Examining the Role of Extrachromosomal DNA in Lung Cancer

Azhar Khandekar[1,2,3,□,] Phuc H. Hoang[1], Jens Luebeck□, Marcos Díaz-Gay[2,3,□,□,] Wei Zhao[1], John P. McElderry[1], Caleb Hartman[1], Mona Miraftab[1], Olivia W. Lee[1], Kara M. Barnao[1], Kristine Jones[1,□,] Amy Hutchinson[1],□, Belynda Hicks[1,□,] Erik N. Bergstrom[2,3,□,] Yang Yang□, Martin A. Nowak□[,1□,] Nathaniel Rothman[1], Angela C Pesatori[11,12], Dario Consonni[12], Robert Homer[13], Marina K. Baine[1]□, Lynette M. Sholl[1]□, Philippe Joubert[1]□, Charles Leduc[1]□, William D. Travis[1]□, Soo-Ryum Yang[1]□, Qing Lan[1], David C. Wedge[1]□[,1□,] Lixing Yang[8,2□,21,] Stephen J. Chanock[1], Tongwu Zhang[1]*, Ludmil B. Alexandrov[2,3,□,22]*, Maria Teresa Landi[1]*

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, [2]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA, [3]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA, [4]Moores Cancer Center, University of California San Diego, La Jolla, CA, USA, [5]Department of Computer Science, University of California San Diego, La Jolla, CA, USA, [6]Digital Genomics Group, Cancer Genomics Program, Spanish National Cancer Research Center (CNIO), Madrid, Spain, [7]Cancer Genomics Research Laboratory, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, MD, USA, [8]Ben May Department for Cancer Research, The University of Chicago, Chicago, IL, USA, [9]Department of Mathematics, Harvard University, Cambridge, MA, USA, [10]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA, [11]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy. [12]Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy, [13]Department of Pathology, Yale School of Medicine, New Haven, CT, USA, [14]Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA, [15]Brigham and Women's Hospital, Boston, MA, USA, [16]Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval, Canada; [17]Centre Hospitalier de l'Université de Montréal, Montreal, Canada, [18]Manchester Cancer Research Centre, The University of Manchester, Manchester, UK, [19]Manchester NIHR Biomedical Research Centre, Manchester, UK, [20]Department of Human Genetics, The University of Chicago, Chicago, IL, USA, [21]The University of Chicago Medicine Comprehensive Cancer Center, The University of Chicago, Chicago, IL, USA, [22]Stanford Stem Cell Institute, University of California San Diego, La Jolla, CA, USA

*Correspondence should be addressed to tongwu.zhang@nih.gov, L2alexandrov@health.ucsd.edu and landim@nih.gov

ABSTRACT

The role of extrachromosomal DNA (ecDNA) in lung cancer, particularly in subjects who never smoked (LCINS), remains unclear. Examination of over 1200 whole-genome-sequenced lung cancers identified ecDNA in 18.9% of patients. Recurrent amplification of *MDM2* and other oncogenes *via* ecDNA possibly drives a LCINS subset. Tumors harboring ecDNA showed worse overall survival than tumors harboring other focal amplifications. A strong association with whole-genome doubling suggests most ecDNA reflects genomic instability in treatment-naïve lung cancer.

MAIN

Extrachromosomal DNA (ecDNA) is a potent mechanism for oncogene amplification in human cancer[1], exhibiting unique characteristics such as a circular structure[2], non-Mendelian inheritance[3,4], and an altered epigenetic and transcriptional landscape[5]. Prior studies have linked ecDNA to worse overall survival in pan-cancer analyses, highlighting its clinical significance[6-8] and potential as a therapeutic target[9]. Recent advances in computational techniques[10] have enabled large-scale detection of ecDNA from whole-genome sequencing (WGS) data. However, only three pan-cancer studies have examined ecDNA in primary non-small cell lung cancer cohorts with over 100 samples, focusing predominantly on patients of European descent[6-8]. Two studies[6,7] analyzed lung cancer samples from the United States, reporting ecDNA prevalence rates of about 21% in 36 lung adenocarcinomas (LUAD) and 27% in 47 lung squamous cell carcinomas (LUSC). Additionally, a UK lung cancer study[8] analyzed 14,778 patients across 39 tumor types, including 718 LUAD tumors and 378 LUSC tumors, detecting ecDNA in 8.2% of LUAD and 22.4% of LUSC cases. All three studies[6-8] reported lung cancer findings as part of broader pan-cancer analyses and did not distinguish between lung cancers in subjects who have never-smoked (LCINS) and subjects who have smoked (LCSS). Notably, based on the high prevalence of the tobacco-associated[11] mutational signature SBS4 in these studies, more than 84% of the analyzed lung cancer cases were from individuals with a history of smoking.

Despite these prior studies, the clinical relevance of ecDNA in lung cancer remains unclear, including its prevalence across histologies, geographical areas, genetic ancestries, and biological sexes. Its role in LCINS, LCSS, and subjects who were exposed to passive smoking also requires further exploration. To address these questions, here, we analyzed WGS data from 1,216

treatment-naïve lung cancers from diverse patient populations (**Fig. 1*a***; **Supplementary Fig. 1**). This large international cohort enabled a comprehensive evaluation of ecDNA's association with genomic alterations, as well as its prevalence across smoking status, biological sexes, ancestries, and 18 areas defined primarily by country of residence. The dataset includes 871 LCINS, sourced from the Sherlock-Lung[12] (*n*=817), the Environment And Genetics in Lung Cancer Etiology (EAGLE[13], *n*=25), the Cancer Genome Atlas (TCGA[14], *n*=15), and other publicly available cohorts (*n*=14)[15-19], representing patients from 17 geographical areas (**Supplementary Fig. 2*a***). Additionally, it comprises 345 LCSS, drawn from EAGLE[13] (*n*=221) and TCGA[14] (*n*=83) and other publicly available cohorts (*n*=41)[15-19], spanning five geographical areas (**Supplementary Fig. 2*b***). The majority of tumors were LUAD (*n*=1,023; 84.1%) or LUSC (*n*=67; 5.5%), with carcinoids (*n*=61; 5%, exclusively in LCINS) and other rarer histological lung cancer subtypes (*n*=65; 5.3%) also represented. Amongst Sherlock-*Lung* LCINS, 250 were exposed to secondhand tobacco smoke, while 208 had documented non-exposure (**Fig. 1*a***).

To detect and reconstruct ecDNA across the cohort, a computational pipeline that accounts for both tumor purity and ploidy was developed (**Methods**). A total of 231 samples harbored at least one ecDNA (18.9%), with similar prevalence between LCINS (17%) and LCSS (23%; q-value: 0.21; **Fig. 1*b***). No statistically significant differences were observed in the prevalence of ecDNA across various histological subtypes, biological sexes, ancestries or age at diagnosis (**Fig. 1*c*&e**), except for carcinoids, which showed a lower prevalence compared to LUAD in the overall analysis (OR=0.12; q=0.01) and when LCINS were considered separately (OR=0.13; q=0.01). ecDNA was present at similar rates across all geographic locations (**Fig. 1*d***; **Supplementary Fig. 2*c*-*d***), as indicated by a one-versus-all multivariate logistic regression for each country.

There was also no difference in ecDNA prevalence between LCSS and LCINS (**Fig.1*c***). ecDNA was not associated with tumor stage (**Supplementary Fig. 3*a&b***), subsequent development of metastases (**Supplementary Fig. 3*c&d***) or passive smoking status (**Supplementary Fig. 3*e***). These findings remained consistent when analyses were restricted to only LUAD cases in either LCSS or LCINS (**Supplementary Fig. 4**).

To examine the association between ecDNA and genomic features previously linked to genome instability[20-23] , we evaluated the relationship of ecDNA with whole-genome doubling (WGD), chromothripsis, tumor mutational burden, mitochondrial copy number, overall genome instability index (wGII), and telomere length tumor/normal ratio. Amongst these, only WGD showed a significant association with ecDNA in both LCSS and LCINS (**Fig. 2*a***), as determined by logistic regression adjusted for age, sex, ancestry, histology, and tumor purity. Specifically, ecDNA was present in 7.1% of LCINS and 8.9% of LCSS without WGD, but its prevalence increased markedly in WGD-positive tumors from LCINS (24.9%, OR=4.0; $q=1.24 \times 10^{-8}$) and LCSS (30.3%, OR=5.85; $q=8.06 \times 10^{-4}$). The same association was observed when only LUAD was examined (**Supplementary Fig. 5*a***). Moreover, this association was further validated in a pan-cancer dataset of 974 samples with high-confidence WGD annotation[14], using logistic regression adjusted for age, purity, sex, and cancer type (OR=2.81; $q=1.63 \times 10^{-6}$; **Supplementary Fig. 5*b***).

In addition to WGD, chromothripsis (OR=1.76; q=0.048; **Fig. 2*a***) and telomere length tumor/normal ratio, (OR=1.76; q=0.01; **Fig. 2*a***) were also significantly associated with ecDNA, but this was exclusively observed in LCINS. Prior studies have indicated that chromothripsis

may contribute to the formation of certain ecDNA[23]. In this study, 115 out of 151 LCINS with ecDNA (76.2%) and 60 out of 80 LCSS with ecDNA (75%) exhibited chromothripsis (**Supplementary Tables 1-2**). Nevertheless, in samples where both ecDNA and chromothripsis were detected, overlap between ecDNA segments and chromothripsis regions was rare, occurring more frequently in LCINS—observed in 31 of 115 LCINS cases compared to 4 of 60 LCSS (OR=5.01, p=0.02 comparing LCINS *vs.* LCSS; **Supplementary Fig. 5*c***). This suggests that there could be potentially different mechanisms of ecDNA formation between LCINS and LCSS. However, while the majority of ecDNA emerges in genomically unstable lung cancers that also exhibit chromothripsis, the actual genomic overlap between ecDNA and chromothripsis regions is limited in both lung cancer types. The association between shortened telomere length and ecDNA persisted even when using an alternative telomere estimation method[24] (**Supplementary Fig. 5d*)***, and whether it was input as a binary or numeric variable (**Supplementary Fig. 5e**). This result is consistent with prior studies that have reported that telomere length is linked with cell division[25] and that cells with higher ecDNA levels divide faster[26].

To evaluate the relationship between ecDNA presence and lung cancer driver genes, we analyzed driver genes altered in at least 10 samples as well as driver hotspot mutations occurring at the same genomic position in at least 10 samples. Using multivariate logistic regression, adjusted for age, sex, tumor purity, ancestry, and histology, no significant associations were found between ecDNA and individual driver gene alterations (**Fig. 2*b***) or between ecDNA and individual driver alterations (**Fig. 2*c***).

Mutational signature analysis provides insights into both the endogenous and exogenous mutational processes that have been active throughout the lineage of a cancer genome[11]. We investigated the association between ecDNA presence and mutational signatures. As expected[27], multiple copy-number and structural variant signatures were associated with ecDNA presence similarly in LCSS and LCINS (**Supplementary Fig. 6*a*-*b***). Additionally, in LCINS, ecDNA-positive samples were enriched for clock-like signatures (SBS1, SBS5, ID1, ID2) and, as previously reported[19,28], APOBEC-associated signatures (SBS2, SBS13; **Supplementary Fig. 6*a***). In LCSS, enrichment was observed only for SBS3 (**Supplementary Fig. 6*b***), a signature previously associated with homologous recombination deficiency[29].

Previous studies have shown that ecDNA can harbor oncogenes, regulatory elements, and immunomodulatory genes in varying configurations[3,8,30,31]. In our cohort, 151 (17%) LCINS and 80 (23%) LCSS harbored ecDNA, with a total of 236 ecDNAs in LCINS and 123 in LCSS, reflecting multiple ecDNAs in some tumors. Among these, ecDNA in LCINS were enriched with oncogenes ($p<0.001$) and immunomodulatory genes ($p=0.008$) compared to LCSS (**Fig. 2*d***; **Supplementary Fig. 7*a*-*b***). Further, ecDNA containing oncogenes exhibited elevated copy numbers (CN) compared to other ecDNA without oncogenes in LCINS ($p=1.5 \times 10^{-3}$) and LCSS (5.8×10^{-3}), suggesting greater selective pressure for these ecDNAs (**Supplementary Fig. 7*c***), as previously reported in pan-cancer[8]. *MDM2* was the most frequently amplified oncogene on ecDNA in both LCINS (n=36) and LCSS (n=5). Given its prominence, we examined the relationship between ecDNA-driven *MDM2* amplification and *TP53* alterations, as MDM2 negatively regulates TP53 through ubiquitination[32]. We found that *MDM2* amplification on ecDNA was mutually exclusive with *TP53* alterations (OR=7.51, p=0.01). In LCINS, the next

four most prevalent oncogenes that did not co-occur with *MDM2* were *TERT* (n=10), *CCND1* (*n*=8), *MYC* (*n*=8), and *EGFR* (*n*=7; **Supplementary Table 3**). In LCSS, no oncogene besides *MDM2* was amplified on ecDNA in more than two samples (**Supplementary Table 4).**

Prior pan-cancer studies have shown that cancers harboring ecDNA amplifications exhibit worse overall survival[6-8]. To evaluate this association in the lung cancer cohort, survival analysis was conducted on 692 LCINS and 333 LCSS with available stage and covariate data. Lung cancers were categorized as harboring ecDNA, other focal amplifications, or no focal amplifications. Multivariate Cox proportional-hazards regression analysis revealed that patients with ecDNA-positive lung cancers had worse overall survival than those with only chromosomal amplifications; this association was statistically significant in LCINS (hazard ratio=1.53; p=0.03; **Fig. 2*e***).

In summary, we examined the full spectrum of ecDNA in the largest dataset of whole-genome sequenced lung cancers to date, from many geographical regions and ancestry, and with unique information on tobacco smoking status. ecDNA was present in 17% of LCINS and 23% of LCSS. ecDNA in LCINS were enriched with oncogenes and immunomodulatory genes in comparison to LCSS. In LCINS, the *MDM2* locus was often amplified through ecDNA, and shortened telomere length was associated with ecDNA. No associations were identified between presence of ecDNA and biological sexes, smoking status, or ancestries. The carcinoid subtype was depleted of ecDNA; no other differences by histology were observed. However, tumors carrying ecDNA showed worse overall survival in comparison to those with focal chromosomal amplification, particularly in LCINS. A strong association was observed between ecDNA and

WGD, and this observation was validated in an independent pan-cancer cohort, suggesting that most ecDNA in treatment-naïve lung cancer are likely to be a product of genomic instability. Notably, ecDNA carrying oncogenes could contribute to driving LCINS and affecting prognosis. Future studies will be needed to dissect the mechanisms of ecDNA formation and maintenance in lung cancer and evaluate its clinical implications in both treatment-naïve and post-treatment settings.

METHODS

Criteria for sample inclusion

Similar to our previous publication[33,34], we used seven rigorous criteria for sample inclusion, as follows. (1) Sequencing coverage. We maintained a minimum average sequencing coverage of more than 40× for tumor samples and more than 25× for normal samples. (2) Contamination and relatedness. Cross-sample contamination was limited to less than 1% by Conpair[35], and detected relatedness was maintained at less than 0.2 by Somalier[36]. (3) Copy-number analysis. Individuals with abnormal copy-number profiles in normal samples were excluded, as determined by Battenberg[37]. (4) Mutational signatures. Tumor samples exhibiting mutational signatures SBS7 (associated with exposure to ultraviolet light[11]) and SBS31 (associated with platinum chemotherapy)[38] were removed from the analysis. (5) Tumor type validation. Tumor samples reported as non-lung cancer or not originating from primary lung cancer were excluded. (6) WGS quality control. Tumor samples with a total genomic alteration count of less than 100, or samples with a total genomic alteration less than 1,000 and a number of reads per clonal copy (NRPCC) less than 10, were excluded as low-quality samples. (7) Multiple-region sequencing. In rare cases in which multiple regions of a tumor were sequenced, only one high-purity tumor sample was included to avoid redundancy. These stringent criteria were applied consistently to ensure the robustness and reliability of the data collected for the Sherlock-Lung study. Following the application of these criteria, a total of 1217 samples were available for analysis, but one sample with unknown smoking status was excluded, resulting in a total of 1216 samples across the cohort.

Determination of sample purity and ploidy

The Battenberg algorithm (v.2.2.9)[37] was used to determine the sample purity and ploidy. Any somatic copy number variations (CNV) profile determined to have low-quality after manual inspection underwent a refitting process. This process required new tumor purity and ploidy inputs, either estimated by Ccube (v.1.0)[39] or recalculated from local copy number status. The Battenberg refitting procedures were iteratively executed until the final CNV profile was established and met the criteria of manual validation check.

ecDNA detection and characterization

CNVKit v.0.9.6[40] was run in tumor-normal mode to call somatic CNVs against the matched normal whole-genome sequenced samples for each patient, using the log ratio of tumor to normal read depths (logR) and the circular binary segmentation (CBS) method. Note that this method, by default, assumes a sample purity of 1 and a sample ploidy of 2 as it does not correct for purity and ploidy. This limitation can lead to high copy number segments being missed due to low purity[41]. Due to the wide range of tumor purities in our dataset, consistent with prior reports in lung cancer[41], we rescaled the total copy number estimate using the purity and ploidy estimated from Battenberg according to the following equation:

$$Equation\ 1.1. TCN = \frac{(\psi * (2^{logR})) - 2 * (1 - \rho)}{\rho}$$

where ψ is the tumor ploidy, ρ is the tumor purity, and logR represents the log ratio of the tumor read depth to the normal read depth (**Supplementary Fig. 8a**). These copy number calls were used to identify putative 'seed' regions. We then used AmpliconArchitect (v.1.3)[10] to reconstruct the architecture of the amplicons and AmpliconClassifier (v.1.3.1) to classify amplicons

according to the most probable mechanism of formation (ecDNA, BFB, complex non-cyclic, or linear). Visual inspection and verification were performed for all ecDNA amplicon regions. This pipeline allowed for samples with low purities to be utilized as it completely removed the effect of tumor purity in ecDNA detection (**Supplementary Fig. 8b**). As a result, more ecDNA^{+} samples were detected (**Supplementary Fig. 8b**) and all 1216 samples could be used for ecDNA analysis (**Supplementary Fig. 8c**). Of the additional ecDNA detected, eight of these were *EGFR*-ecDNA (seven in LCINS and one in LCSS) and each of these ecDNA except for one also harbored a driver mutation (**Supplementary Table 5**). Among the eight EGFR-ecDNA+ samples, five had RNA-seq data, which showed very high EGFR expression (**Supplementary Fig. 8d).**

Determination of whole genome doubling status

As previously done[42], samples were considered to have a whole genome doubling event if the major copy number was ≥3 for more than 50% of copy number segments.

Chromothripsis detection and characterization

To detect chromothripsis regions in genomes, we applied the computational algorithm ShatterSeek[43], which employs a set of statistical criteria to identify chromothripsis given an input of copy number and structural variants. We used the most stringent criteria corresponding to high confidence chromothripsis calls. Specifically, these criteria were: *(i)* a cluster of structural variants (>6 DUP/DEL/h2hINV/t2tINV), *(ii)* oscillating CNV between two states (>7 CNV events), *(iii)* chromosomal enrichment and distribution of DNA breakpoints (p-value<0.05), *(iv)* randomness of fragment joins (p-value>0.05) and/or ≥4 inter-chromosomal rearrangements

between multiple chromosomes. The input CNV calls were generated using CNVkit in tumor-normal mode with default parameters, and the input structural variants calls were generated using a union of Meerkat (v.0.189) and Manta (v.1.6.0) calls with the recommended filtering.

Association of ecDNA with demographic and clinical features

Demographic and clinical features where information was available for all samples (age, sex, ancestry, and tumor histology and purity) served as input variables in a multivariate logistic regression model with ecDNA presence or absence as the dependent outcome variable, and these models were constructed separately for LCINS and LCSS. In addition, separate models for passive smoking, tumor stage, and presence of metastases were run along with the rest of the variables in both LCINS and LCSS.

Determination of telomere length tumor/normal lung tissue ratio

We estimated telomere length (TL) in kilobases using TelSeq (v.0.0.2)[44]. As was previously done[34], we used seven as the threshold for the number of TTAGGG/CCCTAA repeats in a read for the read to be considered telomeric. The TelSeq calculation was done individually for each read group within a sample, and the total number of reads in each read group was used as weight to calculate the average TL for each sample. The telomere length ratio ($\log_2$ scale) between tumor and normal tissue was then calculated for each patient and served as the independent variable in multivariable logistic regressions.

Association of mutational signatures and ecDNA status

Mutational signature attributions were determined as was previously done, based on the methodology in Díaz-Gay et. al[34]. SigProfilerMatrixGenerator[45,46] was used to create the mutational matrix for all types of somatic mutations. The deconvolution of mutational signatures was performed by SigProfilerExtractor (v.1.1.3)[47] using the whole-genome sequencing setting and the COSMIC mutational signatures (v3.2) as reference. Signature extractions were performed separately for LCINS and LCSS. Signature attributions were input into a multivariate logistic regression model as either present or absent if the signature was present in less than 50% of the samples. If the signature was present in ≥50% samples (*e.g.*, SBS1 and SBS5), it was input as either above median or below median. For SBS1 and SBS5, the signature attributions were also input as continuous numeric variables to confirm the significant associations found with inputting them as above median or below median.

Association of genomic features with ecDNA status

Multivariate logistic regression was performed for each feature individually, with adjustment for covariates (age, sex, ancestry, histology, wGII, and tumor purity). All continuous numeric variables were both input as raw numeric values and binarized and input as above median/below median.

Association of driver genes and ecDNA status

The IntOGen pipeline (v.2020.02.0123)[48], which combines seven state-of-the-art computational methods, was employed with default parameters to detect signals of positive selection in the mutational patterns of driver genes across the cohort. The status of each driver gene (altered or

unaltered) was then input into a logistic regression model with covariates age, sex, ancestry, histology, wGII, and tumor purity, and ecDNA status was the output. These sets of genes were considered altered if there was a focal deletion overlapping the gene, an inactivating (stop or missense) mutation, or an activating gain of function mutation.

Association of driver mutations and ecDNA status

To identify driver mutations within the set of identified driver genes, we implemented a rigorous and multifaceted strategy, considering multiple criteria: *(i)* the presence of truncating mutations specifically in genes annotated as tumor suppressors, *(ii)* the recurrence of missense mutations in a minimum of 3 samples, *(iii)* mutations designated as "Likely drivers" with a boostDM score[49] exceeding 0.5, *(iv)* mutations categorized as either "Oncogenic" or "Likely Oncogenic" based on the criteria established by OncoKB[50], an expert-guided precision oncology knowledge base, *(v)* mutations previously recognized as drivers in the TCGA MC3 drivers paper[51], and *(vi)* missense mutations characterized as "likely pathogenic" in genes that are annotated as tumor suppressors, as described in Cheng et al.[52] Any mutation meeting one or more of these criteria was recognized as a potential driver mutation.

Calculation of genome instability index score

To calculate the genome instability index score (wGII), the method devised in Bailey et al.[8] was used. The overall ploidy was calculated as the length-weighted average of all copy number segments. The proportion of segments that were aberrant (where total copy number was greater than the ploidy or less than the ploidy) was computed for each chromosome, these proportions were summed, and the total was divided by 22.

Identifying regulatory regions on ecDNA

A set of lung enhancer regions were downloaded from EnhancerAtlas[53], and each ecDNA region was overlapped with these enhancer regions with the requirement that the entire enhancer had to be fully contained (100% overlap) on the ecDNA. A set of lung cancer promoter regions was extracted using a R script which downloaded regions 1000 base pairs upstream of transcription start sites for all genes.

Defining a set of immunomodulatory genes

ecDNAs were considered immunomodulatory if a gene from the significant gene sets mapped to an ecDNA that did not contain an oncogene, and that significant gene set had an immunomodulatory function (GO terms: 0006968, 0002228, 0042267, 0001906, 0001909, 0002698, 0001910, 0031341, 0002367, 0002695, 0050866, 0051250, 0050777), following the procedure in Bailey et. al[8]**.** This resulted in a set of 376 genes.

Mitochondrial copy number estimation

Mitochondrial copy number was estimated per sample using the ratio of mitochondrial read depth to nuclear genome depth, adjusted for tumor ploidy and purity, following the procedure in Reznik *et al*.[54]

Identification of LINE-1 retrotransposons

To identify putative transposable elements (TE), we utilized a pipeline called TraFiC-mem (v.1.1.0)[55] (Transposome Finder in Cancer) retaining only TE insertions that passed the default filtering criteria. Detailed methods can be found in Zhang et. al[42].

Statistical analysis

All statistical analyses and graphic displays were performed using the R software v4.2.3 1011 (https://www.r-project.org/). Standard logistic regression with adjustment for covariates was used for the enrichment analyses of categorical variables. We applied Firth's penalized logistic regression using the logistf package in R to model the association between ecDNA status and geographic regions, in order to reduce small-sample bias and address potential issues of data separation. For the comparison of numerical variables across groups, we used non-parametric Mann-Whitney (Wilcoxon rank sum) tests. P-values <0.05 were considered statistically significant. Survival analyses were performed using Cox proportional-hazards regression adjusted for age at diagnosis, sex, tumor stage, tumor histology, ancestry, and wGII. If multiple hypothesis testing was required, we used a false-discovery rate correction based on the Benjamini-Hochberg method and reported FDR. FDR <0.05 were considered statistically significant.

FIGURE LEGENDS

Figure 1: Landscape of extrachromosomal DNA in lung cancer across histologies, sex, ancestry, countries, and smoking status. ***a)*** Epidemiological features of the LCINS (top) and LCSS (bottom) cohorts. Total number of samples in each category is indicated. ***b)*** The number and proportion of tumors harboring ecDNA among LCINS and LCSS, stratified by histology. Samples with 1 or more ecDNA were classified as ecDNA$^+$, and samples with no ecDNA were classified as ecDNA$^-$. ***c)*** Forest plot of the full cohort (LCINS and LCSS) from a multivariate logistic regression analysis of key epidemiological and clinical features and the presence of ecDNA in the tumors. ***d)*** Volcano plot showing $\log_2$ odds ratio (x-axis) and $\log_{10}$ FDR corrected q-value (y-axis) for a one-vs-all statistical comparison of ecDNA prevalence by region. ***e)*** Forest plots for multivariate logistic regression models testing key epidemiological and clinical features with the presence of ecDNA in the tumors, stratified by smoking status. NA/EU: North America or Europe region, AS: Asia region EAS: East Asian genetic ancestry super-sample, EUR: European genetic ancestry super-sample, LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma.

Figure 2. Genomic and clinical characteristics associated with ecDNA. ***a)*** Volcano plot of logistic regression models of genomic features in association with ecDNA status for LCINS (left) and LCSS (right). ***b)*** Volcano plot of a logistic regression model of driver gene alterations in a sample in association with ecDNA status for LCINS (left) and LCSS (right). ***c)*** Volcano plot of a logistic regression model of hotspot driver mutations in a sample in association with ecDNA status for LCINS (left) and LCSS (right). In all volcano plots, the x-axes reflect the $\log_2$ odds ratio, and the y-axes correspond to the $\log_{10}$ FDR q-value. An FDR q-value threshold of 0.05 is

indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line. ***d)*** Bar plot indicating the proportion of ecDNA containing a particular genomic element in LCINS vs LCSS. ***e)*** Kaplan–Meier survival curves for 5-year overall survival stratified by the mode of amplification status for lung cancers in LCINS (left) and LCSS (right). P-values for significance and HRs of the difference were calculated using two-sided Cox proportional-hazards regression with adjustment for age, sex, tumor stage, ancestry and histology and are indicated within each plot.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Fig 1. Additional cohort information. Sankey diagrams showing the numbers of samples with clinical variables (stage, metastasis) as well as survival data available for ***a)*** LCINS and ***b)*** LCSS

Supplementary Fig 2. Distribution of ecDNA harboring lung cancers in LCINS and LCSS by country. Maps showing the worldwide prevalence of lung cancers harboring ecDNA ($ecDNA^+$) by country for the ***a)*** LCINS cohort and ***b)*** LCSS cohort. Volcano plot of a multivariate one-vs-all logistic regression model for each country, with ecDNA status as the outcome, for ***c)*** LCINS and ***d)*** LCSS.

Supplementary Fig 3. Analysis of the association between stage, metastasis, and passive smoking with ecDNA status in all LCINS and LCSS. Forest plots of logistic regression model with ecDNA status as the outcome and the variables for stage in **a)** LCINS and **b)** LCSS; metastasis in **c)** LCINS and **d)** LCSS; and passive smoking in **e)** LCINS. The total number of samples with information for all variables is indicated above each plot.

Supplementary Fig 4. Analysis of the association between stage, metastasis, and passive smoking with ecDNA status in lung adenocarcinomas. Forest plots of logistic regression model with ecDNA status as the outcome and the variables for stage in **a)** LCINS and **b)** LCSS; metastasis in **c)** LCINS and **d)** LCSS; and passive smoking in **e)** LCINS. The total number of samples with information for all variables is indicated above each plot.

Supplementary Fig 5. Genomic features associated with the presence of ecDNA. ***a)*** Volcano plot of logistic regression models of genomic features in association with ecDNA status for lung adenocarcinoma in LCINS (left) and lung adenocarcinoma in LCSS (right). In all volcano plots, the x-axes reflect the $\log_2$ odds ratio, and the y-axes correspond to the $\log_{10}$ FDR q-value. An FDR q-value threshold of 0.05 is indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line. ***b)*** Forest plot of multivariate logistic regression modeling the presence or absence of ecDNA in relation to whole-genome doubling in a pan-cancer dataset comprising four different tumor types from the PCAWG project, adjusted for the covariates of age, purity, sex, and tissue type. All significant associations are indicated in red. ***c)*** Bar plot indicating the proportion of samples that contained overlapping chromothripsis and ecDNA regions (red) and the proportion of all samples that did not contain any overlap (gray) for all samples that had both an ecDNA and chromothripsis event detected. ***d)*** Forest plot of multivariate logistic regression modeling the presence or absence of ecDNA in relation to telomere length estimates from Telhunter, whereby the telomere length ratio (tumor/normal) was input as a continuous numeric variable. **e)** Boxplot indicating the distributions and median telomere length ratio (tumor/normal) for ecDNA+ vs ecDNA- samples, both in never-smokers (left) and smokers (right). The p-value from a multivariate logistic regression with age, gender, tumor purity, histology, ancestry and wGII score is also indicated.

Supplementary Fig 6. Association of mutational signatures with ecDNA status. Volcano plots of logistic regression models of the activities of Single base substitution (SBS), Indel (ID), copy number (CN) and structural variant (SV) signatures in association with ecDNA in lung cancers of ***a)*** LCINS and ***b)*** LCSS. The logistic regression models were adjusted for age, sex,

ancestry, histology, and purity. In all volcano plots, the x-axes reflect the $\log_2$ odds ratio, and the y-axes correspond to the $\log_{10}$ FDR q-value. An FDR q-value threshold of 0.05 is indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line.

Supplementary Fig 7. Genomic content of ecDNA

a) Upset plot indicating the number of ecDNA with oncogenes, immunomodulatory genes, promoter or enhancer regions, other genes that are not oncogenes, or unidentified genomic elements (Unknown) or any combination thereof for LCINS and LCSS. **b)** Bar plot indicating the proportion of ecDNA containing a particular genomic element in LCINS vs LCSS. **c)** $\log_2$ median copy number (y-axis) for each class of ecDNA (x-axis) in LCINS and LCSS.

Supplementary Fig 8. Highly sensitive pipeline for ecDNA detection that accounts for tumor purity and ploidy.

a) Schematic depicting the methodology for the rescaled pipeline that considers tumor purity and ploidy. The logR track for the region containing the *EGFR* gene on chr7 and the purity and ploidy computed from Battenberg is indicated in one tumor as example (left). The exact location of the *EGFR* gene is shown by the dashed red lines. The equation that simultaneously accounts for sample purity and ploidy when computing the total copy number is shown, as well as the total copy number estimate when assuming a sample purity of 1.0 and ploidy of 2.0 (Uncorrected CN) *vs*. considering the purity and ploidy (Corrected CN). The Sashimi plot indicates the structural variants (colored arcs) and smoothed coverage around the *EGFR* region, which was annotated as ecDNA by AmpliconClassifier. **b)** Distribution of tumor purities for samples with an ecDNA

annotation for both the default (left) and rescaled (right) pipelines. P-value for a Mann-Whitney test between ecDNA$^+$ and ecDNA$^-$ samples is indicated. The total number of samples in each group is also indicated. **c)** The ecDNA prevalence (y-axis), which is the proportion of ecDNA$^+$ samples over the total number of samples, for all possible tumor purity cutoffs applied to samples (x-axis). The total number of samples available for analysis after each tumor purity cutoff is applied is also indicated. **d)** Gene expression ($\log_2$ CPM) of the EGFR gene for all samples with RNA-seq data available. LCINS samples with EGFR-ecDNA are indicated in red.

SUPPLEMENTARY TABLE LEGENDS

Supplementary Table 1. Co-occurrence of ecDNA genetic cargo in LCINS

Supplementary Table 2. Co-occurrence of ecDNA genetic cargo in LCSS

Supplementary Table 3. Recurrently amplified oncogenes on ecDNA in LCINS across unique genomic loci

Supplementary Table 4. Recurrently amplified oncogenes on ecDNA in LCSS across unique genomic loci

Supplementary Table 5. Additional *EGFR*-ecDNA detected by modified pipeline

Ethics declarations

Since the National Cancer Institute only received de-identified samples and data from collaborating centers, had no direct contact or interaction with the study participants, and did not use or generate identifiable private information, Sherlock-*Lung* has been determined to constitute "Not Human Subject Research (NHSR)" based on the federal Common Rule (45 CFR 46; https://www.ecfr.gov/cgi-bin/ECFR?page=browse).

Data Availabiltiy

Normal and tumor-paired CRAM files for the WGS subjects of the Sherlock-Lung study and the EAGLE study have been deposited in dbGaP under the accession numbers phs001697.v2.p1 and phs002992.v1.p1, respectively.

Code availability

All custom code developed for the analysis and generation of figures in this manuscript is publicly available in the GitHub repository (https://github.com/azhark2/Sherlock-Lung-ecDNA). Copy number variants were called using CNVkit (v.0.9.6) and rescaled for tumor purity and ploidy using estimates from Battenberg (v.2.2.9). Extrachromosomal DNA was detected and classified using AmpliconArchitect (v.1.3) and AmpliconClassifier (v.1.3.1). Chromothripsis was identified using ShatterSeek with structural variants called by Meerkat (v.0.189) and Manta (v.1.6.0). Telomere length was estimated using TelSeq (v.0.0.2). Mutational signatures were extracted using SigProfilerExtractor (v.1.1.3) with matrices generated by SigProfilerMatrixGenerator. Driver genes were identified using IntOGen (v.2020.02.0123).

Additional publicly available software and algorithms used are detailed in the **Reporting Summary** and the **Methods** section. Any additional supporting scripts are available from the corresponding authors upon reasonable request.

Author Contributions

Conceptualization, MTL, TZ, LBA; Methodology, AK, TZ; Formal Analysis, AK, TZ, MDG, JL, WZ, KB, EB, YY; Data Curation, TZ, PH, JM, CH, MM, OWL, MAN; Resources: NR, QL, DCW, LY, SJC, ACP, DC, KJ, AH, BH; Pathology evaluation: RH, MKB, LMS, PJ, CL, WDT, S-RY; Writing – Original Draft, AK; Writing – Editing, LBA, TZ, PHH, MTL; Writing – Review: All authors; Visualization: AK; Supervision, LBA, TZ, MTL.

ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of the National Institute of Health (NIH) (project ZIACP101231 to MTL), and by the Anne Wojcicki Foundation (Grant Number: LC009). The contributions of the NIH author(s) were made as part of their official duties as NIH federal employees, are in compliance with agency policy requirements, and are considered Works of the United States Government. However, the findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services. We want to acknowledge the patients and the INCLIVA Biobank (PT17/0015/0049) integrated in the Spanish National Biobanks Network and in the Valencian Biobanking Network for their collaboration. This study was supported by the Health and Medical Research Fund of Hong Kong SAR, HMRF 03142856. The related studies of Taiwan site were supported by grants from the Ministry of Health and

Welfare, Taiwan DOH97-TD-G-111-026 (C.A.H.), DOH98-TD-G-111-015 (C.A.H.), DOH99-TD-G-111-028 (C.A.H.); DOH97-TD-G-111-029 (C.Y.C.), DOH98-TD-G-111-018 (C.Y.C.), DOH99-TD-G-111-015 (C.Y.C.), DOH97-TD-G-111-028(I.S.C.), DOH98-TD-G-111-017(I.S.C.), DOH99-TD-G-111-014(I.S.C.), and the Ministry of Science and Technology, Taiwan MOST109-2740-B-400-002 (C.A.H.), MOST110-2740-B-400-002 (C.A.H.), MOST111-2740-B-400-002 (C.A.H.). This work has been supported in part by the Tissue Core at the H. Lee Moffitt Cancer Center & Research Institute, a comprehensive cancer center designated by the National Cancer Institute and funded in part by a Moffitt Cancer Center Support Grant (no. P30-CA076292). And, in part, by NIH (NCI) grant # U01CA209414 to the Boston Lung Cancer Survival Study of the Dana-Farber/ Harvard Cancer Center (D.C.C.). The authors would like to thank the team at the IUCPQ site of the Quebec Respiratory Health Network Biobank of the FRQS for their valuable assistance, and would like to thank the staff at Harvard University, Yale University, Roswell Park Cancer Institute and Roswell PI, Instituto Nacional de Cancerologia, Nice University Hospital Centre (Nice UHC) - University Côte d'Azur and the Nice Biobank CRB, Toronto University Health Network, and Mayo Clinic for their assistance providing samples and corresponding clinical data. The computational analyses reported in this manuscript have utilized the NIH high-performance Biowulf Cluster. We thank the study participants and the staff at Westat Inc. for their assistance in collecting samples and corresponding clinical data. Part of this work was supported by the US National Institute of Health grants R01ES032547, R01CA269919, P01CA281819, and U01CA290479 to LBA as well as by LBA’s Packard Fellowship for Science and Engineering. MD-G was awarded with a fellowship within the “Generación D” initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by the European Union NextGeneration

EU funds, through PRTR. The research in this study was also supported by UC San Diego Sanford Stem Cell Institute. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov). The funders had no roles in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

COMPETING INTERESTS

LBA is a co-founder, CSO, scientific advisory member, and consultant for io9, has equity and receives income. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. LBA is a compensated member of the scientific advisory board of Inocras. LBA's spouse is an employee of Hologic, Inc. LBA declares U.S. provisional applications with serial numbers: 63/289,601; 63/269,033; 63/366,392; 63/412,835 as well as international patent application PCT/US2023/010679. LBA is also an inventor of a US Patent 10,776,718 for source identification by non-negative matrix factorization. LBA and MD-G further declare a European patent application with application number EP25305077.7. Soo-Ryum Yang has received speaking fees from Medscape, OncLive, Cure Today, Medical Learning Institute, PRIME Education, AstraZeneca, Roche; and consulting fees from AstraZeneca, AbbVie, Merus, Eli Lilly, Boehringer Ingelheim, Roche, Amgen, Sanofi. All other authors declare no conflicts of interest.

REFERENCES

1 Benner, S. E., Wahl, G. M. & Von Hoff, D. D. Double minute chromosomes and homogeneously staining regions in tumors taken directly from patients versus in human tumor cell lines. *Anticancer Drugs* **2**, 11-25 (1991). https://doi.org/10.1097/00001813-199102000-00002

2 Nielsen, J., Walsh, J. T., Degen, D. R., Drabek, S. M. & McGill, J. R. Evidence of gene amplification in the form of double minute chromosomes is frequently observed in lung cancer. *Cancer Genetics* **65**, 120-124 (1993). https://doi.org/https://doi.org/10.1016/0165-4608(93)90219-C

3 Hung, K. L. *et al.* Coordinated inheritance of extrachromosomal DNAs in cancer cells. *Nature* **635**, 201-209 (2024). https://doi.org/10.1038/s41586-024-07861-8

4 Lange, J. T. *et al.* The evolutionary dynamics of extrachromosomal DNA in human cancers. *Nat Genet* **54**, 1527-1533 (2022). https://doi.org/10.1038/s41588-022-01177-x

5 Wu, S. *et al.* Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* **575**, 699-703 (2019). https://doi.org/10.1038/s41586-019-1763-5

6 Kim, H. *et al.* Mapping extrachromosomal DNA amplifications during cancer progression. *Nat Genet* **56**, 2447-2454 (2024). https://doi.org/10.1038/s41588-024-01949-7

7 Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet* **52**, 891-897 (2020). https://doi.org/10.1038/s41588-020-0678-2

8 Bailey, C. *et al.* Origins and impact of extrachromosomal DNA. *Nature* **635**, 193-200 (2024). https://doi.org/10.1038/s41586-024-08107-3

9 Tang, J. *et al.* Enhancing transcription-replication conflict targets ecDNA-positive cancers. *Nature* **635**, 210-218 (2024). https://doi.org/10.1038/s41586-024-07802-5

10 Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat Commun* **10**, 392 (2019). https://doi.org/10.1038/s41467-018-08200-y

11 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020). https://doi.org/10.1038/s41586-020-1943-3

12 Landi, M. T. *et al.* Tracing Lung Cancer Risk Factors Through Mutational Signatures in Never-Smokers. *Am J Epidemiol* **190**, 962-976 (2021). https://doi.org/10.1093/aje/kwaa234

13 Consonni, D. *et al.* Outdoor particulate matter (PM10) exposure and lung cancer risk in the EAGLE study. *PLoS One* **13**, e0203539 (2018). https://doi.org/10.1371/journal.pone.0203539

14 Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020). https://doi.org/10.1038/s41586-020-1969-6

15 Lee, J. J. *et al.* Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma. *Cell* **177**, 1842-1857 e1821 (2019). https://doi.org/10.1016/j.cell.2019.05.013

16 Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-1120 (2012). https://doi.org/10.1016/j.cell.2012.08.029

17 Lee, J. K. *et al.* Clonal History and Genetic Predictors of Transformation Into Small-Cell Carcinomas From Lung Adenocarcinomas. *J Clin Oncol* **35**, 3065-3074 (2017). https://doi.org/10.1200/JCO.2016.71.9096

18 Leong, T. L. *et al.* Deep multi-region whole-genome sequencing reveals heterogeneity and gene-by-environment interactions in treatment-naive, metastatic lung cancer. *Oncogene* **38**, 1661-1675 (2019). https://doi.org/10.1038/s41388-018-0536-1

19 Bergstrom, E. N. *et al.* Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. *Nature* **602**, 510-517 (2022). https://doi.org/10.1038/s41586-022-04398-6

20 Chen, J. *et al.* MYC-driven increases in mitochondrial DNA copy number occur early and persist throughout prostatic cancer progression. *JCI Insight* **8** (2023). https://doi.org/10.1172/jci.insight.169868

21 Bielski, C. M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet* **50**, 1189-1195 (2018). https://doi.org/10.1038/s41588-018-0165-1

22 O'Sullivan, R. J. & Karlseder, J. Telomeres: protecting chromosomes against genome instability. *Nat Rev Mol Cell Biol* **11**, 171-181 (2010). https://doi.org/10.1038/nrm2848

23 Shoshani, O. *et al.* Chromothripsis drives the evolution of gene amplification in cancer. *Nature* **591**, 137-141 (2021). https://doi.org/10.1038/s41586-020-03064-z

24 Feuerbach, L. *et al.* TelomereHunter – in silico estimation of telomere content and composition from cancer genomes. *BMC Bioinformatics* **20**, 272 (2019). https://doi.org/10.1186/s12859-019-2851-0

25 Karimian, K. *et al.* Human telomere length is chromosome end-specific and conserved across individuals. *Science* **384**, 533-539 (2024). https://doi.org/10.1126/science.ado0431

26 Wang, Y. *et al.* Extrachromosomal DNA Gives Cancer a New Evolutionary Pathway. *bioRxiv*, 2025.2004.2026.650733 (2025). https://doi.org/10.1101/2025.04.26.650733

27 Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature* **606**, 984-991 (2022). https://doi.org/10.1038/s41586-022-04738-6

28 Hadi, K. *et al.* Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs. *Cell* **183**, 197-210 e132 (2020). https://doi.org/10.1016/j.cell.2020.08.006

29 Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med* **23**, 517-525 (2017). https://doi.org/10.1038/nm.4292

30 Hung, K. L. *et al.* ecDNA hubs drive cooperative intermolecular oncogene expression. *Nature* **600**, 731-736 (2021). https://doi.org/10.1038/s41586-021-04116-8

31 Zhu, Y. *et al.* Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer Cell* **39**, 694-707.e697 (2021). https://doi.org/10.1016/j.ccell.2021.03.006

32 Chene, P. Inhibiting the p53-MDM2 interaction: an important target for cancer therapy. *Nat Rev Cancer* **3**, 102-109 (2003). https://doi.org/10.1038/nrc991

33 Zhang, T. *et al.* Genomic and evolutionary classification of lung cancer in never smokers. *Nat Genet* **53**, 1348-1359 (2021). https://doi.org/10.1038/s41588-021-00920-0

34 Diaz-Gay, M. *et al.* The mutagenic forces shaping the genomes of lung cancer in never smokers. *Nature* **644**, 133–144 (2025). https://doi.org/10.1038/s41586-025-09219-0

35 Bergmann, E. A., Chen, B. J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics* **32**, 3196-3198 (2016). https://doi.org/10.1093/bioinformatics/btw389

36 Pedersen, B. S. *et al.* Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Medicine* **12**, 62 (2020). https://doi.org/10.1186/s13073-020-00761-2

37 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012). https://doi.org/10.1016/j.cell.2012.04.023

38 Boot, A. *et al.* In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res* **28**, 654-665 (2018). https://doi.org/10.1101/gr.230219.117

39 Dharanipragada, P. *et al.* Blocking Genomic Instability Prevents Acquired Resistance to MAPK Inhibitor Therapy in Melanoma. *Cancer Discov* **13**, 880-909 (2023). https://doi.org/10.1158/2159-8290.CD-22-0787

40 Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873 (2016). https://doi.org/10.1371/journal.pcbi.1004873

41 Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**, 8971 (2015). https://doi.org/10.1038/ncomms9971

42 Zhang, T. *et al.* Deciphering lung adenocarcinoma evolution and the role of LINE-1 retrotransposition. *Nature* (in press; accepted). Scheduled publication date: December 10, 2025. Preprint available at *bioRxiv* (2025): https://doi.org/10.1101/2025.03.14.643063

43 Cortes-Ciriano, I. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* **52**, 331-341 (2020). https://doi.org/10.1038/s41588-019-0576-7

44 Ding, Z. *et al.* Estimating telomere length from whole genome sequence data. *Nucleic Acids Res* **42**, e75 (2014). https://doi.org/10.1093/nar/gku181

45 Khandekar, A. *et al.* Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator. *BMC Genomics* **24**, 469 (2023). https://doi.org/10.1186/s12864-023-09584-y

46 Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019). https://doi.org/10.1186/s12864-019-6041-2

47 Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom* **2**, 100179 (2022). https://doi.org/10.1016/j.xgen.2022.100179

48 Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* **10**, 1081-1082 (2013). https://doi.org/10.1038/nmeth.2642

49 Muinos, F., Martinez-Jimenez, F., Pich, O., Gonzalez-Perez, A. & Lopez-Bigas, N. In silico saturation mutagenesis of cancer genes. *Nature* **596**, 428-432 (2021). https://doi.org/10.1038/s41586-021-03771-1

50 Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017** (2017). https://doi.org/10.1200/PO.17.00011

51 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034-1035 (2018). https://doi.org/10.1016/j.cell.2018.07.034

52 Cheng, F. *et al.* Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat Genet* **53**, 342-353 (2021). https://doi.org/10.1038/s41588-020-00774-y

53 Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* **48**, D58-D64 (2020). https://doi.org/10.1093/nar/gkz980

54 Reznik, E. *et al.* Mitochondrial DNA copy number variation across human cancers. *Elife* **5** (2016). https://doi.org/10.7554/eLife.10769

55 Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet* **52**, 306-319 (2020). https://doi.org/10.1038/s41588-019-0562-0

Fig 1.

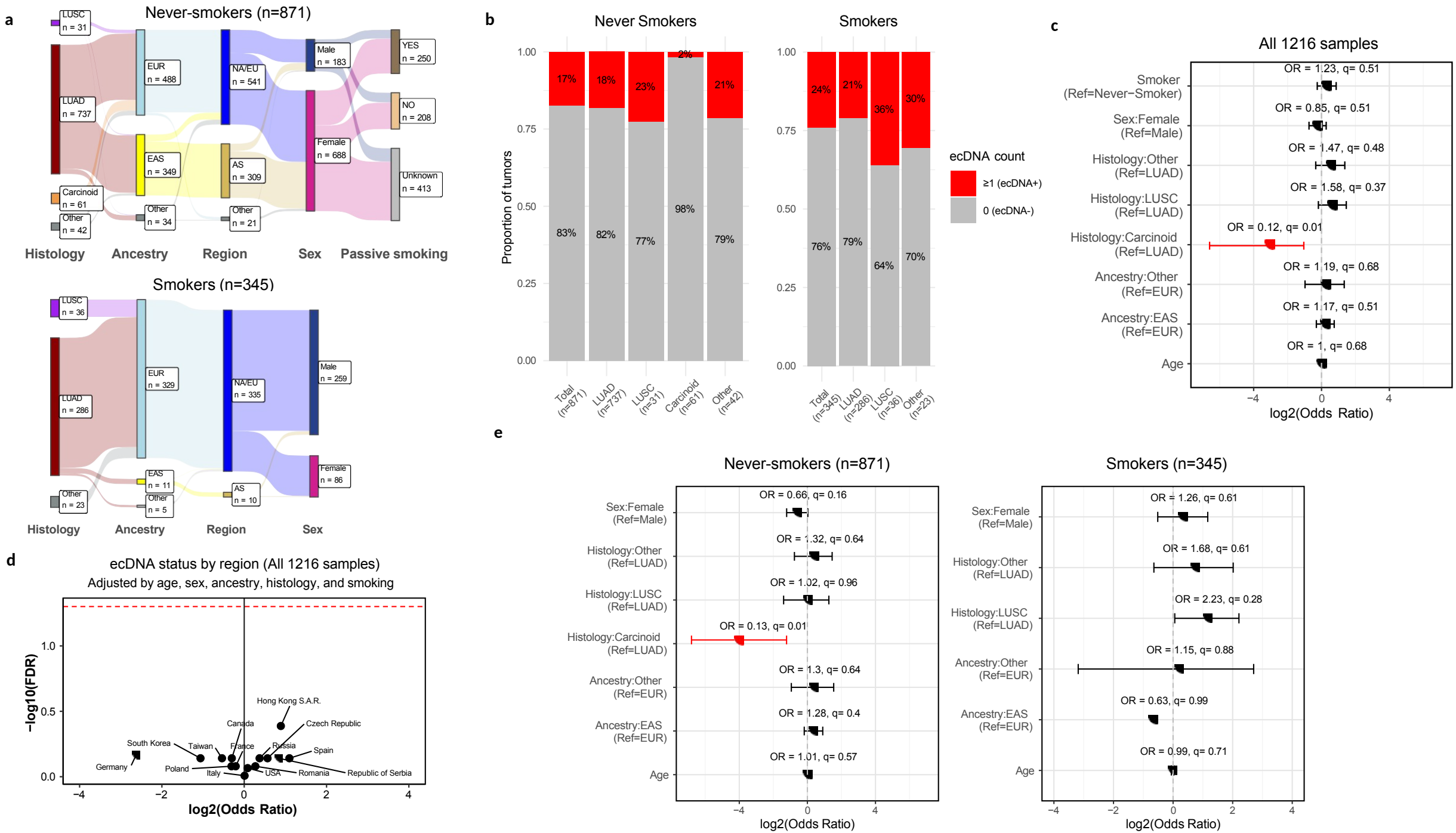
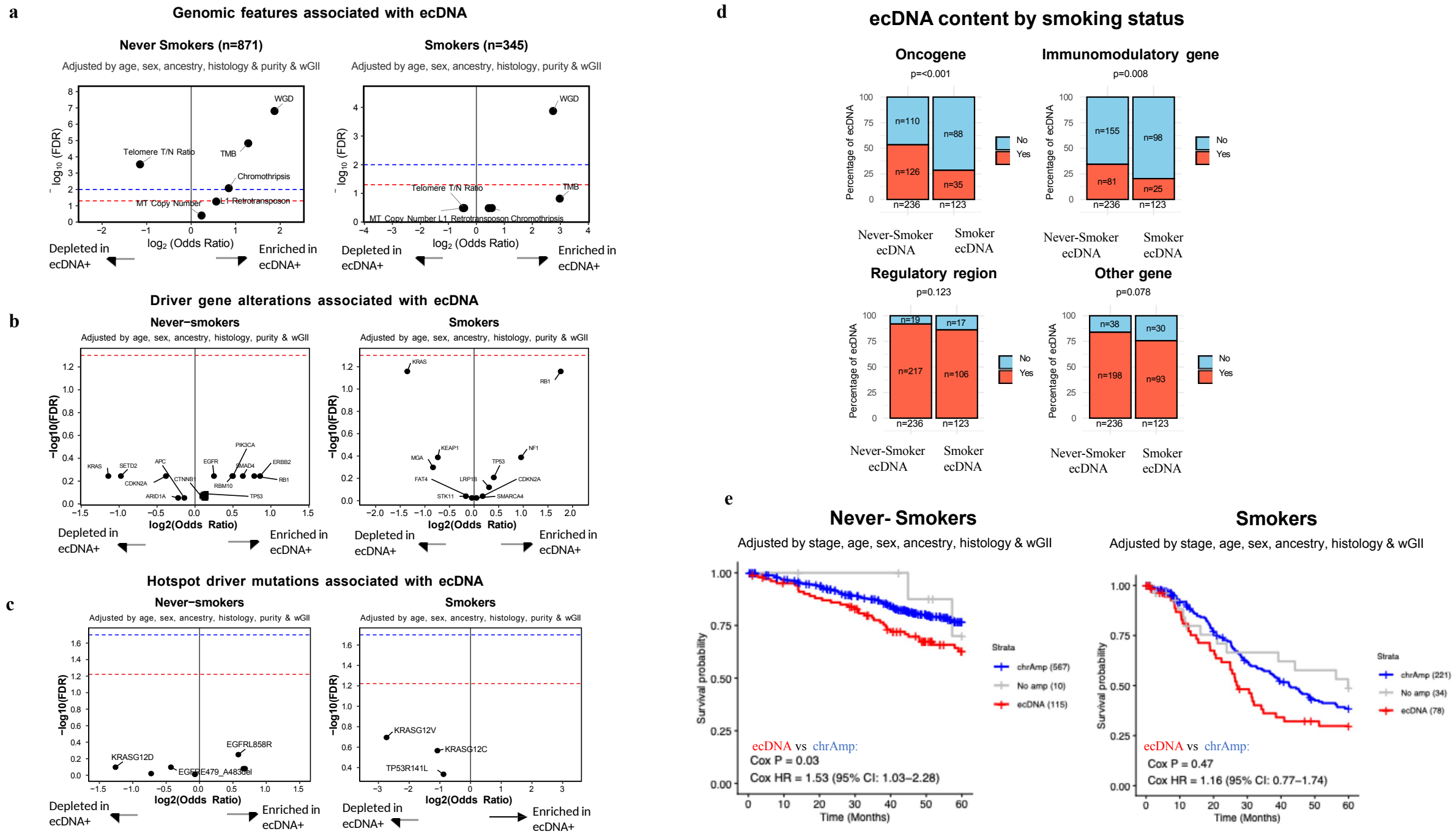

a
Never-smokers (n=871)
LUSC n = 31
LUAD n = 737
Carcinoid n = 61
Other n = 42
EUR n = 488
EAS n = 349
Other n = 34
NA/EU n = 541
AS n = 309
Other n = 21
Male n = 183
Female n = 688
YES n = 250
NO n = 208
Unknown n = 413
Histology
Ancestry
Region
Sex
Passive smoking
Smokers (n=345)
LUSC n = 36
LUAD n = 286
Other n = 23
EUR n = 329
EAS n = 11
Other n = 5
NA/EU n = 335
AS n = 10
Male n = 259
Female n = 86
Histology
Ancestry
Region
Sex
b
Never Smokers
Smokers
Proportion of tumors
1.00
0.75
0.50
0.25
0.00
17%
18%
23%
2%
21%
83%
82%
77%
98%
79%
24%
21%
36%
30%
76%
79%
64%
70%
Total (n=871)
LUAD (n=737)
LUSC (n=31)
Carcinoid (n=61)
Other (n=42)
Total (n=345)
LUAD (n=286)
LUSC (n=36)
Other (n=23)
ecDNA count
≥1 (ecDNA+)
0 (ecDNA-)
c
All 1216 samples
Smoker (Ref=Never−Smoker)
OR = 1.23, q= 0.51
Sex:Female (Ref=Male)
OR = 0.85, q= 0.51
Histology:Other (Ref=LUAD)
OR = 1.47, q= 0.48
Histology:LUSC (Ref=LUAD)
OR = 1.58, q= 0.37
Histology:Carcinoid (Ref=LUAD)
OR = 0.12, q= 0.01
Ancestry:Other (Ref=EUR)
OR = 1.19, q= 0.68
Ancestry:EAS (Ref=EUR)
OR = 1.17, q= 0.51
Age
OR = 1, q= 0.68
−4
0
4
log2(Odds Ratio)
d
ecDNA status by region (All 1216 samples)
Adjusted by age, sex, ancestry, histology, and smoking
−log10(FDR)
1.0
0.5
0.0
Hong Kong S.A.R.
Canada
Czech Republic
South Korea
Taiwan
France
Russia
Spain
Germany
Poland
Italy
USA
Romania
Republic of Serbia
−4
−2
0
2
4
log2(Odds Ratio)
e
Never-smokers (n=871)
Sex:Female (Ref=Male)
OR = 0.66, q= 0.16
Histology:Other (Ref=LUAD)
OR = 1.32, q= 0.64
Histology:LUSC (Ref=LUAD)
OR = 1.02, q= 0.96
Histology:Carcinoid (Ref=LUAD)
OR = 0.13, q= 0.01
Ancestry:Other (Ref=EUR)
OR = 1.3, q= 0.64
Ancestry:EAS (Ref=EUR)
OR = 1.28, q= 0.4
Age
OR = 1.01, q= 0.57
−4
0
4
log2(Odds Ratio)
Smokers (n=345)
Sex:Female (Ref=Male)
OR = 1.26, q= 0.61
Histology:Other (Ref=LUAD)
OR = 1.68, q= 0.61
Histology:LUSC (Ref=LUAD)
OR = 2.23, q= 0.28
Ancestry:Other (Ref=EUR)
OR = 1.15, q= 0.88
Ancestry:EAS (Ref=EUR)
OR = 0.63, q= 0.99
Age
OR = 0.99, q= 0.71
−4
−2
0
2
4
log2(Odds Ratio)

Fig 2.


a
Genomic features associated with ecDNA
Never Smokers (n=871)
Adjusted by age, sex, ancestry, histology & purity & wGII
Smokers (n=345)
Adjusted by age, sex, ancestry, histology, purity & wGII
WGD
Telomere T/N Ratio
TMB
Chromothripsis
MT Copy Number
L1 Retrotransposon
log10 (FDR)
log2 (Odds Ratio)
Depleted in ecDNA+
Enriched in ecDNA+
b
Driver gene alterations associated with ecDNA
Never−smokers
Adjusted by age, sex, ancestry, histology, purity & wGII
Smokers
Adjusted by age, sex, ancestry, histology, purity & wGII
KRAS
SETD2
CDKN2A
APC
CTNNB1
ARID1A
EGFR
PIK3CA
SMAD4
ERBB2
RB1
RBM10
TP53
MGA
KEAP1
FAT4
LRP1B
STK11
SMARCA4
NF1
−log10(FDR)
log2(Odds Ratio)
Depleted in ecDNA+
Enriched in ecDNA+
c
Hotspot driver mutations associated with ecDNA
Never−smokers
Adjusted by age, sex, ancestry, histology, purity & wGII
Smokers
Adjusted by age, sex, ancestry, histology, purity & wGII
KRASG12D
EGFRL858R
EGFRE479_A483del
KRASG12V
KRASG12C
TP53R141L
−log10(FDR)
log2(Odds Ratio)
Depleted in ecDNA+
Enriched in ecDNA+
d
ecDNA content by smoking status
Oncogene
p=<0.001
Immunomodulatory gene
p=0.008
Regulatory region
p=0.123
Other gene
p=0.078
Percentage of ecDNA
n=110
n=126
n=88
n=35
n=155
n=81
n=98
n=25
n=19
n=217
n=17
n=106
n=38
n=198
n=30
n=93
n=236
n=123
No
Yes
Never-Smoker ecDNA
Smoker ecDNA
e
Never- Smokers
Adjusted by stage, age, sex, ancestry, histology & wGII
Survival probability
Time (Months)
Strata
chrAmp (567)
No amp (10)
ecDNA (115)
ecDNA vs chrAmp:
Cox P = 0.03
Cox HR = 1.53 (95% CI: 1.03–2.28)
Smokers
Adjusted by stage, age, sex, ancestry, histology & wGII
chrAmp (221)
No amp (34)
ecDNA (78)
ecDNA vs chrAmp:
Cox P = 0.47
Cox HR = 1.16 (95% CI: 0.77–1.74)