalized using RUVS from RUVseq4. Normalization of reads counts between samples was done with upper-quartile normalization. Differential expression between the groups were determined with EdgeR5.

**EAC gene signature discovery**

Relevant genes were identified using the R package CMA6. All filtered normalised counts per million (cpm) from EdgeR were used in the analysis. For increased stringency, we used all four available splitting rules (leave-one-out- (LOOCV), Monte-Carlo- (MCCV), fivefold-cross-validation (fiveCV) and bootstrap) to generate learning sets. We then applied gene-selection ranking for each method to extract the genes driving the differences between EAC and BE. The top 100 genes from each method and iteration were combined and filtered for genes present in ≥niterations-1 per method.

To increase our confidence in genes identified we took candidates present in ≥2 methodsforward for further evaluation. For further stringency genes were only classified as true potential drivers if they also displayed significant differential expression between EAC and BE in differential gene expression analysis.

**Validation of differentially expressed genes and gene expression signature in independent external datasets**

Two publically available microarray datasets (GSE37203 [n = 46], GSE26886 [n = 41]) using Affymetrix 2.0 HTA arrays containing gene expression data on non-dysplastic Barrett’s esophagus and esophageal adenocarcinoma samples were accessed through ArrayExpress. Similarly, datasets containing data on gastric (GSE19826) and colorectal (GSE23878) cancers and their corresponding normal corresponding tissues were also obtained. Raw data were accessed and normalized per platform using rma-normalisation for the Affymetrix arrays. Random forest analysis was performed with the R package randomForest with 500 trees to predict outcomes. Conditional inference classification trees were visualized using the party package in R. Calculations and plotting of the Area Under the Receiver Operating Characteristic (AUROC) performed using the R package pROC.

**Weighted co-expression gene network analysis.**

Network analysis was conducted with the R package WGCNA7. Briefly, log2 normalised RNA-seq counts (cpm +1) were imported from EdgeR. Only genes with at least one cpm in all samples were used to remove noise. SoftThreshold power was chosen when the line flattened with a value of 10. Modules 90% or more similar to each other were collapsed into one. Gene Ontology enrichment of each module was conducted with the GOenrichmentAnalysis, and only the top 10 terms for each module was plotted. Cytsocape8 was used to visualize the protein-protein interactions. The networks were imported with the cytoscape app PINA4MS (HC Lee, M Pinese, L Bourbon, I Rooman, RJ Daly, AV Biankin, J Wu, manuscript under preparation. URL: http://apps.cytoscape.org/ apps/pina4ms). For the brown module only the genes with ≥10 connections were plotted.

**Exploratory functional analysis of identified target genes**

Identified genes from both the gene signature and network analyses were explored functionally using EAC cell line data provided through the Broad and Dana Faber Institute’s Project Achilles9. This dataset is designed to detect tumor cell line vulnerabilities/gene dependencies by performing genome-wide, lentiviral-based pooled shRNA screens across hundreds of cancer cell lines. Screening of gene dependency for tumor cell viability and/or proliferation is performed with a library of over 50,000 shRNAs targeting over 11,000 genes with experiments performed in quadruplicate of each tested cell line. The quantile normalized shRNA level log fold change score was used to determine relative gene dependency. This score normalizes the shRNA enrichment/depletion values per replicates of the cell lines to the same scale. It is indicative of the relative amount of viable tumor cells left following shRNA transduction/gene knockdown. Higher values occurring through shRNA sequence enrichment indicate low tumor cell gene dependency due to no alterations in proliferation rates. Consequently, low levels indicate high gene essentiality for tumor cell viability.

**Generation and validation of EAC-specific gene signature**

For EAC gene signature discovery a machine-learning approach using the R package CMA6 with four cross-validation techniques was first performed. The resulting RNA-seq derived gene signature was then taken forward for external validation using two publicly available micro-array datasets of Barrett’s and EAC tissue samples (total n = 87)10,11. One study consisted of 31 Barrett's esophagus and 15 EAC samples10, while the other consisted of 20 Barrett's esophagus and 21 EAC samples11. A random-forest analysis was subsequently performed using this RNA-seq derived gene signature for key-driver detection and in an attempt to reduce overfitting whilst maintaining model parsimony. The resulting key-driver gene signature was then subject to a final external validation, using a third publically available micro-array dataset of Barrett’s and EAC tissues (n=84)12.

**Correlation analysis of lncRNAs**

Neighboring lncRNA-mRNA genes were found using the GenomicRanges13 packages for R. Briefly, all lncRNAs, downloaded as a separate gtf file, were removed from the main Gencode v18 gtf file where only protein coding genes were kept. All genes were collapsed to their respective transcription start site (TSS), and to find the neighboring mRNA for lncRNAs the nearest() function was used on the lncRNAs TSS, irrespectively of strand. Only TSS <10,000bp from each other were classified as neighboring genes. The neighbor pair was then extracted from the normalized EdgeR count file, and correlation was tested with the R package Hmisc. Similarly, the 15 available lncRNA-mRNA pairs from the TCGA EAC data set was analysed the same way for validation.

K-means clustering was performed using pheatmap package. All genes differentially expressed between EAC and NE/NDBE/LGD was used and collapsed into 10 clusters based on their corresponding expression profile. These clusters where then checked for number of lncRNAs present, and further enriched for biological functions.

**Single-nucleotide variants (SNV) calling**

For mutational analysis, Samtools mpileup14 and bcftools were used to acquire the consensus sequence of the uniquely mapped reads, while vcfutilis and vcffilter was utilised to call SNVs. The produced VCF file was analyzed using VEP towards RefSeq genes to identify and classify the alterations. Furthermore, VEP was utilized to identify mutations seen in the COSMIC database15. To study previously reported EAC gene mutations the read coverage was set to 10 or more reads. Only mutations present in exons of each gene were counted, and all mutations present in normal tissues were discarded to avoid potential RNA-editing sites. Percentage of mutations present in cbioportal EAC broad institute data set was downloaded using cgdsr. For transcriptome wide mutation signature calling the coverage was set to 20 or more reads for stricter filtering. The mutational signature analysis was performed using the R package SomaticSignature 16

**Immunohistochemistry (IHC) and scoring**

Protein expression of 5 genes (PRKDC, CTSL, COL17A1, KLF4 and E2F3) identified through our EAC gene signature analysis pipeline as well as through WGCNA was examined by immunohistochemistry (IHC). In total, 30 tissue sections from two patient cohorts were stained. The first cohort consisted of patients who provided samples for the RNA-sequencing analysis (NE = 5, NDBE = 5, EAC = 5). As all of the biopsies used for RNA-extraction were used, the samples for IHC used different biopsies from the same index endoscopy that contained a high proportion of the same disease defining epithelium. Subsequently, a separate independent cohort of patients was used for validation of the protein expression findings (NE = 5, NDBE = 5, intramucosal cancers [IMC] = 2, EAC = 4; see Methods and Supplementary Table 1). The stained slides were scored by two investigators blinded to expression information. For statistical analysis data from both cohorts and IMC as well as EAC sample data were combined.

All staining procedures were optimized and performed using the Leica Bond RX platform. The following antibodies, dilutions and incubation times for epitope binding were used: Anti-CTSL antibody (Abcam mouse monoclonal, ab197278; 1:1000; 30mins incubation), Anti-PRKDC antibody (Sigma-Aldrich, rabbit polyclonal, HPA035174; 1:500, 30mins incubation), Anti-E2F3 antibody (Abcam rabbit polyclonal, ab50917; 1:1000; 60mins incubation), Anti-KLF4 antibody (Sigma-Aldrich, rabbit polyclonal, HPA002926; 1:100; 60mins incubation) and Anti-COL17A1 antibody (Sigma-Aldrich, rabbit polyclonal, HPA043673; 1:100; 60mins incubation). All specimens used for IHC were formalin fixed paraffin embedded sectioned at 5µm thickness with positive controls placed on the superfrost plus/polysine slides and subsequently incubated in a 60°C oven for 2 hours to ensure maximum adhesion. Following dewaxing using Bond Dewax Solution (AR9222), a heat-induced epitope retrieval protocol (HIER) was used with either EDTA or citrate based epitope retrieval buffers (ERB): Anti-CTSL (HIER 20mins with ERB1 [Citrate, pH6] at 100°C), anti-E2F3 (HIER 20min with ERB2 [EDTA, pH9] at 100°C), anti-PRKDC (HIER 20mins with ERB1 at 100°C), anti-COL17A1 (HIER 20mins with ERB2 at 100°C), anti-KLF4 (HIER 30min with ERB1 at 100°C). Peroxidase blocking was performed for 5mins with two subsequent washes using Bond Wash solution (10x concentrate, AR9590). Then primary antibody incubation (times see above) with subsequent washes (3x) and post primary and polymer incubation at 15mins each with washing in between each step (3x each) occurred. Counter-staining was performed with DAB for a total of 10mins each and stained slides then washed with both Bond Wash Solution and deionized water. Scoring of slides then occurred by two investigators blinded to gene expression information. Staining of each element (nucleus, cytoplasm, cell-membrane and stroma) was evaluated using a four step scale as described previously

17: (0) no staining or equal to background, (1) weak diffuse staining, (2) moderate staining in at least 10% of cells, (3) strong immunostaining in a majority of cells. Where applicable scores were tallied for quantitative statistical analyses. In cases of disagreement consensus was reached after combined re-analysis on a multi-headed microscope.

**Verification and analysis of differentially expressed lncRNAs**

Following standard protocols, 500ng of RNA from normal squamous tissues (n = 13), NDBE (n = 13), LGD (n = 6) and EAC (n = 9) from the sequenced samples was used to generate cDNA using the Superscript III cDNA synthesis kit (Life Technologies/ThermoFisher Scientific, Scoresby, Victoria, Australia). After retrotranscription, qRT-PCR was performed in duplicate using LightCycler 480 SYBR Green I MasterMix (Roche Diagnostics, Castle Hill, NSW, Australia) on a LightCycler 480 II PCR Instrument (Roche Diagnostics, Castle Hill, NSW, Australia). Relative quantification was performed using the ΔCt method and normalized to the geometric mean of the endogenous control genes TBP and C1orf4318. in our RNA-seq dataset to determine the control genes with most stable and high expression across the analyzed tissue types. All primer sequences are provided in Supplementary Table 7. Following normality and variance distribution analysis, one-way ANOVA was used with post-hoc pairwise t-testing using Benjamini-Hochberg p-value corrections to adjust for multiple comparisons. Differences between means were regarded as statistically significant if p ≤ 0.05 (\*p ≤ 0.05, \*\*p ≤ 0.01, \*\*\*p ≤ 0.001, #p > 0.05). All validation barcharts represent mean ± s.e.m.

1. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2012;29:15–21.

2. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 2014.

3. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications 2013;4.

4. Risso D, Ngai J, Speed TP, et al. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol 2014;32:896–902.

5. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–140.

6. Slawski M, Daumer M, Boulesteix A-L. CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. BMC Bioinformatics 2008;9:439.

7. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.

8. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 2003;13:2498–2504.

9. Cowley GS, Weir BA, Vazquez F, et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Scientific Data, Published online: 30 September 2014; | doi:10.1038/sdata.2014.35 2014;1:140035.

10. Silvers AL, Lin L, Bass AJ, et al. Decreased Selenium-Binding Protein 1 in Esophageal Adenocarcinoma Results from Posttranscriptional and Epigenetic Regulation and Affects Chemosensitivity. Clinical Cancer Research 2010;16:2009–2021.

11. Wang Q, Ma C, Kemmner W. Wdr66 is a novel marker for risk stratification and involved in epithelial-mesenchymal transition of esophageal squamous cell carcinoma. BMC Cancer 2013;13:137.

12. Kim SM, Park Y-Y, Park ES, et al. Prognostic Biomarkers for Esophageal Adenocarcinoma Identified by Analysis of Tumor Transcriptome. PLoS ONE 2010;5:e15074. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0015074.

13. Lawrence M, Huber W, Pagès H, et al. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 2013;9:e1003118.

14. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–2079.

15. Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 2004.

16. Gehring JS, Fischer B, Lawrence M, et al. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics 2015:btv408.

17. Buskens CJ, Van Rees BP, Sivula A, et al. Prognostic significance of elevated cyclooxygenase 2 expression in patients with adenocarcinoma of the esophagus. 2002;122:1800–1807. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12055587.

18. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29:569–574.