nature genetics

Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma

Jia-Jie Hao[1,10], De-Chen Lin[2,3,10], Huy Q Dinh[4,10], Anand Mayakonda[5,10], Yan-Yi Jiang[5,10], Chen Chang[1], Ye Jiang[1], Chen-Chen Lu[1], Zhi-Zhou Shi[6], Xin Xu[1], Yu Zhang[1], Yan Cai[1], Jin-Wu Wang[7], Qi-Min Zhan[1], Wen-Qiang Wei[8], Benjamin P Berman[4], Ming-Rong Wang[1] & H Phillip Koeffler[2,5,9]

Esophageal squamous cell carcinoma (ESCC) is among the most common malignancies, but little is known about its spatial intratumoral heterogeneity (ITH) and temporal clonal evolutionary processes. To address this, we performed multiregion whole-exome sequencing on 51 tumor regions from 13 ESCC cases and multiregion global methylation profiling for 3 of these 13 cases. We found an average of 35.8% heterogeneous somatic mutations with strong evidence of ITH. Half of the driver mutations located on the branches of tumor phylogenetic trees targeted oncogenes, including *PIK3CA*, *NFE2L2* and *MTOR*, among others. By contrast, the majority of truncal and clonal driver mutations occurred in tumor-suppressor genes, including *TP53*, *KMT2D* and *ZNF750*, among others. Interestingly, phyloepigenetic trees robustly recapitulated the topological structures of the phylogenetic trees, indicating a possible relationship between genetic and epigenetic alterations. Our integrated investigations of spatial ITH and clonal evolution provide an important molecular foundation for enhanced understanding of tumorigenesis and progression in ESCC.

Esophageal carcinoma is among the most common human cancers, causing over 400,000 deaths worldwide annually[1,2]. The areas with highest risk are located in eastern Asia, as well as eastern and southern Africa, and the most prevalent type is ESCC[1,2]. The 5-year survival rates for patients with ESCC undergoing surgery are below 30% because a large proportion of tumors are unresectable or have already metastasized before diagnosis[3].

Recently, several large-scale genomic studies have characterized ESCC genomes as having hundreds of somatic mutations and copy number alterations (CNAs) and have identified significantly mutated genes, including *TP53*, *PIK3CA* and *ZNF750*, among others[4–9]. The APOBEC signature is a predominant mutational spectrum and contributes to the mutagenic processes of ESCCs[6,8]. However, the genomic alterations identified in all of these studies were obtained using only single samples representing individual cases, and little is known about the spatial ITH and temporal clonal evolutionary processes of the mutational spectrum in ESCC. Moreover, although alterations in DNA methylation have been observed in ESCC, the degree of ITH for these epigenetic changes is still unknown, and whether such heterogeneity correlates with genetic architecture remains unexplored.

Precise understanding of both the genomic and epigenomic architectures of primary ESCC tumors is crucial for the development of personalized patient treatment and molecular-based biomarkers[10]. Furthermore, an integrated investigation of the genomic and epigenomic evolutionary trajectories of ESCC may also provide new insights into the relationship between the genome and epigenome. In the present study, we address these critical issues through integrative molecular approaches, including multiregion whole-exome sequencing (M-WES) and global methylation profiling, as well as phylogenetic and phyloepigenetic tree construction.

RESULTS

Spatial ITH of ESCC

M-WES was performed on genomic DNA from 13 patients with primary ESCC; the clinicopathological parameters of these patients are listed in **Supplementary Table 1**. In total, 51 tumor regions and 13 matched morphologically normal esophageal tissue samples (4 tumor regions and 1 matched normal tissue sample per case, with the exception of ESCC04, for which samples were obtained from only 3 tumor regions) were sequenced, with a mean coverage depth of 150×. A total of 1,610 non-silent somatic mutations (affecting 1,427 genes) and

[1]State Key Laboratory of Molecular Oncology, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [2]Division of Hematology/Oncology, Cedars-Sinai Medical Center, UCLA School of Medicine, Los Angeles, California, USA. [3]Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Medical Research Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China. [4]Center for Bioinformatics and Functional Genomics, Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, California, USA. [5]Cancer Science Institute of Singapore, National University of Singapore, Singapore. [6]Faculty of Medicine, Kunming University of Science and Technology, Kunming, China. [7]Department of Pathology, Linzhou Cancer Hospital, Henan, China. [8]Department of Cancer Epidemiology, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [9]National University Cancer Institute, National University Hospital Singapore, Singapore. [10]These authors contributed equally to this work. Correspondence should be addressed to M.-R.W. (wangmr2015@126.com), B.P.B. (benjamin.berman@csmc.edu), D.-C.L. (dchlin11@gmail.com) or W.-Q.W. (weiwq2006@126.com).

Received 3 March; accepted 31 August; published online 17 October 2016; doi:10.1038/ng.3683

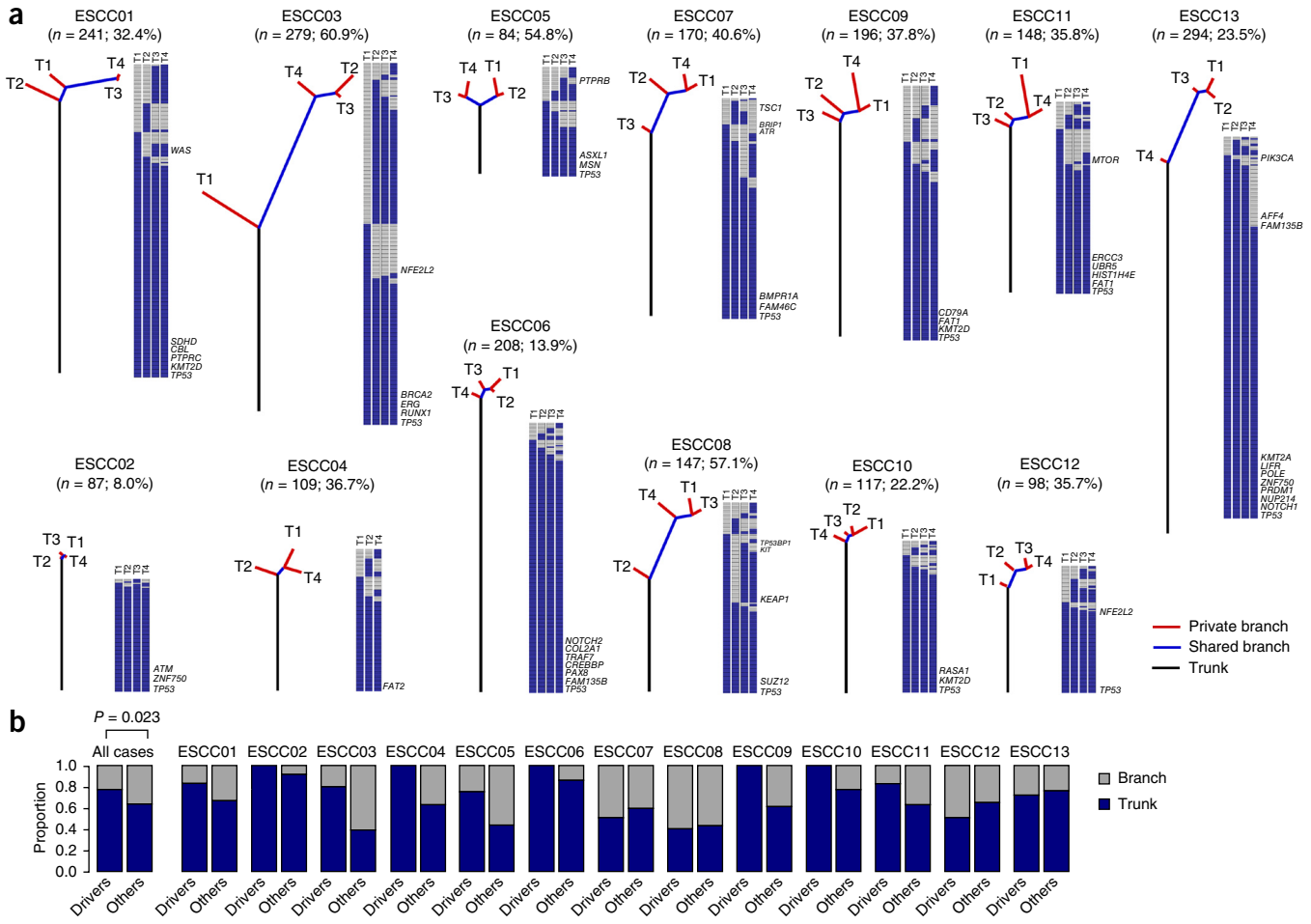

Figure 1 ITH of somatic mutations in 13 ESCCs generated by M-WES. (**a**) Phylogenetic trees were constructed from all somatic mutations by the Wagner parsimony method using PHYLIP (Online Methods). Lengths of trunks and branches are proportional to the numbers of mutations acquired. Heat maps show the presence (blue) or absence (gray) of a somatic mutation in each tumor region (T). Each gene is arranged in a row, and cancer-related genes with putative driver mutations are indicated. The total number of mutations (*n*) and proportion of branched mutations in each case are provided above each tree. (**b**) Bar plots show the proportions of putative driver mutations versus other mutations on the trunks and branches. Statistical differences of truncal and branched proportions, between driver and other mutations across all cases, were analyzed using a χ^2 test, and a significant *P* value is shown.

568 silent mutations were identified, with a validation rate of 90% (**Supplementary Tables 2** and **3**).

To explore ITH and the genomic evolution of ESCC, phylogenetic trees were constructed on the basis of somatic mutations (both silent and non-silent) identified in each tumor region. The trunk, 'shared' branches and 'private' branches of each tree represent mutations in all tumor regions, in some but not all tumor regions, and in only one tumor region, respectively. The phylogenetic trees varied extensively among the different cases (**Fig. 1a** and **Supplementary Fig. 1**), and all 13 of the ESCCs showed evidence of spatial ITH, with an average of 35.8% (780/2,178; range, 8.0–60.9%) of somatic variants having spatial heterogeneity.

Characterization of the relative timing of mutations affecting driver genes with possible biological relevance is essential for identifying the evolutionary processes of the cancer genome, as well as for further improving precision medicine strategies. To address this, we identified potential driver mutations according to recent large-scale ESCC sequencing data[4–8], the Catalog of Somatic Mutations in Cancer (COSMIC) cancer gene census[11] and pan-cancer analysis[12]; these mutations were then traced within the phylogenetic trees (Online Methods). Overall, driver mutations were significantly more enriched in trunks than passenger mutations were (77.8% versus 63.8%; $P = 0.023$; **Fig. 1b**). This indicates that driver mutations are relatively early events during the evolutionary process of the tumors, in accordance with previous findings in other tumor types[13]. We next separated putative driver mutations into those occurring in oncogenes or tumor-suppressor genes (TSGs). Notably, half of the driver mutations (50.0%) that mapped to branches were in oncogenes, including *PIK3CA*, *KIT*, *NFE2L2*, *MTOR* and *FAM135B*. In comparison, only 22.4% of the driver mutations located on trunks affected oncogenes, and the rest were in TSGs. For example, *TP53* mutations were present in 12 of the 13 cases and were truncal in all of the mutated cases, in agreement with recent reports[14,15]. It is worthwhile to note that potentially actionable mutations, such as those targeting *PIK3CA* and *MTOR*, tended to be oncogenic branch events. These findings highlight the additional caution needed when considering the inhibition of these mutants in ESCC, given previous studies showing that the suppression of subclonal drivers leads to growth acceleration for non-mutated subpopulations[16].

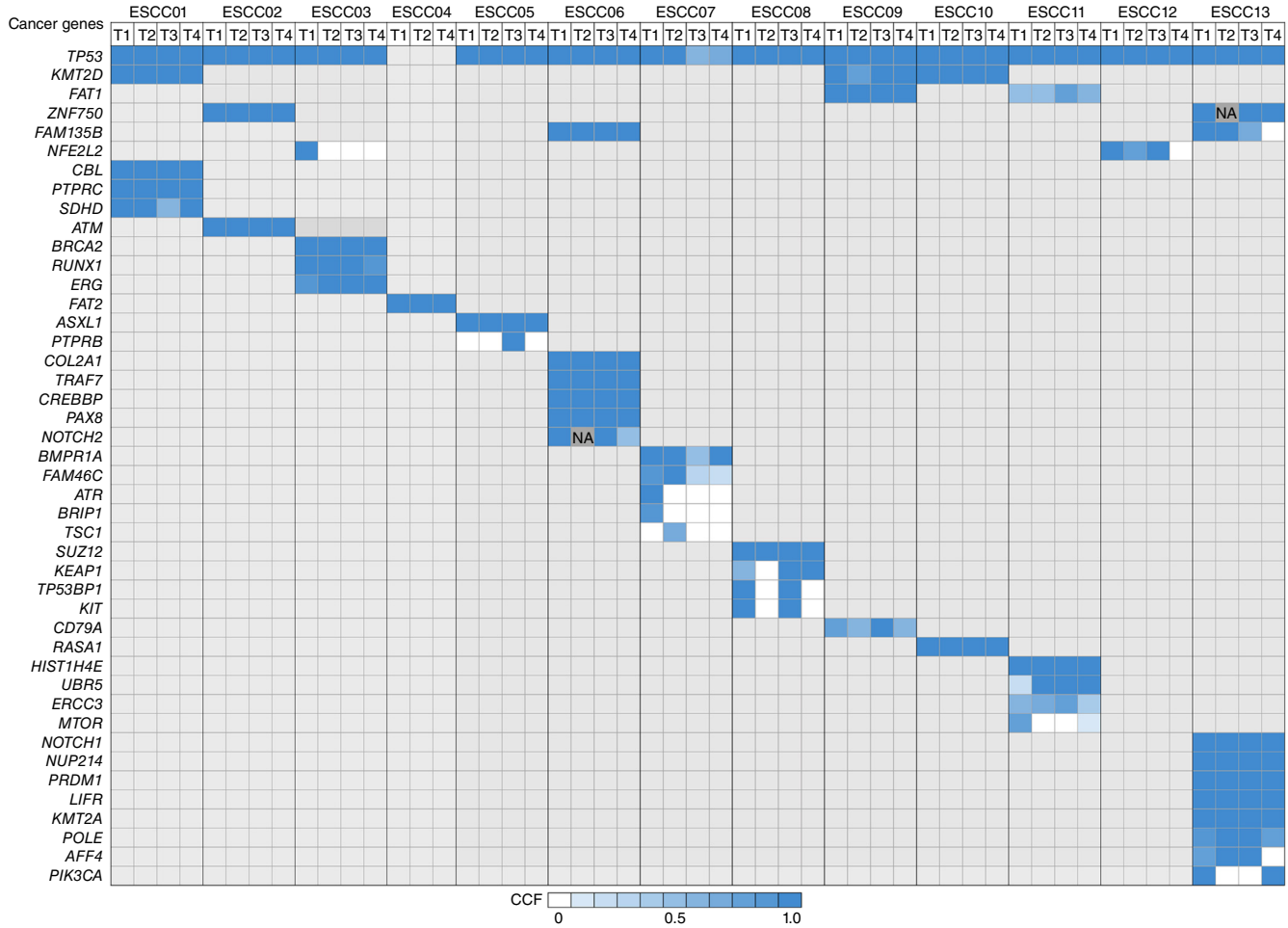

Figure 2 Clonal status of putative driver mutations in ESCC tumors. A heat map displays the CCF of driver mutations in each region of the ESCC tumors. Genomic regions with no segmentation data available are shown as NA.

Clonal status of putative driver mutations

We next investigated the clonal status of somatic mutations within individual regions. The cancer cell fraction (CCF) for each tumor region was calculated as described previously through integrative analysis of local copy number, variant allele frequency (VAF) and tumor cell purity[16,17]. Several driver mutations were subclonal and possibly occurred as late events in ESCC, including mutations in *MTOR*, *KEAP1*, *PTPRB* and *FAM135B*. In contrast, cancer-related genes on the trunks, such as *TP53*, *NOTCH1*, *CREBBP*, *KMT2D* and *ZNF750*, were predominantly mutated in a fully clonal manner (**Fig. 2**), further verifying our earlier phylogenetic tree analysis showing that these mutations were possibly early lesions during ESCC development. Of note, a number of driver variants detected as clonal within some individual tumor regions were absent in others from the same individual, producing an 'illusion' of clonal dominance. For example, a *PIK3CA* hotspot mutation (p.Met1043Ile) was undetectable in tumor regions T2 and T3 in case ESCC13 but was clonally dominant in the other two regions. Likewise, a hotspot mutation in the *KIT* gene (p.Glu601Lys) was present in 100% of tumor cells from regions T1 and T3 in case ESCC08, yet was absent in the rest of the tumor regions. Such clonal dominance was also observed for mutations in *NFE2L2* in case ESCC12. Our results suggest that driver mutations can have mixed and complex intratumoral clonal status in ESCC and that the current single-sampling approach may misinterpret these critical genomic lesions because of the illusion of clonal dominance. We further investigated all non-silent variants in genes and related pathways that could potentially be targeted therapeutically. Mutations affecting components of the phosphoinositide 3-kinase (PI3K)–mammalian target of rapamycin (mTOR) pathway (*KIT*, *AURKA* and *CCND2*) were always late events (branched/subclonal) (**Supplementary Fig. 2**). By contrast, variants in *ERBB4*, *FGFR2*, *BRCA2*, *ATM* and *TP53* were mutated as early events (truncal/clonal), suggesting their potential as candidate actionable targets for ESCC.

ITH of copy number alterations

We next analyzed ITH at the copy number level (**Supplementary Table 4**). First, recurrent CNAs that involve important cancer-related genes in ESCC were identified on the basis of our previous results[6], and we confirmed that the present cohort harbored these recurrent CNAs at similar frequencies (**Supplementary Fig. 3**). Although CNAs were generally more similar within cases than between different cases, we found extensive ITH for CNAs, with 90% (9/10) of all recurrent CNAs being spatially heterogeneous. For example, in ESCC08, amplification of chromosome 7p11.2 (encompassing *EGFR*) was observed in regions T1 and T4, but not in regions T2 and T3. Similarly, deletions of chromosome 9p21.3 (harboring *CDKN2A* and *CDKN2B*) were ubiquitous in some cases but also occurred as heterogeneous aberrations in other samples. The only driver CNA

Table 1 Prevalence of non-silent mutations in ESCC (within patient versus within region)

Cancer-related gene	Prevalence (number of patients with mutation) in previous studies[a]	Within-region prevalence (number of regions with mutation) $n = 51$ regions	Within-patient prevalence (number of patients with mutation) $n = 13$ cases	Within patient/within region[b]
TP53	78.9% (430)	94.1% (48)	92.3% (12)	0.98
KMT2D	13.8% (63)	23.5% (12)	23.1% (3)	0.98
NOTCH1	12.8% (70)	21.6% (11)	23.1% (3)	1.07
FAT1	11.2% (51)	15.7% (8)	15.4% (2)	0.98
ZNF750	5.7% (26)	15.7% (8)	15.4% (2)	0.98
FAM135B	6.4% (29)	13.7% (7)	15.4% (2)	1.12
NFE2L2	5.7% (26)	7.8% (4)	15.4% (2)	1.97
PTPRB	2.9% (13)	7.8% (4)	15.4% (2)	1.97
ATM	1.8% (8)	7.8% (4)	7.7% (1)	0.98
BRCA2	3.1% (14)	7.8% (4)	7.7% (1)	0.98
CREBBP	4.2% (19)	7.8% (4)	7.7% (1)	0.98
KMT2A	1.1% (5)	7.8% (4)	7.7% (1)	0.98
NOTCH2	3.3% (18)	7.8% (4)	7.7% (1)	0.98
FAT2	6.4% (29)	5.9% (3)	7.7% (1)	1.31
KEAP1	1.8% (8)	5.9% (3)	7.7% (1)	1.31
MTOR	1.1% (5)	3.9% (2)	7.7% (1)	1.96
TP53BP1	0.9% (4)	3.9% (2)	7.7% (1)	1.96
KIT	0.7% (3)	3.9% (2)	7.7% (1)	1.96
PIK3CA	9.0% (41)	3.9% (2)	7.7% (1)	1.96
ATR	1.1% (5)	2.0% (1)	7.7% (1)	3.92
BRIP1	0.9% (4)	2.0% (1)	7.7% (1)	3.92
TSC1	1.1% (5)	2.0% (1)	7.7% (1)	3.92

[a]Summary of published data from Agrawal *et al.*[4], Song *et al.*[5], Lin *et al.*[6], Gao *et al.*[7] and Zhang *et al.*[8]. The total number of cases is 545 for *TP53*, *NOTCH1* and *NOTCH2* mutations and is 456 for the rest of gene mutations. [b]Fold change when the prevalence was analyzed using individual cases instead of individual tumor regions.

Table 2 Prevalence of subclonal mutations in ESCC

Case	Within-region prevalence (%) T1	T2	T3	T4	Within-patient prevalence (%)
ESCC01	10.1	16.1	26.7	15.4	40.0
ESCC02	14.7	8.2	10.4	14.9	20.5
ESCC03	13.6	7.2	8.4	24.1	33.2
ESCC04	10.7	5.8	NA	1.2	13.3
ESCC05	27.3	21.4	3.6	33.3	48.8
ESCC06	6.9	28.3	5.4	6.1	33.3
ESCC07	6.1	21.1	92.4	61.1	86.1
ESCC08	11.6	12.7	15.4	16.2	31.7
ESCC09	30.4	41.3	5.7	20.0	56.5
ESCC10	21.2	2.0	3.1	6.1	27.0
ESCC11	42.3	35.5	36.0	41.7	66.4
ESCC12	1.4	38.6	6.1	46.3	62.5
ESCC13	29.5	3.0	14.4	29.5	50.0

Within-patient prevalence was derived by dividing the number of subclonal mutations by the number of total mutations in each patient.

found to be consistently ubiquitous was copy number gain of 11q13, which encompasses a number of oncogenes, including *CCND1*, *ANO1* (refs. 18–20) and *CTTN*[21,22], highlighting the importance of this aberration as a founder genomic lesion in the development of ESCC. These results suggest that, similar to somatic mutations, CNAs also show notable spatial ITH, concordant with observations in several other types of cancer[23–25].

The within-patient mutational rate (mean = 168 mutations per case) was higher than the within-region mutational rate (mean = 139 mutations per region; **Supplementary Table 5**), highlighting the improved resolution of our multiple-biopsy approach for genomic interrogation. In particular, in the case of cancer-related genes on branches, the current M-WES approach markedly increased the sensitivity of the detection rate (**Table 1**). For example, *ATR* and *TSC1* mutations, which were detected in only 2% of tumor regions (in agreement with previous results), occurred in 7.7% of cases. In addition, the proportion of subclonal mutations detected in each tumor region was much lower than that in each case (**Table 2** and **Supplementary Fig. 4**). These results again signify that the analysis of sequencing data obtained from a single biopsy will likely underestimate the prevalence of mutations, especially for those acquired late in the mutational process[24].

Temporal dissection of mutational spectra and signatures

To determine the temporal dynamics of the mutagenic processes in ESCC, the mutational spectra of mutations on both trunks and branches were analyzed using deconstructSigs[26], which identifies the linear combination of predefined signatures that most accurately reconstructs the mutational profile of a single tumor sample. The overall mutational spectra were similar for trunk and branch mutations, with very strong enrichment of signature 1 substitutions (associated with age) and more subtle but enriched representation of APOBEC-associated signatures 2 and 13 substitutions (C>G and C>T substitutions in a TCW context, where W = A, G, C or T) (**Fig. 3a**). We next calculated the contributions of individual mutational signatures to each tumor (**Fig. 3b**) and identified several signatures within the tumors tested, including signature 1 (age), signatures 2 and 13 (APOBEC), and signatures 6 and 15 (DNA mismatch repair), in agreement with previous results in esophageal squamous and other squamous-type cancers[6,8,27]. Interestingly, we noticed that a number of tumors displayed a prominent decrease in the relative contribution of signature 1 in branch as compared to trunk mutations, although this decrease did not reach statistical significance owing to the relatively small number of tumors analyzed. In some of these cases, we also observed an increase in the contribution of signatures associated with DNA damage (including signatures 3 and 15) among the branch mutations (such as in ESCC10 and ESCC12; **Fig. 3c**,**d** and **Supplementary Fig. 5**). Interpreting these temporal differences in mutational spectra within the same tumor will require further investigations, but the data indicate that various mutational processes might have important roles in subclonal diversification during the progression of ESCC.

ITH of DNA methylation in ESCC

As with other cancers, epigenetic abnormalities have been associated with the development and pathogenesis of ESCC[28–30]. To decipher ESCC ITH at the epigenetic level and its potential relationship with subclonal gene mutations, the global methylation levels of 14 M-WES–profiled tumor and normal tissue pairs from three ESCC cases (ESCC01, ESCC03 and ESCC05) were obtained using the Illumina HumanMethylation450 (HM450) BeadChip. We first identified CpG probes that showed significant differences in methylation between tumor regions and normal tissue samples from the same case (except for ESCC01, for which a matched normal tissue sample was not available) and then divided these differentially methylated probes into those with shared changes (consistent within all tumor regions from the same case) and those with private changes (present in one or more of the regions, but not all). We used the probes with private changes to infer tumor evolution and constructed phyloepigenetic trees for each case on the basis of the Euclidean pairwise distances between methylation profiles[31,32] (Online Methods). Topological similarities were tested between the phyloepigenetic and phylogenetic trees for all three cases by determining Robinson–Foulds (RF)

distance relative to unrooted trees[33] (**Fig. 4a**). Notably, in accordance with a recent report on glioma[32], the RF distances (zero for all three cases) suggest high concordance between the genetic and epigenetic tree topologies for all three cases (Online Methods). As the distinction of private versus shared methylation changes is dependent on the probe selection cutoff used, we further tested four different cutoffs and noted that the phyloepigenetic trees were robust to cutoff selection and showed highly similar topological structures for all cutoff values (**Supplementary Fig. 6**). Moreover, to alleviate confounding effects resulting from contamination with non-tumor DNA, two different methods were applied to account for and mitigate the potential influence of immune cells (the major non-cancer cell

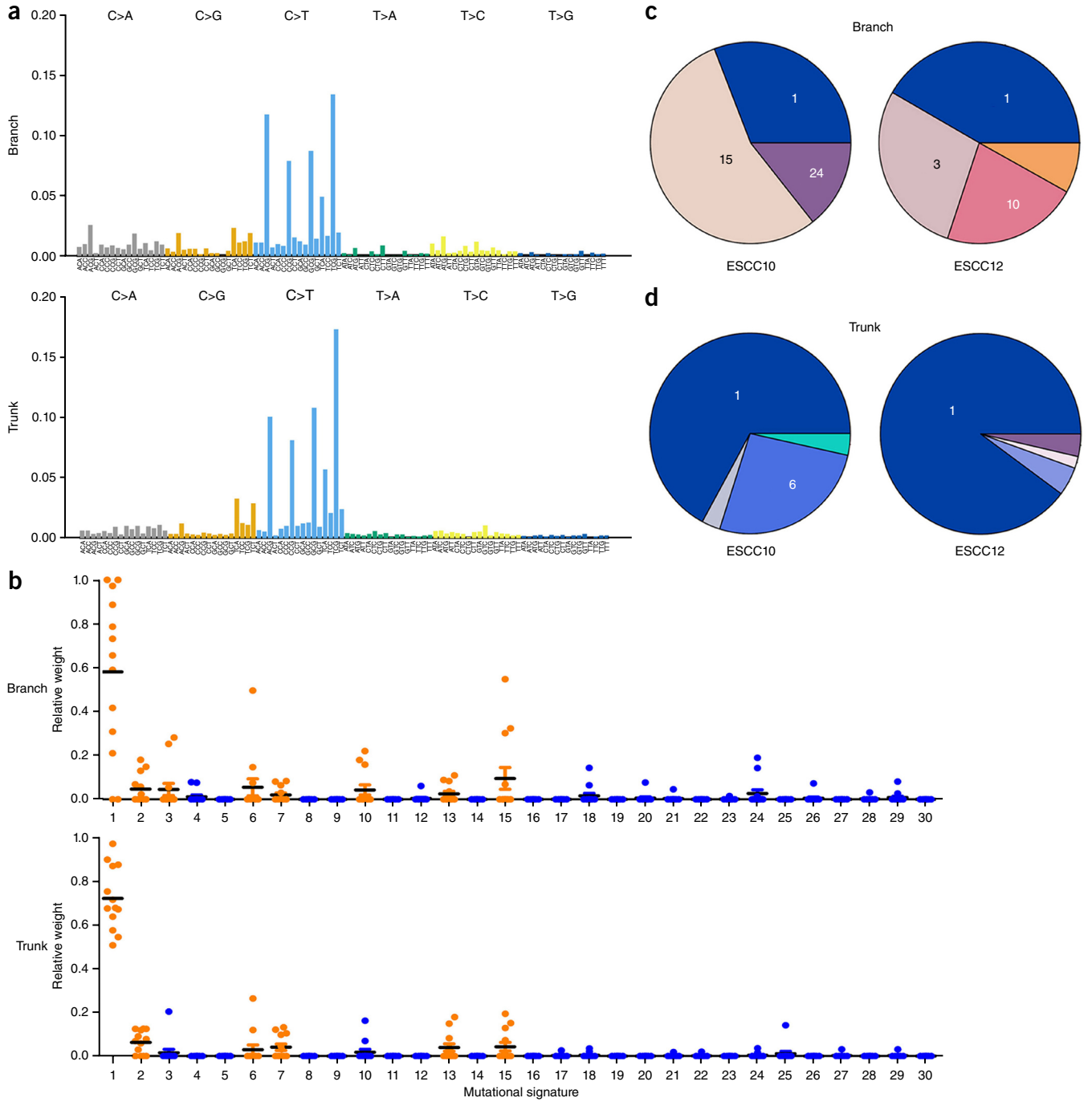

Figure 3 Temporal dissection of mutational signatures in ESCC tumors. (**a**) The 96 trinucleotide mutational spectra of truncal (bottom) and branched (top) mutations across all regions was inferred by deconstructSigs. (**b**) Dot plots display the contributions of individual mutational signatures to individual cases, with each dot representing one case. Signatures 1–30 were based on the Wellcome Trust Sanger Institute COSMIC Mutational Signature Framework. Inferred signatures included signature 1 (associated with age), signatures 2 and 13 (associated with APOBEC), signatures 6 and 15 (associated with DNA mismatch repair), signature 3 (associated with DNA double-strand break repair) and signature 7 (associated with UV exposure in squamous cancer). Bars represent mean values. (**c**,**d**) Pie charts display the truncal and branch mutational signatures in cases ESCC10 and ESCC12; only signatures with contributions over 10% are indicated.

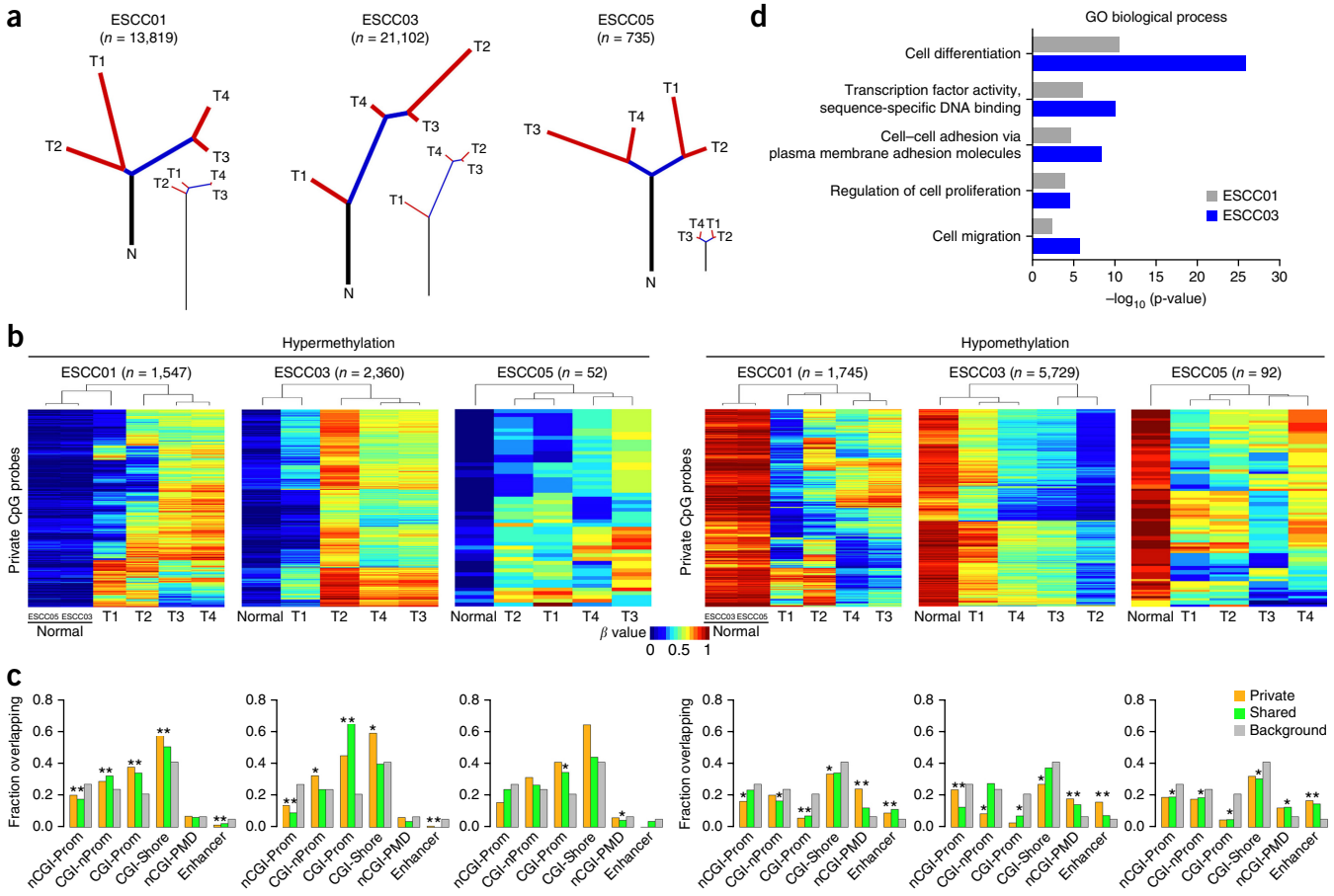

Figure 4 Epigenetic ITH in ESCC. (**a**) Phyloepigenetic trees of three ESCC cases. Lengths of trunks and branches were inferred using a phylogenetic approach, based on Euclidean distances between different tumor regions using private probes (Online Methods). The total number of probes (*n*) is provided above each tree. For comparison, the phylogenetic trees from **Figure 1** are reproduced below each phyloepigenetic tree. (**b**) Heat maps show the β values of private probes for each case, separated into hyper- and hypomethylation. (**c**) Overlap between each probe set from **b** and a variety of functional genomic contexts: non-CGI promoters (nCGI-Prom), non-promoter CGI (CGI-nProm), CGI promoters (CGI-Prom), CGI shore (CGI-Shore), partially methylated domain excluding CGIs (nCGI-PMD) and enhancers. Overlapping frequencies of private probes from **b** are shown in yellow, shared probes (**Supplementary Fig. 9**) are shown in green and gray shows the frequency for the entire set of probes on the array. The hypergeometric test ($*P < 1 \times 10^{-5}$) was used to compare the frequency of each private and shared probe set category to that of array background (Online Methods). (**d**) Enriched GO Biological Processes for the genes associated with privately hypermethylated promoters in ESCC01 and ESCC03 (case ESCC05 was excluded because of the lack of sufficient privately hypermethylated promoters).

component in these samples; **Supplementary Fig. 7**); again, similar results were observed for the trees when using uncorrected methylation values or the values obtained with either correction method (Online Methods and **Supplementary Fig. 8**). These findings suggest a possible relationship between genomic and epigenomic alterations during the clonal evolution of ESCC cells and are indicative of the presence of multiple epigenetically distinct, subclonal cell populations, as recently observed in prostate cancer[31], glioma[32] and hepatocellular carcinoma (D.-C.L., A.M., H.Q.D., P. Huang and L. Lin *et al.*, unpublished data).

We observed that a number of TSGs, including *EPHA7* (refs. 34,35), *PCDH10* (refs. 36,37) and *DOK1* (refs. 38,39), among others, were hypermethylated at their promoters within some but not all tumor regions from the same case, indicating that their expression might be differentially suppressed in different tumor regions. Notably, some TSGs were mutated and acquired promoter hypermethylation, such as *ASXL1* and *EPHA7*. Interestingly, *ASXL1* was subject to both truncal/clonal mutation and shared hypermethylation at its promoter, suggesting that this gene was disrupted early during both the genomic and epigenomic evolutionary processes.

To explore the potential biological relevance of ITH for DNA methylation in ESCC, we next sought to determine whether the differentially methylated CpG loci in each case were enriched in particular functional genomic categories. We first divided the CpG probes into those where methylation was higher in the tumor than in adjacent normal tissues (hypermethylated) and those where methylation was lower (hypomethylated). Shared probes were selected for their relatively consistent changes in different tumor regions (**Supplementary Fig. 9**), whereas the remaining (private) probes exhibited prominent differences among the tumor regions (**Fig. 4b**) and reflect the extensive ITH seen in the phyloepigenetic trees. We next compared shared and private probes by assigning them to various relevant functional genomic categories, including CpG islands (CGIs), CGI shores, promoters and enhancers, among others, and compared the frequencies of the probes in each category to the background frequencies, determined from all probes on the array (**Fig. 4c**). As expected, shared CpG sites showed several methylation patterns commonly seen across cancer types[40,41], including strong enrichment of hypermethylated probes in CGI promoter regions and depletion of these probes in both long-range partially methylated domains (PMDs) and enhancer

regions (after removing CGIs). Shared hypomethylated probes showed an inverse distribution: they were markedly depleted in CGI promoters, whereas they were enriched in PMDs as well as enhancer regions (**Fig. 4c**). Strikingly, the distribution of private CpG sites for the most part resembled that of their shared counterparts (**Fig. 4c**). In light of the known contribution of tumor-specific methylation to cancer biology[42,43], our results suggest that intratumoral methylation heterogeneity might have a role in the subclonal diversification of ESCC tumors. In support of this view, Gene Ontology (GO) analysis of the genes with private hypermethylated probes in their promoters showed that they were significantly enriched in cancer-related processes, including cell proliferation, differentiation, migration, adhesion and transcriptional regulation (**Fig. 4d**). In addition, we noticed that private hypermethylated probes were even more enriched in CGI shores than shared hypermethylated probes (**Fig. 4c**). Given previous observations that (i) cancer-specific differentially methylated regions occur more frequently within CGI shores than within CGIs[44,45] and (ii) CGI shore methylation correlates with the expression of associated genes[44], our observations further suggest the potential involvement of heterogeneity of DNA methylation in the evolutionary biology of ESCC cells.

DISCUSSION

ESCC is one of the most common malignancies, with relatively low overall 5-year survival rates. The main cause leading to unfavorable prognosis of patients with ESCC is the lack of effective therapies. Currently, none of the targeted therapies has been established for clinical management of ESCC[46]. Hundreds of genomic alterations, including somatic mutations and CNAs, have recently been identified in ESCC[4–9], but these data have not been translated into clinical applications. In addition, the genomic and epigenomic ITH and clonal evolution of ESCC tumors have not yet been characterized. In light of the evidence that ITH is the major cause of drug resistance and treatment failure[47], deciphering the genomic diversity and clonal evolution of ESCC tumors will provide both a theoretical and translational basis for identifying new targets and designing personalized medicine strategies.

In the present study, the genomic ITH of 13 ESCC cases, as well as the epigenetic ITH of 3 of these individuals, were investigated through a variety of molecular approaches, and concordant tumor evolutionary trajectories were found as inferred from both DNA mutations and methylation. A very recent study of two ESCC cases reported that the ITH rate for somatic mutations was approximately 90% (ref. 48), whereas the rates in our study were much lower, with an average of 35.8%. The discrepancy may well be due to differences in sequencing depth between the two studies (50× versus 150×). Although the true extent of ITH is difficult to define, high sequencing coverage in our study offers improved resolution to decipher the spatial heterogeneity and clonal evolution of ESCC.

Although phylogeny analysis based on M-WES is not able to completely resolve the true temporal ordering of all somatic variants, we calculated that an average of 93.5% (range of 87.8 to 97.7%) of somatic mutations were compatible with the present phylogenic trees (**Supplementary Fig. 1**). For example, in case ESCC13, 282 of 294 variants (95.9%) were compatible with the evolution model based on the topological structure of the phylogenetic tree, and only 12 mutations, including ones in *PIK3CA*, were incompatible with the phylogenetic tree (**Supplementary Table 6**). Therefore, the phylogeny method correctly resolves the temporal order of the vast majority of somatic mutations. Moreover, the evolutionary models inferred from the M-WES–based phylogeny are strongly supported by our DNA methylation phylogeny in all three cases (**Fig. 4a**). Hence, this reconstruction of the phylogenetic topologies, from a completely independent molecular event, strongly reinforces the validity of these evolutionary models.

Resolving the clonal status of driver mutations will help to distinguish early from late events, and targeting clonally dominant driver mutations (early events) conceivably represents an optimized therapeutic strategy[10,49]. In this study, despite driver mutations having a tendency to be truncal/clonal in comparison with passenger mutations, approximate 40% of driver mutations were branched or subclonal. This observation suggests that these driver mutations were relatively late events during tumor evolution and contributed to the emergence of distinct subclonal expansions after the founding clones were established. Notable examples included *KIT*, components of the PI3K–mTOR pathway (*PIK3CA* and *MTOR*) and components of the NFE2L2 pathway (*NFE2L2* and *KEAP1*). These examples, most of which are oncogenes, were frequently mutated as late events in ESCC. Furthermore, evidence of 'parallel evolution' was noted in some cases. For example, ESCC13 contained branch *PIK3CA* mutations derived in two spatially separated tumor regions, both harboring the p.Met1043Ile variant, which corresponds to a hotspot mutation. Similar parallel evolution was also observed in *NFE2L2* mutations in ESCC12. Interestingly, *PIK3CA*, *KIT* and *NFE2L2* mutations were fully clonal in some tumor regions but were completely absent in others, giving an illusion of clonal dominance. In addition, the number of within-patient mutations was higher than the number of within-region mutations. These results strongly suggest that the prevalence of these driver events and the rate of subclonality overall are likely underestimated when using a single biopsy to represent an individual patient.

Although alterations in DNA methylation in ESCC have been profiled using single-sampling approaches, their intratumoral diversity and the relationship to genetic lesions remain unknown. In the present study, we found a number of TSGs with private hypermethylation at the promoters, some of which have been associated with either tumorigenesis or progression of other cancer types, such as *EPHA7* (refs. 34,50), *ABCB4* (ref. 51), *PCDH10* (refs. 52,53) and *DOK1* (refs. 38,39). This finding suggests that these genes might be differentially inactivated in different tumor regions from the same individuals. We found profound epigenetic ITH in ESCC through global methylation analysis. Notably, subclonal evolutions inferred from DNA methylation closely recapitulated phylogenetic trees, indicating a possible relationship between genetic and epigenetic alterations in ESCC. Therefore, integrative analysis of both phylogenetic and phyloepigenetic trees may generate an enhanced understanding of clonal architecture, and identify the basis for subclonal epigenetic driver events. These features of epigenetic and genetic ITH shown by our study may have important implications in ESCC biology.

URLs. BWA-MEM, http://arxiv.org/abs/1303.3997v2; fpFilter Perl script, https://github.com/ckandoth/variant-filter; Bam-readcount, https://github.com/genome/bam-readcount; PHYLIP, http://evolution.genetics.washington.edu/phylip.html.

METHODS

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

Accession codes. Digital sequencing and HM450 BeadChip files have been deposited in the Sequence Read Archive (SRA) under SRP072112 and the Gene Expression Omnibus (GEO) under GSE79366, respectively.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank H. Shen and D. Weisenberger as well as A.D. Jeyasekharan for their kind help on analysis and discussion. This work was funded by the Singapore Ministry of Health's National Medical Research Council (NMRC) through its Singapore Translational Research (STaR) Investigator Award to H.P.K., an NMRC Individual Research Grant (NMRC/1311/2011) and the NMRC Centre Grant awarded to the National University Cancer Institute of Singapore, the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiatives to H.P.K. D.-C.L. was supported by the American Society of Hematology Fellow Scholar Award, the National Natural Science Foundation of China (81672786) and National Center for Advancing Translational Sciences UCLA CTSI Grant UL1TR000124. M.-R.W. was supported by the National Natural Science Foundation of China (81330052, 81520108023 and 81321091). Y.Z. was supported by the Beijing Natural Science Foundation (7151008). This study was partially supported by a generous donation from the Melamed family and NIH/NCI grant 1U01CA184826 as well as institutional support from the Samuel Oschin Comprehensive Cancer Institute to B.P.B. and H.Q.D.

AUTHOR CONTRIBUTIONS

M.-R.W., D.-C.L., B.P.B. and H.P.K. conceived and designed the experiments. J.-J.H., D.-C.L., H.Q.D., W.-Q.W., B.P.B., M.-R.W. and H.P.K. wrote the manuscript. J.-J.H., D.-C.L., Y.J., C.C., C.-C.L., X.X. and Y.C. performed the experiments. J.-J.H., H.Q.D., A.M., B.P.B. and Z.-Z.S. performed statistical analysis. J.-J.H., D.-C.L., H.Q.D., Y.-Y.J., B.P.B. and H.P.K. analyzed the data. X.X. contributed reagents. W.-Q.W. contributed materials. J.-W.W. and J.-J.H. read slides with hematoxylin and eosin staining. D.-C.L., Y.Z., Q.-M.Z. and H.P.K. jointly supervised research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Torre, L.A. *et al.* Global cancer statistics, 2012. *CA Cancer J. Clin.* **65**, 87–108 (2015).
2. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
3. Enzinger, P.C. & Mayer, R.J. Esophageal cancer. *N. Engl. J. Med.* **349**, 2241–2252 (2003).
4. Agrawal, N. *et al.* Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discov.* **2**, 899–905 (2012).
5. Song, Y. *et al.* Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* **509**, 91–95 (2014).
6. Lin, D.C. *et al.* Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat. Genet.* **46**, 467–473 (2014).
7. Gao, Y.B. *et al.* Genetic landscape of esophageal squamous cell carcinoma. *Nat. Genet.* **46**, 1097–1102 (2014).
8. Zhang, L. *et al.* Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. *Am. J. Hum. Genet.* **96**, 597–611 (2015).
9. Cheng, C. *et al.* Whole-genome sequencing reveals diverse models of structural variations in esophageal squamous cell carcinoma. *Am. J. Hum. Genet.* **98**, 256–274 (2016).
10. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* **27**, 15–26 (2015).
11. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
12. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
13. McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).
14. Durinck, S. *et al.* Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov.* **1**, 137–143 (2011).
15. Shah, S.P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
16. Lohr, J.G. *et al.* Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell* **25**, 91–101 (2014).
17. Landau, D.A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).
18. Shi, Z.Z. *et al.* Consistent and differential genetic aberrations between esophageal dysplasia and squamous cell carcinoma detected by array comparative genomic hybridization. *Clin. Cancer Res.* **19**, 5867–5878 (2013).
19. Shang, L. *et al.* ANO1 protein as a potential biomarker for esophageal cancer prognosis and precancerous lesion development prediction. *Oncotarget* **7**, 24374–24382 (2016).
20. Britschgi, A. *et al.* Calcium-activated chloride channel ANO1 promotes breast cancer progression by activating EGFR and CAMK signaling. *Proc. Natl. Acad. Sci. USA* **110**, E1026–E1034 (2013).
21. Luo, M.L. *et al.* Amplification and overexpression of *CTTN* (*EMS1*) contribute to the metastasis of esophageal squamous cell carcinoma by promoting cell migration and anoikis resistance. *Cancer Res.* **66**, 11690–11699 (2006).
22. Lu, P. *et al.* Genome-wide gene expression profile analyses identify CTTN as a potential prognostic marker in esophageal cancer. *PLoS One* **9**, e88918 (2014).
23. de Bruin, E.C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
24. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
25. Murugaesu, N. *et al.* Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discov.* **5**, 821–831 (2015).
26. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
27. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
28. Toh, Y., Egashira, A. & Yamamoto, M. Epigenetic alterations and their clinical implications in esophageal squamous cell carcinoma. *Gen. Thorac. Cardiovasc. Surg.* **61**, 262–269 (2013).
29. Agarwal, R. *et al.* Epigenomic program of Barrett's-associated neoplastic progression reveals possible involvement of insulin signaling pathways. *Endocr. Relat. Cancer* **19**, L5–L9 (2012).
30. Alvarez, H. *et al.* Widespread hypomethylation occurs early and synergizes with gene amplification during esophageal carcinogenesis. *PLoS Genet.* **7**, e1001356 (2011).
31. Brocks, D. *et al.* Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep.* **8**, 798–806 (2014).
32. Mazor, T. *et al.* DNA methylation and somatic mutations converge on the cell cycle and define similar evolutionary histories in brain tumors. *Cancer Cell* **28**, 307–317 (2015).
33. Robinson, D.F. & Foulds, L.R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
34. Oricchio, E. *et al.* The Eph-receptor A7 is a soluble tumor suppressor for follicular lymphoma. *Cell* **147**, 554–564 (2011).
35. López-Nieva, P. *et al.* *EPHA7*, a new target gene for 6q deletion in T-cell lymphoblastic lymphomas. *Carcinogenesis* **33**, 452–458 (2012).
36. Yu, J. *et al.* Methylation of protocadherin 10, a novel tumor suppressor, is associated with poor prognosis in patients with gastric cancer. *Gastroenterology* **136**, 640–651. e1 (2009).
37. Zhao, Y. *et al.* A novel Wnt regulatory axis in endometrioid endometrial cancer. *Cancer Res.* **74**, 5103–5117 (2014).
38. Saulnier, A. *et al.* Inactivation of the putative suppressor gene *DOK1* by promoter hypermethylation in primary human cancers. *Int. J. Cancer* **130**, 2484–2494 (2012).
39. Mercier, P.L. *et al.* Characterization of *DOK1*, a candidate tumor suppressor gene, in epithelial ovarian cancer. *Mol. Oncol.* **5**, 438–453 (2011).
40. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.* **20**, 274–281 (2013).
41. Quante, T. & Bird, A. Do short, frequent DNA sequence motifs mould the epigenome? *Nat. Rev. Mol. Cell Biol.* **17**, 257–262 (2016).
42. Baylin, S.B. & Jones, P.A. A decade of exploring the cancer epigenome —biological and translational implications. *Nat. Rev. Cancer* **11**, 726–734 (2011).
43. Lay, F.D. *et al.* The role of DNA methylation in directing the functional organization of the cancer epigenome. *Genome Res.* **25**, 467–477 (2015).
44. Irizarry, R.A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
45. Doi, A. *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* **41**, 1350–1353 (2009).
46. Gharwan, H. & Groninger, H. Kinase inhibitors and monoclonal antibodies in oncology: clinical implications. *Nat. Rev. Clin. Oncol.* **13**, 209–227 (2016).
47. Gerlinger, M. *et al.* Cancer: evolution within a lifetime. *Annu. Rev. Genet.* **48**, 215–236 (2014).
48. Cao, W. *et al.* Multiple region whole-exome sequencing reveals dramatically evolving intratumor genomic heterogeneity in esophageal squamous cell carcinoma. *Oncogenesis* **4**, e175 (2015).
49. Yap, T.A., Gerlinger, M., Futreal, P.A., Pusztai, L. & Swanton, C. Intratumor heterogeneity: seeing the wood for the trees. *Sci. Transl. Med.* **4**, 127ps10 (2012).
50. Wang, J. *et al.* Downregulation of EphA7 by hypermethylation in colorectal cancer. *Oncogene* **24**, 5637–5647 (2005).
51. Kiehl, S. *et al.* *ABCB4* is frequently epigenetically silenced in human cancers and inhibits tumor growth. *Sci. Rep.* **4**, 6899 (2014).
52. Jao, T.M. *et al.* Protocadherin 10 suppresses tumorigenesis and metastasis in colorectal cancer and its genetic loss predicts adverse prognosis. *Int. J. Cancer* **135**, 2593–2603 (2014).
53. Narayan, G. *et al.* *PCDH10* promoter hypermethylation is frequent in most histologic subtypes of mature lymphoid malignancies and occurs early in lymphomagenesis. *Genes Chromosom. Cancer* **52**, 1030–1041 (2013).

ONLINE METHODS

Patients and specimens. Tissue samples from 13 patients with ESCC, including primary esophageal tumors and matched morphologically normal esophageal epithelial margins, were collected at the Linzhou Esophagus Cancer Hospital, Henan province, China. All the samples used in this study were residual specimens collected after diagnosis sampling. All patients received no treatment before surgery and signed separate informed consent forms for sampling and molecular analyses. We also considered clinicopathological parameters when selecting these patients with ESCC, including sex, pathological tumor (pT) stage, regional lymph node metastasis and tumor differentiation, to avoid bias toward particular pathological characteristics (**Supplementary Table 1**). Specifically, the male/female ratio in the current cohort was similar to that reported in the latest publication[54]. The number of patients with relatively early (pT1b or pT2) and late (pT3) tumor stage was five and eight, respectively. The status of lymph node metastasis (negative, $n = 4$; positive, $n = 9$), as well as tumor differentiation (G1, $n = 1$; G2, $n = 6$; G2/3, $n = 2$; G3, $n = 4$), was also taken into account. This study has been approved by the ethics committee or institutional review board of the Cancer Hospital/Institute, Peking Union Medical College and the Chinese Academy of Medical Sciences (approval NCC2013-066). The collection and publication of Chinese human genetic data used in the present study has been approved by the Ministry of Science and Technology. In 12 of 13 cases, four spatially separated tumor specimens were obtained from each individual, with each section at least 0.5 cm away from the others. In the case of ESCC04, three tumor regions were sampled. We carefully reviewed the hematoxylin and eosin slides for each tumor region before subjecting them to whole-exome sequencing analysis, to make sure that the tumor cell contents of the selected regions were comparable and were at least greater than 60% (representative hematoxylin and eosin images are provided in **Supplementary Fig. 10**).

Multiregional whole-exome sequencing. For each individual, genomic DNA of cells from different tumor regions and one matched normal epithelial tissue sample at the surgical margins was sequenced. Genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue kit according to the manufacturer's instructions. For cases ESCC01 and ESCC02, whole-exome capture of genomic DNA was performed by BGI, using the BGI Exome Enrichment kit, and massively parallel sequencing of captured genomic DNA was performed and results were analyzed by BGI using the Complete Genomics platform. For the 11 other cases, the Agilent SureSelect Human All Exon v4 (51 Mb) kit was used for whole-exome capture of genomic DNA, and the captured DNA was sequenced by BGI using the Illumina HiSeq 4000 sequencing platform, with 150-bp paired-end sequencing.

Alignment of sequencing reads and somatic variant detection. 150-bp paired-end fastq files were aligned to the human reference genome (build hg19) using the BWA-MEM aligner in default mode (see URLs). Alignments were then filtered for duplicate reads using Samblaster[55], and BAM files were indel realigned and base quality scores were recalibrated according to GATK best practices[56].

Somatic variants were detected using VarScan2 (ref. 57). Tumor and matched normal pileup files were generated using the SAMtools 'mpileup' command and fed into the VarScan 'somatic' command[58]. Reference genome positions covered by at least 10 reads in the normal sample and 14 reads in tumor samples were considered for variant calling. Variants with VAF less than 0.07 were discarded. Raw somatic variants were filtered using the VarScan 'processSomatic' command with arguments --min-tumor-freq 0.07, --max-normal-freq 0.02 and --p-value set to 0.05. The resulting high-quality somatic variants were filtered for false positives using the fpFilter Perl script (see URLs). The filtered variants were annotated with ANNOVAR[59] and filtered against the dbSNP135 database for commonly occurring SNPs[59]. Disease-associated variants annotated in the ClinVar database and the COSMIC database were retained.

Phylogenetic tree construction. For mutations that were detected from at least one tumor region, a method described by Stachler *et al.*[60] was used to increase the sensitivity of detecting these mutations in other tumor regions from the same individual with low VAF. In brief, Bam-readcount (see URLs) was used to obtain read counts for unique somatic variants across all tumor regions. A variant was considered to be absent if either its VAF was less than 0.02 or there were fewer than three reads. The VAFs across all the tumor regions for each individual were then used to generate a binary table. Phylogenetic trees were constructed on the basis of the binary tables using Discrete Character Parsimony, implemented in the PHYLogeny Inference Package (see URLs), with the matched morphologically normal epithelial margins as outgroup roots. On the basis of calculated branch/trunk lengths inferred from mutation counts, final trees were drawn manually.

Cancer cell fraction analysis. Copy number analysis from whole-exome sequencing data was performed using ReCapSeg, which is implemented as part of GATK (v4). Briefly, read counts for each of the exome targets were extracted from all samples and were divided by the total number of reads to generate proportional coverage. A panel of normal controls[14] was created using proportional coverage from all of the normal samples. Each of the tumor samples was compared to a panel of normal controls (PoN), followed by tangent normalization. The normalized coverage profiles were then segmented using circular binary segmentation[61]. Variants on the sex chromosomes (X and Y) were excluded from this analysis.

Tumor cellularity was determined on the basis of VAF and segmented copy number data using ABSOLUTE[62], to determine the CCF of each mutation, as was previously described by McGranahan *et al.*[13]. Clonal status was defined according to the confidence interval of CCF. Mutations were classified as subclonal if the upper bound of their 95% confidence interval was less than 1.

Identification of putative driver mutations. We first identified putative cancer driver genes on the basis of recent large-scale ESCC sequencing data[4–8], the COSMIC cancer gene census (August 2015)[11] and pan-cancer analysis[12]. Next, non-silent variants in these genes were evaluated, and putative driver mutations were identified if they met one of the following requirements: (i) either the exact mutation, the same mutation site or at least three mutations located within 15 bp of the variant were found in COSMIC and (ii) if the candidate gene was marked as recessive in COSMIC and the variant was predicted to be deleterious, including stop-gain, frameshift and splicing mutation, and had a SIFT score <0.05 (ref. 63) or a PolyPhen score >0.995 (refs. 64,65).

Mutational signature analysis. Both silent and non-silent somatic mutations were classified as either truncal or branch as described earlier, and the mutational signatures of these variants were generated separately. We performed a multiple regression approach, deconstructSigs[26], to extract signatures based on the Wellcome Trust Sanger Institute Mutational Signature Framework[27] and to statistically quantify the contribution of each signature for each tumor.

DNA methylation analysis and construction of phyloepigenetic trees. The DNA methylation profiles of 12 tumor regions and 2 matched normal esophageal epithelial tissue samples from 3 ESCC cases examined by M-WES (ESCC01, ESCC03 and ESCC05) were generated using the Illumina Infinium HumanMethylation450 platform at the University of Southern California Norris Comprehensive Cancer Center Genomics Core. We performed basic data processing of the HM450 data using many of the same processing steps that we performed previously for The Cancer Genome Atlas (TCGA) data analysis, which is based on the Methylumi R package[66] with several additional quality control steps. Probes with detection P values greater than 0.01 in any of the samples were removed, as were probes overlapping with dbSNP SNPs and probes on the X or Y chromosome.

For intratumoral analysis, we defined a probe as private if the difference in β values for any single pair of tumor regions was at least 0.3; we defined a probe as shared if the differences in β values for all pairs of tumor regions were less than 0.1. Only private probes were used for construction of phyloepigenetic trees. For each tumor, pairwise Euclidean distances were calculated between all tumor regions using the complete set of private probes.

Phyloepigenetic trees were constructed from these pairwise distances, using the minimal evolution method implemented by the fastme.bal function in the R package ape[67]. Different probe selection cutoffs for calling private and shared probes produced similar results, with only minimal changes in cases ESCC01 (at the cutoff of 0.5) and ESCC05 (at the cutoff of 0.2; **Supplementary Fig. 6**).

Topological comparison for phylogenetic versus phyloepigenetic trees and other tree pairs was performed using the RF.dist function in the CRAN R package phangorn. Comparison in case ESCC01 was carried out on the basis of only tumor samples because of the lack of a matched normal sample in DNA methylation data (for visualization in **Fig. 4a**, we used normal samples from the other two cases as the root).

To mitigate the confounding effects of non-cancer cells in phyloepigenetic tree reconstruction, we performed additional bioinformatic analyses, as follows.

The major source of nonmalignant DNA contamination in esophageal tumors is immune cells[68]; this has been shown by TCGA to be the case for most solid tumors[62,69]. We confirmed this by review of all of our methylation-profiled samples through immunostaining of the leukocyte common antigen (LCA)/CD45 (representative immunohistochemistry images are shown in **Supplementary Fig. 7**), which is a common marker of immune cells and is widely used in distinguishing infiltration of immune cells[70–73]. To precisely determine the extent of immune cell contamination, we estimated the fraction of leukocytes in each sample using profiles of immune-specific methylation probes[74], as described previously[62,69]. Using this method, we noted that case ESCC01 was highly pure (estimated immune cell fraction = 7.1%, ranging from 1 to 14% in various tumor regions) and cases ESCC03 and ESCC05 contained an average of 20% and 32% immune cells, respectively (**Supplementary Table 7**).

We recalculated each phyloepigenetic tree using one of two methods to model the mixture of cancer and immune cells within the samples.

(1) As demonstrated in several TCGA marker papers, performing analysis using only the subset of Infinium probes unmethylated in purified leukocytes and dichotomizing/binarizing the tumor β value for these probes with a minimum β-value cutoff could minimize the influence of contaminating leukocytes[75–77]. We used HM450 profiles from purified leukocyte populations[74] and selected probes with a maximum β value less than 0.3 across all leukocyte samples. We then binarized the tumor β values as 1 if they were >0.3 and 0 otherwise. We computed pairwise distances between binarized values using the Jaccard index (dist function in R) and performed tree construction using these pairwise distances. The resulting trees are labeled "dichotomized" in **Supplementary Figure 8**.

(2) In an independent approach, we modeled tumor β value as a linear combination of DNA from a mixture of cancer cells and leukocytes. The mixing ratio was estimated for each sample on the basis of methylation of leukocyte-specific probes, as described above and previously[62,69]. For each probe, we used the fixed mixing ratio, the average β value of the probe in purified leukocytes[74] and our measured β value in the tumor to estimate the methylation β value of the cancer cells alone. This method was used to reconstruct phyloepigenetic trees for each case, and the resulting trees are labeled "immune cell adjusted" in **Supplementary Figure 8**.

Trees for both methods 1 and 2 were compared to the original trees using the RF method; RF values are listed in **Supplementary Figure 8**.

Determining the genomic context of shared versus private methylation patterns. Shared versus private probes were identified on the basis of heterogeneity between different tumor regions. The groups were further divided into 'hypermethylated' and 'hypomethylated' probes, on the basis of comparisons of methylation in tumor samples and adjacent normal tissue. For hypermethylated probes from a specific case, we selected all probes with a methylation β value <0.3 in the adjacent normal sample (or a maximum β value of two other normal samples <0.3 for ESCC01) and a mean β value across all tumor regions that was at least 0.3 higher than the mean of normal sample(s). Similarly, for hypomethylated probes, we selected probes with ≥0.6 in the normal sample and at least 0.3 higher in the tumor than the mean of normal samples. For ESCC01, with no matched normal sample, we averaged the β values from the other two normal samples. Hyper- and hypomethylated probe sets are shown in **Figure 4b–d** and **Supplementary Figure 9**. For the enrichment analysis in **Figure 4c**, promoters were defined as 1.5-kb regions up- and downstream of the RefSeq transcription start site, CGIs were taken from the HMM-defined set[78], and CGI shores and enhancers were defined using the standard Illumina 450K annotation manifest. PMDs were called using the Roadmap[79] normal esophagus sample (E079), using an HMM-based segmentation method[80]. Enrichment/depletion *P* values for the enrichment of private versus shared probes in each genomic context were computed on the basis of a hypergeometric test, where null model frequencies were calculated on the basis of all probes present on the array (shown as "background" in **Fig. 4c**).

Immunohistochemistry. Formalin-fixed and paraffin-embedded tissue slides were deparaffinized using xylene, rehydrated using xylene and ethanol, and then immersed in 3% hydrogen peroxide solution for 10 min, heated in citrate at 95 °C for 25 min and cooled at room temperature for 60 min. Slides were incubated overnight at 4 °C with antibody to LCA/CD45 (Cell Marque, 145M-96; diluted 1:100) and visualized using the PV-9000 Polymer Detection System following the manufacturer's instructions (Beijing Zhongshan Golden Bridge Biotechnology). Counterstaining was carried out with hematoxylin.

54. Chen, W. *et al.* Cancer statistics in China, 2015. *CA Cancer J. Clin.* **66**, 115–132 (2016).
55. Faust, G.G. & Hall, I.M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
56. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
57. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
58. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
59. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
60. Stachler, M.D. *et al.* Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat. Genet.* **47**, 1047–1055 (2015).
61. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
62. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
63. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
64. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
65. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7, Unit 7.20 (2013).
66. Triche, T.J. Jr., Weisenberger, D.J., Van Den Berg, D., Laird, P.W. & Siegmund, K.D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41**, e90 (2013).
67. Desper, R. & Gascuel, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* **9**, 687–705 (2002).
68. Takahashi, T. *et al.* Estimation of the fraction of cancer cells in a tumor DNA sample using DNA methylation. *PLoS One* **8**, e82302 (2013).
69. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
70. Pagès, F. *et al. In situ* cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer. *J. Clin. Oncol.* **27**, 5944–5951 (2009).
71. de Miranda, N.F. *et al.* Infiltration of Lynch colorectal cancers by activated immune cells associates with early staging of the primary tumor and absence of lymph node metastases. *Clin. Cancer Res.* **18**, 1237–1245 (2012).
72. Punt, S. *et al.* Whole-transcriptome analysis of flow-sorted cervical cancer samples reveals that B cell expressed TCL1A is correlated with improved survival. *Oncotarget* **6**, 38681–38694 (2015).
73. Gorter, A., Prins, F., van Diepen, M., Punt, S. & van der Burg, S.H. The tumor area occupied by Tbet$^+$ cells in deeply invading cervical cancer predicts clinical outcome. *J. Transl. Med.* **13**, 295 (2015).
74. Reinius, L.E. *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One* **7**, e41361 (2012).
75. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
76. Ceccarelli, M. *et al.* Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016).
77. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
78. Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A. & Feinberg, A.P. Redefining CpG islands using hidden Markov models. *Biostatistics* **11**, 499–514 (2010).
79. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
80. Song, Q. *et al.* A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* **8**, e81148 (2013).

doi:10.1038/ng.3683