Oesophageal cancers are prominent worldwide; however, there are few targeted therapies and survival rates for these cancers remain dismal. Here we performed a comprehensive molecular analysis of 164 carcinomas of the oesophagus derived from Western and Eastern populations. Beyond known histopathological and epidemiologic distinctions, molecular features differentiated oesophageal squamous cell carcinomas from oesophageal adenocarcinomas. Oesophageal squamous cell carcinomas resembled squamous carcinomas of other organs more than they did oesophageal adenocarcinomas. Our analyses identified three molecular subclasses of oesophageal squamous cell carcinomas, but none showed evidence for an aetiologic role of human papillomavirus. Squamous cell carcinomas showed frequent genomic amplifications of CCND1 and SOX2 and/or TP63, whereas ERBB2, VEGFA and GATA4 and GATA6 were more commonly amplified in adenocarcinomas. Oesophageal adenocarcinomas strongly resembled the chromosomally unstable variant of gastric adenocarcinoma, suggesting that these cancers could be considered a single disease entity. However, some molecular features, including DNA hypermethylation, occurred disproportionately in oesophageal adenocarcinomas. These data provide a framework to facilitate more rational categorization of these tumours and a foundation for new therapies.

Oesophageal cancers have 5-year survival rates of 12–20% in Western populations1,2 and cause the deaths of over 400,000 people worldwide annually3. Oesophageal cancer is classified by histology as adenocarcinoma (EAC) or squamous cell carcinoma (ESCC)4. EAC incidence has increased several fold in Western countries in recent decades5, occurs predominantly in the lower oesophagus near the gastric junction, and is associated with obesity, gastric reflux and a precursor state termed Barrett's oesophagus. Rising EAC rates are paralleled by increasing incidences of proximal stomach cancer5. ESCCs predominate in the upper and mid-oesophagus and are associated with smoking and alcohol exposure in Western populations. In non-Western countries, risk factors for ESCCs are less established.

The appropriate demarcation between gastric and oesophageal adenocarcinomas and the classification of adenocarcinomas spanning the gastroesophageal junction (GEJ) remain unresolved6–9, and there is debate regarding the utility of histological distinctions10. To improve oesophageal cancer classification, we performed a comprehensive molecular analysis of 164 oesophageal tumours, 359 gastric adenocarcinomas and 36 additional adenocarcinomas at the GEJ. We evaluated approaches for categorizing oesophageal tumours and identified molecular features and candidate pathways that define molecular subgroups and offer potential therapeutic targets.

Sample collection and molecular characterization
We addressed the challenge of clinically distinguishing oesophageal and gastric adenocarcinomas through review of adenocarcinomas originating near the GEJ, using anatomic data and histopathologic criteria, to categorize tumours by oesophageal, gastric or indeterminate origin (Fig. 1a, Supplementary Table 1, Supplementary Fig. 1.1). We identified 90 ESCCs, 72 EACs (61 definite oesophageal and 11 probable oesophageal), 36 GEJ carcinomas of indeterminate origin, 63 gastric GEJ carcinomas (15 definite gastric and 48 probable gastric), 140 gastric carcinomas of the fundus or body, and 143 gastric antral or pyloric carcinomas. We were unable to localize 13 gastric adenocarcinomas more narrowly within the stomach, and 2 oesophageal tumours were undifferentiated carcinomas.

Fresh-frozen tumour samples from patients who were not previously treated with chemotherapy or radiation therapy were obtained from multiple countries with informed consent and local Institutional Review Board approval. Germline DNA was collected from blood or nonmalignant oesophageal mucosa. Genetic material was subjected to whole-exome sequencing, single-nucleotide polymorphism (SNP) array profiling to evaluate somatic copy-number alterations (SCNAs), DNA methylation profiling and mRNA and microRNA sequencing. DNA from 51 oesophageal cancers was subjected to low-pass (6–8× coverage) whole-genome sequencing. Reverse-phase protein array proteomic analysis was performed on 113 tumours.

Molecular separation of ESCC and EAC
We evaluated the 164 oesophageal carcinomas using integrated clustering of SCNA, DNA methylation, mRNA and microRNA expression data using iCluster10. Both independent and integrated analyses from each molecular platform revealed separation between squamous cancers and adenocarcinomas (Fig. 1b; Extended Data Fig. 1a–e). Gene expression analysis (Extended Data Fig. 2) revealed that EACs showed increased E-cadherin (CDH1) signalling and upregulation of ARF6 and FOXA pathways, which regulate E-cadherin11. By contrast, ESCCs exhibited upregulation of Wnt, syndecan and p63 pathways, the latter being essential for squamous epithelial cell differentiation12. These data suggest the presence of lineage-specific alterations that drive progression in EACs and ESCCs.

Somatic genomic alterations in oesophageal cancer
We evaluated somatic genomic alterations in ESCC and EAC using MutSig13 to search for genes with significantly recurring mutations (Extended Data Fig. 3a, b). In ESCC, we identified significantly mutated genes, TP53, NFE2L2, MLL2, ZNF750, NOTCH1 and TGFBR2, consistent with previous studies14–20. In EAC, we identified significant mutations in TP53, CDKN2A, ARID1A, SMAD4 and ERBB2, as reported previously21. These findings are consistent with the prominence of CDKN2A and TP53 mutations in dysplastic Barrett’s oesophagus, a precursor to EAC. Similarly, we analysed SCNA data
We found frequent alterations of receptor tyrosine kinases and downstream signalling mediators, particularly in EAC. In ESCCs, we identified amplification or mutation of *EGFR* in 19% of tumours and alterations of *PIK3CA*, *PTEN* or *PIK3R1*, all of which are believed to activate the PI3K pathway, in 24% of tumours. EACs had a wider range of potentially oncogenic amplifications, most commonly of *ERBB2*, which was altered in 32% of EACs, but in only 3% of ESCCs. Although clinical trials that led to approval by the US Food and Drug Administration of the *ERBB2*-directed antibody trastuzumab were limited to gastric and GEJ adenocarcinomas, ERBB2-positive EACs are routinely treated off-label with trastuzumab. Notably, we found mutations of *ERBB2* in four tumours lacking *ERBB2* amplification, suggesting that more patients may benefit from *ERBB2*-directed therapy. Transcriptome data identified six cases with *ERBB2* amplification that expressed a fusion transcript in which exon 12 of *ERBB2* was fused to the 3′ untranslated region of neighbouring gene *JUP* (Supplementary Fig. 3.1; Supplementary Table 3). Because this fusion transcript omits the *ERBB2* transmembrane and tyrosine kinase domains, its potential functionality is unclear. Other EACs showed amplification of *KRAS*, *EGFR*, *IGF1R* or *VEGFA*.

Additional analysis identified dysregulation of the TGF-β pathway and less frequent *CTNNB1* (β-catenin) activation, both more common in EAC than ESCC. We found that 6% of ESCCs (but no EACs) had inactivating alterations of *PTCH1*, as previously described, suggesting activated hedgehog signalling. ESCC tumours, like other squamous cancers, had amplifications of chromosome 3q, focused on the *SOX2* locus. Genes that encode *SOX2* or squamous transcription factor p63, also on chromosome 3p, were amplified in 48% of ESCCs. Moreover, mutations in *ZNF750* and *NOTCH1* in ESCCs may similarly modulate squamous cell maturation. In EACs, however, we found frequent amplifications of genes that encode GATA4 and GATA6 developmental factors, as described in gastric adenocarcinomas, and (for GATA6), experimentally validated in EAC.

Both EAC and ESCCs showed alterations of chromatin-modifying enzymes. Integrative clustering of ESCC data using iCluster revealed two classes, denoted iCluster 1 and iCluster 2 (Fig. 3a). Within iCluster 2, we identified a group of tumours with shared features including mutations in *SMARCA4* (encoding the SWI/SNF factor BRG1), increased DNA methylation (Fig. 3a, rightmost samples) and relatively unaltered SCNA profiles (Fig. 3b). We designated the distinct set of tumours with these features as subtype ESCC3, thus dividing ESCCs into three molecular subtypes: ESCC1 (n = 50), ESCC2 (n = 36) and ESCC3 (n = 4).

ESCC1 was characterized by alterations in the NRF2 pathway, which regulates adaptation to oxidative stressors including some carcinogens and some chemotherapy agents. Mutations in *NFE2L2* (NRF2), are associated with poor prognosis and resistance to chemotherapeutics. Alterations were seen in *NFE2L2*, in genes encoding proteins that degrade NRF2 (*KEAP1* and *CUL3*), and in *ATG7*, encoding an NRF2 pathway autophagy factor. ESCC1 had a higher frequency of *SOX2* and/or *TP53* amplification (Fig. 3c, Extended Data Fig. 5). ESCC1 gene expression resembled the classical subtype described in The Cancer Genome Atlas (TCGA) studies of lung SCC and head and neck SCC (HNSCC) (Extended Data Fig. 6), which possess similar somatic alterations. ESCC1 showed higher rates of

**Figure 1** | Major subdivisions of gastroesophageal cancer. a. 559 oesophageal and gastric carcinoma tumours were categorized into sample sets. CIN, chromosomal instability; EBV, Epstein–Barr virus; GEJ, gastroesophageal junction; GS, genomically stable; MSI, microsatellite instability. UC, undifferentiated carcinoma. b. Integrated clustering of four molecular platforms shows that oesophageal carcinomas fall into two molecular subtypes (iCluster 1 and iCluster 2) that are virtually identical to histological classes ESCC and EAC. Clinical (top) and molecular data (bottom) from 164 tumours profiled with all four platforms are depicted.
YAP1 (11q22.1) amplification and VGLL4/ATG7 deletion, suggesting activation of Hippo.

ESCC2 showed higher rates of mutation of NOTCH1 or ZNF750 (Extended Data Fig. 5), more frequent inactivating alterations of KDM6A and KDM2D, CK6 amplification, and inactivation of PTEN or PIK3R1. We found greater leukocyte infiltration of ESCC2 tumours and higher levels of cleaved Caspase-7 protein (Extended Data Fig. 7), the latter implying enhanced potential for XIAP-directed agents to facilitate apoptosis. The gene with the lowest P value for the methylation difference between ESCC1 and ESCC2 was the immunomodulatory molecule BST2 (ref. 35) (P = 3 × 10⁻⁴, Fisher’s exact test; Supplementary Table 4), which showed less methylation and higher expression in ESCC2 (Extended Data Fig. 7), suggesting potential for BST2 inhibition.

ESCC3 tumours showed no evidence for genetic deregulation of the cell cycle and had TP53 mutations in only one of four samples. All samples in ESCC3, however, sustained alterations predicted to activate the PI3K pathway (Extended Data Fig. 5), and three of four possessed somatic alterations of PIK3R1 (Extended Data Fig. 7), suggesting potential for PI3K pathway activation.

Mutations and SCNAs for selected genes and pathways are provided in Supplementary Table 4. For ESCC3, however, the predilection for C > A transversions at AA dinucleotides seen in EAC (Supplementary Table 5) was significantly higher in our Vietnamese population (P = 0.013, Wilcoxon), suggesting a role for tobacco chewing. The APOBEC signature was overrepresented in ESCC2 (Fig. 3d, P = 0.03, Kruskal–Wallis test) and enriched in patients from Ukraine and Russia (P = 0.01, Wilcoxon rank-sum test). ESCC tumours lacked the predilection for A > C transversions at AA dinucleotides seen in EAC (Supplementary Table 5).

We evaluated whether the human papilloma virus (HPV), which has a pathogenic role in cervical SCC and HNSCC, also contributes to ESCC, as has been reported. Comparison of ESCC mRNA sequencing data to TCGA HNSCC data found that ESCC HPV transcript levels resembled HPV-negative HNSCC tumours (Fig. 3e). These data do not support an aetiologic role for HPV in ESCC.

**EAC in relation to gastric cancer**

Given the uncertainty regarding appropriate demarcations of EAC relative to both gastric cancer and ESCC, we analysed both EAC and ESCC relative to the cancer types that occur nearest to the oesophagus, HNSCC and gastric adenocarcinoma. Analysis of mRNA expression, DNA methylation and SCNA data demonstrated that ESCC had a stronger resemblance to HNSCC than to EAC (Fig. 4a). Similarly, EACs more closely resembled gastric cancer than they did ESCC. In our previous TCGA study, we classified gastric tumours into four subtypes on the basis of having (1) Epstein–Barr virus (EBV) infection, (2) microsatellite instability (MSI), (3) chromosomal instability (CIN) and (4) genomic stability (GS), a group largely comprised of the diffuse histologic type. When we evaluated EACs jointly with gastric cancers, we observed that EACs and MSI gastric cancers jointly formed a group distinct from EBV, MSI or GS tumours (Extended Data Fig. 8). When we evaluated EACs jointly with gastric cancers, we observed that EACs and CIN gastric cancers jointly formed a stronger resemblance to HNSCC than to ESCC (Fig. 4a). Similarly, EACs more closely resembled gastric cancer than they did ESCC. In our previous TCGA study, we classified gastric tumours into four subtypes on the basis of having (1) Epstein–Barr virus (EBV) infection, (2) microsatellite instability (MSI), (3) chromosomal instability (CIN) and (4) genomic stability (GS), a group largely comprised of the diffuse histologic type. When we evaluated EACs jointly with gastric cancers, we observed that EACs and MSI gastric cancers jointly formed a group distinct from EBV, MSI or GS tumours (Extended Data Fig. 8). When we evaluated EACs jointly with gastric cancers, we observed that EACs and CIN gastric cancers jointly formed a group distinct from EBV, MSI or GS tumours (Extended Data Fig. 8). When we evaluated EACs jointly with gastric cancers, we observed that EACs and CIN gastric cancers jointly formed a group distinct from EBV, MSI or GS tumours (Extended Data Fig. 8).
The enrichment of CIN in EAC suggested that comparisons of EAC with gastric cancers would be confounded by non-CIN tumours nearly exclusively in the stomach. We therefore sought to find features that could differentiate EAC from CIN gastric cancers by analysis of the 288 CIN GEAs (GEA-CIN; Fig. 1a). We found clear similarity between chromosomal aberrations in gastric CIN tumours and EAC (Fig. 4c), with stronger similarity between EAC and CIN gastric cancers than between those of EAC and ESCC. Clustering of GEA-CIN data from individual platforms (Extended Data Fig. 9) and by integrative clustering revealed no consistent separation of EACs and CIN gastric cancers, thus arguing against classifying these as distinct diseases (Extended Data Fig. 10). As misannotation of tumours near the GEJ could enhance the apparent similarity of EACs and CIN gastric tumours, we repeated our analysis after excluding equivocal GEJ cases, but saw no definitive separation of EAC and CIN gastric adenocarcinomas (Supplementary Fig. 7.1). However, clustering of DNA methylation data revealed a progression of DNA methylation features from proximal to distal GEA-CIN tumours (Fig. 5a). Samples in cluster 1, those with the most frequent hypermethylation, were enriched in the oesophagus or proximal stomach/GEJ (Fig. 5b). The proportion of cancers showing more frequent DNA hypermethylation (that is, clusters 1 or 2) was significantly higher among EACs than among gastric CIN cancers (70% versus 30%, respectively; \( P = 1.0 \times 10^{-8} \), Fisher’s exact test). By contrast, cluster 4, with the lowest rates of hypermethylation, included more distal stomach cancers (Fig. 5b). Unlike hypermethylated gastric CpG island methylator phenotype tumours, no GEA-CIN tumours exhibited epigenetic silencing of MLH1, consistent with their MSI-negative status, but they showed a higher propensity for epigenetic silencing of CDKN2A, (Supplementary Table 6, Fig. 5c). Additional genes silenced in cluster 1 included MGMT and CHFR, for which methylation has been associated with responses to alkylating agents and microtubule inhibitors, respectively\(^{39,40}\).

We evaluated the GEA-CIN tumours for somatic features that could differentiate EACs from gastric CIN tumours (Fig. 5c). EACs had higher rates of mutation of SMARCA4 and deletion of tumour suppressor RUNX1, but lower APC mutation rates relative to gastric tumours, suggesting a less prominent role for Wnt/\(\beta\)-catenin in EAC. Copy-number analysis revealed higher rates of deletions of putative fragile site genes FHIT or WWOX, suggestive of differences in the underlying genomic instability between distal and proximal GEA-CIN tumours. Analysis of oncogenes identified subtle distinctions, with VEGFA and MYC amplifications being more common in EACs. Although additional samples will be required to refine understanding of the progressive gradations of features from the distal stomach to the oesophagus, these data indicate that gastric and oesophageal CIN tumours lack absolute dichotomizing features and do not appear to be distinct tumour types.

**Discussion**

These analyses call into question the premise of envisioning oesophageal carcinoma as a single entity. These molecular data show that histological subtypes of EAC and ESCC are distinct in their molecular characteristics across all platforms tested. ESCC emerges as a disease more reminiscent of other SCCs than of EAC, which itself bears striking analogies to colorectal carcinomas, whereby CIN prevalence increases from proximal to distal GEAs (Fig. 5c), with increasing prevalence of the CIN phenotype proximally, to the point that EACs appear to represent a disease of chromosomal instability. This CIN gradient is analogous to colorectal carcinomas, whereby CIN prevalence increases distally towards the rectum\(^{41}\). EAC has been considered separate from gastric cancer according to a model whereby EAC originates from Barrett's oesophagus and thus is not of gastric origin. Although the origin of Barrett's oesophagus remains controversial, recent mouse models suggest that Barrett's oesophagus and EAC might originate from proximal gastric cells or embryonic remnant cell populations at the GEJ\(^{42,43}\). The notable molecular similarity between EACs and CIN gastric cancers provides indirect support for gastric origin of Barrett's oesophagus and EAC and indicates that we may view GEA as a singular entity, analogously to colorectal adenocarcinoma. However, these similarities between EAC and CIN gastric cancers do not indicate that all CIN GEAs are indistinguishable. Indeed, differences in more proximal GEAs should be expected, given their distinct epidemiology, rapid increase in Western countries, and inverse association with Helicobacter pylori. Continued exploration of the molecular characteristics of EAC might not absolutely differentiate them from CIN...
## Figure 4 | Similarity of oesophageal adenocarcinoma and CIN variant of gastric cancer.

(a) Molecular profiles of head and neck, oesophageal, and gastric cancers subdivided by molecular subtypes. 
(b) Distribution of gastric molecular subtypes by anatomic location across gastroesophageal adenocarcinomas. 
(c) Composite copy number profiles for ESCC, EAC, gastric-CIN and gastric non-CIN tumours with gains in red and losses in blue and grey highlighting differences between ESCC and EAC.

### Table 1: DNA methylation cluster frequencies by anatomic location

| Anatomic Location | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|-------------------|-----------|-----------|-----------|-----------|
| Antrum/pylorus    | 2        | 21        | 81        | 100       |
| Fundus/body       | 24       | 23        | 32        | 31        |
| Gastric GEJ       | 4         | 23        | 23        | 31        |
| Indeterminate GEJ | 32        | 23        | 4         | 23        |
| EAC/oesophageal GEJ| 34        | 23        | 8         | 23        |

### Table 2: Frequency of alterations in selected genes along the anatomic axis

| Cancer Type | APC | ITGAV | RUNX1 | SMARCA4 | SMAD4 | CDKN2A (epigenetic silencing) | CDKN2A (total) | FHit | WWOX | CD44 | ErBB2 | MYC | VEGFA |
|-------------|-----|-------|-------|---------|-------|-----------------------------|----------------|------|------|------|-------|-----|-------|
| Antrum/pylorus | 16% | 14%   | 0%    | 5%      | 21%   | 6%                          | 38%            | 48%  | 48%  | 34%  | 7%    | 31% | 22%   |
| Fundus/body   | 21% | 14%   | 3%    | 2%      | 18%   | 5%                          | 24%            | 28%  | 37%  | 12%  | 23%   | 23% | 10%   |
| Gastric GEJ   | 11% | 11%   | 4%    | 2%      | 15%   | 26%                         | 52%            | 39%  | 34%  | 6%   | 26%   | 19% | 13%   |
| Indeterminate GEJ | 12% | 8%    | 0%    | 8%      | 44%   | 38%                         | 70%            | 58%  | 59%  | 4%   | 38%   | 29% | 21%   |
| EAC/oesophageal GEJ | 6% | 4%    | 18%   | 10%     | 24%   | 48%                         | 81%            | 80%  | 77%  | 3%   | 32%   | 31% | 30%   |

---

**Figure 5**: Molecular features of CIN gastroesophageal adenocarcinomas by anatomic location. 
(a) Heat map representation of consensus clustering of DNA methylation of GEA-CIN tumours with molecular and clinical features shown above and methylation profiles of normal oesophagus (n = 2) and stomach (n = 13) on the left. 
(b) Fraction of tumours belonging to each methylation subgroup by anatomic location (top right) and distribution of tumour anatomic location by methylation cluster (bottom). 
(c) Frequency of alterations in selected genes along the anatomic axis with tumour suppressor inactivation in blue and oncogene activation in red.
gastrointestinal cancers, but may reveal additional features that are enriched in this variant of GEA.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 March 2016; accepted 20 November 2016.

Published online 4 January 2017.

1. De Angelis, R. et al. Cancer survival in Europe 1999–2007 by country and age: results of EUROCARE—5—a population-based study. Lancet Oncol. 15, 23–34 (2014).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. CA Cancer J. Clin. 66, 7–30 (2016).
3. Torre, L. A. et al. Global cancer statistics, 2012. CA Cancer J. Clin. 65, 87–108 (2015).
4. Sievert, J. R. & Ott, K. Are squamous and adenocarcinomas of the esophagus the same disease? Semin. Radiat. Oncol. 17, 38–44 (2007).
5. Brown, L. M., Devesa, S. S. & Chow, W. H. Incidence of adenocarcinoma of the stomach. The widths of the colour bands represent the proportion of the subtypes present within anatomic regions. Key features of subtypes are indicated in associated text.

Figure 6 | Gradients of molecular subclasses of gastroesophageal carcinoma. Schematic representing shifting proportion of subtypes of gastroesophageal carcinoma from the proximal esophagus to the distal stomach. The widths of the colour bands represent the proportion of the subtypes present within anatomic regions. Key features of subtypes are indicated in associated text.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 March 2016; accepted 20 November 2016.

Published online 4 January 2017.

1. De Angelis, R. et al. Cancer survival in Europe 1999–2007 by country and age: results of EUROCARE—5—a population-based study. Lancet Oncol. 15, 23–34 (2014).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. CA Cancer J. Clin. 66, 7–30 (2016).
3. Torre, L. A. et al. Global cancer statistics, 2012. CA Cancer J. Clin. 65, 87–108 (2015).
4. Sievert, J. R. & Ott, K. Are squamous and adenocarcinomas of the esophagus the same disease? Semin. Radiat. Oncol. 17, 38–44 (2007).
5. Brown, L. M., Devesa, S. S. & Chow, W. H. Incidence of adenocarcinoma of the stomach. The widths of the colour bands represent the proportion of the subtypes present within anatomic regions. Key features of subtypes are indicated in associated text.
Genome-characterization Centers and Genome Data Analysis Centers. All data were released through the Data Coordinating Center. The National Cancer Institute and National Human Genome Research Institute project teams coordinated project activities. The following TCGA investigators of the Oesophageal Analysis Working Group contributed substantially to the analysis and writing of this manuscript. Project leaders: Adam J. Bass, Peter W. Laird, Ilya Shmulevich; data coordinator: Vésteinn Thorsson; analysis coordinators: Vésteinn Thorsson, Francisco Sánchez-Vega; manuscript coordinator: Barbara G. Schneider; graphics coordinator: Toshinori Hinoue; DNA sequence analysis: Andrew Dunford, Jaegil Kim, Michael D. McLellan, Angeliki Pantazi, Carrie Cibulskis, Melanie Kucherlapati, Peter J. Park, Lixing Yang; Samuel R. Meier; mRNA analysis: Reanne Bowlby, Andrew J. Mungall; miRNA analysis: Reanne Bowlby; DNA methylation analysis: Toshinori Hinoue, Peter W. Laird; Copy-number analysis: Andrew D. Cherniack, Juliann Shih; protein analysis: Ju-Seog Lee, Apurva Hegde, Rehan Akbani; pathway/integrated analysis: Francisco Sánchez-Vega, Varsha Dhankani, Christine Curtis, Jose Antonio Seoane, Ronglai Shen, Hsin-Ta Wu, Benjamin J. Raphael, Alexandra M. Wong, Vésteinn Thorsson, Nikolaus Schultz, Arsh Arora; pathology expertise and clinical data: Alex Boussioutas, Barbara G. Schneider, David Kelsen, Robert D. Odze, Shannon J. McCall, Kenneth Wang, Arjun Pennathur, Joseph E. Willis, Margaret L. Gulley, Katherine S. Garman, M. Blanca Piazuelo, Sarah Derks, Kristen M. Leraas, Tara M. Lichtenberg, John A. Demchok, David G. Beer, Brian J. Reid, Zhongren Zhou, Laura Tang, Jihun Kim, Jaffer A. Ajani; microbiome analysis: Charles S. Rabkin, Margaret L. Gulley, Reanne Bowlby, Chandra Sekhar Pedamallu, Sara Sadeghi, Akinwumi I. Ojesina, Susan Bullman, Karen Mungall.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Adam J. Bass (adam_bass@dfci.harvard.edu) or Vésteinn Thorsson (Vesteinn.Thorsson@systemsbiology.org).

Reviewer Information Nature thanks S. Macgregor and the other anonymous reviewer(s) for their contribution to the peer review of this work.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Specimen collection and staging. Tissue source sites (TSS) are listed in Supplementary Information S1.1. Oesophageal tumours were collected and shipped to a central BioSpecimen Core Resource (BCR) between 1 December 2011 and 23 December 2013. Samples were obtained from patients who had received no previous chemotherapy or radiotherapy for their disease. Each frozen primary tumour specimen had a companion normal tissue specimen (blood or blood components, including DNA extracted at the TSS). Adjacent nontumorous oesophageal tissue was also submitted for a subset of patients.

Cases were staged according to the American Joint Committee on Cancer 7th edition staging system28. Pathology quality control was performed on each tumour and adjacent normal tissue specimen (if available) from a frozen section slide to confirm that the tumour specimen was histologically consistent with oesophageal cancer and that the adjacent tissue specimen contained no tumour cells. Tumour samples with ≥ 80% tumour nuclei and ≤ 20% necrosis were submitted for nucleic acid extraction.

Nucleic acid processing and qualification. DNA and RNA were co-isolated, and quality was assessed at the central BCR as described previously (supplementary S1.1 in ref. 27). A custom Sequenom SNP panel and the AmpFISTR Identifier (Applied Biosystems) was used to verify that tumour DNA and germline DNA representing a case were derived from the same patient. RNA was analysed through the RNA6000 Nano assay (Agilent) to determine a RNA Integrity Number, and only analytes with an integrity number ≥ 7.0 were included. Only cases yielding a minimum of 6.9 μg of tumour DNA, 5.15 μg of RNA and 4.9 μg of germline DNA were included.

The BCR received tumour samples with germline controls from a total of 322 oesophageal cancer cases, of which 185 qualified, on the basis of BCR pathology review and molecular characteristics. Distribution and quality control of cases is shown in Supplementary Fig. 1.1. Of the 185 cases that qualified, 171 cases were used for genomic analysis, as 14 cases were excluded after independent pathology review (described in ‘Expert pathology review’, below) or discovery of clinical or molecular disqualifiers.

Of the 171 qualifying cases, matched nontumorous oesophageal tissue was available for 58 cases. Samples with residual tumour tissue after extraction of nucleic acids were considered for proteomics analysis. When available, a 10–20 mg piece of snap-frozen tumour adjacent to the piece used for molecular sequencing and characterization was submitted for reverse-phase protein array analysis. We compared these 171 oesophageal adenocarcinomas to 388 similarly characterized gastric adenocarcinomas (Supplementary Fig. 1.1).

Microsatellite instability assay. Microsatellite instability (MSI) in qualified oesophageal adenocarcinoma tumour-derived DNA samples was evaluated by the BCR at Nationwide Children’s Hospital, Columbus, Ohio, USA. MSI-monodinucleotide assay was performed to test a panel of four mononucleotide repeat loci (polyadenylic tandem BAT25, BAT26, BAT40 and transforming growth factor receptor type II) and three dinucleotide repeat loci (CA repeats in D2S123, D5S346 and D17S250) as previously described27.

Expert pathology review. All cancers included in this study were secondarily reviewed by an Expert Pathologists’ Committee that consisted of seven experienced gastrointestinal pathologists (R.O., S.McC., Z.Z., J.K., L.T., M.B.P. and J.W.). A centralized virtual pathology review system was constructed using an Aperio slide scanner housed at the BCR at Nationwide Children’s Hospital. Typically, two frozen sections flanking the tumour tissue from which all material was extracted for this study and one additional high-quality formalin-fixed paraffin-embedded tissue section were scanned and reviewed. Two committee members reviewed all cases before inclusion into the study. For cases with discrepant results, a tiebreaker reviewer was assigned.

All oesophageal cancers were categorized as squamous or adenocarcinoma, according to the World Health Organization Classification of Tumours of the Digestive System, 4th edition45. Nine cases were excluded on the basis of pathology review, including four cases where quality control identified inadequate material for analysis, two cases where only noninvasive neoplasm was observed, and two cases where the neoplasm was unclassifiable on the basis of the material available for review. As part of this review, an additional 77 gastric adenocarcinomas that had not undergone pathology review as part of this group’s original published analysis were also subject to pathology re-review as performed previously27.

Clinical staging was assessed27 by two reviewers according to criteria for each tumour type (ESCC or EAC). T, N and M status and tumour grade (0, 1, 2 or 3) were based on pathology reports from the TSS.

Anatomic subclassification of adenocarcinomas involving the GEJ. All adenocarcinomas (oesophageal or gastric) from the TCGA collections that had a potential origin near the GEJ were further reviewed to refine their anatomic location. Pathology reports were obtained from the TSSs with the original gross pathology description of the tumour at resection or endoscopic biopsy. Two independent clinical reviewers reviewed each TSS pathology report. Tumours were classified as oesophageal, probable oesophageal, indeterminate, probable gastric or gastric, according to criteria outlined in Supplementary Information S1.2. For downstream analyses, the oesophageal and probable oesophageal were grouped together, as were the gastric and probable gastric.

Somatic copy-number analysis. Analysis of SCNs was performed on the basis of DNA profiling of each tumour or germline sample on Affymetrix SNP 6.0 at the Genome Analysis Platform of the Broad Institute as previously described46. As part of this process of copy-number assessment and segmentation, regions corresponding to germline copy-number alterations were removed by applying filters generated from either the TCGA germline samples from our ovarian cancer analysis or from samples in this collection. Analysis of recurrent broad and focal SCNs was performed with the GISTIC 2.0 algorithms27 with clustering performed in R, on the basis of Euclidean distance using thresholded copy number at recurring alteration peaks from GISTIC analysis using Ward’s method, both as previously reported27. Allelic copy number and purity and ploidy estimates were calculated using the ABSOLUTE algorithm27. Tumours were classified as having high chromosomal instability, SCNA-high, if they possessed at least one arm-level loss (apart from that of 18p, 18q or 21, which were recurrent in tumours of both low and high copy-number events) and otherwise as SCNA-low. Chromosomal arms were considered altered if at least 80% of the arm was lost or gained with a relative log; copy ratio change of at least 0.15 (Shih et al., unpublished observations). This method of classifying copy number instability has 93% concordance with previously described copy-number clustering27.

DNA methylation. Genomic DNA (1 μg per sample) was bisulphite-modified, subjected to quality control, and analysed using the Illumina Infinium DNA methylation platform, HumanMethylation450, as detailed in Supplementary Information S2. Data files generated are listed in Supplementary Information S2.3.

CDKN2A epigenetic silencing calls. CDKN2A (also known as p16INK4) epigenetic silencing calls were made using both DNA methylation and RNA-seq data. CDKN2A DNA methylation status was assessed in each sample based on the probe (cg13601799) located in the p16INK4 promoter CpG island. p16INK4 expression was determined by the log2(RPKM +1 ) level of its first exon (chr9: 21974403–21975132). The epigenetic silencing calls for each sample were made by evaluating a scatterplot showing an inverse association between DNA methylation and expression as described in Supplementary Information S2.

DNA sequence analysis. Exome and full-coverage whole-genome sequencing was split between two sequencing centres. Samples that were submitted to TCGA as stomach adenocarcinomas (that is, STAD, as labelled by the TSS) were sent for sequencing at the Broad Institute. Samples labelled as oesophageal cancers (that is, ESCA) were sequenced at Washington University. Each centre was responsible for generating BAM files from both tumour and normal DNA samples with additional filtering to remove likely artefacts of the sequencing process. From these BAM files, four different TCGA analysis sites performed distinct mutation and insertion/ deletion detection procedures. The results of these distinct mutation-calling efforts were integrated to generate a common mutation annotation file for subsequent analysis. See Supplementary Section S3.1.

Broad Institute sequencing. Whole-exome sequencing of 0.5 to 3 μg of DNA from tumour and normal blood samples was performed as previously described26 using the Agilent SureSelect Human All Exon V5 kit, followed by 2 × 76-bp paired-end sequencing on the Illumina HiSeq platform. Whole-genome sequencing, 2 × 101-bp reads were sequenced on the same platform. Read alignment and processing were performed using the Burrows–Wheeler Aligner (BWA) and Picard at the Broad Institute (http://broadinstitute.github.io/picard/) as previously published27. Alignments were first subjected to quality control using ContEst27 to avoid misannotation of tumour and germline DNA samples, or cross-contamination between tumour samples. Only samples with less than 5% estimated cross-contamination were analysed further.

Washington University sequencing. Whole-exome sequencing and whole-genome Illumina libraries were constructed as described previously46 using Nimblegen’s SeqCap EZ Hull Exome Library v3.0 combined with additional 120-mer IDT custom probes, targeting DNA from cancer-related viruses (for example, HPV, EBV) and sequenced in multiple lanes of Illumina HiSeq 2000 flow cells to achieve a minimum coverage of 20× across 80% of coding target exons. Each lane or sub-lane of data was aligned using BWA v0.5.9. to
and removed. In addition, for the normal germline BAM, putative variants and secretory by the LUSC classification. Of the 839 genes used for the HNSCC centroid gene expression profiles were required as a minimum. Filtering of putatively spurious mutation calls due to 8-xoxygenate artefacts was performed to remove candidate mutations attributed to these sequencing artefacts. Further filtering removed candidate mutations that had been identified through sequencing of cohorts of non-neoplastic DNA samples to remove alternative artefacts or unfiltered germline calls. Read counts were generated for all remaining novel putative variants, and these variants were incorporated into the final mutation annotation file if they met the same minimum coverage, maximum coverage, and variant allele fraction requirements described above.

**Mutation annotation and significance analysis.** Functional annotation of mutations was performed with Oncotator (http://www.broadinstitute.org/cancer/sgc/oncotator) using Gencode V18. Significantly recurrently mutated genes were identified using the MutSigCV2.0 algorithm34.

**Mutation signature analysis.** Mutation signature discovery was performed using Bayesian non-negative matrix factorization algorithm for mutation signature analysis as described in Supplementary Information S1.2.

**Low-pass whole-genome sequencing for rearrangement identification.** Genomic DNA (500–700 ng per sample) was sheared into 250-bp fragments using a Covaris E220 ultrasonicator, then converted to a paired-end Illumina library using KAPA Bio kits with Caliper (PerkinElmer) robotic NGS Suite (Partek Genomics) according to manufacturers’ protocols. All libraries were sequenced on a HiSeq2000 using one sample per lane, with a paired-end 2 × 51-bp read length. Tumour DNA and its matching normal DNA were usually loaded on the same flow cell. Raw data were converted to the FASTQ format, and BWA alignment (to hg19) was used to generate BAM files as previously described (supplementary S3.6 in ref. 27). Detection of structural rearrangements was performed using two algorithms, BreakDancer50 and Merkrek51. The set of structural variant calls from each tumour sample was filtered by the calls from its matched normal DNA to remove germline variants. Data were then re-examined using the Merkrek algorithm, which necessitated the identification of at least two discordant read pairs, with one read covering the actual breakpoint junction. Alterations found in simple or satellite repeats were also excluded. (Candidate fusion genes from this analysis are shown in Supplementary Table 3 with more detailed listing of structural alterations in Supplementary Table 7.)

**mRNA sequencing and analysis methods.** mRNA sequence data were generated as described previously (supplementary S5.1 in ref. 27). For combined clustering of oesophageal cancer and oesophageal cancer data with their MapSplice/RSEM pipeline32. We generated candidate fusion events from mRNA sequence data as described previously (supplementary S5.4 in ref. 27), except that we used TransABYS v1.4.8 (http://www.bcgsc.ca/platform/bioinfo/software/trans-abys/releases/1.4.8/).

To identify subtypes within our various cohorts, we used hierarchical clustering with heatmap v1.0.2 in R. The input in each case was a reads per kilobase of exon per million (RPKM) data matrix for the top 25% most variable genes with mean greater than 10 RPKM. We transformed each row of the matrix by log_{10}(RPKM+1), then used heatmap to scale the rows. We used ward.D2 for the clustering method and correlation and Euclidean distance measures for clustering the columns and rows, respectively. For analyses comparing oesophageal with gastric and head and neck cancers, we used the top 25% (~300) most variable 5p or 3p mature strand microRNAs35 within the oesophageal carcinoma sample set with mean expression greater than 1,000 RSEM-normalized counts. We transformed each row of the matrix by log_{10}(RSEM+1), then used heatmap to scale the rows. We used ward.D2 and a Euclidean distance measure.

**Reverse-phase protein array.** Proteins isolated from tumours were used to prepare reverse-phase protein arrays with 187 validated primary antibodies by methods described previously (supplementary S7 in ref. 27). Data were normalized, and clustering analysis was performed as detailed in Supplementary Section S4.

**Pathogen analysis.** We used two tools to examine whole-exome and RNA sequence data for the presence of microbial sequences: BBT (BioBloomTools, v1.2.4b1) and PathSeq. Details of these analyses are provided in Supplementary section S5. MicroRNA data were analysed using an in-house pipeline as previously described (supplementary S9.2 in ref. 27).

**Pathway analysis of mRNA.** We performed pathway-level analysis of gene expression to compare EAC and ESCC samples. Pathways, as gene-sets, were obtained from the National Cancer Institute’s pathway interaction database (NCI-PID)35. A P value, comparing EAC with ESCC using Kruskal–Wallis one-way analysis of variance by ranks, was obtained for each gene. For each of the 224 pathways, the gene-level P values were log-transformed and summed using an approach based on Fisher’s combined statistic to yield a pathway-level composite score. The statistical significance of this score was then estimated empirically by similarly scoring 10,000 randomly generated pathways for each NCI-PID pathway, with matched pathway size.

**Integrative clustering.** To discover which tumour samples shared molecular signatures across platforms, the following four integrative clustering approaches were used: iCluster, Multiple Kernel Learning k-means (MKL k-means), SuperCluster, and Clustering of Cluster Assignments (COCA). In the iCluster method36,56,57, subgroups were discovered through their representation as latent variables in joint multivariate regression. MKL k-means combines the k-means clustering algorithm with the use of kernels that encode the similarity between the samples, to define features for classifying the tumours. SuperCluster and COCA both use clusters derived from individual molecular platforms to form an overall categorical description of each sample, but they differ in details, such as the metric used to compare those samples. SuperCluster performs a variance adjustment such that each molecular platform receives equal weight, whereas the implementation of COCA in R (http://www.broadinstitute.org/cancer/software/annotation/coca) uses a weighting method that takes into account the granularity of the divisions within each platform-specific category. Further details on these methods are given in Supplementary Section S7. Data availability. The primary and processed data used to generate the analyses presented here can be downloaded from the TCGA manuscript publication page, (https://tcga-data.nci.nih.gov/docs/publications/esca_2016), and from the Genomic Data Commons (https://gdac-portal.broadinstitute.org/legacy-archive).
53. Wilkerson, M. D. et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin. Cancer Res. 16, 4864–4875 (2010).
54. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 42, D68–D73 (2014).
55. Schaefer, C. F. et al. PID: the Pathway Interaction Database. Nucleic Acids Res. 37, D674–D679 (2009).
56. Shen, R. et al. Integrative subtype discovery in glioblastoma using iCluster. PLoS One 7, e35236 (2012).
57. Mo, Q. et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl Acad. Sci. USA 110, 4245–4250 (2013).
Extended Data Figure 1 | Platform-specific unsupervised clustering analyses of oesophageal cancers. a–e, Unsupervised clustering of oesophageal cancers based on DNA hypermethylation (a), SCNA (b), gene expression profiles (c), microRNA profiles (d) and reverse-phase protein array data (e) revealed strong separation between EAC and ESCC in multiple molecular platforms. Samples are displayed as columns. EAC, oesophageal adenocarcinoma; ESCC, oesophageal squamous cell carcinoma; UC, undifferentiated carcinoma.
Extended Data Figure 2 | Pathways with significant expression differences between EAC and ESCC. a, NCI PID pathways in which expression differs significantly between EAC and ESCC ($P_s < 10^{-3}$, where $P_s$ is the statistical significance of the pathway score (see Methods)) are listed. The colour scale shows the median (log 2) expression value of significantly differentially expressed genes ($P < 10^{-3}$) in the corresponding pathway, normalized to unit range. b, TP63ΔN transcript levels were measured in EAC, solid tissue normal, and ESCC samples. c, Median gene expression values of genes in the NCI-PID pathway 'Validated transcriptional targets of the ΔN p63 isoforms' in EAC and ESCC. Each point represents one sample, and the value is the median expression value of the 46 genes in the pathway.
Extended Data Figure 3 | MutSig analyses of significantly mutated genes in EAC and ESCC. **a**, Plot of significantly mutated genes from the MutSigCV2 computational analysis of whole-exome sequencing data from EAC samples. Genes are ordered by level of significance (q value as plotted at right). At left is the prevalence of each mutation in the sample set. The coloured boxes show samples with specific mutations, with the type of mutation labelled by box colour, with legend at upper right. The top plot shows the number of mutations per sample with synonymous (Syn.) and non-synonymous (Non syn.) mutations plotted separately. The bottom plot shows the distribution of allelic fraction of mutations for the samples sequenced. **b**, The MutSig plot for ESCC is shown the same as for the EAC samples above.
Extended Data Figure 4 | GISTIC analysis of foci of recurrent amplification and deletion. These figures demonstrate foci of significantly recurrent focal amplification and deletion as determined from GISTIC 2.0 analysis of somatic copy number data from SNP arrays. Separate plots are shown for CIN-gastric cancer (left), EAC (middle) and ESCC (right). Each plot arrays the chromosomes from 1 (top) to X (bottom) and shows foci of significant amplification (left, red with scale at bottom) or deletion (right, blue with scale at top). Candidate targets of each focus of amplification or deletion are shown in the label for the respective peak. Peaks without clear targets are labelled by chromosome band. The number in parentheses indicates the number of genes in each peak as calculated by GISTIC. Genes marked with asterisks are likely drivers located adjacent to peak areas defined by GISTIC.
Extended Data Figure 5 | Comparison of somatic alterations in ESCC and HNSC subtypes. Mutations and copy-number changes for selected genes in selected signalling pathways are shown for the three ESCC subtypes identified in our study and the HPV-negative (n = 243) and HPV-positive (n = 36) subtypes that had previously been identified by TCGA in the HNSC study. Amplifications and deep deletions indicate a change of more than half of the baseline gene copies. Missense mutations were included if they were found in the COSMIC repository. Alteration frequencies are expressed as percentage of altered cases within each molecular subtype. Bottom panels show percentage of altered cases per signalling pathway for each molecular subtype and percentage of altered cases per molecular subtype for each signalling pathway.
Extended Data Figure 6 | Distinct clusters of ESCC. Columns indicate Pearson correlation between each of the mRNA profiles of 90 ESCC tumours with the centroids of the mRNA expression profiling subtypes that were developed for lung squamous cell carcinoma (LUSC, top) and head and neck squamous cell carcinoma (HNSC, bottom) gene expression analyses. Samples are in ESCC cluster order as in Fig. 3a.
Extended Data Figure 7 | Characterization of ESCC subtypes. a, We identified genes exhibiting epigenetic silencing in individual samples and compared the number of samples where each gene was silenced in ESCC1 and ESCC2. Genes that showed statistical associations between number of silenced samples and ESCC subtypes are shown in the table (P < 0.01, Fisher’s exact test). Two genes remained significant after Bonferroni correction. The panel on the right shows DNA methylation versus gene expression for BST2 and SH3TC1. b, A detailed analysis of BST2 DNA methylation in ESCC samples and non-cancer controls. c, d, The plots of (c) estimated leukocyte fraction and (d) levels of cleaved caspase-7 protein show the median, 25th and 75th percentile values (horizontal bar, bottom and top bounds of the box), and the highest and lowest values within 1.5 times the interquartile range (top and bottom whiskers, respectively).
Extended Data Figure 8 | EACs are more similar to CIN-type gastric adenocarcinomas than to other gastric subtypes. a, b. Integrative clustering of platform-specific clusters for gastroesophageal adenocarcinomas (GEA) was performed using the SuperCluster method (a) and Clustering of Cluster Assignments (COCA) (b).
Extended Data Figure 9 | Platform-specific unsupervised clustering analyses of GEA-CIN tumours. a–d. Shown are heat map representations of gene expression (a), microRNA (b), SCNAs (c), and reverse-phase protein array profiles of GEA-CIN tumours (columns) (d).
Extended Data Figure 10 | Integrative clustering of GEA-CIN samples.

a, Integrative clustering by Multiple Kernel Learning, $k$-means (MKL $k$-means) yielded a four cluster solution, in which Cluster 4 is enriched for EAC. b, Clustering of Cluster Assignments (COCA), was performed for the 267 samples for which complete platform-specific cluster information (see Fig. 5a, Extended Data Fig. 8) was available for gene expression, microRNA expression, DNA methylation and somatic copy number alteration (SCNA), and yielded three integrative clusters. Details of the methods can be found in Supplementary section S10.2. c, Frequency of EAC in four integrative clustering methods. Integrated clustering with iCluster and SuperCluster was performed as described in Methods.