nature genetics

Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity

Austin M Dulak[1,2,13], Petar Stojanov[1–3,13], Shouyong Peng[1,2], Michael S Lawrence[2], Cameron Fox[1], Chip Stewart[2], Santhoshi Bandla[4], Yu Imamura[1], Steven E Schumacher[1,2], Erica Shefler[2], Aaron McKenna[2], Scott L Carter[2], Kristian Cibulskis[2], Andrey Sivachenko[2], Gordon Saksena[2], Douglas Voet[2], Alex H Ramos[2], Daniel Auclair[2], Kristin Thompson[2], Carrie Sougnez[2], Robert C Onofrio[2], Candace Guiducci[2], Rameen Beroukhim[1,2,5,6], Zhongren Zhou[4], Lin Lin[7], Jules Lin[7], Rishindra Reddy[7], Andrew Chang[7], Rodney Landrenau[8], Arjun Pennathur[8], Shuji Ogino[1,6,9,10], James D Luketich[8], Todd R Golub[1,2,6,11], Stacey B Gabriel[2], Eric S Lander[2,3,6], David G Beer[7], Tony E Godfrey[4], Gad Getz[2,12,14] & Adam J Bass[1,2,5,6,14]

The incidence of esophageal adenocarcinoma (EAC) has risen 600% over the last 30 years. With a 5-year survival rate of ~15%, the identification of new therapeutic targets for EAC is greatly important. We analyze the mutation spectra from whole-exome sequencing of 149 EAC tumor-normal pairs, 15 of which have also been subjected to whole-genome sequencing. We identify a mutational signature defined by a high prevalence of A>C transversions at AA dinucleotides. Statistical analysis of exome data identified 26 significantly mutated genes. Of these genes, five (*TP53*, *CDKN2A*, *SMAD4*, *ARID1A* and *PIK3CA*) have previously been implicated in EAC. The new significantly mutated genes include chromatin-modifying factors and candidate contributors *SPG20*, *TLR4*, *ELMO1* and *DOCK2*. Functional analyses of EAC-derived mutations in *ELMO1* identifies increased cellular invasion. Therefore, we suggest the potential activation of the RAC1 pathway as a contributor to EAC tumorigenesis.

In recent decades, the incidence of EAC has increased markedly in the United States and other Western countries[1,2]. The increasing frequency and poor prognosis of this cancer represent a substantial health concern. EAC does not develop from the native esophageal epithelium but rather originates from intestinal metaplasia of the esophageal epithelium (Barrett's esophagus) that develops in response to chronic gastroesophageal reflux. Although the reason for the marked rise in these cancers is unknown, factors influencing the rising rates include gastroesophageal reflux disease (GERD), Barrett's esophagus and obesity[3]. There is great urgency to elucidate the genomic alterations underlying EAC to enhance understanding of these tumors, aid in early diagnosis and identify therapeutic targets.

Knowledge of the somatic mutations in EAC has been limited to studies in small collections of tumors. These studies have identified frequent mutations in *TP53* (ref. 4) and *CDKN2A*[5]. Beyond these two genes, small, focused studies have noted sporadic mutations in *APC*[6], *BRAF*[7], *CDH1* (ref. 8), *CTNNB1* (ref. 6), *EGFR*[9,10], *KRAS*[7], *PIK3CA*[11], *PTEN*[12] and *SMAD4* (ref. 12). Although comparative whole-exome sequencing has been reported for 11 EACs and esophageal squamous cell carcinomas, no clear contributors to EAC were identified at the gene level[13].

Here, we describe the landscape and spectrum of genomic alterations in 149 fresh-frozen, surgically resected cases of EAC, including adenocarcinomas arising in the gastric-esophageal junction (GEJ), not treated with chemotherapy or radiation before surgery. All cases were subjected to whole-exome sequencing, with 16 tumor-normal pairs also analyzed by whole-genome sequencing. Examination of the somatic alterations in sample pairs identified a high frequency of mutations and rearrangements. Additionally, we identify a mutational signature defined by A>C transversions at AA dinucleotide sites (with the second adenine denoting the site of the mutation). Through systematic analysis of the mutated genes, we identify many genes not previously associated with this cancer. These include *ELMO1* and *DOCK2*, upstream modulators of the RAC1 GTPase, and characterize the presence

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Cancer Program, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Department of Surgery, University of Rochester, Rochester, New York, USA. [5]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [6]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [7]Department of Surgery, University of Michigan, Ann Arbor, Michigan, USA. [8]Department of Cardiothoracic Surgery, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA. [9]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [10]Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA. [11]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [12]Massachusetts General Hospital Cancer Center and Department of Pathology, Boston, Massachusetts, USA. [13]These authors contributed equally to this work. [14]These authors jointly directed this work. Correspondence should be addressed to A.J.B. (adam_bass@dfci.harvard.edu) or G.G. (gadgetz@broadinstitute.org).

Received 26 November 2012; accepted 1 March 2013; published online 24 March 2013; doi:10.1038/ng.2591

of mutations affecting signal transduction pathways. These results provide a foundation for the further study of the mutagenic exposures and signaling pathways contributing to EAC tumorigenesis.

RESULTS

Landscape of EAC mutations and rearrangements

To identify somatic alterations in EAC, we performed whole-exome sequencing on tumor-normal pairs from 149 cases and whole-genome sequencing on 16 pairs (**Supplementary Note**). Fifteen of the samples from whole-genome sequencing had matched whole-exome sequencing data, and 14 samples from whole-genome sequencing were evaluated on mRNA expression arrays (**Supplementary Fig. 1** and **Supplementary Table 1**). One tumor on which whole-genome sequencing was performed lacked matched whole-exome sequencing data owing to sequencing failure for this sample. Somatic mutations were identified using the MuTect and Indelocator tools[14–17].

For whole-genome sequencing, tumors were sequenced to an average depth of 49× coverage, and matched germline DNA samples were sequenced to 30× coverage, with paired 101-bp reads on Illumina HiSeq instruments (**Supplementary Table 2**). We identified a median of 26,161 mutations across the genome per tumor (range of 18,881–66,225), corresponding to a median mutation frequency of 9.9 mutations/Mb (range of 7.1–25.2 mutations/Mb) relative to a haploid genome. The median mutation frequency was highest in intergenic regions (13.9 mutations/Mb), intermediate in intronic regions (8.7 mutations/Mb) and lowest in coding exons (6.6 mutations/Mb) (**Supplementary Table 3**). This stepwise decrease in mutation frequency was consistently seen in other cancers[16]. Compared to other cancer types, EAC has a high overall mutation frequency that is exceeded only in lung cancer[18,19] and melanoma[20], diseases that emerge from clear mutagens. By contrast, analogous sequencing of colorectal adenocarcinomas (CRCs) identified a mutation frequency of 5.6 mutations/Mb[21] across the genome. The high mutation frequency in EAC suggests that these tumors may be exposed to damaging mutagens, perhaps attributable to the harsh environment created by gastric refluxate and inflammation[22].

We also analyzed whole-genome sequencing data using the dRanger algorithm[21] to identify chromosomal rearrangements. A total of 2,952 candidate rearrangements were identified, with a median of 172 per tumor (range of 77–402) (**Supplementary Table 4**). Consistent with array data showing a higher degree of structural alterations in EAC compared to CRC[23], the number of rearrangements was much greater than observed with a comparable analysis of CRC genomes[21]. No correlation was observed between the numbers of mutations and rearrangements ($R^2 = 0.0046$). Of the rearrangements identified, 20% were interchromosomal translocations. Among the intrachromosomal alterations identified, a majority (55%) involved aberrant fusions of two sequences located within 1 Mb of each other. To identify potential fusion gene products that might contribute to the pathogenesis of EAC, we examined data for predicted in-frame gene fusions. Thirty-eight such events were identified (**Supplementary Table 5**), but no recurrent gene fusions were detected.

High frequency of A>C transversions at AA sites in EAC

Epithelial cancers often have variable mutation spectra pointing to particular mutagenic stimuli. Therefore, we analyzed the spectrum of mutations in EAC detected by whole-genome sequencing. Earlier exome sequencing of EACs noted A>C transversions to be more common in EAC compared to squamous esophageal carcinoma[13]. Evaluating the whole-genome sequencing data, we found that A>C base changes comprised an average of 34% of the total mutations (**Supplementary Table 6**). To comprehensively characterize the mutation spectra, we measured the frequencies of base mutations in different trinucleotide contexts and observed a preponderance of C>T transitions (39.2 mutations/Mb), as seen in most epithelial cancers. We further investigated the high frequency of A>C transversions (or equivalent T>G transversions on the complementary strand). These events showed preference (20.2 mutations/Mb) for the context of AA dinucleotides—that is, adenines flanked by a 5′ adenine and any 3′ nucleotide (**Fig. 1a**). In total, 84% of A>C mutations were flanked by a 5′ adenine. Expanding on these findings, we found that A>C transversions at AA dinucleotides were most pronounced when the 3′ base was a guanine (49.3 mutations/Mb) and was lower when it was an adenine (8.0 mutations/Mb), cytosine (16.8 mutations/Mb) or thymidine (6.7 mutations/Mb) (**Supplementary Table 7**). Validating these results, genotyping of randomly selected AA>AC mutations from intragenic regions showed a concordance rate of 100% (25/25). The high frequency of AA>AC transversions seems to be unique to EAC, as equivalent events have not been identified in other cancer types[15,16,18–20,24,25].

Overall, A>C transversions at AA sites accounted for 29% of the total mutations (**Fig. 1a**). Within individual tumors, these transversions at AA sites accounted for 5–48% of mutations (**Supplementary Table 7**), and the event number was correlated with the overall mutation frequency ($R^2 = 0.92$) (**Supplementary Fig. 2**). When we excluded transversions at AA sites, the mutation frequency was 8.5 mutations/Mb, still higher than in most tumor types. Thus, AA>AC mutations do not fully explain the elevated mutation frequency in EAC relative to other cancers.

We next characterized the distribution of mutations in different genomic regions. Although A>C transversions remained notable at AA sites, the percentage of all mutations consisting of these transversions was significantly lower in exons (16%) than across the entire genome (AAG, $P = 0.001$; AAT, $P = 0.0006$; AAC, $P = 0.0007$; AAA, $P = 0.0006$; two-tailed Student's t test) (**Fig. 1a**). These results were consistent across the coding regions of all 16 cases evaluated by whole-genome sequencing (**Supplementary Tables 8** and **9**). In contrast, the attenuation of C>T transitions at CG dinucleotides in coding regions (39.2 mutations/Mb versus 25.4 mutations/Mb) was smaller than that seen for transversions at AA dinucleotides.

The lower frequency of transversions at AA dinucleotides in coding areas relative to intergenic regions suggested that these mutations may be less likely to occur in transcribed regions or may be repaired effectively by transcription-coupled repair. To evaluate the potential impact of gene expression on mutation rates, we compared sample-specific frequencies of mutation at AA dinucleotides within gene boundaries at varying levels of gene expression in 14 whole-genome sequencing samples from which mRNA was available for microarray expression profiling. Higher expression was associated with lower global mutation frequency. Additionally, the impact of gene expression on attenuating mutation rates was threefold greater at AA sites than for mutations in other nucleotide contexts (**Fig. 1b** and **Supplementary Table 10**). This finding shows a strong effect of local gene expression on the development of transversions at AA sites in EAC.

Given the impact of transcription on these mutations, we analyzed transversions at AA sites for strand bias. The mutation rates for transversions at AA dinucleotides in introns and exons were calculated separately, according to whether the adenine base was on the transcribed or non-transcribed strand. The results indicated that AA>AC mutations were more common when the AA site was located on the non-transcribed strand (12.4 mutations/Mb versus 11.2 mutations/Mb; $P = 0.0016$, Student's t test, paired). When evaluating mutations at all base

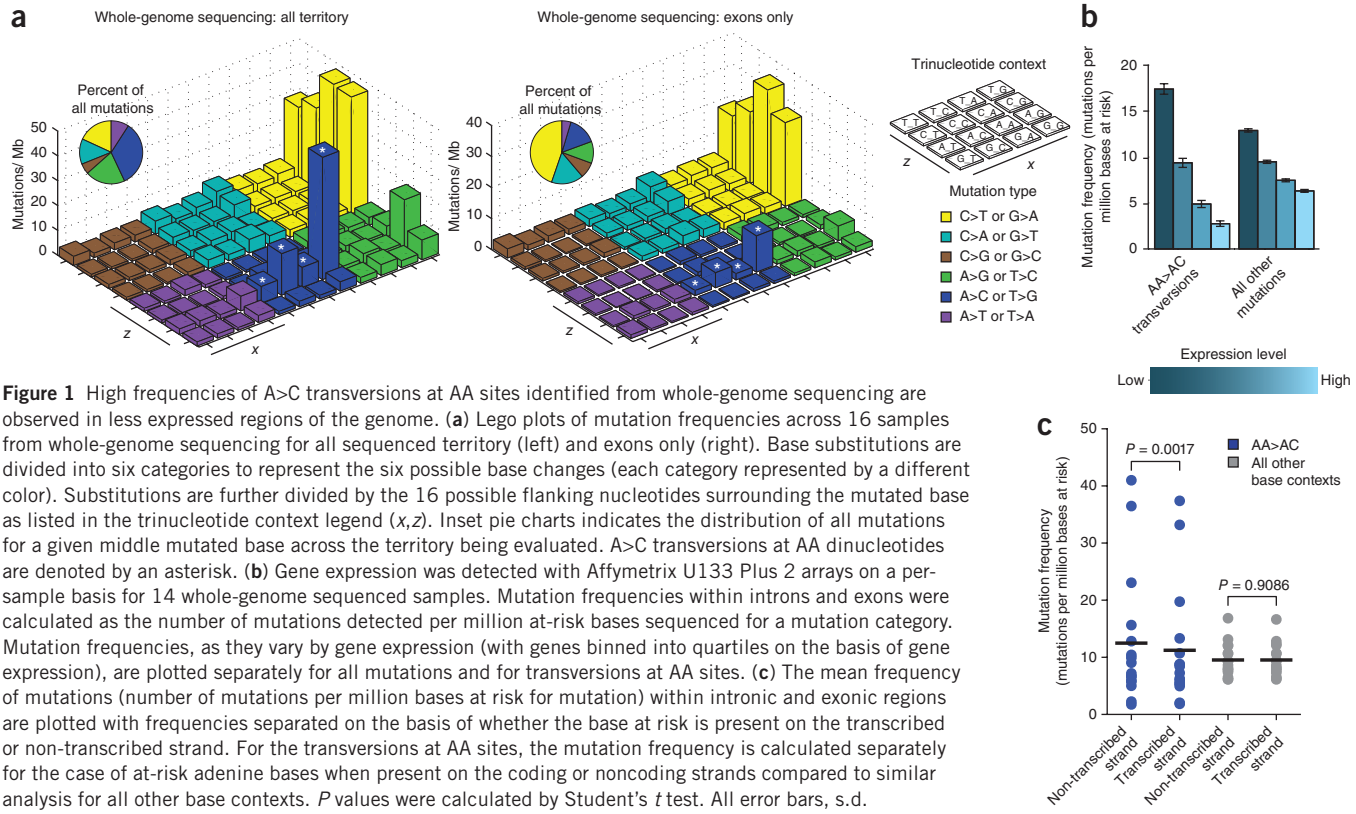

Figure 1 High frequencies of A>C transversions at AA sites identified from whole-genome sequencing are observed in less expressed regions of the genome. (**a**) Lego plots of mutation frequencies across 16 samples from whole-genome sequencing for all sequenced territory (left) and exons only (right). Base substitutions are divided into six categories to represent the six possible base changes (each category represented by a different color). Substitutions are further divided by the 16 possible flanking nucleotides surrounding the mutated base as listed in the trinucleotide context legend (*x*,*z*). Inset pie charts indicates the distribution of all mutations for a given middle mutated base across the territory being evaluated. A>C transversions at AA dinucleotides are denoted by an asterisk. (**b**) Gene expression was detected with Affymetrix U133 Plus 2 arrays on a per-sample basis for 14 whole-genome sequenced samples. Mutation frequencies within introns and exons were calculated as the number of mutations detected per million at-risk bases sequenced for a mutation category. Mutation frequencies, as they vary by gene expression (with genes binned into quartiles on the basis of gene expression), are plotted separately for all mutations and for transversions at AA sites. (**c**) The mean frequency of mutations (number of mutations per million bases at risk for mutation) within intronic and exonic regions are plotted with frequencies separated on the basis of whether the base at risk is present on the transcribed or non-transcribed strand. For the transversions at AA sites, the mutation frequency is calculated separately for the case of at-risk adenine bases when present on the coding or noncoding strands compared to similar analysis for all other base contexts. *P* values were calculated by Student's *t* test. All error bars, s.d.

contexts other than AA dinucleotides, a strand bias was not detected (9.5 mutations/Mb versus 9.5 mutations/Mb; $P = 0.9086$, Student's *t* test, paired) (**Fig. 1c** and **Supplementary Table 11**). These results suggest that transversions at AA sites may be more effectively recognized and repaired when the mutated adenine is located on the transcribed strand.

Mutations identified by exome sequencing

We next analyzed whole-exome sequencing data from 149 tumor-germline pairs (**Supplementary Table 12**). A mean coverage depth of 83.3× was achieved in neoplastic DNA, and 85.9× coverage was achieved in the non-cancerous tissue. Of the exons, 89% were covered at 8× or greater depth for normal samples and at 14× or greater depth for tumor samples, a threshold at which MuTect is powered to detect a mutation with an allele fraction of above or equal to 0.3 (refs. 14,16,17,26). We evaluated mutation calling by comparing candidate coding mutations identified by whole-exome sequencing to calls from whole-genome sequencing for the same tumor. Concordance of 85.1% (2,200/2,585) was observed for all events, and 90% concordance was seen for mutations present at greater than 0.1 allele fraction (**Supplementary Table 13**).

Four tumors had markedly higher coding mutation frequencies than other cases (14.6–50.9 mutations/Mb). This pattern resembled that of CRC where a subset of tumors were hypermutated, which was largely attributable to microsatellite instability (MSI). Similarly, MSI-positive tumors have been reported to represent 7% of EACs[27]. The four cases with the highest mutation rates were found to be MSI positive, with the tumor with the highest mutation frequency having mutations in two mismatch repair genes, *MSH6* and *MSH3* (**Supplementary Table 14**). By contrast, none of the 24 EAC samples with the next highest mutation frequencies (greater than 5 mutations/Mb) scored positive for MSI. To avoid a potential confounding effect on statistical analysis, we omitted the MSI-positive cases from the final analysis, leaving 145 tumors.

A total of 17,383 mutations, consisting of 16,516 nonsilent mutations and 1,954 insertion-deletion and/or null mutations were detected in the 145-sample cohort for a median of 104 nonsilent coding mutations per tumor (**Fig. 2**). The overall nonsilent median mutation frequency was 3.51 mutations/Mb (range of 0.97–10.8 mutations/Mb). We investigated whether the fraction of transversions at AA sites was associated with clinical variables, including age, disease stage, sex and tumor location. Notably, a trend was seen in which EACs developing within the tubular esophagus harbored a greater fraction of transversions at AA sites compared to tumors in the GEJ ($P = 0.076$, two-tailed Student's *t* test), a noteworthy result given the possibility that gastric refluxate in the lower esophagus serves as a mutagenic insult (**Fig. 2** and **Supplementary Fig. 3**). No other significant associations were identified.

Genes significantly mutated in esophageal adenocarcinoma

We observed mutations in 8,331 genes, of which 3,639 were mutated in 2 or more samples (**Supplementary Table 15**). Of these genes, 199 were mutated in 5% or more of the tumors, including 33 genes mutated in over 10% of cases. To identify genes showing evidence of positive selection for mutation, we used the mutation significance algorithm MutSig[14–16,26]. This tool compares the mutation occurrence in each gene to that which would be expected by chance given a background mutation frequency model that factors in the mutation spectra, presence of silent mutations, mutation frequencies and regional mutation frequencies along the genome[19]. We found 26 genes to be significantly mutated (false discovery rate (FDR) $q < 0.1$), with 2 known tumor suppressors in EAC, *TP53* and *CDKN2A*, being the most significant (**Fig. 2**). With the exception of *ARID1A*, *PIK3CA*

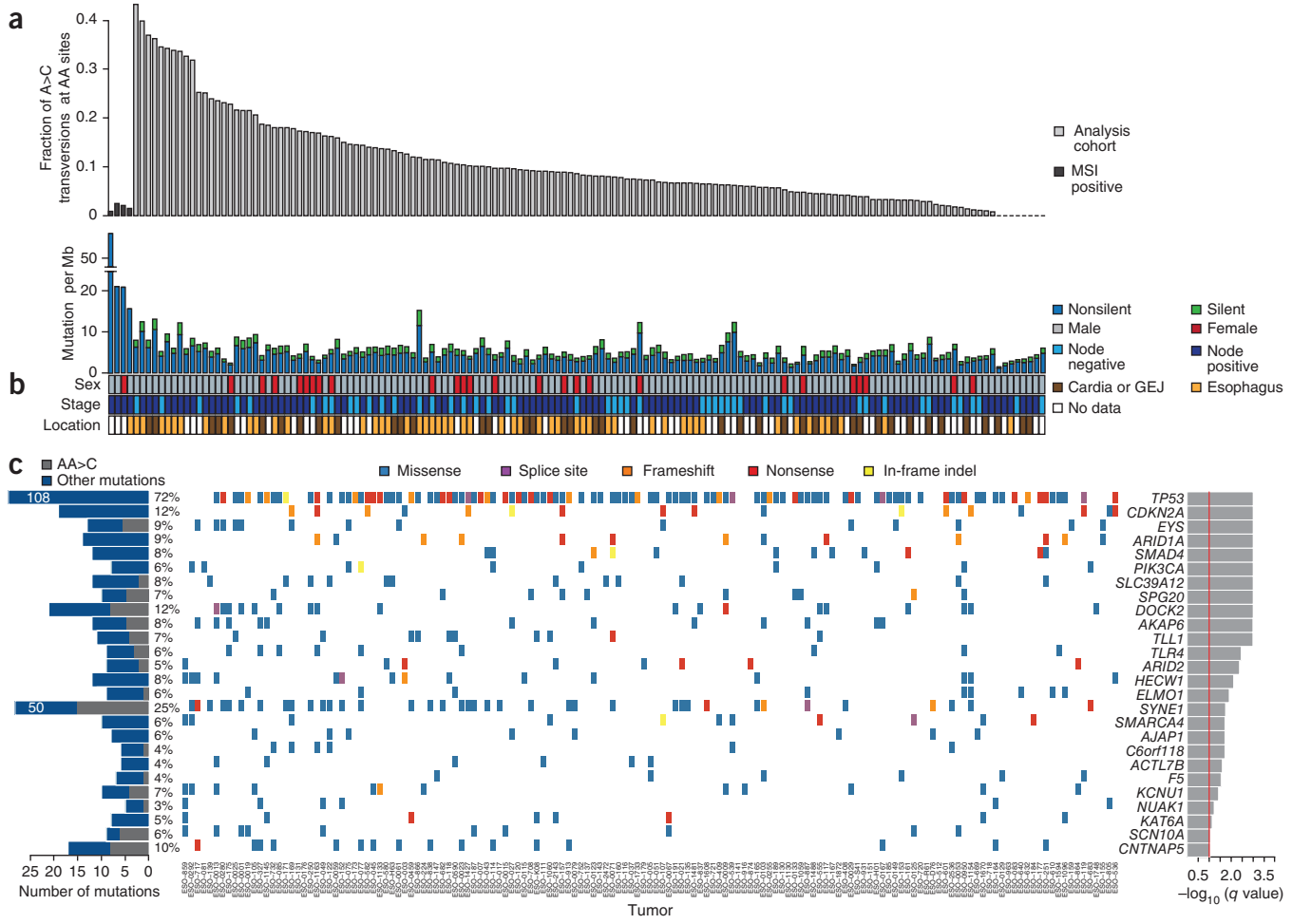

Figure 2 Mutation frequencies and significantly mutated genes in EAC as identified by whole-exome sequencing. (**a**) Mutation frequency of a cohort of 149 primary EACs is sorted by the fraction of mutations consisting of A>C transversions at AA sites. MSI-positive samples were not included in mutation significance analysis. (**b**) Key clinical parameters described in **Supplementary Table 1**. (**c**) Center, mutations in significantly mutated genes colored by the type of coding mutation. Each column denotes an individual tumor, and each row represents a gene. Left, number and percentage of samples with mutations in a given gene. The gray bar represents the number of transversions at AA sites in a gene. Listed numbers within bars represent values exceeding the scale. Right, $-\log_{10}$ (q values) for the significance level of mutated genes shown for all genes with FDR $q < 0.1$.

and *SMAD4*, no other significantly mutated gene had previously been implicated in EAC, although several had been implicated in other cancers.

Notably, two significantly mutated genes, *ELMO1* and *DOCK2*, encode dimerization partners and intracellular mediators of the Rho family GTPase, RAC1 (refs. 28,29). Because aberrant RAC1 activation has been implicated in malignant transformation in other cancer types, mainly by enhancing cellular motility[30–34], recurrent mutations in these genes may be functionally important. Although no *RAC1* or *RAC2* mutations were identified, *ELMO1* or *DOCK2* was mutated in 25 EAC samples (17%), with 2 samples having mutations in both genes and 2 samples having 2 independent mutations in *DOCK2* (**Fig. 3** and **Supplementary Table 15**). Notably, a single amino acid, Lys312 of ELMO1, was affected by mutation in three tumors, which suggests a gain-of-function phenotype. DOCK2 is a guanine nucleotide exchange factor (GEF) that activates RAC1 directly through GTP loading[28,35]. To fully activate RAC1, DOCK2 and ELMO1 interact to relieve mutual autoinhibition[29]. In cancer models, ELMO1 and other DOCK family members have been associated with enhanced migration and invasion[36,37]. Mutations were also present in other RAC1 GEFs (*TRIO*, *TIAM1*, *VAV2* and *ECT2*) (**Supplementary Fig. 4**). Furthermore, we previously observed focal copy number gain of the 11q13 locus containing the serine-threonine kinase *PAK1*, which encodes a principal downstream effector of RAC1 that has been shown to be oncogenic in breast cancer[23,38]. The aberrant activation of genes related to RAC1 suggests that the motility pathway might be important in EAC.

To examine the relevance of *ELMO1* mutations, wild-type and mutant *ELMO1* constructs were generated and introduced into NIH/3T3 cells. On the basis of studies in glioblastoma showing a correlative increase in cellular invasion with the overexpression of wild-type ELMO1 (ref. 37), we hypothesized that *ELMO1* mutations would enhance cell invasion. Compared to green fluorescent protein (GFP) control, wild-type ELMO1 increased invasion by sevenfold ($P = 0.0040$, Student's t test, unpaired) (**Fig. 3c**). ELMO1 alterations (p.Phe59Leu, p.Lys312Glu, p.Lys312Thr, p.Lys349Arg and p.Thr421Asn) resulted in further significant increases (two- to sevenfold) in invasion compared to wild-type ELMO1 (**Fig. 3c**). These results suggest that *ELMO1* mutations can increase invasiveness and potentially contribute to tumorigenesis in EAC.

Additional significantly mutated genes included members of the SWI/SNF family of chromatin-remodeling factors: *ARID1A*, *SMARCA4* and *ARID2*. Together, these genes were mutated in 20% of tumors. The enzymatic subunit of the chromatin-remodeling

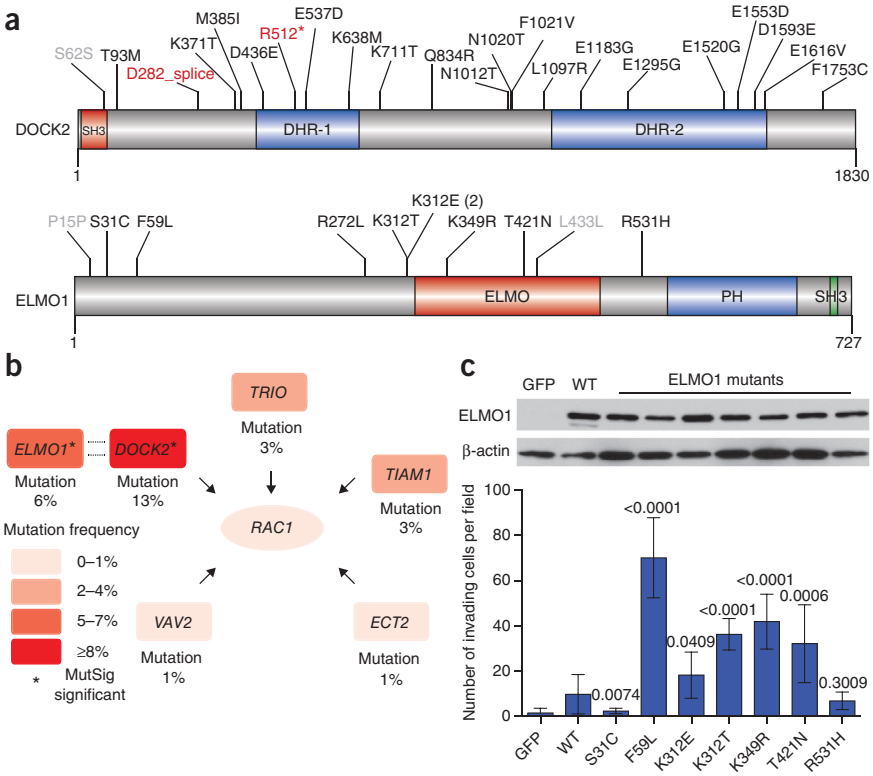

Figure 3 Recurrent somatic alterations in ELMO1, DOCK2 and other RAC1 GEFs. (**a**) Schematics of protein alterations in DOCK2 and ELMO1 detected by whole-exome sequencing. Coding alterations in EAC are colored either black (missense) or red (splice site or nonsense); silent mutations are depicted in gray. Conserved domain mapping is from UniProt; SH3, SRC homology 3; DHR, Dlg homologous region, ELMO, engulfment and cell motility; PH, Pleckstrin homology. (**b**) Sample mutational frequency of candidate *ELMO1* and *DOCK2* as well as other RAC1-activating GEFs in 145 whole-exome sequenced EACs. (**c**) Wild-type or mutant ELMO1 proteins (or GFP control) were expressed in NIH/3T3 cells using retroviral transduction with the pBabe vector. Protein expression was confirmed by immunoblot analysis. Cells were plated in Matrigel invasion chambers with medium containing full serum in the lower chamber only, and invading cells from four fields were counted. The numbers of invading cells from three independent replicates are shown. Error bars, s.d. *P* values compare mutant ELMO1 to wild-type protein, Student's *t* test.

complex, *SMARCA4*, has been established as a putative tumor suppressor[14,39]. Likewise, *ARID1A* and *ARID2* have been implicated as tumor suppressors in cancers including gastric cancer[40–43]. Notably, a candidate protein fusion identified by whole-genome sequencing also targeted *SMARCA4*. The predicted fusion between exon 11 of *SMARCA4* and exon 14 of *DNM2* might point to an alteration that results in a loss-of-function phenotype (**Supplementary Table 5**). Mutations were also found in other chromatin-modifying enzymes, including *PBRM1* (ref. 44) and *JARID2*. Taken together, 24% (35/145) of EACs harbored mutations in genes encoding chromatin-modifying factors (**Supplementary Figs. 4** and **5**).

Another noteworthy gene was *SPG20*, which was mutated in 7% of EACs, with five of the mutations generated by transversions at AA dinucleotides (**Supplementary Fig. 6**). Spartin, the gene product of *SPG20*, was reportedly mutated in Troyer syndrome[45], a genetic disorder characterized by progressive muscle stiffness and limb paralysis. The functions of Spartin include the endosomal trafficking of growth factor receptors, inhibition of bone morphogenetic protein signaling and ubiquitin targeting[46]. More recently, *SPG20* hypermethylation has been linked to colon cancer progression[47].

TLR4 was mutated in 6% of EACs. Germline polymorphisms in *TLR4* correlate with risk of *Helicobacter pylori*–mediated gastric carcinoma[48]. In a lung cancer model, TLR4 inactivation through mutation contributes to greater inflammation and tumorigenesis[49]. TLR4 activates the innate immune response to pathogen exposure through heterodimerization with MD-2 (ref. 50). Notably, the mutations in *TLR4* affect residues between amino acids Asp379 and Phe487, a region critical for MD-2 interaction[51]. One mutation affects Glu439, a site essential for the hydrogen bonding of TLR4 with MD-2 (**Supplementary Fig. 6**). These mutations suggest disruption of the TLR4–MD-2 complex as a potential driver of tumor progression in EAC.

We also identified other significantly mutated candidate genes, including the protein kinase A–anchoring factor *AKAP6* (mutated in 8% of samples), the E3 ubiquitin ligase *HECW1* (8%) and *AJAP1* (6%), which mediates signaling at adherens junctions and increases invasiveness in cancer cell lines[52]. *NUAK1* (also known as *ARK5*) was mutated in 3% of cases, which is notable given that *MYC*-overexpressing hepatocellular carcinoma models are dependent on NUAK1 (ref. 53).

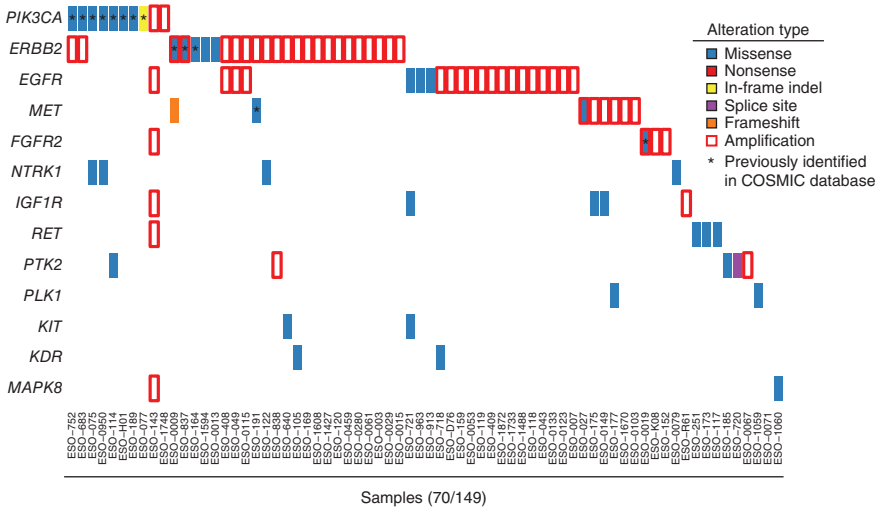

Figure 4 Somatic mutations in frequently altered pathways in cancer, putative therapeutic targets and treatment biomarkers. Potential therapeutic targets or treatment biomarkers are listed by sample. Each column denotes an individual tumor, and each row represents a gene. Mutations are colored by the type of mutation event, and genes with amplification of greater than four copies relative to diploid baseline are marked by red outlines.

The lysine acetyltransferase *KAT6A* (also known as *MYST3*), recurrently targeted by translocation in leukemia[54], was also mutated in seven specimens (5%) (**Supplementary Fig. 6**).

Additional candidate genes

Beyond the genes mutated at a statistically significant frequency, we queried the data for mutations of biological relevance given their recurrence in other cancers. We identified mutations in EAC that had been seen 2 or more times across all cancers in the Catalogue of Somatic Mutations in Cancer (COSMIC) database[55]: we found 22 such genes (**Supplementary Table 16**). Additionally, ten genes were significantly mutated (FDR $q < 0.1$) in limited analysis of COSMIC gene territory, including *KRAS*, *CTNNB1* and *ERBB2* (**Supplementary Table 17**). These results indicate that genes not reaching statistical significance in the cohort may harbor mutations of biological relevance in individual tumors.

Mutations targeting therapeutically relevant genes

We queried the data for mutations in genes encoding the therapeutic targets of inhibitors approved for clinical use or in preclinical development[56]. Mutations in actionable genes were discovered in 23% of tumors, with *PIK3CA* being the most frequently mutated (**Fig. 4**). When also evaluating the amplification status of genes, 48% of tumors in this cohort had a genomic alteration in a gene with a targeted agent. The high frequency of focally amplified therapeutic targets[23] exceeded that of mutation in these same genes in EAC. Therefore, determining how to effectively treat tumors with amplified targets, especially receptor tyrosine kinases (RTKs), should be considered a priority.

Somatic alterations in signal transduction pathways

To explore the functional impact of the mutations, we performed unbiased Gene Ontology (GO) term enrichment in the overall ranked MutSig list using the 8,356 genes with at least 1 nonsilent mutation[57,58]. GO processes related to cell adhesion and chemotaxis ranked as enriched near the top of the list (**Supplementary Table 18**). These findings support the hypothesis that enhanced cellular motility and invasiveness has an important role in EAC disease progression.

We also studied how cancer-associated pathways were disrupted by mutation in EAC. Cell cycle control was altered by point mutation in 14% of EACs, with most of the mutations occurring in *CDKN2A*

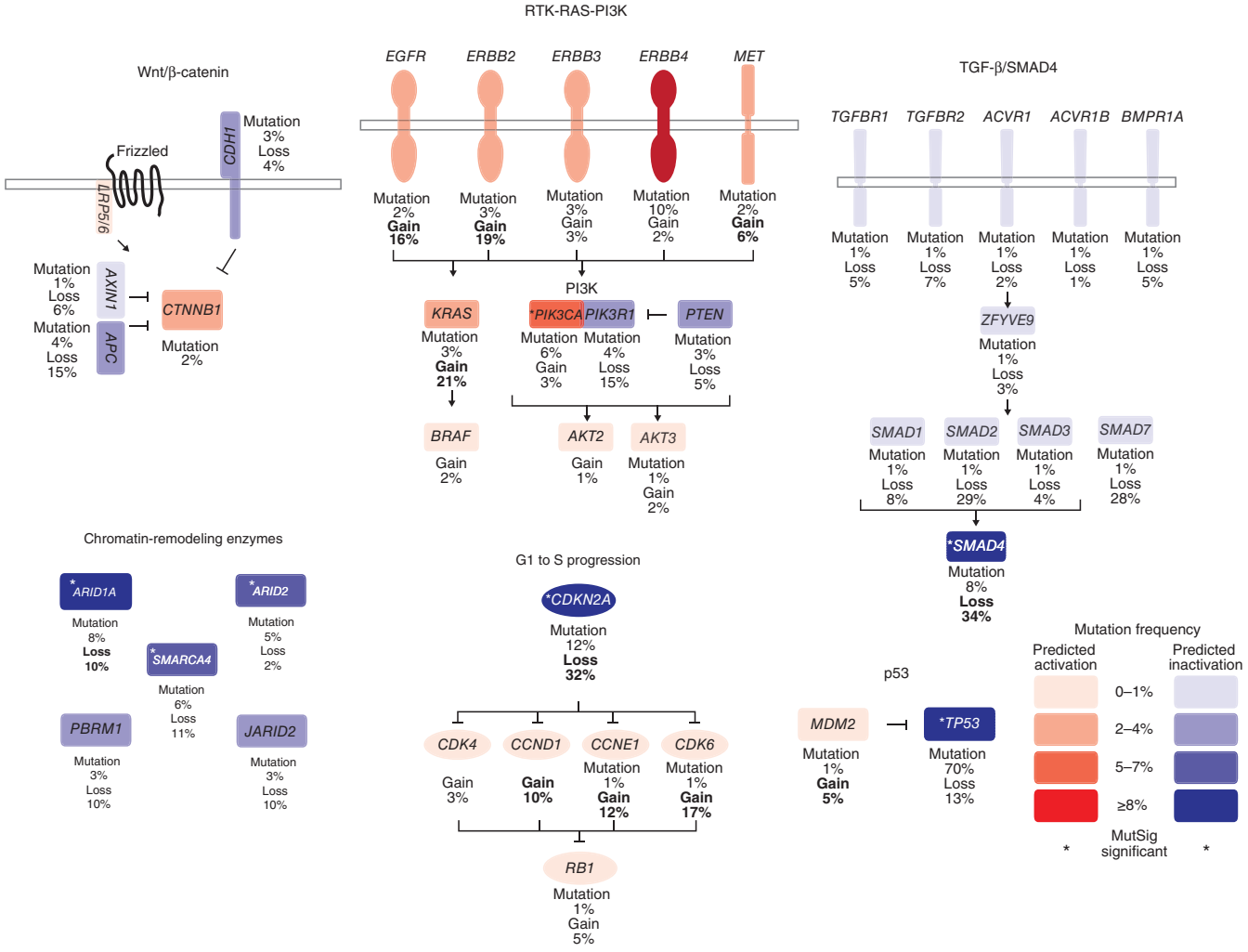

Figure 5 Genetic alterations identified by whole-exome sequencing across 145 EACs affecting the Wnt/β-catenin, RTK-RAS-PI3K, TGF-β/SMAD4, chromatin-remodeling enzyme, G1 to S progression and p53 pathways. Percentages represent the number of mutations in a given gene across the cohort. Genes that are predicted to have gain or loss of function are depicted in red and blue, respectively, with color intensity based on the mutation frequency of a given gene. Frequencies of alteration by mutation or copy number variation are shown. Genes subject to significant focal gain or loss in EAC[23] have copy number frequencies marked in bold.

(**Fig. 5** and **Supplementary Fig. 4**). This process was also frequently affected by amplifications at the *CCND1*, *CCNE1* and *CDK6* loci[23]. Although activation of β-catenin signaling is ubiquitous in CRC, mutations in this pathway were found in only 9% of EACs, with two tumors having *APC* mutations that coincided with mutation in either *CDH1* or *AXIN1* (**Fig. 5** and **Supplementary Fig. 4**). Moreover, a potential *AXIN1* fusion was identified by whole-genome sequencing in sample ESO-1060 spanning exon 5 of *AXIN1* and exon 2 of *GALNT7*, which might alter normal gene function (**Supplementary Table 5**). As in other cancer types, the TGF-β–SMAD signaling pathway was mutated in 18% of EAC tumors. The most recurrently altered gene in this pathway was *SMAD4*, which was mutated in ten samples and was also subject to frequent copy number loss (**Fig. 5** and **Supplementary Fig. 4**).

We evaluated the frequency and manner of somatic alterations in mitogen-activated protein kinase (MAPK) and phosphatidylinositol 3-kinase (PI3K) signaling, two common pathways required for the proliferation and survival of cancer cells. Unlike in other epithelial tumor types where MAPK pathway mutations are common, no *BRAF* mutations were observed in EAC, and *NF1* and *KRAS* mutations were seen in only three (2%) and five (3%) tumors, respectively. Three of the five *KRAS* mutations altered Gly12; however, one EAC harbored a *KRAS* c.351A>C (p.Lys117Asn) event, a mutation caused by a transversion at an AA dinucleotide that was previously observed in CRC[59]. The PI3K pathway was the most frequently altered oncogenic pathway affected by mutation (13%). *PIK3CA* was mutated in seven tumors, and *PIK3R1* and *PTEN* were mutated in five and four tumors, respectively (**Fig. 5b** and **Supplementary Fig. 4**).

We explored mutations in the ErbB family of RTKs, which are important therapeutic targets in many cancer types. Although three samples harbored *EGFR* mutations, these alterations were not previously annotated in other tumors. Moreover, two of these alterations, p.Ser447Tyr and p.Ser1153Ile, were predicted by PolyPhen-2 score[60] to not be deleterious to normal function and were thus of questionable biological relevance. By contrast, *ERBB2* mutations were present in five tumors. Three mutations affected the kinase domain, including two c.351A>C (p.Asp769Tyr) mutations and one c.2327G>T (p.Gly776Val) mutation. These alterations have been observed previously in other cancers[61–63].

DISCUSSION

Here, through mutation analysis, we provide insight into somatically altered genes and signaling pathways, as well as confirm a high rate of A>C transversions in EAC[13]. We further establish that the rates of these mutations are highest in noncoding regions and, within coding areas, are over-represented in less expressed genes. Additionally, we demonstrate the context specificity of these mutations, showing that A>C transversions are most common when the mutated adenine follows a 5′ flanking adenine (AA), especially at AAG trinucleotides.

This mutational spectrum seems to be unique to EAC, suggesting that these mutations are attributable to gastroesophageal reflux, where the gastric and duodenal contents travel into the lower esophagus, creating an environment of inflammation[22]. Previous studies have linked particular substances such as bile acids, nitrosamines and reactive oxygen species to the development of metaplasia and carcinomas[64], but the precise mutagen(s) remain poorly understood. Experiments in *Escherichia coli* exploring the mutagenic potential of an oxidatively damaged DNA precursor, 8-hydroxydeoxyguanosine triphosphate, showed that it preferentially induces A>C transversions[65]. These data suggest that A>C transversions in EAC might arise from the oxidative damage induced by GERD; however, experimental evidence is necessary to identify a culprit stimulus. The identification of this mutational signature enables future studies to define specific carcinogen(s) that contribute to EAC that potentially aid in the explanation of the rising incidence of this cancer.

Statistical analysis also enabled a comprehensive assessment of mutated genes in EAC and identified mutations in cancer-related genes such as *TP53*, *CDKN2A*, *SMAD4* and *PIK3CA*. It was notable that most well-annotated cancer genes were not affected by transversions at AA sites. In many cases, it is impossible to generate hotspot mutations, such as those affecting KRAS Gly12 or PIK3CA Glu545, with a transversion at an AA dinucleotide. Additionally, given the base composition of stop codons, it is difficult to generate nonsense events from transversions at AA sites and impossible to create a stop mutation from an A>C transversion when it occurs in an AAG trinucleotide context, the most common context for mutations at AA sites in our data set. Of the 2,570 coding mutations caused by these events, none is a predicted nonsense mutation. Moreover, the data suggest that transversions at AA sites accumulate in genes with lower expression, thus reducing their prevalence in the genes contributing to oncogenesis. Despite these caveats, it is likely that the mutations caused by transversions at AA dinucleotides do affect genes relevant for these tumors. For example, a known transforming mutation in *KRAS* (c.351A>C; p.Lys117Asn) is created by a transversion at an AA dinucleotide.

Consistent with previous reports[39–44], loss-of-function mutations in chromatin-remodeling enzymes are common in EAC. Previous gene studies have also suggested frequent activation of the MAPK, PI3K and β-catenin pathways. The data presented here verify the presence of frequent mutations in the PI3K cascade but argue against wide-reaching mutations in these pathways, thus drawing contrasts between EAC and CRC, where β-catenin activation and missense mutations of *KRAS* and *BRAF* are highly prevalent[66].

For the first time, we detected EAC-relevant mutations in regulators of invasion and motility, including significantly recurrent mutations in *DOCK2* and *ELMO1*. These mutations may increase tumor fitness through alteration of cytoskeletal structure, increase in invasive properties or mitogenesis. We show that *ELMO1* mutations augment cellular invasiveness, thus suggesting one mechanism by which these events contribute to tumorigenesis. Given that EAC is a highly invasive tumor that is prone to early metastasis, alterations in the RAC1 pathway may contribute to this phenotype.

Although we identified potentially actionable genomic alterations in 48% of samples, the trastuzumab antibody to ERBB2 (also known as HER2) is the only targeted agent used in the treatment of GEJ adenocarcinomas, which include EAC adenocarcinomas, with its use guided by the overexpression and genomic amplification of *ERBB2* (ref. 67). Currently, *ERBB2* mutation assessment is not performed for EAC, despite *ERBB2* being altered by both amplification and mutation in 3% of samples. The current data point to a potential use of mutation as an additional biomarker to guide the use of ERBB2-targeting agents.

Limited knowledge of the genomic aberrations underlying EAC has hindered the development of new therapies. Numerous candidates, not previously implicated in this disease, have emerged from the current analysis. Functional study of these genes will be required to validate and understand their roles in tumorigenesis and to identify the etiology of the unique spectrum of observed transversions at AA dinucleotides. These data provide an enhanced road map for the study of EAC and the much-needed development of new therapies for this deadly cancer.

URLs. MutSig Algorithm, http://www.broadinstitute.org/cancer/cga/mutsig; CCDS, http://www.ncbi.nlm.nih.gov/CCDS/; Broad Institute

Picard Sequencing Pipeline, http://picard.sourceforge.net/; Broad Institute Firehose Pipeline, http://www.broadinstitute.org/cancer/cga; ABSOLUTE algorithm, http://www.broadinstitute.org/cancer/cga/ABSOLUTE.

METHODS

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

Accession codes. Binary sequence alignment/map (BAM) files were deposited in the database of Genotypes and Phenotypes (dbGaP) under accession phs000598.v1.p1. Raw mRNA expression data for 14 EAC samples have been deposited at the Gene Expression Omnibus (GEO) under accession GSE42363.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Meyerson for helpful discussions and review of the manuscript and members of the Broad Institute Biological Samples Platform, Genetic Analysis Platform and Genome Sequencing Platform for their assistance. We are also grateful for the physicians and hospital staff whose efforts in collecting these samples are essential to this research. This work was supported by the US National Human Genome Research Institute (NHGRI) Large-Scale Sequencing Program (U54 HG003067 to the Broad Institute, E.S.L.), the National Cancer Institute (K08 CA134931 to A.J.B.), the DeGregorio Family Foundation (A.J.B.), the Karin Grunebaum Cancer Research Foundation (A.J.B.), the Target Cancer (A.J.B.) and Connecticut Conquers Cancer (A.J.B.). S.O. and Y.I. are supported by the National Cancer Institute (R01 CA151993 to S.O.) and the Dana-Farber/Harvard Cancer Center GI Cancer Specialized Programs of Research Excellence (US National Institutes of Health (NIH) grant P50 CA127003). D.G.B. is supported by NIH grants CA163059 and CA46592. J.D.L. is supported by NIH grant CA090665. T.E.G. is supported by NIH grant CA130853.

AUTHOR CONTRIBUTIONS

P.S., S.P., M.S.L., C.F., C. Stewart, S.E.S., A.M., K.C., A.S., S.L.C., G.S., D.V., A.H.R. and R.B. performed computational analyses. E.S., D.A., K.T., C. Sougnez, R.C.O., C.G. and S.B.G. processed samples and supervised exome sequencing. A.M.D., S.B., D.Z., L.L., J.L., R.R., A.C., R.L., J.D.L., A.P., D.G.B., T.E.G. and A.J.B. coordinated sample acquisition, processing, pathological review and analysis. Y.I. and S.O. performed MSI testing. A.M.D., P.S., T.R.G., S.B.G., E.S.L., G.G. and A.J.B. designed the study. A.M.D., P.S., S.P., M.S.L., G.G. and A.J.B. analyzed the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Holmes, R.S. & Vaughan, T.L. Epidemiology and pathogenesis of esophageal cancer. *Semin. Radiat. Oncol.* **17**, 2–9 (2007).
2. Pohl, H. & Welch, H.G. The role of overdiagnosis and reclassification in the marked increase of esophageal adenocarcinoma incidence. *J. Natl. Cancer Inst.* **97**, 142–146 (2005).
3. Wu, A.H., Wan, P. & Bernstein, L. A multiethnic population-based study of smoking, alcohol and body size and risk of adenocarcinomas of the stomach and esophagus (United States). *Cancer Causes Control* **12**, 721–732 (2001).
4. Chung, S.M., Kao, J., Hyjek, E. & Chen, Y.T. p53 in esophageal adenocarcinoma: a critical reassessment of mutation frequency and identification of 72Arg as the dominant allele. *Int. J. Oncol.* **31**, 1351–1355 (2007).
5. Hardie, L.J. *et al.* p16 expression in Barrett's esophagus and esophageal adenocarcinoma: association with genetic and epigenetic alterations. *Cancer Lett.* **217**, 221–230 (2005).
6. Choi, Y.W., Heath, E.I., Heitmiller, R., Forastiere, A.A. & Wu, T.T. Mutations in β-catenin and APC genes are uncommon in esophageal and esophagogastric junction adenocarcinomas. *Mod. Pathol.* **13**, 1055–1059 (2000).
7. Sommerer, F. *et al.* Mutations of *BRAF* and *KRAS2* in the development of Barrett's adenocarcinoma. *Oncogene* **23**, 554–558 (2004).
8. Wijnhoven, B.P., de Both, N.J., van Dekken, H., Tilanus, H.W. & Dinjens, W.N. E-cadherin gene mutations are rare in adenocarcinomas of the oesophagus. *Br. J. Cancer* **80**, 1652–1657 (1999).
9. Pühringer-Oppermann, F.A., Stein, H.J. & Sarbia, M. Lack of *EGFR* gene mutations in exons 19 and 21 in esophageal (Barrett's) adenocarcinomas. *Dis. Esophagus* **20**, 9–11 (2007).
10. Guo, M., Liu, S. & Lu, F. Gefitinib-sensitizing mutations in esophageal carcinoma. *N. Engl. J. Med.* **354**, 2193–2194 (2006).
11. Phillips, W.A. *et al.* Mutation analysis of *PIK3CA* and *PIK3CB* in esophageal cancer and Barrett's esophagus. *Int. J. Cancer* **118**, 2644–2646 (2006).
12. Boonstra, J.J. *et al.* Mapping of homozygous deletions in verified esophageal adenocarcinoma cell lines and xenografts. *Genes Chromosom. Cancer* **51**, 272–282 (2012).
13. Agrawal, N. *et al.* Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discov.* **2**, 899–905 (2012).
14. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
15. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
16. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
17. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
18. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
19. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
20. Berger, M.F. *et al.* Melanoma genome sequencing reveals frequent *PREX2* mutations. *Nature* **485**, 502–506 (2012).
21. Bass, A.J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion. *Nat. Genet.* **43**, 964–968 (2011).
22. Orlando, R.C. Mucosal defense in Barrett's esophagus. in *Barrett's Esophagus and Esophageal Adenocarcinoma* (ed. Sharma, P.) 60–72 (Blackwell Publishing, Oxford, 2006).
23. Dulak, A.M. *et al.* Gastrointestinal adenocarcinomas of the esophagus, stomach and colon exhibit distinct patterns of genome instability and oncogenesis. *Cancer Res.* **72**, 4383–4393 (2012).
24. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
25. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
26. Barbieri, C.E. *et al.* Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
27. Farris, A.B. III *et al.* Clinicopathologic and molecular profiles of microsatellite unstable Barrett esophagus–associated adenocarcinoma. *Am. J. Surg. Pathol.* **35**, 647–655 (2011).
28. Sanui, T. *et al.* DOCK2 regulates Rac activation and cytoskeletal reorganization through interaction with ELMO1. *Blood* **102**, 2948–2950 (2003).
29. Hanawa-Suetsugu, K. *et al.* Structural basis for mutual relief of the Rac guanine nucleotide exchange factor DOCK2 and its partner ELMO1 from their autoinhibited forms. *Proc. Natl. Acad. Sci. USA* **109**, 3305–3310 (2012).
30. Gómez del Pulgar, T., Benitah, S.A., Valeron, P.F., Espina, C. & Lacal, J.C. Rho GTPase expression in tumourigenesis: evidence for a significant link. *Bioessays* **27**, 602–613 (2005).
31. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
32. Kissil, J.L. *et al.* Requirement for Rac1 in a K-ras–induced lung cancer in the mouse. *Cancer Res.* **67**, 8089–8094 (2007).
33. Pan, Y. *et al.* Expression of seven main Rho family members in gastric carcinoma. *Biochem. Biophys. Res. Commun.* **315**, 686–691 (2004).
34. Sander, E.E. *et al.* Matrix-dependent Tiam1/Rac signaling in epithelial cells promotes either cell-cell adhesion or cell migration and is regulated by phosphatidylinositol 3-kinase. *J. Cell Biol.* **143**, 1385–1398 (1998).
35. Nishihara, H. *et al.* Non-adherent cell–specific expression of DOCK2, a member of the human CDM-family proteins. *Biochim. Biophys. Acta* **1452**, 179–187 (1999).
36. Sanz-Moreno, V. *et al.* Rac activation and inactivation control plasticity of tumor cell movement. *Cell* **135**, 510–523 (2008).
37. Jarzynka, M.J. *et al.* ELMO1 and Dock180, a bipartite Rac1 guanine nucleotide exchange factor, promote human glioma cell invasion. *Cancer Res.* **67**, 7203–7211 (2007).
38. Shrestha, Y. *et al. PAK1* is a breast cancer oncogene that coordinately activates MAPK and MET signaling. *Oncogene* **31**, 3397–3408 (2012).
39. Medina, P.P. *et al.* Frequent *BRG1*/*SMARCA4*-inactivating mutations in human lung cancer cell lines. *Hum. Mutat.* **29**, 617–622 (2008).
40. Zang, Z.J. *et al.* Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat. Genet.* **44**, 570–574 (2012).
41. Fujimoto, A. *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44**, 760–764 (2012).
42. Jones, S. *et al.* Somatic mutations in the chromatin remodeling gene *ARID1A* occur in several tumor types. *Hum. Mutat.* **33**, 100–103 (2012).
43. Guan, B., Wang, T.L. & Shih Ie, M. ARID1A, a factor that promotes formation of SWI/SNF-mediated chromatin remodeling, is a tumor suppressor in gynecologic cancers. *Cancer Res.* **71**, 6718–6727 (2011).
44. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene *PBRM1* in renal carcinoma. *Nature* **469**, 539–542 (2011).

45. Patel, H. *et al. SPG20* is mutated in Troyer syndrome, an hereditary spastic paraplegia. *Nat. Genet.* **31**, 347–348 (2002).
46. Bakowska, J.C., Jupille, H., Fatheddin, P., Puertollano, R. & Blackstone, C. Troyer syndrome protein spartin is mono-ubiquitinated and functions in EGF receptor trafficking. *Mol. Biol. Cell* **18**, 1683–1692 (2007).
47. Lind, G.E. *et al. SPG20*, a novel biomarker for early detection of colorectal cancer, encodes a regulator of cytokinesis. *Oncogene* **30**, 3967–3978 (2011).
48. Garza-Gonzalez, E. *et al.* Assessment of the toll-like receptor 4 Asp299Gly, Thr399Ile and interleukin-8-251 polymorphisms in the risk for the development of distal gastric cancer. *BMC Cancer* **7**, 70 (2007).
49. Bauer, A.K. *et al.* Toll-like receptor 4 in butylated hydroxytoluene–induced mouse pulmonary inflammation and tumorigenesis. *J. Natl. Cancer Inst.* **97**, 1778–1781 (2005).
50. Kennedy, M.N. *et al.* A complex of soluble MD-2 and lipopolysaccharide serves as an activating ligand for Toll-like receptor 4. *J. Biol. Chem.* **279**, 34698–34704 (2004).
51. Park, B.S. *et al.* The structural basis of lipopolysaccharide recognition by the TLR4–MD-2 complex. *Nature* **458**, 1191–1195 (2009).
52. Schreiner, A. *et al.* Junction protein shrew-1 influences cell invasion and interacts with invasion-promoting protein CD147. *Mol. Biol. Cell* **18**, 1272–1281 (2007).
53. Liu, L. *et al.* Deregulated MYC expression induces dependence upon AMPK-related kinase 5. *Nature* **483**, 608–612 (2012).
54. Pelletier, N., Champagne, N., Stifani, S. & Yang, X.J. MOZ and MORF histone acetyltransferases interact with the Runt-domain transcription factor Runx2. *Oncogene* **21**, 2729–2740 (2002).
55. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
56. Garnett, M.J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
57. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
58. Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.* **3**, e39 (2007).
59. Smith, G. *et al.* Activating K-Ras mutations outwith 'hotspot' codons in sporadic colorectal tumours—implications for personalised cancer medicine. *Br. J. Cancer* **102**, 693–703 (2010).
60. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
61. Ikediobi, O.N. *et al.* Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. *Mol. Cancer Ther.* **5**, 2606–2612 (2006).
62. Lee, J.W. *et al. ERBB2* kinase domain mutation in the lung squamous cell carcinoma. *Cancer Lett.* **237**, 89–94 (2006).
63. Lee, J.W. *et al.* Somatic mutations of *ERBB2* kinase domain in gastric, colorectal, and breast carcinomas. *Clin. Cancer Res.* **12**, 57–61 (2006).
64. Badreddine, R.J. & Wang, K.K. Barrett esophagus: an update. *Nat. Rev. Gastroenterol. Hepatol.* **7**, 369–378 (2010).
65. Inoue, M. *et al.* Induction of chromosomal gene mutations in *Escherichia coli* by direct incorporation of oxidatively damaged nucleotides. New evaluation method for mutagenesis by damaged DNA precursors *in vivo*. *J. Biol. Chem.* **273**, 11069–11074 (1998).
66. MacConaill, L.E. *et al.* Profiling critical cancer gene mutations in clinical tumor samples. *PLoS ONE* **4**, e7887 (2009).
67. Bang, Y.J. *et al.* Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet* **376**, 687–697 (2010).

ONLINE METHODS

DNA extraction and sample collection. All samples were obtained under institutional review board (IRB) approval and with documented informed consent. All samples were fresh-frozen primary resections from individuals with EAC not treated with previous chemotherapy or radiation. Slides stained with hematoxylin and eosin were examined by a board-certified pathologist to select cases with estimated carcinoma content of >70%. DNA was extracted using salt precipitation or phenol-chloroform extraction. DNA was quantified using PicoGreen dsDNA Quantitation Reagent (Invitrogen).

Whole-exome sequencing. Whole-exome capture libraries were constructed from 100 ng of tumor and normal DNA after shearing, end repair, phosphorylation and ligation to barcoded sequencing adaptors[68]. Ligated DNA was size selected for lengths between 200 and 350 bp and subjected to exonic hybrid capture using SureSelect v2 Exome bait (Agilent Technologies). Samples were multiplexed and sequenced on multiple Illumina HiSeq flow cells to average target exome coverage of 83.3× in neoplastic DNA and 85.9× in non-cancerous tissue.

Whole-genome sequencing. Samples were chosen for whole-genome sequencing on the basis of purity and ploidy estimates from ABSOLUTE[69]. Whole-genome sequencing library construction was carried out with 500 ng of native DNA from primary tumor and germline samples for each individual. DNA was sheared to a range of 101–700 bp using the Covaris E210 Instrument and then phosphorylated and adenylated according to the Illumina protocol. Adaptor-ligated purification was performed by preparatory gel electrophoresis (4% agarose at 85 V for 3 h), and size was selected by excision of two bands (at 500–520 bp and 520–540 bp), yielding two libraries per sample with average sizes of 380 and 400 bp, respectively[15,16,21]. Qiagen Min-Elute column-based cleanup was performed after each step. For a subset of samples, gel electrophoresis and extraction were performed using the automated Pippin Prep system (Sage Science). Libraries were then sequenced with the Illumina Genome Analyzer IIx or the Illumina HiSeq sequencer with 101-bp reads, achieving an average of ~30× coverage depth.

Identification of rearrangements. The dRanger algorithm was used to detect genomic rearrangements by identifying instances where the two read pairs mapped to distinct regions of the genome or mapped in a manner that suggested another structural event such as an inversion. Candidate somatic rearrangements were queried in both the matched normal genome and a panel of non-tumor genomes to remove germline events. The final scorings of these somatic reads were then calculated by multiplying the number of supporting read pairs by the estimated quality of the candidate rearrangement (0 to 1). This metric is generated by taking into account the ability to align of the two regions joined by the putative rearrangement and the chance of detecting such a read pair given the library fragment size distribution. Events with scores of ≥4 (observed in at least 4 read pairs) were included in this analysis.

Validation of selected mutations by mass spectrometry genotyping. A total of 45 intergenic AA>AC mutations were selected for validation in tumor and germline sample using mass spectrometry genotyping (Sequenom). Mutations were randomly selected across six samples, and sites chosen all had estimated mutation allelic fractions exceeding 30%, thus enabling mutation detection[20,26,70]. Of those assays performed, 25 yielded interpretable data, with others failing owing to lack of PCR amplification or probe hybridization in the tumor and/or germline samples.

Sequencing data processing and quality control. The processing and analysis of exome and whole-genome sequencing data were performed using Broad Institute pipelines[15,16,26,70]. A BAM file aligned to the hg19 human genome build was generated from Illumina sequencing reads for each tumor and normal sample by the Picard pipeline. The Firehose pipeline was used to manage input and output files and submit analyses for execution in GenePattern[71].

Quality control modules in Firehose were used to compare the genotypes derived from Affymetrix arrays and sequencing data to ensure concordance. Genotypes from SNP arrays were also used to monitor for low levels of cross-contamination between samples from different individuals in sequencing data using the ContEst algorithm[72]. One tumor-normal pair (ESO-774) analyzed by whole-genome sequencing was not included in the exome analysis, as the exome sequencing from that case failed quality control metrics.

Mutation calling. The MuTect algorithm was used to identify somatic mutations in targeted exons and whole-genome data[14,16,17,26]. MuTect identifies candidate somatic mutations by Bayesian statistical analysis of bases and their qualities in the tumor and normal BAM files at a given genomic locus. We required a minimum of 14 reads covering a site in the tumor and 8 in the normal sample to declare that a site was adequately covered for mutation calling. We determined the lowest allelic fraction at which somatic mutations could be detected on a per-sample basis, using estimates of cross-contamination from the ContEst pipeline[72]. Small somatic insertions and deletions were detected using the Indelocator algorithm, after local realignment of tumor and normal sequences[14]. All somatic mutations detected by whole-exome sequencing were analyzed for potential false positive calls by performing a comparison to mutation calls from a panel of 2,500 germline DNA samples. Mutations found in 2% of the germline samples or 2% of sequencing reads were removed from analysis. MutSig significant mutations, except for all *TP53* mutants, were reviewed manually in the respective BAM files using the Integrative Genomics Viewer.

Mutation annotation. Somatic point, insertion and deletion mutations were annotated using information from publicly available databases, including the UCSC Genome Browser's UCSC Genes track[73], miRBase release 15 (ref. 74), dbSNP build 132 (ref. 75), UCSC Genome Browser's ORegAnno track[76], UniProt release 2011_03 (ref. 77) and COSMIC v51 (ref. 55).

Mutation significance analysis. For the purpose of discovering recurrently mutated genes, we used the MutSig algorithm, as described[19]. In short, this method builds a background model of mutational processes, which takes into account the genome-wide variability in mutation rates. We achieved this by considering different covariates that have been shown to affect mutation rate: GC content (measured on 100-kb windows), local relative replication time[78,79], open versus closed chromatin status as determined by HiC (fine-scale mapping of the three-dimensional DNA contacts in the nucleus[80]), gene expression[16] and, finally, local gene density measured in a 1-Mb window. For each gene, we defined a set of nearest neighbors according to these covariates and estimated the background mutation rate from noncoding (in flanking sequences and introns) and silent mutations of these neighbors. We then assigned a score based on the ratio between the nonsilent coding mutation rate of the gene and the noncoding and silent mutation rate of the given gene and its neighbors. Furthermore, we performed an independent significance analysis that was restricted to events that had been previously reported in the COSMIC database.

MSI testing. MSI analysis was performed using ten microsatellite markers (D2S123, D5S346, D17S250, BAT25, BAT26, BAT40, D18S55, D18S56, D18S67 and D18S487) as described previously[23].

Copy number calling from whole-exome sequencing. Copy number ratios were calculated as the ratio of tumor read depth to the average read depth observed in normal samples for that region using the CapSeg DNA sequencing–based tool (A.M., S.L.C., B. Hernandez, M. Meyerson, G.G. *et al.*, unpublished data).

Processing of Affymetrix expression arrays. Raw data were processed using the gene chip robust multiarray averaging[81] (RMA) approach to provide normalized expression data for each probe set on the arrays.

Cell lines and culture conditions. NIH/3T3 and 293 cells were obtained from the American Type Culture Collection. All cells were maintained in DMEM supplemented with 10% FBS at 37 °C in 5% CO_2.

***ELMO1* site-directed mutagenesis.** Full-length *ELMO1* cDNA was obtained from Open Biosystems–Thermo Scientific and cloned into the EcoRI site of pBabe(puro). Mutants were generated by site-directed mutagenesis using the QuikChange II Site-Directed Mutagenesis kit (Agilent Technologies) according to the manufacturer's instructions. All mutations were verified by sequencing.

***ELMO1* retrovirus production and cell infection.** pBabe(puro) vector encoding wild-type ELMO1, ELMO1 mutants or GFP (1 μg) was cotransfected with 1 μg of pCL-Eco into 293 cells with Fugene HD (Roche) overnight. Growth medium was replaced with new full-serum medium after 24 h. After an additional 24 h, retroviral supernatants were harvested, and fresh medium was added. Retroviral supernatants were filtered and incubated with target NIH/3T3 cells in the presence of 5 μg/ml polybrene (hexadimethrine bromide). This procedure was repeated again after 24 h. Stably infected cells were selected for under puromycin (1 μg/ml) pressure for 2 weeks. Expression of the ELMO1 constructs was confirmed by protein blotting with an antibody to ELMO1 (ab2239, Abcam).

Matrigel invasion assays. Transwell chambers (BD Biosciences) coated with Growth Factor–Reduced Matrigel were activated in serum-free medium at 37 °C for 2 h. NIH/3T3 cells (1×10^4) were plated in Matrigel invasion chambers with full-serum medium in the lower chamber only. After 24 h, non-invading cells in the top chamber were removed by cotton swab, and invading cells were fixed and stained using Diff-Quik staining solutions according to the manufacturer's instructions (VWR International). The number of invading cells from each of four fields was counted at 20× magnification.

68. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
69. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
70. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
71. Reich, M. *et al.* GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).
72. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601–2602 (2011).
73. Fujita, P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876–D882 (2011).
74. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, D152–D157 (2011).
75. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
76. Griffith, O.L. *et al.* ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* **36**, D107–D113 (2008).
77. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39**, D214–D219 (2011).
78. Chen, C.L. *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.* **20**, 447–457 (2010).
79. Stamatoyannopoulos, J.A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
80. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
81. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).