Supplementary information, Data S1

Materials and Methods

Human small cell esophageal specimens

The Institutional Review Board (IRB) of Sun Yat-sen University Cancer Center, Guangzhou, China, approved this study. Human samples were obtained from patients under IRB-approved protocols following the provision of written informed consent. The samples were primary tumors diagnosed as Stage I-IV SCCEs. All samples were reviewed by at least two independent expert pathologists and the diagnosis of SCCE was histomorphologically confirmed by haematoxylin and eosin (H&E) staining and immunohistochemistry (IHC) for synaptophysin, chromogranin A, NSE, CD56 and Ki67. Any sample with squamous or adenocarcinoma differentiation were excluded. These tumor samples were pathologically assessed to have a purity of at least 60% and minimal necrosis. Additionally, adjacent non-tumorigenic esophageal tissue was provided as matching normal samples and was confirmed to be free of tumor contaminants by pathological assessment. Furthermore, mass spectrometric fingerprint genotyping of 21 common SNPs was used to verify that both tumor and normal DNA were derived from the same patient.

WES was performed on 55 SCCE tumor samples and matched normal tissues that passed the pathology assessment and DNA quality controls. In addition, we profiled the copy number alterations of a total of 24 patients using the OncoScan FFPE CNV Assay and performed ultra-deep targeted sequencing of 20 of the 55 SCCE samples. Clinical

information including sex, age at diagnosis, stage, family history, survival status, smoking status, alcohol consumption status, tumor location, chemotherapy and radiotherapy was collected (Supplementary information, Table S1). The median follow-up time for this cohort of 55 patients with SCCE was 26.4 months, and 14.5% (8/55) of the patients were alive at the time of final follow-up.

DNA extraction

Nucleic acids were extracted from fresh-frozen tissue specimens that had been cut into 20-30 sections 20 μm thick on a cryostat (Leica Microsystems GmbH, Wetzlar, Germany) at a constant temperature of -80 °C (Leica). In the case of FFPE samples, 6-10 sections 10 μm thick were prepared. DNA was extracted from fresh-frozen tissues using a DNeasy Blood & Tissue Kit according to the manufacturer's protocol. For FFPE tissues, DNA was extracted using a GeneRead DNA FFPE Kit following the manufacturer's protocol. The quantity and quality of DNA were determined using a Qubit Fluorometer and agarose gel electrophoresis.

Whole-exome sequencing

The qualified genomic DNA sample was randomly fragmented by Covaris (Covaris, Woburn, MA, USA) into fragments averaging 200-250 base pairs (bp) in length. The fragments were end repaired, and an extra A base was added to the 3' end. Illumina adapters (Illumina, San Diego, CA, USA) were ligated to both ends of the resulting fragments and proper cycles of PCR amplification were applied to each sample. After

each step, an Agilent SureSelect Human All Exon 51 Mb Kit (Agilent Technologies, Santa Clara, CA, USA) was used for whole exome capture according to the standard manufacturer's protocol. The final library was quantitated in two ways: determining the average molecule length using the Agilent 2100 bioanalyzer instrument (Agilent DNA 1000 Reagents), and quantifying the library by qRT-PCR (TaqMan Probe; Thermo Fisher Scientific). The qualified libraries were then loaded on Hiseq2000 platform (Illumina) and the sequences of each library were generated as 2 × 90 bp paired-end reads.

Data processing

Sequencing reads were discarded if they contained:

- adaptor reads;
- low-quality reads, with too many Ns (> 10%);
- low-quality bases (> 50% bases with quality < 5).

High quality paired-end reads were then subjected to gapped alignment to the UCSC human reference genome (hg19) using BWA-MEM (v0.7.12)[1]. Picard (v1.84; http://broadinstitute.github.io/picard/) was used to sort and mark duplicate reads caused by PCR. Then local realignment and base quality score recalibration of the BWA-aligned reads were conducted using the Genome Analysis Toolkit (GATK; v3.4, http://www.broadinstitute.org/gatk)[2].

Somatic mutation detection

Using the default parameters, MutTect[3] (v1.1.4) was used to detect somatic substitutions. Somatic indels were first detected by GATK using the default parameter. A candidate indel was retained if the following criteria were met:

- The median/mad of indel offsets from the starts or ends of the reads were ≥ 5 bp;
- The depth of the site in both tumor and normal samples was ≥ 5×;
- The average mapping qualities of the reads supported reference and indel in tumor samples was ≥ 20;
- They were not located in simple repeat regions (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/); and
- They passed two statistical tests (Fisher's exact test of strand bias and Fisher's exact test of the reads supported reference between tumor and normal tissues).

All SNVs and indels were annotated using ANNOVAR [4] and then eliminated if they were registered in dbSNP release 142. In addition, we filtered out mutations with a frequency of > 0.01 in the following databases: 1000 Genome Project April 2015 release; the National Heart, Lung, and Blood Institute (NHLBI) Grand Opportunity (GO) Exome Sequencing Project (ESP) ESP6500SI-V2 release; and The Exome Aggregation Consortium (ExAC) database release 0.3. Mutations in *TP53* and *RB1* were reviewed manually in the respective Binary Alignment/Map (BAM) files without the filtering steps.

Somatic mutation validation

Somatic mutations were validated by Sanger sequencing or ultra-deep targeted sequencing. For Sanger sequencing validation, PCR primers for putative somatic variants were designed using Primer3 (https://sourceforge.net/projects/primer3/) and used to amplify the source DNA from the tumor and matched normal samples. PCR was performed on a 96-well PTC-200 PCR System (Bio-Rad Laboratories, Inc., Hercules, CA, USA), and 20 ng of template DNA from each sample was used per reaction. The products were sequenced using a 3730×l DNA Analyzer (Applied Biosystems, Foster City, CA, USA). All sequences were analysed by novoSNP[5]. A mutation was determined to be successfully validated if it was confirmed in the tumor sample and absent in the normal sample.

Ultra-deep target region sequencing validation was performed on 118 selected genes, including genes frequently mutated in patients with SCCE or reported in ESCC, HNSCC and EAC. One microgram of genomic DNA from each sample for validation was used for hybrid capture and library construction. Libraries were then sequenced on HiSeq 4000 platform with 2 × 100-bp paired-end reads. Sequenced reads were processed as WES data as described above. SNVs were validated by at least three reads supporting the mutant allele presented in WES. In addition, the Pearson correlation coefficient was calculated to estimate consistency of mutation frequency identified in WES and ultra-deep targeted sequencing. Indels were manually validated using the SAMtools (http://samtools.sourceforge.net/) ‘tview’ command. We calculated the frequency of each pileup site to see whether the site was detected by target sequencing.

Mutational signature analysis

The percentage of somatic mutations was calculated for each type of substitution to generate a mutation counts matrix. This matrix contained mutation counts along 96 tri-nucleotide mutation contexts (rows) across 55 samples (columns). Then, the mutational signature analysis was performed using a BayesNMF algorithm[6, 7]. Default parameters were applied, except for the following: the parameter (reduce the effect of hyper-mutant samples in the signature discovery) was set to FALSE (default TRUE); and (L2KL [half-normal priors]) was chosen as the priors for W and H (default L1KL [exponential priors] is recommended). Using this matrix, we identified three significant signatures. We used the cosine similarity to compare our three signatures with thirty reported COSMIC signatures (http://cancer.sanger.ac.uk/cosmic/signatures).

Analysis of significantly mutated genes

A method previously reported[8-10] was used to compute the significance of the observed mutations in each gene. Both mutation prevalence and functional impact were taken into consideration. Functional impact was evaluated as mutation score assigned in the following order: missense < in-frame indel < mutation in splice sites < frameshift indel = nonsense. Also, different types of missense mutations were assigned different scores based on the BLOSUM80 matrix. The *P*-value for each gene was calculated from the background distribution of mutation score for each gene and the test statistics from the observed mutation scores across samples. *P*-values were adjusted using the Benjamini-Hochberg method. Finally, the significantly mutated genes were selected by a threshold

q-value ≤ 0.01. We discarded two significantly mutated genes *CSMD3* and *OR52L1* that belong to the CSMD and olfactory receptor gene families. Genes in these two families were frequently mutated in all types of cancers and likely to represent background noise[11,12].

Somatic copy number alteration calling from whole-exome sequencing

Using default parameters, EXCAVATOR[13](v1.1.2) was applied to determine the SCNAs of each pair of matched normal and tumor samples. To infer significantly amplified or deleted genomic regions, we implemented the GISTIC2[14] algorithm using copy numbers in 100-kilo bases (kb) windows instead of SNP array probes as markers. Parameters were set as follows: -genegistic 1 -broad 1 -brlen 0.7 -conf 0.99 -armpeel 1 -js 4. The thresholds for gene copy number alterations were: amplifications, GISTIC score = 2; gains, GISTIC score = 1; losses, GISTIC score = -1; deletions, = -2. We compared SCCE SCNA data to ESCC, EAC, HNSCC and SCLC. SCNA data for ESCC, EAC and HNSCC were downloaded from The Cancer Genome Atlas (TCGA, http://gdac.broadinstitute.org/). For SCLC, SCNA data were obtained from George *et al*. [15].

Analysis of Affymetrix OncoScan® CNV FFPE Assay data

Genomic DNA was quantified using a Qubit™ Fluorometer. At least 80 ng of genomic DNA with a DNA concentration ≥ 12 ng/μl was required for each sample. DNA integrity was verified by agarose gel electrophoresis and then DNA samples were

processed according to the OncoScan® CNV FFPE Assay Kit protocol. Array fluorescence intensity (CEL) files were generated automatically from DAT files using Affymetrix® GeneChip® Command Console® (AGCC) software (v4.1.2; Affymetrix). For FFPE and frozen samples, OSCHP files were generated by the OncoScan Console (v1.2; Affymetrix) from fluorescence intensity (CEL) files using their in-house workflow FFPE Analysis NA33 and REF103 Analysis NA33, respectively. The quality of the data was evaluated by two metrics: single nucleotide polymorphism quality control (SNPQC); and median absolute value pairwise difference (MAPD). Samples that did not pass the thresholds (MAPD $\leq$ 0.3 and SNPQC $\geq$ 20) were excluded from downstream analysis. Nexus Express software for OncoScan (v1; BioDiscovery) was used to call SCNAs using the SNP-FASST2 algorithm with default parameters. All segments were then exported. Segments that spanned < 100 kb or with < 25 probes were removed.

Pathway enrichment analysis

WebGestalt[16] (http://www.webgestalt.org) was employed to perform pathway enrichment analysis to investigate the distribution of genes affected by SNVs, indels and CNAs within the Kyoto encyclopedia of genes and genomes (KEGG) database. Enrichment was determined to be informative if the adjusted *P*-value was $\leq$ 0.01 (Benjamini-Hochberg method).

RNA isolation and mRNA quantification

Total RNA was isolated from tissues and cells using TRIzol regent (Life Technologies, Carlsbad, USA). Quantitative real-time PCR was performed using GoTaq qPCR Master Mix (Promega Corp., Madison, WI, USA) according to the manufacturer's instructions. The relative mRNA expression levels of *PDE3A*, *DVL3, LGR5*, *SNAIL*, *OCT4*, *TWIST*, *SOX2* and *AXIN2* were calculated using *ACTB* as a reference. The primers used are listed in Supplementary information, Table S13.

Immunohistochemical analysis

IHC was performed on human tumor or normal FFPE samples to analyze the protein expression of PDE3A, β-catenin and DVL3 according to the methods described previously[17]. The tumor type and histological features were characterized by two pathologists. A representative field was photographed using an Olympus BX-51TF microscope (Olympus Corp., Tokyo, Japan).

Genomic comparison analysis

Genomic data of ESCC[18], HNSCC[19], EAC[20], GA-CIN[21] and SCLC[15] were obtained from corresponding supplementary materials. For comparison among the five cancer type including SCCE, the mutation frequency profile of 96 possible mutation types in each cancer type was applied to hierarchical clustering.

Statistical Analysis

The specific statistic methods used are indicated in the figure legend and were performed mainly using GraphPad Prism version 6.0. For comparison between tumor and para-normal samples, paired student's *t*-test was applied. For survival analysis, log-rank test was used to compare survival outcome between groups. In addition, Wilcox rank sum test were used for testing the correlation between mutation rates and clinical information of patients with SCCE, and it was carried out with R version 3.3.3. All tests were two-tailed with an alpha level of 0.05.

References

1 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

2 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

3 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

4 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010).

5 Weckx, S. *et al.* novoSNP, a novel computational tool for sequence variation discovery. *Genome Res* **15**, 436-442, doi:10.1101/gr.2754005 (2005).

6 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169-175, doi:10.1038/nature20805 (2017).

7 Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48**, 600-606, doi:10.1038/ng.3557 (2016).

8 Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27**, 175-181, doi:10.1093/bioinformatics/btq630 (2011).

9 Kan, Z. *et al.* Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res* **23**, 1422-1433, doi:10.1101/gr.154492.113 (2013).

10 Wu, K. *et al.* Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas. *Nat Commun* **6**, 10131, doi:10.1038/ncomms10131 (2015).

11 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

12 Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat Genet* **46**, 573-582, doi:10.1038/ng.2983 (2014).

13 D'Aurizio, R. *et al.* Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res* **44**, e154, doi:10.1093/nar/gkw695 (2016).

14 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).

15 George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47-53, doi:10.1038/nature14664 (2015).

16 Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* **41**, W77-83, doi:10.1093/nar/gkt439 (2013).

17 Han, J. *et al.* Reduced expression of p21-activated protein kinase 1 correlates with poor histological differentiation in pancreatic cancer. *BMC cancer* **14**, 650, doi:10.1186/1471-2407-14-650 (2014).

18 Song, Y. *et al.* Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* **509**, 91-95, doi:10.1038/nature13176 (2014).

19 Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-1160, doi:10.1126/science.1208130 (2011).

20 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, doi:10.1038/ng.2591 (2013).

21 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202-209, doi:10.1038/nature13480 (2014).